TESIS DOCTORAL

Análisis Bayesiano de datos con respuesta categórica

Bayesian analysis of categorical response data

Lizbeth Naranjo Albarrán

Departamento de Matemáticas

2014

UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

Análisis Bayesiano de datos con respuesta categórica

Bayesian analysis of categorical response data

Autora:
Lizbeth Naranjo Albarrán

Departamento de Matemáticas

Conformidad de los Directores:

Fdo: Dr. Carlos Javier Pérez Sánchez Fdo: Dr. Jacinto Ramón Martín Jiménez

2014

Dedicado a Reberiano, Bertha y Beatriz

Agradecimientos

Quiero agradecer a mis directores, Dr. Carlos Javier Pérez Sánchez y Dr. Jacinto Ramón Martín Jiménez, por su tiempo, esfuerzo y dedicación que en estos años me han prestado, así como su conocimiento y guía para el enriquecimiento de esta tesis, y por todo el apoyo que me han dado para mi desarrollo profesional.

También quiero agradecer a la Universidad de Extremadura por el financiamiento recibido durante estos cuatro años, ya que hizo posible el desarrollo de esta tesis doctoral. En especial quiero agradecer a los profesores del Departamento de Matemáticas de la Facultad de Ciencias por su acogida y amistad.

Agradecezco al Profesor Emmanuel Lesaffre del *Department of Biostatistics* de *Erasmus Medical Center* (Rotterdam), por sus sugerencias en el desarrollo de mi investigación y por la oportunidad de colaborar con él, y a los miembros de dicho departamento por su hospitalidad durante mi estancia.

Mi más sincero agradecimiento a mis profesores de la Universidad Nacional Autónoma de México por mi instrucción académica, y en especial, a los profesores del Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Dr. Eduardo Arturo Gutiérrez Peña, Dra. Silvia Ruiz-Velasco Acosta y Dr. Federico O'Reilly Togno, por su apoyo y guía en mi desarrollo profesional y laboral.

Finalmente, gracias a mis padres Reberiano y Bertha, a mis hermanos Karina e Israel, y a mis sobrinos Julio y Sahara, por su amor y apoyo incondicional. Gracias a Beatriz por su lealtad, ayuda y comprensión durante estos años en España. También quiero agradecer a mis amigos y a todos aquellos que de alguna manera hicieron posible esta tesis doctoral.

A todos gracias.

Resumen

Esta tesis tiene como objetivo estudiar dos temas importantes en Estadística. El primero es el desarrollo de modelos robustos de regresión utilizando funciones de enlace flexibles y, el segundo es la extensión de modelos de datos categóricos para abordar la clasificación incorrecta. Con este fin, se han desarrollado e implementado varios métodos Bayesianos.

Las distribuciones potencial-exponencial asimétricas (AEP) pueden tratar con éxito la simetría/asimetría y las colas ligeras/pesadas de forma simultánea. Aún más, las distribuciones pueden ajustar cada cola por separado. La familia AEP incluye la distribución potencial-exponencial (EP) como un caso particular. Esto proporciona distribuciones flexibles, con colas más ligeras o más pesadas que la distribución normal. La gran flexibilidad de estas distribuciones motiva el desarrollo de nuevos modelos.

Se han propuesto métodos robustos de regresión que utilizan distribuciones potencial-exponencial simétricas y asimétricas en tres contextos diferentes. En primer lugar, se desarrolla una exploración de la distribución a posteriori mediante el uso de la distribución AEP para el modelo de datos. En segundo lugar, se considera un modelo de regresión lineal donde la distribución AEP se utiliza para la variable de error. Finalmente, se presentan dos modelos de regresión binaria donde la inversa de la función de distribución acumulativa EP/AEP se utiliza como la función de enlace. Los tres modelos comparten algunas características comunes, a pesar del hecho de que se han construido para diferentes propósitos. La representación de la mixtura y el enfoque computacional son los puntos comunes de las aproximaciones. La idea de utilizar representaciones de mixtura de escala de distribuciones uniformes para las distribuciones EP y AEP se ha utilizado para obtener algoritmos de muestreo Gibbs eficientes. Esto ha permitido evitar las dificultades computacionales que han hecho que estos modelos sean poco prácticos. Se han realizado comparaciones con modelos competitivos que muestran el buen funcionamiento de las aproximaciones propuestas.

Los modelos anteriores consideran datos libres de error. Sin embargo, muchas veces los procesos de generación/obtención de datos no están exentos de error cuando se recogen en situaciones reales. Incluso una pequeña proporción de datos clasificados incorrectamente puede producir un impacto importante sobre las inferencias, porque la cantidad efectiva de información puede reducirse drásticamente. Si el error de clasificación en un proceso de generación de datos no se modela adecuadamente, la información puede percibirse como más precisa de lo que realmente es, dando lugar,

en muchos casos, a una toma de decisiones no óptima. Este hecho y la escasez de aproximaciones que abordan este tema desde un punto de vista Bayesiano motivan el desarrollo de nuevas aproximaciones.

Se han considerado modelos lineales generalizados para describir la dependencia de los datos sobre las variables explicativas cuando las respuestas categóricas están sujetas a errores de clasificación. Se han propuesto modelos de regresión con funciones de enlace probit y t para datos binarios mal clasificados, y se han aplicado en el contexto de la pérdida auditiva causada por la exposición al ruido durante la jornada laboral. Las dificultades computacionales se han evitado mediante el uso de esquemas de aumento de datos con dos tipos diferentes de variables latentes. Esto permite obtener algoritmos eficientes de muestreo de Gibbs y *Expectation-Maximization*. Esta aproximación se ha extendido a datos con respuestas politómicas mal clasificadas, con especial énfasis en el caso ordinal. Por último, motivado por un estudio longitudinal de salud oral, se ha desarrollado un tipo diferente de modelo ordinal que considera la clasificación incorrecta. Este modelo explota la estructura multinivel de los datos.

Summary

This thesis aims at studying two important topics in Statistics. The first one is the development of robust regression models by using flexible link functions and, the second one is the extension of categorical data models to address misclassification. Several approaches have been developed in this thesis under Bayesian methodology.

The asymmetrical exponential power (AEP) distributions can successfully handle both symmetry/asymmetry and light/heavy tails in a simultaneous way. Even more, the distributions can fit each tail separately. The AEP family includes the exponential power (EP) distribution as a particular case. It provides more flexible distributions with lighter and heavier tails compared to the normal one. The great flexibility of these distributions motivates the development of new models.

Robust regression approaches that use symmetrical and asymmetrical exponential power distributions have been proposed in three different contexts. Firstly, a posterior distribution exploration is developed by using the AEP distribution for data models. Secondly, a linear regression model is considered where the AEP distribution is used for the error variable. Finally, binary regression models are presented where the inverse of the EP/AEP cumulative distribution function is used as the link function. All three models share some common characteristics in spite of the fact that they have been built for different purposes. The mixture representation and the computational issue are the linking points of all the approaches. The idea of using scale mixture of uniforms representations of the EP and AEP distributions has been exploited to derive efficient Gibbs sampling algorithms. This has allowed to avoid the computational difficulties that have made these models impractical. Comparisons with competing models show the good performance of the proposed approaches.

All previous approaches consider error-free data. However, many times data-generating processes are not error-free when data are collected in real situations. Even a small proportion of misclassified data can produce an important impact on inferences, because the effective amount of information can be dramatically reduced. If the misclassification in a data-generating process is not properly modelled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non optimal decision making. This fact and the scarcity of approaches addressing this topic from a Bayesian viewpoint motivate the development of new approaches.

Generalized linear models have been considered to describe the dependence of data on explanatory variables when the categorical outcome is subject to misclassification.

Probit and t-link regression models for misclassified binary data have been proposed, and they have been applied in the context of hearing loss caused by exposure to occupational noise. The computational difficulties have been avoided by using data augmentation frameworks with two different types of latent variables. This allows to derive efficient Gibbs sampling and Expectation-Maximization algorithms. This approach has been extended to misclassified polychotomous response data, with special emphasis in the ordinal case. Finally, motivated by a longitudinal oral health study, a different type of ordinal model considering misclassification has been developed. This model exploits the multilevel structure of the data.

Diffusion

This thesis corresponds to a collection of the following original publications:

Naranjo, L., Pérez, C. J., and Martín, J. (2012). Bayesian analysis of a skewed exponential power distribution. In *Proceedings of COMPSTAT 2012, 20th International Conference on Computational Statistics*, pages 641–652, Limassol, Cyprus.

Naranjo, L., Martín, J., and Pérez, C. J. (2014). Bayesian binary regression with exponential power link. *Computational Statistics and Data Analysis*, 71:464–476.

Naranjo, L., Pérez, C. J., and Martín, J. (2014). Bayesian analysis of some models that use the asymmetric exponential power distribution. *Statistics and Computing*. In Press.

Naranjo, L., Martín, J., Pérez, C. J., and Rufo, M. J. (2014). Addressing misclassification for binary data: probit and t-link regressions. *Journal of Statistical Computation and Simulation*. In Press.

Naranjo, L., Pérez, C. J., Martín, J. R., and Lesaffre, E. (2014). A Bayesian approach for misclassified ordinal response data. Preprint 155, Universidad de Extremadura, Badajoz, Spain.

Naranjo, L., Mutsvari, T., Pérez, C. J., Martín, J. R., and Lesaffre, E. (2014). A Bayesian multilevel model for misclassified ordinal response data: an application to caries experience lesion severity. Preprint 154, Universidad de Extremadura, Badajoz, Spain.

The approaches proposed in this thesis have been presented as communications in the following conferences:

Martín, J., Naranjo, L., and Pérez, C. J. (2011). A Bayesian binary regression model with exponential power link. In *ISI 2011, 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland. Poster.

Naranjo, L., Martín, J., and Pérez, C. J. (2011). Regression models for misclassified binary data. In *Workshop Métodos Bayesianos 11*, Madrid, Spain. Talk.

Naranjo, L., Pérez, C. J., and Martín, J. (2011). A skewed probit link-based regression model for misclassified binary data. In *CEIB 2011, XIII Conferencia Española y III Encuentro Iberoamericano de Biometría*, Barcelona, Spain. Poster.

Pérez, C. J., Naranjo, L., and Martín, J. (2011). Bayesian regression models for misclassified binary data. In *Conference Proceedings of the Finance and Economics Conference 2011*, Frankfurt am Main, Germany. Talk.

Naranjo, L., Pérez, C. J., and Martín, J. (2012). Un modelo Bayesiano de regresión binaria con función de enlace potencial-exponencial asimétrica. In *SEIO 2012, XXXIII Congreso Nacional de Estadística e Investigación Operativa*, Madrid, Spain. Poster.

Naranjo, L., Pérez, C. J., and Martín, J. (2012). Bayesian analysis of a skewed exponential power distribution. In *COMPSTAT 2012, 20th International Conference on Computational Statistics*, Limassol, Cyprus. Poster.

Naranjo, L., Pérez, C. J., and Martín, J. (2012). Bayesian analysis of misclassified polychotomous response data. In *CFE-ERCIM 2012, 5th International Conference of the ERCIM Working Group on Computing & Statistics, 6th CSDA International Conference on Computational and Financial Econometrics*, Oviedo, Spain. Talk.

Martín, J., Naranjo, L., and Pérez, C. J. (2013). Bayesian analysis of misclassified polychotomous response data. In *EMS 2013, 29th European Meeting of Statisticians*, Budapest, Hungary. Poster.

Contents

List of Tables

List of Figures

Part I

General introduction

Chapter 1

Introduction

1.1 Bayesian analysis

Over the last years there has been a significant upsurge of interest in the development of Bayesian methods to make inferences. The reason is that Bayesian methodology provides a complete paradigm for statistical inference under uncertainty that allows to combine information derived from observations with information elicited from experts (see Berger (1985), Bernardo and Smith (1994) and Robert (1994) for an introduction to Bayesian Statistics). A great advantage of Bayesian approaches is that initial information can be used. This initial information is incorporated into the model through the prior distributions (see, for instance, O'Hagan et al. (2006) for eliciting prior distributions). This can be very useful in specific situations where expert knowledge or historical information can be obtained.

The Bayesian paradigm has long been recognized as conceptually appealing, however, its implementation in practice is far for simple due to computational difficulties. Essentially, high dimensional integration is required to calculate the posterior distribution. Bayesian methods have become popular with the advent of new computational algorithms that tackle this integration in a direct way. Although the posterior distribution is usually difficult (or most time impossible) to explicitly calculate, the development of Markov chain Monte Carlo (MCMC) methods has allowed to provide numerical solutions for problems based on truly complex models (see, e.g., Gilks et al. (1996)). Sometimes auxiliary or latent random variables are included in the model to make the generating process easier, in spite of the fact that the dimensionality is increased (see, e.g., Tanner and Wong (1987) and Tan et al. (2010)). MCMC methods have been decisive in the growing of the Bayesian literature in general (see, e.g., Chen et al. (2000b) and Gamerman and Lopes (2006)) and, in particular, for the development of new methods for categorical data analysis (see, e.g., Johnson and Albert (1999) and Congdon (2005)) and generalized linear models (see, e.g. Dey et al. (2000)).

The implementation of the algorithms plays a fundamental role for the Bayesian methodology. In order to implement the algorithms, the software should be flexible

enough at the same time as powerful for calculations. R is one of the most widely used programming languages for statistical computing and graphics (R Development Core Team (2008)). In fact, it is a language and an environment. It provides a wide variety of statistical and graphical techniques, and is highly extensible through packages. There are many books considering R as the software for Bayesian analysis. For example, Albert (2009) presented an introduction to Bayesian modelling by the use R-based computation software. Another interesting project is BUGS (Bayesian inference Using Gibbs Sampling). It is concerned with flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods. WinBUGS and its open-source version OpenBUGS are part of this project and they are widely used by the Bayesian community (Lunn et al. (2000)). Also, there is an important number of books addressing Bayesian methods with WinBUGS (see, e.g., Ntzoufras (2009) and Lunn et al. (2012)).

1.2 Objective and motivation

The general objective of this thesis is the development and implementation of robust models for categorical response data from a Bayesian viewpoint. This thesis aims at developing new approaches that remove unrealistic assumptions. It also aims at extending models that do not consider misclassification. This would allow advancing in a research line that has a significant potential.

Most models in statistics have been developed under the assumption that the errors follow normal distributions. However, many authors provide data that can not be properly fitted by using a normal distribution. This mainly happens because the shapes of the tails and the symmetry make the model not flexible enough to handle some types of data. Many robust models have been developed by using distributions more flexible than the normal one, in order to handle asymmetry and heavy/light tails. The interest to find more flexible methods to properly represent features of the data and to reduce unrealistic assumptions is increasing. This motivates the development and implementation of models that use symmetric/asymmetric exponential power distributions.

Many times, the data collection is often imperfect and it can not be exactly observed. When a categorical variable is subject to measurement error, misclassifications occur. This is very usual, for example, in election surveys (voters are reluctant to provide their true opinion), in medical diagnostics (test failures) or in consumer surveys for marketing research (consumers may not remember their behavior or they misunderstand survey questions). Misclassification produces an effective loss of information and distortions the reality. When misclassified data are obtained, the models that do not consider misclassification produce estimation errors, both in the estimated value and in its accuracy. In these contexts, additional parameters are necessary to correct the bias yielded by the use of misclassified data. This motivates extensions of the error-free models to more appropriate models.

Finally, there is a great amount of problems in different knowledge areas where new approaches could be applied. Also, there is a lack of approaches to address the previously discussed topics from a Bayesian perspective. This justifies the research

made in this thesis. The great applicability of these approaches in real situations is an added value of this research.

1.3 Manuscript structure

This document is divided into four parts. Part I contains two chapters. Chapter 1 provides an introduction to the Bayesian methodology, the objective of the thesis and the manuscript structure. Chapter 2 presents a general overview of the state-of-the art related to the topics addressed in the following chapters. Specifically, the asymmetric exponential power family, categorical data models, and misclassification are considered.

Parts II and III correspond to the main contributions. They consist of either published or submitted articles. Part II contains two chapters. Chapter 3 proposes a flexible Bayesian approach to a binary regression model that uses the inverse of the exponential power cumulative distribution function as the link function. This approach contains the probit model and an approximation to the logit one as particular cases. Chapter 4 exploits the idea of using a scale mixture of uniform representation of the asymmetric exponential power distribution to derive efficient Gibbs sampling algorithms in three different Bayesian contexts: posterior exploration, linear regression models, and binary regression models. All three models share some common characteristics in spite of the fact that they have been built for different purposes.

Part III contains three chapters. Chapter 5 considers generalized linear models to describe the dependence of data on explanatory variables when the binary outcome is subject to misclassification. Both probit and t-link regressions are proposed. In Chapter 6, the previous approach is extended to misclassified polychotomous response data, with special emphasis in the ordinal case. Motivated by a longitudinal oral health study, a different type of ordinal model considering misclassification is developed in Chapter 7. This model has been designed to exploit the multilevel structure of the data.

Part IV contains one chapter, providing the main conclusions derived from this research and some possible future research issues.

Finally, references that have been cited through the manuscript are presented at the end.

Chapter 2

Background

This chapter presents a summary of the state-of-the art related to the topics addressed in the following chapters. Specifically, the asymmetric exponential power family, categorical data models, and misclassification are considered.

2.1 The asymmetric exponential power family

An introduction to symmetric and asymmetric exponential power distributions is presented.

Robust models

In the statistical literature, most of the research has been developed under the assumption that the errors follow normal distributions. However, many authors provide data that can not be properly fitted by using a normal distribution. This mainly happens because the shapes of the tails and the symmetry make the model not flexible enough to handle some types of data. Box and Tiao (1962) commented that, in many problems, the particular physical set-up is such that the errors involved might behave like a linear aggregate of component errors and, consequently, a central limit effect would operate. Of course, the central limit theorem does not imply that a linear aggregate of a finite number of component errors would be exactly normal. For example, Blattberg and Gonedes (1974) considered a family of symmetric distributions that can also account for heavy tail distributions. Several robust models have been defined, see, for instance, Huber (1981) and Tiku et al. (1986). In addition, statistical inference based on the normal distribution is known to be vulnerable to outliers, what has increased the need to develop procedures directed at detecting outliers. Lange et al. (1989) illustrated the ability of models based on the t distribution to handle outliers in a wide range of settings.

In the recent statistical literature, some multivariate models have been considered to handle asymmetry, see, for instance, Azzalini and Dalla-Valle (1996) and Arnold and Beaver (2000). Arellano-Valle and Genton (2005) discussed about the increasing

interest in finding more flexible methods to represent features of the data as adequately as possible and to reduce unrealistic assumptions. They highlight the book of Genton (2004), which provided a collection of applications in several knowledge areas where skew distributions are considered.

In Part II of this thesis, some models have been defined by using symmetric and asymmetric exponential power distributions in order to provide robustness and flexibility to the analyzed data.

Exponential power distribution

The exponential power (EP) distribution with mean μ, scale parameter σ and shape parameter (or power parameter) $\theta \in (0, 2]$ is denoted by $Y \sim \text{EP}(\mu, \sigma, \theta)$. Its probability density function (pdf) is given by

$$f_{EP}(y|\mu, \sigma, \theta) = \frac{1}{2^{\theta/2+1}\Gamma(1+\theta/2)\sigma} \exp\left\{ -\frac{1}{2}\left|\frac{y-\mu}{\sigma}\right|^{2/\theta}\right\}.$$

When $\theta = 1$, the normal distribution is recovered. For $\theta < 1$, the distributions are platykurtic. For $\theta > 1$, the distributions are leptokurtic. When $\theta = 2$, the Laplace or double exponential distribution is recovered, and if $\theta \to 0$, then it approaches to the uniform distribution. Subbotin (1923), Vianelli (1963), Box and Tiao (1973) and Gómez et al. (1998) considered different reparameterizations for the same family of distributions.

The main reason why the EP distribution has not been used as often as desired is purely computational, i.e., because most standard statistical software did not contain procedures based on this distribution. This problem is overcome by using the mixture representation proposed by Walker and Gutiérrez-Peña (1999), which makes this family of distributions more tractable for computational purposes. In Chapter 3 this mixture representation of the EP distribution is used in a binary regression context.

Asymmetric exponential power distributions

The distributions of the EP family and its reparameterizations are symmetrical. Considering asymmetry can be useful when fitting experimental data. Different skewed exponential power (SEP) or asymmetric exponential power (AEP) distributions have been defined by Azzalini (1986), Fernández et al. (1995), Arellano-Valle et al. (2005), Zhu and Zinde-Walsh (2009), and Bottazzi and Secchi (2011), among others. These families include the EP distribution as a particular case and can provide symmetric or asymmetric distributions with lighter or heavier tails. The SEP and AEP families of distributions have been analyzed from a frequentist viewpoint, however, up to the authors' knowledge, they have not been considered from a Bayesian viewpoint. Possibly, the main reason has been the intractability of the posterior distribution. In this thesis an AEP distribution has been considered to develop approaches addressed with Bayesian methodology.

In Chapter 4 three Bayesian approaches that use the rescaled AEP distribution proposed by Zhu and Zinde-Walsh (2009) are proposed. The rescaled AEP distribution has location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, skewness parameter

$\alpha \in (0,1)$, and left and right tail parameters (shape parameters or power parameters controlling the kurtosis) $\theta_1 > 0$ and $\theta_2 > 0$. When $Y \sim \text{AEP}(\mu, \sigma, \alpha, \theta_1, \theta_2)$, the pdf is given by

$$f_{AEP}(y|\mu,\sigma,\alpha,\theta_1,\theta_2) = \begin{cases} \frac{1}{\sigma}\exp\left(-\left|\frac{y-\mu}{\alpha\sigma/\Gamma(1+1/\theta_1)}\right|^{\theta_1}\right) & \text{if } y \le \mu, \\ \frac{1}{\sigma}\exp\left(-\left|\frac{y-\mu}{(1-\alpha)\sigma/\Gamma(1+1/\theta_2)}\right|^{\theta_2}\right) & \text{if } y > \mu. \end{cases}$$

Gibbs sampling algorithms in three different Bayesian contexts (posterior exploration, linear regression and binary regression) have been obtained by using a scale mixture of uniform representation of the AEP distribution. The models have been built in such a way that they share some full conditional distributions to sample from their respective posterior distributions.

2.2 Categorical data

An introduction to generalized linear models when the response is categorical is now presented.

Generalized linear models

A Generalized linear model (GLM) is an extension of a linear model, encompasses non-normal response distributions and models functions of the mean (see McCullagh and Nelder (1989) and Dey et al. (2000)). The GLMs are specified by three components: the random component, the systematic component and the link function. The random component identifies the response variable Y and its probability distribution. The components of $\mathbf{Y}$, given by $Y_1, \ldots, Y_n$, are independent with distribution given by $\mathcal{F}$, where $\mathcal{F}$ may come from an exponential family and $\text{E}(Y) = \mu$. The systematic component specifies explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_k$ used in a linear predictor function given by $\boldsymbol{\eta} = \mathbf{x}\boldsymbol{\beta}$, where $\mathbf{x}$ denotes the $n \times k$ design matrix and $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients. The link function connects the random and systematic components. The link function $g(\cdot)$, where $\eta = g(\mu)$, may be any monotonic differentiable function.

Binary data

For binary data, where the response variable takes values in $\{0, 1\}$, $\text{E}(Y) = \pi$ and $\text{p}(Y = 1) = \pi$. The link function relates the response probability π and the covariate vector $\mathbf{x}$, so it satisfies the condition that it maps the interval $(0, 1)$ on to the whole real line $(-\infty, \infty)$, i.e. $g(\pi) = \mathbf{x}^T\boldsymbol{\beta}$ (see Cox (1971), Collett (1991), and Agresti (2002)).

A wide choice of link functions $g(\cdot)$ is available. The most common ones are the logit (canonical link), probit and the complementary log-log. The logit link is $g(\pi) = \log[\pi/(1-\pi)]$, that is, the inverse of the cumulative distribution function (cdf) of a logistic distribution, $\pi = g^{-1}(\eta) = \exp(\eta)/[1 + \exp(\eta)]$. The probit link

is the inverse of the cdf of a normal distribution, $g(\pi) = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution. The complementary log-log link is given by $g(\pi) = \log[-\log(1-\pi)]$. The logit and probit links are symmetrical in the sense that $g(\pi) = -g(1-\pi)$, i.e. $g^{-1}(\eta)$ has a symmetrical form around 0.5.

A seminal paper in this topic has been written by Albert and Chib (1993). They used the idea of data augmentation (see Tanner and Wong (1987)) to develop methods for modeling binary and polychotomous response data from a Bayesian perspective. The probit regression model for binary outcomes proposed by Albert and Chib (1993) considers an underlying normal regression structure on latent continuous data. Values of the latent data can be simulated from suitable truncated normal distributions. Then, if the latent data are known, the posterior distribution of the parameters can be computed using standard results for normal linear models, and finally, draws from this posterior distribution are used to sample new latent data. This process is iterated by using Gibbs sampling (see Gelfand and Smith (1990) and Gilks et al. (1996)). Specifically, Albert and Chib (1993) introduced n latent variables $Z_1, \ldots, Z_n$, where the Z_i are independent and distributed as $\mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, 1)$. They defined

$$Y_i = \begin{cases} 1 & \text{if} \quad Z_i > 0 \\ 0 & \text{if} \quad Z_i \leq 0 \end{cases}.$$

By this way, Y_i are independent Bernoulli random variables with $\pi = \mathrm{p}(Y_i = 1)$.

The data augmentation of Albert and Chib (1993), by introducing the Z_i's into the probit regression model, provides a general framework for analyzing binary regression models. This approach has been extended to allow other link functions. Albert and Chib (1993) also generalized the probit link to belong to the t-link family by defining Z_i as independently distributed from Student's t distributions with location parameter $\mathbf{x}_i^T\boldsymbol{\beta}$, scale parameter 1, and degrees of freedom ν. Haro-López et al. (2000) proposed to choose an arbitrary link function from a set of different inverse platykurtic cdfs. They defined their model as $Z_i \sim \mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, h(\lambda_i))$, and $\lambda_i \sim \Pi(\lambda_i|\alpha)$, where $\Pi(\lambda_i|\alpha)$ is the scale mixture parameter distribution for $\lambda_i \in \Re^+$, $\alpha \in (a, b)$ is the shape parameter, and $h(\cdot)$ is a positive function. These links are symmetrical. However, they do not always provide the best fit for a given dataset, and the overall fit could be significantly improved using asymmetric links.

In order to describe a link, Chen et al. (1999a) considered the rates at which the probabilities of a given binary response approach 1 or 0. Under this notion, a link is symmetric if the rates are the same, otherwise the link is skewed or asymmetric. A skewed link can be characterized as positively skewed if the rate approaching 1 is faster than the rate approaching 0, otherwise it is negatively skewed. Chen et al. (1999a) suggested that an asymmetric link-based model may be more appropriate than a symmetric link-based one when the number of 1's and 0's are much different. Czado and Santner (1992) showed that a misspecification of the link function can result in a substantial increase in bias and mean square error of the success probability and regression parameter estimates.

Some asymmetric link functions have been defined by using the data augmentation scheme described by Albert and Chib (1993). Chen et al. (1999a) proposed a class of skewed link model, where the underlying latent variable has a mixed-effect model

structure. They defined the skewed link model by $Z_i = \mathbf{x}_i^T\boldsymbol{\beta} + \delta W_i + \epsilon_i$, $W_i \sim \mathcal{G}$ and $\epsilon_i \sim \mathcal{F}$, where Z_i and ϵ_i are independent, $\mathcal{G}$ is the cdf of a skewed distribution, $\mathcal{F}$ is the cdf of a symmetric distribution, $\delta \in (-\infty, \infty)$ is a skewness parameter, and $\mathcal{G}$ and $\mathcal{F}$ are known to ensure the identificability of model parameters. However, the model has the limitation that the intercept term is confounded with the skewness parameters. Later, Kim et al. (2008) introduced a link based on a generalized t distribution, which overcomes the problem of Chen et al. (1999a). They proposed the skewed generalized t link model by $Z_i = \mathbf{x}_i^T\boldsymbol{\beta} + \delta[W_i - \mathrm{E}(W)] + \epsilon_i$, and $\epsilon_i \sim p_{gt,\nu_1,\nu_2=1}$, where $0 < \delta \leq 1$, $\nu_1 > 1$ and p_{gt,ν_1,ν_2} is the probability density function (pdf) of the generalized t distribution introduced by Arellano-Valle and Bolfarine (1995). Other link functions have been defined by Basu and Mukhopadhyay (2000b) and Bazán et al. (2010), among others.

In this thesis the data augmentation scheme introduced by Albert and Chib (1993) has been adapted to derive two binary regression approaches based on symmetrical and asymmetrical links. Specifically, the inverse of the EP cdf and the inverse of the AEP cdf have been used as the link functions. In Chapter 3 the EP-link model is defined by $Z_i \sim \mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \theta)$, where θ is the shape parameter that allows to model platykurtic or leptokurtic shapes. In Chapter 4 the AEP-link model is defined by $Z_i \sim \mathrm{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, 0.5, \theta_1, \theta_2)$ (the rescaled AEP distribution proposed by Zhu and Zinde-Walsh (2009)), where θ_1 and θ_2 are the shape parameters that allow to model symmetry/asymmetry and light/heavy tails. Even more, the distributions can fit each tail separately leading to a flexible approach.

Ordinal data

For ordinal data, the response variable Y takes one of J categories (see Johnson and Albert (1999)), and thus $Y \sim \mathrm{Categorical}(\boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is the vector of model probabilities. Since the outcome is ranked, usually the ordinal regression model is defined by considering cumulative probabilities conditioned on the covariates $\mathbf{x}$, i.e. $\theta_j(\mathbf{x}) = \mathrm{p}(Y \leq j|\mathbf{x}) = \pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})$, and the link function is given by $g(\theta_j) = \gamma_j + \mathbf{x}\boldsymbol{\beta}$, for $j = 1, \ldots, J$, where $\gamma_1, \ldots, \gamma_J$ are the cutpoints. Also in this case, the most common link functions are logit, probit and complementary log-log.

Albert and Chib (1993) also used a data augmentation framework to model ordered categories. The model links the cumulative response probabilities with the linear regression structure. They assumed that there exist latent continuous random variables Z_i distributed $\mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, 1)$, and $Y_i = j$ if $\gamma_{j-1} < Z_i < \gamma_j$ for $j = 1, \ldots, J$, where $\gamma_0 = -\infty$ and $\gamma_J = \infty$.

The binary regression models presented in Chapters 3 and 4 can be extended to model ordered categories by adapting the data augmentation framework proposed by Albert and Chib (1993). Specifically, latent continuous random variables $Z_1, \ldots, Z_n$ are introduced having either EP or AEP distribution with location $\mathbf{x}_i^T\boldsymbol{\beta}$, i.e. $\mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \theta)$ or $\mathrm{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \alpha, \theta_1, \theta_2)$, and defining $Y_i = j$ if $\gamma_{j-1} < Z_i < \gamma_j$ for $j = 1, \ldots, J$. Gibbs sampling algorithms can be obtained in a similar way as those in Chapters 3 and 4.

2.3 Misclassification

In this section, an introduction to misclassification is shown.

Measurement error

Measurement error occurs whenever one or more of the variables from a model of interest can not be exactly observed. When the true and observed values are both categorical, then measurement error is more specifically referred to as misclassification. There are many reasons why such errors occur. Two of the most common errors are instrument and sampling errors. However, there are many other reasons why misclassification may occur, for example, in election surveys, voters are reluctant to provide their true opinion. The impact of measurement error results in a weaker estimation between the explanatory and the outcome variables. When the variable of interest can not be properly measured, and the data are analyzed by pretending that the surrogate is actually the variable of interest, biased inferences can be obtained.

An analysis which uses the mismeasured variable as the variable of interest is referred to as naive. One can glean a sense of how well naive analysis performs with some simple computer-simulated examples. Gustafson (2003) and Buonaccorsi (2010) have studied models with measurement errors. Gustafson (2003) provides Bayesian approaches for dealing with measurement error and misclassification in numerous settings. Buonaccorsi (2010) describes the impacts of measurement error on naive analyses that ignore them and presents ways to correct for them across a variety of statistical models.

Buonaccorsi (2010) pointed out that there are three main ingredients in a measurement error problem: a model for the true value (which can be essentially any statistical model), a measurement error model (which involves specification of the relationship between the true and observed values), and extra data, information or assumptions that may be needed to correct for measurement error. This extra information can be typically the following: knowledge about some of the measurement error parameters or their functions, replicate values, estimated standard errors attached to the error prone variables, validation (internal or external) in which both true and mismeasured values are obtained on a set of units, or instrumental variables. Moreover, sometimes measurement error may depend on other variables, which themselves may or may not be measured with error. The measurement error model is said to be differential when the measurement error does depend on other variables, otherwise it is nondifferential. Additionally, two general objectives in a measurement error problem are related to the consequences for naive analyses which ignore the measurement error, and the way to correct for measurement error.

Misclassified categorical data

The effects of ignoring misclassification were first noted by Bross (1954), who showed that classical estimators base sampling on a dichotomous process under the assumption of known noise parameters. These parameters are needed to correct the bias resulting from estimation based on the observed proportion. Pérez et al. (2007)

presented a review of the existing literature about the problem of inference with misclassified multinomial data. In the last years the effect of misclassification has been analyzed with generalized linear models, see, for instances, Albert et al. (1997), Gustafson (2003), Mwalili et al. (2005) and Roy and Banerjee (2009), among others.

This thesis analyzes the misleading effect of misclassification. It also shows how to make statistical inferences that reflect or adjust for misclassification at play. Assume that the true classification status for an observation is denoted by Y^{true}. Let Y be the outcome subject to error. The classification error model, which specifies the behavior of Y given Y^{true}, is specified by $\lambda_{y|y^{true}} = \mathrm{p}(Y = y|Y^{true} = y^{true})$. The probabilities $\lambda_{y|y^{true}}$ are referred to as misclassification probabilities or misclassification rates. Measurement errors are characterized in terms of misclassification probabilities, i.e. given the true classification, how likely a correct classification is. Bayesian inference can provide separate information for each group of parameters (regression parameters and misclassification probabilities) through their respective posterior distributions. Besides, information on misclassification can be included into the model through the prior distribution of the misclassification parameters. For example, Paulino et al. (2003), McInturff et al. (2004) and McGlothlin et al. (2008) use Bayesian methods and misclassification probabilities to characterize measurement error.

In Part III of this thesis, some models have been addressed to describe the dependence of data on explanatory variables when the categorical outcome is subject to misclassification. In Chapter 5 both probit and t-link regression models for misclassified binary data are presented. The idea of using a data augmentation framework is exploited to derive efficient Gibbs sampling and Expectation-Maximization algorithms. This data augmentation scheme allows to model binomial data (grouped data) in an easy way. In Chapter 6 both probit and logit ordinal regression models incorporating misclassification are proposed. The model considering unordered categories is also presented. This approach generalizes the binary regression model addressing misclassification presented in Chapter 5. Moreover, the models presented in Chapter 5 can be extended by introducing link functions based on asymmetric distributions. Finally, a different type of ordinal model considering misclassification is developed in Chapter 7. This model has been designed to exploit the multilevel structure of the data in hand.

Part II

EP/AEP distribution-based approaches

Chapter 3

Bayesian binary regression with exponential power link

Naranjo, L., Martín, J., and Pérez, C. J. (2014). Bayesian binary regression with exponential power link. *Computational Statistics and Data Analysis*, **71**:464-476.

Abstract

A flexible Bayesian approach to a generalized linear model is proposed to describe the dependence of binary data on explanatory variables. The inverse of the exponential power cumulative distribution function is used as the link to the binary regression model. The exponential power family provides distributions with both lighter and heavier tails compared to the normal distribution, and includes the normal and an approximation to the logistic distribution as particular cases. The idea of using a data augmentation framework and a mixture representation of the exponential power distribution is exploited to derive efficient Gibbs sampling algorithms for both informative and noninformative settings. Some examples are given to illustrate the performance of the proposed approach when compared with other competing models.

Keywords: Bayesian methods; Binary data; Data augmentation; Exponential power distribution; Generalized linear models; Gibbs sampling.

3.1 Introduction

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed, where Y_i is Bernoulli distributed with success probability $\mathrm{p}(Y_i = 1|\boldsymbol{\beta}, \mathbf{x}_i) = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$. $\boldsymbol{\beta}$ is a k vector of unknown parameters, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known covariates, and Ψ is a known nonnegative function ranging between 0 and 1. The standard approach to modeling the dependence of binary data on explanatory variables under the generalized linear model setting is performed through a cumulative density function (cdf) Ψ. For instance, the probit model is obtained when Ψ is the standard normal cdf and the logit model when Ψ is the logistic cdf. See, for example, Cox (1971), McCullagh and Nelder (1989), and Collett (1991).

In Bayesian learning, the approach starts with a prior probability distribution for the unknown model parameters. Specifically, one denotes by $\pi(\boldsymbol{\beta})$ the (proper or improper) prior density function for the unknown parameter vector $\boldsymbol{\beta}$. Then the posterior density of $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Psi(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1 - \Psi(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i}}{\int \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Psi(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1 - \Psi(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i} d\boldsymbol{\beta}},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the vector of binary data and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the matrix of explanatory variables. In general, this distribution is largely intractable.

Bayesian approaches to binary regression models have been extensively developed in the literature. Zellner and Rossi (1984) examined the generalized linear models considering link functions such as the probit or logit. They commented on the inaccuracy of the normal approximation for small samples. For a small number of parameters (k small), they summarized the posterior distribution by using numerical integration. For large models (k large), they computed posterior moments by Monte Carlo integration (importance sampling with a multivariate Student's t as importance function).

With the upsurge of MCMC methods in the nineties (Gelfand and Smith, 1990), the new simulation tools allowed one to efficiently apply Bayesian regression methods. For example, Dellaportas and Smith (1993) used the Gibbs sampling procedure to compute the posterior distribution of the regression parameters for log-concave densities by using the adaptive rejection algorithm. A key paper on this topic was written by Albert and Chib (1993). It proposed a data augmentation framework for the Bayesian probit model by using Gibbs sampling. It also proposed extensions of the probit link by using mixtures of normal or Student's t distributions. The logit regression is studied as a particular case. Holmes and Held (2006) highlight a technique to improve performance in probit regression simulation by jointly updating the regression coefficients and the auxiliary variables. They show that the proposed approach is also possible for logistic regression by using a scale mixture of normal distributions for the noise process. Meza et al. (2009) proposed a stochastic approximation of the EM algorithm to analyze binary data under a two stage probit normal model with random effects. Haro-López et al. (2000) proposed a binary regression model that chooses an arbitrary link function from a set of different inverse platykurtic cdfs. The Bayesian analysis (focused on robustness) is implemented using cdfs of scale mixtures of normal distributions. Another remarkable result is that of Eyheramendy and Madigan

(2007). They presented a class of sparse generalized linear models that include probit and logit regressions as special cases and offer some extra flexibility. They provided an EM algorithm for learning the regression parameter. Nott and Leng (2010) proposed a Bayesian approach to variable selection in generalized linear models. This approach has been implemented and validated for a logistic binary regression model.

A different kind of approach to modeling binary data is based on nonparametric models. Diaconis and Freedman (1993) studied a nonparametric Bayesian binary regression model from a theoretical viewpoint. They reviewed the relationship between consistency of Bayes estimates and rules for model selection, as well as sieves and orthogonal series estimation. Newton et al. (1996) proposed a semiparametric regression model for binary response data that places no structural restrictions on the link function other than monotonicity and known location and scale. By modifying the Dirichlet process they obtained a prior measure over the semiparametric model, and used Polya sequence theory to formulate the proposed measure in terms of a finite number of unobservables variables. When there is only one predictor, the method in Newton et al. (1996) is fully nonparametric. Qian et al. (2000) introduced a nonparametric Bayesian binary regression model with a single predictor variable that is usually more flexible than the commonly used logistic or probit models. The model presented by Qian et al. (2000) is a special case of the semiparametric binary regression model of Newton et al. (1996), in which a simpler computing algorithm is provided. However, the extension to more than one predictor variable is not obvious. Wood and Kohn (1998) proposed a Bayesian approach to binary nonparametric regression which assumes that the argument of the link is an additive function of the explanatory variables, and uses splines for the calculations. Although the nonparametric binary regression models offer robustness because of the flexibility of the link, the methods (except in special cases) are very difficult to implement and the results are not interpretable in terms of parameters.

A Bayesian approach to a binary regression model is proposed here. The inverse of the exponential power cdf is used as the link function. The exponential power (EP) family (Box and Tiao, 1973) includes the normal distribution and incorporates additional shapes, including platykurtic and leptokurtic ones. This means that distributions with both lighter and heavier tails compared to the normal case can be achieved, which is always an advantage when analyzing robustness. These distributions allow the modeling of kurtosis, providing, in general, more flexible fits to experimental data than the normal distribution. The proposed approach contains, among others, the probit model and an approximation to the logistic model as special cases.

The implementation of the approach is based on two main ideas. The first is to develop a data augmentation framework by introducing latent variables in a similar way to Albert and Chib (1993). The second is to use the mixture representation for the EP distribution suggested by Walker and Gutiérrez-Peña (1999). These two ideas are exploited to derive efficient Gibbs sampling algorithms for both informative and noninformative settings. All the full conditional distributions can be easily generated. The applicability of the approach is illustrated through some examples that show its good performance when compared with other competing models.

The outline of the paper is as follows. In section 3.2, the EP distribution and some properties are presented. Section 3.3 presents the proposed binary regression model, including the derivation of the full conditional distributions that are necessary to apply Gibbs sampling. Section 3.4 illustrates the performance of the proposed approach through some examples. Finally, section 3.5 presents the conclusions.

3.2 Exponential power distribution

The EP family (see, for example, Box and Tiao (1973), and Gómez et al. (1998)) includes the normal distribution and incorporates additional shapes, including platykurtic (lighter tails compared to the normal) and leptokurtic ones (heavier tails compared to the normal). These distributions allow the modeling of kurtosis, providing, in general, more flexible fits to experimental data than the normal distribution.

Let Z be a random variable distributed as EP with location parameter μ, scale parameter σ, and shape parameter θ with density given by

$$f(z|\mu,\sigma,\theta) = w(\theta)\sigma^{-1}\exp\left\{-c(\theta)\left|\frac{z-\mu}{\sigma}\right|^{2/\theta}\right\},$$

where $-\infty < z < +\infty$, $-\infty < \mu < +\infty$, $\sigma > 0$, $0 < \theta \leq 2$,

$$w(\theta) = \frac{\{\Gamma(\frac{3}{2}\theta)\}^{1/2}}{\theta\{\Gamma(\frac{1}{2}\theta)\}^{3/2}} \text{ and } c(\theta) = \left\{\frac{\Gamma(\frac{3}{2}\theta)}{\Gamma(\frac{1}{2}\theta)}\right\}^{1/\theta}.$$

The parameters μ and σ are the mean and the standard deviation, respectively. The parameter θ can be regarded as a measure of kurtosis. When $\theta = 1$, the normal distribution is recovered, for $\theta < 1$ the distributions are platykurtic, for $\theta > 1$ they are leptokurtic. When $\theta = 2$, the Laplace or double exponential distribution is recovered.

In order to simplify the presentation without loss of generality, consider the following representation of the density

$$f(z|\mu,\sigma,\theta) = \frac{1}{2^{\theta/2+1}\Gamma(1+\theta/2)\sigma}\exp\left\{-\frac{1}{2}\left|\frac{z-\mu}{\sigma}\right|^{2/\theta}\right\}.$$

Now, $\mathrm{E}(Z) = \mu$,

$$\mathrm{Var}(Z) = \sigma^2 2^{\theta}\frac{\Gamma(\frac{3}{2}\theta)}{\Gamma(\frac{1}{2}\theta)}, \quad \text{and} \quad \mathrm{E}[(Z-\mu)^s] = \begin{cases} 0 & \text{if } s \text{ is odd} \\ \sigma^s 2^{s\theta/2}\frac{\Gamma\left(\frac{1+s}{2}\theta\right)}{\Gamma\left(\frac{1}{2}\theta\right)} & \text{if } s \text{ is even} \end{cases}.$$

Figure 3.1 shows the pdfs and the cdfs, respectively, for some parameter values.

If $\theta = 1$, then the EP distribution, $\mathrm{EP}(\mu,\sigma,1)$, is the normal distribution, $\mathrm{N}(\mu,\sigma^2)$. The logistic distribution can be approximated by using the EP distribution. The standard logistic distribution $\mathrm{L}(0,1)$ has variance $\pi^2/3$ and kurtosis 1.2. Then the EP parameters σ and θ are obtained by minimizing the maximum difference between the

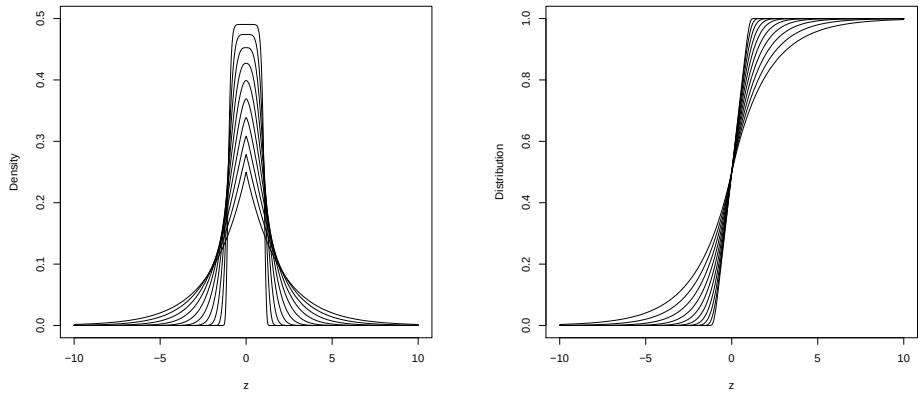

Figure 3.1: Pdf and cdf for EP$(0, 1, \theta)$, with $\theta = 0.2\,i,\ i = 1, 2, \ldots, 10$.

following cdfs: the standardized logistic (L$(0, \sqrt{3}/\pi)$) and EP distributions. Then, the EP distribution EP$(0, 0.7597641, 1.277734)$ is approximately the standardized logistic distribution L$(0, \sqrt{3}/\pi)$, with variance 0.9628842 and kurtosis 0.6208674. Figure 3.2 shows this approximation and some Student $t(\nu)$ distribution approximations, with $\nu = 7$, 8, and 9 used by Liu (2004), Albert and Chib (1993), and Mudholkar and George (1978), respectively.

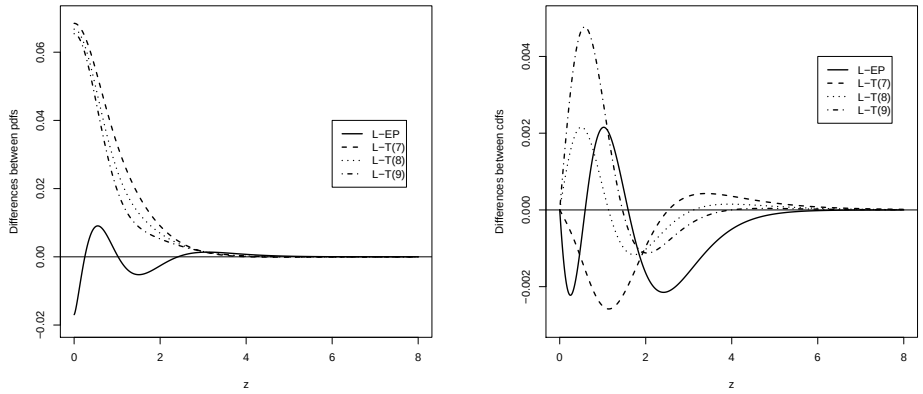

Figure 3.2: Differences between the logistic and the EP and the Student-t densities for $z > 0$, denoted by $L(z; 0, \sqrt{3}/\pi) - EP(z; 0, 0.7597641, 1.277734)$ and $L(z; 0, \sqrt{3}/\pi) - T(z; \nu, \sqrt{(\nu-2)/\nu})$, $\nu = 7, 8, 9$.

Walker and Gutiérrez-Peña (1999) suggested the following mixture representation for the EP distribution. If $Z|U = u \sim \mathrm{U}(\mu - \sigma u^{\theta/2}, \mu + \sigma u^{\theta/2})$ and $U \sim \mathrm{Ga}(1 + \theta/2, 1/2)$, then Z is distributed as EP(μ, σ, θ). This representation will allow us to derive an efficient simulation framework for a binary regression model in the next

section.

3.3 Binary regression model

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed together with predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where Y_i is Bernoulli distributed with success probability $\mathrm{p}(Y_i = 1|\boldsymbol{\beta}, \mathbf{x}_i) = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$. The EP cdf is considered to be Ψ.

The posterior distribution of $\boldsymbol{\beta}$ is computed by developing a Gibbs sampling algorithm that uses the idea of data augmentation proposed by Albert and Chib (1993) and the mixture representation of Walker and Gutiérrez-Peña (1999). Independent latent variables $Z_1, \ldots, Z_n$ are introduced, where Z_i is distributed as $\mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma, \theta)$ and one defines $Y_i = 1$ if $Z_i > 0$, and $Y_i = 0$ if $Z_i \leq 0$. By using random variables U_i, the distribution of Z_i is expressed as the following mixture

$$\begin{aligned} Z_i|\boldsymbol{\beta}, \sigma, \theta, u_i &\sim \mathrm{U}\left(\mathbf{x}_i^T\boldsymbol{\beta} - \sigma u_i^{\theta/2}, \mathbf{x}_i^T\boldsymbol{\beta} + \sigma u_i^{\theta/2}\right), \\ U_i|\theta &\sim \mathrm{Ga}\left(1 + \theta/2, 1/2\right). \end{aligned}$$

Note that if $Z_i \sim \mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma, \theta)$ where σ is unknown, the following equality holds

$$\begin{aligned} \mathrm{p}(Y_i = 1) &= \mathrm{p}(Z_i > 0) = \mathrm{p}\left(\frac{Z_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma} > -\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right) \\ &= \mathrm{p}\left(Z_i^* - \mathbf{x}_i^T\boldsymbol{\beta}^* > -\mathbf{x}_i^T\boldsymbol{\beta}^*\right) = \mathrm{p}\left(Z_i^* > 0\right), \end{aligned}$$

where $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sigma$, $Z_i^* = Z_i/\sigma$ and $Z_i^* \sim \mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}^*, 1, \theta)$. Without loss of generality, in fact, for identifiability, σ will be considered equal to 1. Hence $Z_i \sim \mathrm{EP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \theta)$.

The following step is to define the prior distributions. A usual approach for the binary regression models assumes a multivariate normal distribution for the regression parameters in the informative case and an improper uniform distribution in the noninformative one (see, e.g. Zellner and Rossi (1984)). Many ways to elicit the parameters have been proposed. One of them is the empirical Bayesian approach. This proposes obtaining the mean vector and covariate matrix for the multivariate normal distribution by using historical data or a randomly selected small portion of the current data (see, e.g., Carlin and Louis (1996)). Another form of prior distribution for the regression parameters is based on the idea of Bedrick et al. (1996). They proposed a method to induce a prior probability distribution on the regression vector by using the so-called conditional means prior (CMP). Applications of this method to real data can be found in Bedrick et al. (1997) and Paulino et al. (2003). This method is used for the EP regression.

3.3.1 Multivariate normal and improper prior distributions

In this subsection, the prior distribution $\boldsymbol{\beta} \sim \mathcal{F}$ is considered, where $\mathcal{F}$ is the multivariate normal distribution family $\mathrm{N}_k(\mathbf{b}, \mathbf{B})$. Also, an improper prior distribution $\pi(\boldsymbol{\beta}) \propto 1$ is considered. The prior distribution for the parameter θ is $\mathrm{U}(0, 2]$. This parameter is allowed to vary over its range, providing flexibility to fit different shapes.

Some scenarios can be considered: (1) when no information is obtained for θ, the prior distribution of this parameter should be $\mathrm{U}(0,2]$, (2) if the interest is only focused on platykurtic links, then $\theta \sim \mathrm{U}(0,1)$, and (3) when the interest is only focused on leptokurtic links, then $\theta \sim \mathrm{U}(1,2]$.

The joint posterior density of the unobservable variables, $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{U} = (U_1, \ldots, U_n)$, and unknown parameters, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ and θ, given the data $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\begin{aligned}
\mathrm{p}(\mathbf{z}, \mathbf{u}, \boldsymbol{\beta}, \theta | \mathbf{y}) \quad \propto \quad & \prod_{i=1}^{n} \frac{1}{2u_i^{\theta/2}} I\left[z_i \in \left(\mathbf{x}_i^T\boldsymbol{\beta} - u_i^{\theta/2}, \mathbf{x}_i^T\boldsymbol{\beta} + u_i^{\theta/2}\right)\right] \\
\times \quad & \prod_{i=1}^{n} \left\{I\left[z_i > 0\right] I\left[y_i = 1\right] + I\left[z_i \leq 0\right] I\left[y_i = 0\right]\right\} \\
\times \quad & \prod_{i=1}^{n} \frac{u_i^{\theta/2}}{\Gamma(1+\theta/2)2^{1+\theta/2}} e^{-u_i/2} I\left[u_i > 0\right] \\
\times \quad & \pi(\boldsymbol{\beta}) \\
\times \quad & \frac{1}{2} I\left[\theta \in (0,2]\right],
\end{aligned}$$

where $I[\cdot]$ denotes the indicator function.

Note that this joint posterior distribution is complicated in the sense that it is difficult to normalize and sample from directly. However, it is possible to generate from the posterior distribution of $\boldsymbol{\beta}$ by using a Gibbs sampling algorithm. The full conditional distributions of $\mathbf{Z}$, $\mathbf{U}$, $\boldsymbol{\beta}$ and θ are:

- The full conditional distributions of $Z_1, \ldots, Z_n$ are independent, with

$$Z_i | \mathbf{y}, \mathbf{u}, \boldsymbol{\beta}, \theta \sim \begin{cases} \mathrm{U}\left(\max\left\{0, \mathbf{x}_i^T\boldsymbol{\beta} - u_i^{\theta/2}\right\}, \mathbf{x}_i^T\boldsymbol{\beta} + u_i^{\theta/2}\right) & \text{if } y_i = 1 \\ \mathrm{U}\left(\mathbf{x}_i^T\boldsymbol{\beta} - u_i^{\theta/2}, \min\left\{0, \mathbf{x}_i^T\boldsymbol{\beta} + u_i^{\theta/2}\right\}\right) & \text{if } y_i = 0 \end{cases}. \tag{3.1}$$

- The full conditional distributions of $U_1, \ldots, U_n$ are independent, with

$$U_i | \mathbf{y}, \mathbf{z}, \boldsymbol{\beta}, \theta \sim \mathrm{Exp}\left(\tfrac{1}{2}\right) I\left[u_i > \left|z_i - \mathbf{x}_i^T\boldsymbol{\beta}\right|^{2/\theta}\right]. \tag{3.2}$$

- Since the distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution, then the conditional distribution of β_j given $\boldsymbol{\beta}_{(-j)}$ is a normal distribution, where $\boldsymbol{\beta}_{(-j)}^T = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$. The subscript $(-j)$ denotes that the jth element has been removed. Then, the posterior distribution of β_j conditioned on $\mathbf{Z}$, $\mathbf{U}$, $\boldsymbol{\beta}_{(-j)}$ and θ is given by

$$\beta_j | \mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}_{(-j)}, \theta \sim \mathrm{N}(\mathrm{b}_j^*, \mathrm{B}_j^*) I\left[\beta_j \in \left(\underline{\beta}_j, \overline{\beta}_j\right)\right], \tag{3.3}$$

for $j = 1, \ldots, k$, where

$$\begin{aligned}
\underline{\beta}_j &= \max_{\{i:x_{ij}\neq 0\}} \left\{ \frac{z_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)} - \text{sign}(x_{ij}) u_i^{\theta/2}}{x_{ij}} \right\}, \\
\overline{\beta}_j &= \min_{\{i:x_{ij}\neq 0\}} \left\{ \frac{z_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)} + \text{sign}(x_{ij}) u_i^{\theta/2}}{x_{ij}} \right\}, \\
\mathrm{b}_j^* &= \mathrm{b}_j - \mathrm{B}_{j(-j)} \mathrm{B}_{(-j)(-j)}^{-1} \left(\boldsymbol{\beta}_{(-j)} - \mathrm{b}_{(-j)} \right), \\
\mathrm{B}_j^* &= \mathrm{B}_{jj} - \mathrm{B}_{j(-j)} \mathrm{B}_{(-j)(-j)}^{-1} \mathrm{B}_{(-j)j}.
\end{aligned}$$

If $\pi(\boldsymbol{\beta}) \propto 1$ is considered, then

$$\beta_j | \mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}_{(-j)}, \theta \sim \mathrm{U}\Big(\underline{\beta}_j, \overline{\beta}_j \Big).$$

- The posterior density of θ given $\mathbf{Z}$, $\mathbf{U}$ and $\boldsymbol{\beta}$ is given by

$$\mathrm{p}(\theta | \mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}) \quad \propto \quad \frac{1}{\Gamma(1+\theta/2)^n 2^{n\theta/2}} I\left[\theta \in \left[\underline{\theta}, \overline{\theta} \right] \right], \tag{3.4}$$

where

$$\begin{aligned}
\underline{\theta} &= \max\left\{ 0, \max_{i\in\Theta^+} \left\{ \tfrac{2\log(|z_i - \mathbf{x}_i^T\boldsymbol{\beta}|)}{\log(u_i)} \right\} \right\}, \quad \Theta^+ = \{i : \log(u_i) > 0\}, \\
\overline{\theta} &= \min\left\{ 2, \min_{i\in\Theta^-} \left\{ \tfrac{2\log(|z_i - \mathbf{x}_i^T\boldsymbol{\beta}|)}{\log(u_i)} \right\} \right\}, \quad \Theta^- = \{i : \log(u_i) < 0\}.
\end{aligned}$$

The proposed order to iterate from the Gibbs algorithm is $\mathbf{Z}$, $\mathbf{U}$, $\boldsymbol{\beta}$ and θ, and the initial points are proposed to be set: $u_i = 1$ for all i, $\boldsymbol{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$, and $\theta = 1$. Note that generation from the full conditional distributions is easy. Specifically, the full conditional distributions (3.1), (3.2), and (3.3) are standard, whereas the full conditional distribution (3.4) has log-concave density for all n, and can be easily generated using the adaptive rejection sampling method given in Gilks and Wild (1992). The generated sample is usually highly correlated, so a detailed check of chain convergence must be carried out (see, e.g. Cowles and Carlin (1996)).

3.3.2 Conditional means prior

In this subsection, the interest is on building a prior distribution for $\boldsymbol{\beta}$ based on the expert prior elicitation framework proposed by Bedrick et al. (1996), and deriving an algorithm to sample from the posterior distribution. Bedrick et al. (1996) proposed a method to induce a prior probability distribution on the regression vector by using the so-called conditional means prior (CMP) on $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_k)^T$, where, in binomial regression, $\tilde{p}_l = \mathrm{E}(\tilde{y}_l | \tilde{\mathbf{x}}_l)$ is the success probability for a potentially observable response $\tilde{y}_l$ at covariate vector $\tilde{\mathbf{x}}_l$. Prior knowledge from experts or previous studies is used to

specify uncertainty about probabilities of the present condition, given various specified covariate configurations.

Assuming k regression coefficients (including the intercept), prior probabilities $\tilde{p}_l$ are elicited in the predictor space, $l = 1, \ldots, k$, for selected locations $\tilde{\mathbf{x}}_l$. The covariate vectors $\tilde{\mathbf{x}}_l$ are chosen subjected to expert opinion in the predictor variable range in order to make it reasonable to assume prior independence among the quantities $\tilde{p}_l$. With k linearly independent sets of covariate values, a 1-1 transformation between $\boldsymbol{\beta}$ and $\tilde{\mathbf{p}}$ is obtained, namely $\boldsymbol{\beta} = \tilde{\mathbf{x}}^{-1}\Psi^{-1}(\tilde{\mathbf{p}})$, where $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_1^T, \ldots, \tilde{\mathbf{x}}_k^T)^T$. Uncertainty about $\tilde{p}_l$ is modeled with independent distributions $\text{Be}(a_l, b_l)$. The hyperparameters a_l and b_l are determined (in general indirectly) from expert prior judgements.

The independence CMP

$$\pi(\tilde{\mathbf{p}}) \propto \prod_{l=1}^{k} \tilde{p}_l^{a_l - 1}(1 - \tilde{p}_l)^{b_l - 1},$$

induces a prior on $\boldsymbol{\beta}$ given by

$$\pi(\boldsymbol{\beta}) \propto \prod_{l=1}^{k} \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})^{a_l - 1}[1 - \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})]^{b_l - 1}\psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta}), \tag{3.5}$$

where Ψ is the cdf and ψ is the pdf of the $\text{EP}(\mu = 0, \sigma = 1, \theta)$ distribution. Then, the posterior distribution is

$$\pi(\boldsymbol{\beta}, \theta|\mathbf{y}, \mathbf{x}) \propto \pi(\boldsymbol{\beta})\pi(\theta)L(\boldsymbol{\beta}, \theta|\mathbf{y}, \mathbf{x}), \tag{3.6}$$

where $L(\boldsymbol{\beta}, \theta|\mathbf{y}, \mathbf{x})$ is the likelihood function.

The Metropolis-Hastings algorithm can be used for generation from (3.6). However, the proportion of acceptance is usually low and highly dependent on the proposal distribution used. As an efficient alternative, it is proposed to derive a Gibbs sampling algorithm by using the same latent variable scheme defined previously. Now, the full conditional distributions are the same as those presented in the previous subsection, except that of $\boldsymbol{\beta}$. This full conditional distribution is

$$\beta_j|\mathbf{y}, \mathbf{z}, \mathbf{u}, \boldsymbol{\beta}_{(-j)}, \theta \sim \pi(\boldsymbol{\beta})I\left[\beta_j \in \left(\underline{\beta}_j, \overline{\beta}_j\right)\right].$$

Generation from this distribution is easy because the support is bounded for each j. This fact allows an efficient implementation of the conditional means prior for this regression model.

The next section illustrates the performance of the proposed approaches by way of some examples.

3.4 Examples

An empirical study using a wide range of datasets showed us that the proposed approach is useful and has good performance when compared with competing models.

Two examples are presented in this section. Firstly, a simulated example is considered to show the performance when multiple datasets are generated. Then, a real data set (Rocker et al., 1988) is used to illustrate the applicability of the proposed approach for both noninformative and informative settings. In this example, the results are analyzed in detail.

In order to compare several competing models, the deviance information criterion (DIC) is used. This criterion was proposed by Spiegelhalter et al. (2002) and is useful to assess the performance of models with different amounts of partial information. DIC is designed for complex hierarchical models with possibly improper prior distributions. It overcomes the problem of having to identify the number of parameters in the model, which is required for the calculation of the Akaike Information Criterion (AIC) (see Akaike (1973)). The calculation of DIC is straightforward:

$$DIC = \overline{D(\eta)} + \widehat{\rho_D},$$

where $D(\eta) = -2 \log L(\eta)$ is the deviance of the model (that includes the likelihood $L(\eta)$). In the DIC approach, the *fit* of a model is summarized by the posterior mean of the deviance $\overline{D(\eta)} = \mathrm{E}(D(\eta)|\text{data})$, while the *complexity* of a model is captured by $\widehat{\rho_D} = \overline{D(\eta)} - D(\bar{\eta})$, where $D(\bar{\eta})$ is the deviance at the posterior means of the parameters of interest, $\bar{\eta} = \mathrm{E}(\eta|\text{data})$.

Spiegelhalter et al. (2002) named $\widehat{\rho_D}$ the effective number of parameters and stated that this approximately matches the number of real parameters in models when the underlying distribution is normal. They obtained negative values for $\widehat{\rho_D}$ in some particular cases when non-logconcave likelihoods were used, when there was substantial conflict between prior and data or when the posterior distribution for a parameter was extremely asymmetric, or symmetric and bimodal. They argued that a negative effective number of parameters is indicative of a possibly poor fit between the model and the data. Although, Spiegelhalter et al. (2002) mostly focused on generalized linear models, they discussed the possibility of extending the notion of DIC to models like mixtures of distributions. Zhu and Carlin (2000) applied the DIC to model selection for hierarchical models in medical applications. Celeux et al. (2006) proposed versions of DIC to deal with mixture and missing data models. They showed that the DIC and the corresponding effective number of parameters allow for a wide range of interpretations and extensions outside exponential families. McGrory and Titterington (2007) showed how the DIC can be extended to variational methods applied to the Bayesian analysis of mixtures of Gaussian distributions. Huang (2008) used the DIC to compare the effects of drug adherence with drug resistance in a Bayesian nonlinear mixed-effects model investigated for estimating dynamic parameters by fitting the model to viral load data from an AIDS clinical trial.

For binary data, the deviance is

$$D(\eta) = -2 \sum_i y_i \log \eta_i - 2 \sum_i (1 - y_i) \log(1 - \eta_i).$$

Models with smaller DIC should be preferred over models with larger DIC. Models are penalized both by $\overline{D(\eta)}$, which favours a good fit, and by the effective number of parameters. Since $\overline{D(\eta)}$ will decrease as the number of parameters in a model increases, $\widehat{\rho_D}$ compensates for this effect by favouring models with fewer parameters.

3.4.1 Simulation study

In this example, multiple binary response data are generated and the model's performance is analyzed by using DIC. The following general model is considered

$$\Psi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \quad i = 1, \ldots, 100,$$

where $\boldsymbol{\beta} = (2, -2, 1)^T$ and Ψ is the cdf of a N$(0, 1)$ (probit), L$(0, 1)$ (logit), T$(\nu = 12)$ (t-link), EP$(0, 1, \theta = 0.5)$ (platykurtic EP-link) or EP$(0, 1, \theta = 1.5)$ (leptokurtic EP-link). A covariate set (the same for all the models) is generated from the specification given by $x_{i1} \sim$ U$(0, 3)$ and $x_{i2} \sim$ U$(0, 2)$, $i = 1, \ldots, 100$. Note that the covariate set is randomly dispersed over $[0, 3] \times [0, 2]$, so there is no separation criterion. Finally, for each link the outcomes are randomly generated by using the following process: (1) generate $w_i \sim$ U$(0, 1)$, (2) if $w_i \leq p_i$, then $y_i = 1$, else $y_i = 0$.

Probit, t-link and EP-link models are fitted. Specifically, the following nine models are considered: (1) normal, N$(0, 1)$, (2) Student's t with eight degrees of freedom, T$(\nu = 8)$ (approximation to the logistic distribution), (3) Student's t with degrees of freedom distributed in $\{6, 7, 8, 9, 10\}$ with probability $\{0.1, 0.2, 0.4, 0.2, 0.1\}$, T$(\nu \in \{6 - 10\})$, (4) Student's t with twelve degrees of freedom T$(\nu = 12)$, (5) Student's t with degrees of freedom distributed in $\{10, 11, 12, 13, 14\}$ with probability $\{0.1, 0.2, 0.4, 0.2, 0.1\}$, T$(\nu \in \{10 - 14\})$, (6) Student's t with degrees of freedom uniformly distributed in $\{1, 2, \ldots, 20\}$, T$(\nu \in \{1 - 20\})$, (7) EP$(0, 1, \theta)$ with $\theta \sim$ U$(0, 1)$, EP$(\theta \in (0, 1))$, (8) EP$(0, 1, \theta)$ with $\theta \sim$ U$(1, 2)$, EP$(\theta \in (1, 2))$, and (9) EP$(0, 1, \theta)$ with $\theta \sim$ U$(0, 2)$, EP$(\theta \in (0, 2))$.

A weakly informative prior distribution for $\boldsymbol{\beta}$ is used for all the fitted regression models, i.e. the prior distribution is a multivariate normal with parameters $\mathbf{b}^T = (0, 0, 0)$ and $\mathbf{B} = \text{diag}(10, 10, 10)$. For the t-link (EP-link) models, the prior distributions for the degrees of freedom ν (shape parameter θ) are the discrete (continuous uniform) ones presented in the previous paragraph. A total of $220,000$ iterations of MCMC are run. Then, a burn-in of $20,000$ is considered and one out of 20 values is saved. With this specification all the chains seem to have converged.

Table 3.1 shows the mean and the standard deviation of the 100 DIC values obtained for each link and each fitted model, i.e., for each link the generation process previously described is performed 100 times with the same covariate set and the 100 DICs obtained are averaged to give a general measure for each fitted model.

Table 3.1: Estimated DICs for all the data generation processes.

Model	**DIC mean (sd)**	$\overline{\boldsymbol{D}}$ **mean (sd)**	$\widehat{\boldsymbol{\rho_D}}$ **mean (sd)**
	Ψ is the cdf of a N$(0, 1)$ (probit)		
N$(0, 1)$	54.809 (12.363)	52.293 (12.098)	2.516 (0.370)
T$(\nu = 8)$	55.108 (12.368)	52.537 (12.086)	2.571 (0.379)
T$(\nu \in \{6 - 10\})$	55.126 (12.375)	52.546 (12.091)	2.580 (0.388)
T$(\nu = 12)$	55.005 (12.368)	52.447 (12.090)	2.558 (0.381)
T$(\nu \in \{10 - 14\})$	55.016 (12.363)	52.452 (12.088)	2.563 (0.375)

Model	**DIC mean (sd)**	$\bar{D}$ **mean (sd)**	$\widehat{\rho_D}$ **mean (sd)**
$\mathrm{T}(\nu \in \{1-20\})$	55.264 (12.461)	52.688 (12.149)	2.576 (0.392)
$\mathrm{EP}(\theta \in (0,1))$	54.432 (12.484)	51.979 (12.205)	2.453 (0.425)
$\mathrm{EP}(\theta \in (1,2))$	54.985 (12.332)	52.615 (12.040)	2.370 (0.368)
$\mathrm{EP}(\theta \in (0,2))$	54.539 (12.634)	52.136 (12.351)	2.402 (0.375)
Ψ is the cdf of a $\mathrm{L}(0,1)$ (logit)			
$\mathrm{N}(0,1)$	90.492 (15.728)	87.623 (15.523)	2.869 (0.342)
$\mathrm{T}(\nu = 8)$	90.553 (15.708)	87.635 (15.503)	2.918 (0.331)
$\mathrm{T}(\nu \in \{6-10\})$	90.545 (15.703)	87.631 (15.500)	2.913 (0.328)
$\mathrm{T}(\nu = 12)$	90.532 (15.720)	87.626 (15.517)	2.906 (0.331)
$\mathrm{T}(\nu \in \{10-14\})$	90.523 (15.716)	87.621 (15.514)	2.902 (0.332)
$\mathrm{T}(\nu \in \{1-20\})$	90.560 (15.731)	87.561 (15.507)	2.999 (0.361)
$\mathrm{EP}(\theta \in (0,1))$	90.423 (15.948)	87.568 (15.726)	2.854 (0.407)
$\mathrm{EP}(\theta \in (1,2))$	90.332 (15.737)	87.570 (15.494)	2.762 (0.330)
$\mathrm{EP}(\theta \in (0,2))$	90.197 (16.057)	87.389 (15.824)	2.808 (0.332)
Ψ is the cdf of a $\mathrm{T}(\nu = 12)$ (t-link)			
$\mathrm{N}(0,1)$	60.308 (13.482)	57.665 (13.191)	2.642 (0.420)
$\mathrm{T}(\nu = 8)$	60.463 (13.379)	57.775 (13.087)	2.688 (0.411)
$\mathrm{T}(\nu \in \{6-10\})$	60.481 (13.364)	57.786 (13.074)	2.696 (0.411)
$\mathrm{T}(\nu = 12)$	60.413 (13.401)	57.733 (13.107)	2.680 (0.415)
$\mathrm{T}(\nu \in \{10-14\})$	60.414 (13.379)	57.733 (13.097)	2.681 (0.404)
$\mathrm{T}(\nu \in \{1-20\})$	60.524 (13.331)	57.825 (13.025)	2.699 (0.419)
$\mathrm{EP}(\theta \in (0,1))$	60.101 (13.780)	57.483 (13.458)	2.618 (0.503)
$\mathrm{EP}(\theta \in (1,2))$	60.202 (13.179)	57.765 (12.906)	2.437 (0.369)
$\mathrm{EP}(\theta \in (0,2))$	59.931 (13.649)	57.424 (13.352)	2.507 (0.389)
Ψ is the cdf of a $\mathrm{EP}(0,1,\theta = 0.5)$ (platykurtic EP-link)			
$\mathrm{N}(0,1)$	36.490 (9.094)	34.389 (8.741)	2.101 (0.421)
$\mathrm{T}(\nu = 8)$	36.921 (9.134)	34.789 (8.749)	2.132 (0.456)
$\mathrm{T}(\nu \in \{6-10\})$	36.934 (9.133)	34.799 (8.747)	2.136 (0.451)
$\mathrm{T}(\nu = 12)$	36.769 (9.118)	34.639 (8.744)	2.130 (0.446)
$\mathrm{T}(\nu \in \{10-14\})$	36.775 (9.122)	34.644 (8.746)	2.131 (0.448)
$\mathrm{T}(\nu \in \{1-20\})$	37.051 (9.157)	34.928 (8.770)	2.123 (0.450)
$\mathrm{EP}(\theta \in (0,1))$	35.760 (9.110)	33.766 (8.777)	1.994 (0.443)
$\mathrm{EP}(\theta \in (1,2))$	36.913 (9.100)	34.958 (8.748)	1.955 (0.406)
$\mathrm{EP}(\theta \in (0,2))$	36.037 (9.249)	34.059 (8.922)	1.978 (0.409)
Ψ is the cdf of a $\mathrm{EP}(0,1,\theta = 1.5)$ (leptokurtic EP-link)			
$\mathrm{N}(0,1)$	83.260 (15.235)	80.402 (15.020)	2.858 (0.353)
$\mathrm{T}(\nu = 8)$	83.250 (15.156)	80.351 (14.950)	2.899 (0.337)
$\mathrm{T}(\nu \in \{6-10\})$	83.251 (15.166)	80.350 (14.954)	2.901 (0.334)
$\mathrm{T}(\nu = 12)$	83.249 (15.194)	80.358 (14.989)	2.891 (0.342)
$\mathrm{T}(\nu \in \{10-14\})$	83.265 (15.166)	80.368 (14.966)	2.898 (0.332)
$\mathrm{T}(\nu \in \{1-20\})$	83.161 (15.134)	80.200 (14.913)	2.961 (0.352)

Model	**DIC mean (sd)**	$\bar{D}$ **mean (sd)**	$\widehat{\rho_D}$ **mean (sd)**
EP($\theta \in (0,1)$)	83.302 (15.470)	80.430 (15.220)	2.872 (0.448)
EP($\theta \in (1,2)$)	82.893 (15.089)	80.203 (14.863)	2.690 (0.312)
EP($\theta \in (0,2)$)	82.817 (15.373)	80.065 (15.156)	2.751 (0.326)

The estimated DICs show that the EP-links produce the best models in all the cases. Some remarks will now be presented. When data are generated by using a probit link, EP($\theta \in (0,1)$) and EP($\theta \in (0,2)$) are the best ones, followed by the probit model, which is better than EP($\theta \in (1,2)$). In the logit case, EP($\theta \in (0,2)$) is the best model, followed by EP($\theta \in (1,2)$) and EP($\theta \in (0,1)$). The EP models are also the best ones when data are generated by using the Student's t link with 12 degrees of freedom. In the case where the platykurtic link EP$(0,1,0.5)$ is considered, the best model is EP($\theta \in (0,1)$), which enhances platykurtosis, and the second is EP($\theta \in (0,2)$) which gives flexibility to the parameter θ. They are followed by the probit model, which is better than EP($\theta \in (1,2)$) because this latter enhances leptokurtosis. For the leptokurtic link EP$(0,1,1.5)$, something similar is the case by enhancing leptokurtosis.

EP-link models with very narrow supports that include the true values of θ have been considered. Although some interesting results have been obtained, this does not guarantee better performance than allowing the parameter θ to vary over a wider range. Besides, in practice, it is difficult to obtain information about a narrow range for θ. For example, when data were generated by using a standard normal, the average DIC for EP($\theta \in (0.9,1.1)$) was 54.836, which is greater than the values for EP($\theta \in (0,1)$) (=54.432) and EP($\theta \in (0,2)$) (=54.539). It is also of the same magnitude as the average DIC for the probit (=54.809) and smaller than that for EP($\theta \in (1,2)$) (=54.985). Another example is when data are generated by using an EP$(0,1,0.5)$ distribution. In this case, the average DIC for EP($\theta \in (0.4,0.6)$) is 35.833, is greater than the one of EP($\theta \in (0,1)$) (=35.760), but smaller than the ones for EP($\theta \in (0,2)$) (=36.037), probit (=36.490) and EP($\theta \in (1,2)$) (=36.913). This happens because the data were generated by using an EP$(0,1,0.5)$ that enhances platykurtic links. And finally, when data are generated by using an EP$(0,1,1.5)$ distribution, the average DIC for EP($\theta \in (1.4,1.6)$) is 82.996, which is greater than the ones for EP($\theta \in (0,2)$) (=82.817) and EP($\theta \in (1,2)$) (=82.893), but smaller than the ones for the probit (=83.260) and the EP($\theta \in (0,1)$) (=83.302). This is the case because the data were generated by using an EP$(0,1,1.5)$, that enhances leptokurtic links.

In the EP-link model the effective number of parameters $\widehat{\rho_D}$ tends to be lower than the total number of parameters in the model. This effect has been reported in other works as, for example, Zhu and Carlin (2000), Celeux et al. (2006), McGrory and Titterington (2007) and Huang (2008). Zhu and Carlin (2000) and Huang (2008) interpreted that this is due to the 'borrowing of strength' across individual-level parameters in hierarchical models.

We would emphasize that, in general, the models based on an EP-link with θ as a random variable allow great flexibility. Generally, they produce better predictions, and the inclusion of the additional parameter does not make the generation process difficult.

3.4.2 Adult respiratory distress syndrome

Adult respiratory distress syndrome (ARDS) is a complication in many critically ill patients. The usual diagnosis of ARDS is based on clinical findings of refractory respiratory failure and X-rays of the lungs showing fluid accumulation. Rocker et al. (1988) used lung images obtained after labeling the plasma protein transferrin in patients meeting the clinical criteria for ARDS and in patients who did not, to calculate a lung protein accumulation index (P) that is greater the greater the amount of protein in the lungs. They also recorded other characteristics of these patients, including their sex (S), age (A), X-ray lung fluid score (R), and amount of oxygen in their blood (O). Figure 3.3 shows the data ($n = 44$ patients). The light gray dots correspond to the patients with absence of ARDS ($y_i = 0$), and the dark gray dots to the patients with presence of ARDS ($y_i = 1$).

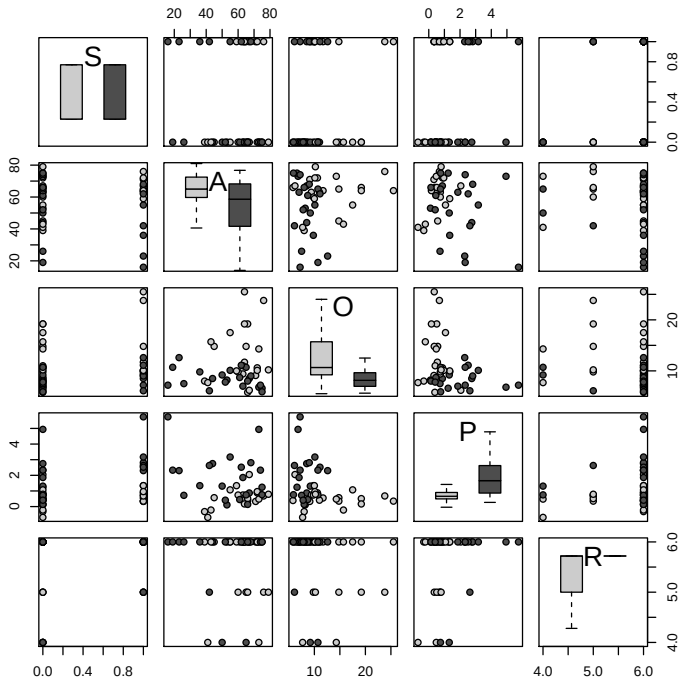

Figure 3.3: ARDS data: The light (dark) gray dots correspond to the patients with absence (presence) of ARDS.

Firstly, a model of interest is given by

$$\eta_i = \Psi^{-1}(p_i) = \beta_0 + \beta_1 S_i + \beta_2 A_i + \beta_3 O_i + \beta_4 P_i + \beta_5 R_i,$$

where the observed binary outcome y_i is the presence or absence of ARDS in the ith patient. A model comparison framework is considered by using different distribution functions Ψ.

An EP-link regression model with θ as a random variable in $(0, 2)$ is considered. For the regression parameters, a multivariate normal prior distribution for $\boldsymbol{\beta}$ is used, i.e., $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{b}, \mathbf{B})$. In a weakly informative setting, $\mathbf{b}$ is a vector of zeros and $\mathbf{B}$ is the diagonal matrix whose elements are 100. The prior distribution for the shape parameter θ is $\mathrm{U}(0, 2)$. A total of $200,000$ iterations are run. Then, a burn-in of $50,000$ is considered and one out of 30 observations is saved. With these MCMC specifications, the chain seems to have converged. This model showed that the 90% probability regions for the regression parameters of both S and R contained 0, so it was decided to remove them. Then, simplified models are given by

$$\eta_i = \Psi^{-1}(p_i) = \beta_0 + \beta_1 A_i + \beta_2 O_i + \beta_3 P_i,$$

with $i = 1, \ldots, 44$. Probit, t-link, and EP-link models are fitted. Specifically, the following six models are considered: (1) normal, $\mathrm{N}(0, 1)$, (2) Student's t with eight degrees of freedom, $\mathrm{T}(\nu = 8)$ (approximation to the logistic distribution), (3) Student's t with degrees of freedom uniformly distributed in $\{1, 2, \ldots, 20\}$, $\mathrm{T}(\nu \in \{1 - 20\})$, (4) $\mathrm{EP}(0, 1, \theta)$ with $\theta \sim \mathrm{U}(0, 1)$, $\mathrm{EP}(\theta \in (0, 1))$, (5) $\mathrm{EP}(0, 1, \theta)$ with $\theta \sim \mathrm{U}(1, 2)$, $\mathrm{EP}(\theta \in (1, 2))$, and (6) $\mathrm{EP}(0, 1, \theta)$ with $\theta \sim \mathrm{U}(0, 2)$, $\mathrm{EP}(\theta \in (0, 2))$.

A weakly informative prior distribution for $\boldsymbol{\beta}$ is used for all the regression models, i.e., the prior distribution is the multivariate normal distribution with parameters $\mathbf{b}^T = (0, 0, 0)$ and $\mathbf{B} = \mathrm{diag}(100, 100, 100)$. For the t-link (EP-link) models, the prior distributions for the degrees of freedom ν (shape parameter θ) are the discrete (continuous uniform) ones presented above. Under the same previous MCMC specifications for the total number of iterations, burn-in and storage, all the chains seem to have converged.

Table 3.2 presents the estimated DICs in this weakly informative setting. The values show that $\mathrm{EP}(\theta \in (1, 2))$ is the best model. This is followed by the other two EP-based models.

Table 3.2: Estimated DICs for models fitted to ARDS data in a weakly informative setting.

Model	DIC	$\overline{D}$	$\widehat{\rho_D}$
$\mathrm{N}(0, 1)$	39.388	35.930	3.459
$\mathrm{T}(\nu = 8)$	39.944	36.091	3.853
$\mathrm{T}(\nu \in \{1 - 20\})$	39.510	35.777	3.734
$\mathrm{EP}(\theta \in (0, 1))$	39.265	35.930	3.335
$\mathrm{EP}(\theta \in (1, 2))$	38.965	35.702	3.263
$\mathrm{EP}(\theta \in (0, 2))$	39.353	35.865	3.488

Following the method of constructing an informative prior distribution proposed by Bedrick et al. (1996), the prior elicitation was carried out as in Paulino et al. (2003). In order to induce the prior distribution on $\boldsymbol{\beta}$, four covariate configurations are chosen. These are the quartiles and means of the data. For each configuration, the 25%, 50%, and 75% quantiles for the ARDS probabilities are chosen. This is an overspecification of the beta distributions, because only two quantiles are necessary. Specifically, two quantiles are used to calculate the prior hyperparameters and the

third is used to check the choice. The complete set of prior hyperparameters elicited from the quantiles is presented in Table 3.3.

Table 3.3: Configurations and hyperparameters.

Configurations $\tilde{\mathbf{x}} = (1, A, O, P)$	Hyperparameters	
$\tilde{\mathbf{x}}_1^T = (1, 48.75,\ 7.775, 0.410)$	$a_{11} = 6.935$	$a_{21} = 3.66$
$\tilde{\mathbf{x}}_2^T = (1, 63.00,\ 9.600, 0.755)$	$a_{12} = 7.600$	$a_{22} = 13.60$
$\tilde{\mathbf{x}}_3^T = (1, 57.55, 10.718, 1.218)$	$a_{13} = 6.420$	$a_{23} = 6.83$
$\tilde{\mathbf{x}}_4^T = (1, 67.25, 11.250, 1.907)$	$a_{14} = 3.340$	$a_{24} = 3.00$

Table 3.4 presents the estimated DICs in this informative setting for several models. Note that the initial information is the same for all the models, so that they are comparable.

Table 3.4: Estimated DICs for models fitted to ARDS data in an informative setting.

Model	DIC	$\bar{D}$	$\widehat{\rho_D}$
N(0, 1)	36.611	34.355	2.255
T($\nu = 8$)	36.454	34.207	2.247
T($\nu \in \{1 - 20\}$)	36.115	34.008	2.108
EP($\theta \in (0, 1)$)	36.691	34.503	2.187
EP($\theta \in (1, 2)$)	35.801	33.928	1.872
EP($\theta \in (0, 2)$)	35.634	33.884	1.751

The estimated DICs show that the EP($\theta \in (0, 2)$) and EP($\theta \in (1, 2)$) models give the best performances. The EP($\theta \in (0, 1)$) model (that supports platykurtic distributions) and the probit model are the poorest, having DIC values of the same order of magnitude. This means that the data are better supported by values of θ leading to leptokurtic distributions, as was the case in the noninformative setting.

The sample generated by the EP($\theta \in (0, 2)$) model was used to evaluate the posterior distribution. Figure 3.4 displays the posterior distributions with the 90% and 95% highest density intervals (HDP), and the simulated parameters are summarized in Table 3.5. The intervals and the plots were obtained by using the R package hdrcde. The R function hdr.cde provides a smoothing which usually makes the intervals slightly larger. This explains why the upper limit of the 95% HDP interval for θ is greater than the maximum value allowed for the parameter, while all the generated values for θ are less than or equal to 2.

The coefficient for protein (P) is positive, indicating that high lung protein is associated with a high probability of having ARDS. Both age (A) and oxygen (O) have negative effects. Thus, higher oxygen is associated with lower probability of ARDS, which makes sense because ARDS is clinically defined in terms of low oxygenation. Similarly, greater age is associated with a lower probability of ARDS. This is because the ARDS patients tend to be younger in this sample.

A very interesting result was obtained for the parameter θ, since its posterior mean is 1.444 and its 95% HDP interval is $(0.676, 2.084)$. In the case of the EP($\theta \in (1, 2)$)

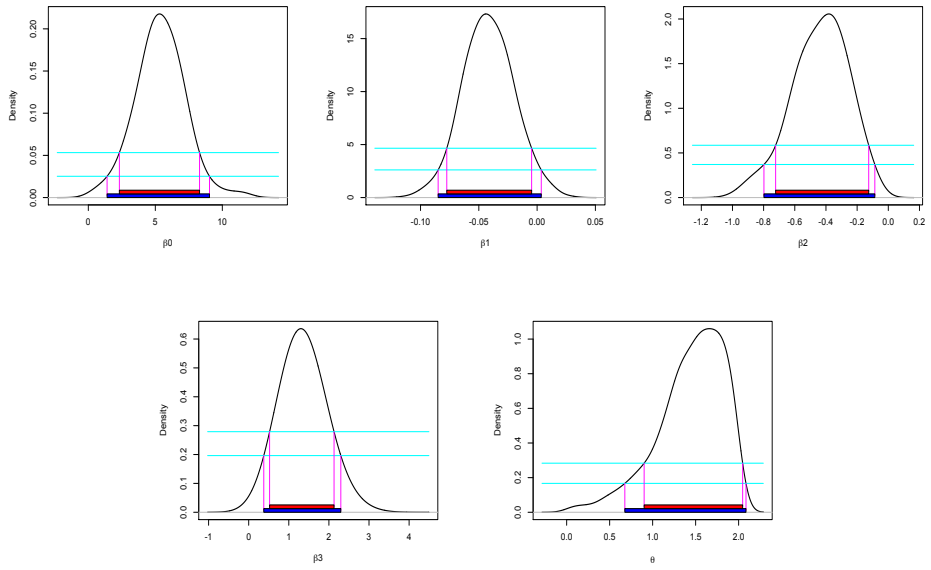

Figure 3.4: Estimated posterior distributions with 90% and 95% HDP intervals for the parameters of the EP($\theta \in (0, 2)$) model.

Table 3.5: Summary of the posterior estimates for the parameters of the EP($\theta \in (0, 2)$) model.

Variable	Mean	Median	Standard deviation	95% HDP interval
β_0	5.389	5.375	1.868	$(1.403, 9.043)$
β_1 (A)	-0.041	-0.042	0.022	$(-0.084, 0.003)$
β_2 (O)	-0.439	-0.422	0.182	$(-0.798, -0.086)$
β_3 (P)	1.353	1.328	0.494	$(0.376, 2.297)$
θ	1.444	1.510	0.388	$(0.676, 2.084)$

model the posterior mean of θ is 1.626 and its 95% HDP interval is $(1.135, 2.016)$. The flexibility of these EP-link models with θ as a random variable has allowed us to choose the values of θ for the best fits. Note that these value ranges (with the posterior distributions of θ) give mainly leptokurtic cdfs. This explains why EP($\theta \in (0, 2)$) and EP($\theta \in (1, 2)$) are the best models.

Finally, a Bayesian residual analysis is performed as in Albert and Chib (1995). Figure 3.5 presents the marginal posterior densities of the residuals $r_i = y_i - p_i$ for the EP($\theta \in (0, 2)$) model. The shape of the residual distribution depends on its location. The distributions that are concentrated near the endpoints of the support region show substantive skewness.

Figure 3.6 shows the boxplots of posterior distributions of the residuals against the fitted probabilities $\mathrm{E}(p_i|y_i)$ for the EP-link model. The middle section of the

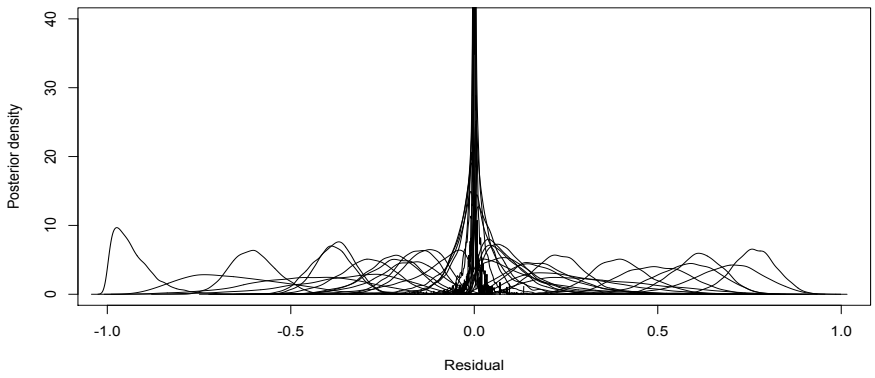

Figure 3.5: Residual posterior densities for the EP($\theta \in (0, 2)$) model.

boxplot corresponds to the quartiles, and the extreme values correspond to the 5th and 95th percentiles of the distribution. The light gray boxplots correspond to the observations where $y_i = 0$ (absence of ARDS), whose support is $(-1, 0)$, and the dark gray boxplots correspond to the observations where $y_i = 1$ (presence of ARDS), whose support is $(0, 1)$.

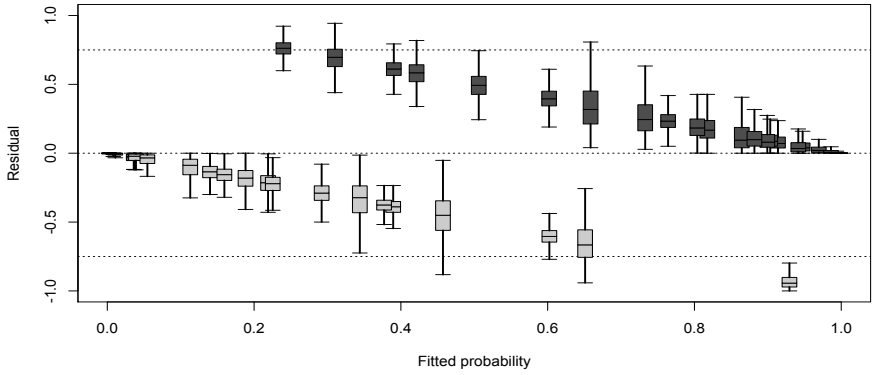

Figure 3.6: Boxplots of posterior distributions for residuals against the fitted probabilities of the EP($\theta \in (0, 2)$) model.

Outlier observations correspond to densities of residuals that have locations away from zero, concentrated at extremes of the possible value range of the residuals and showing strong asymmetry. One way of gauging the relative sizes of these residuals is to compute the posterior probabilities that r_i exceed in absolute value some positive constant ϵ. Albert and Chib (1995) used $\epsilon = 0.75$. Parallel horizontal lines are drawn at residual values -0.75 and 0.75. If the boxplot for a particular residual does not

cross these lines, then the outlying probability is under 0.05. Residuals with large outlying probabilities correspond to boxplots that significantly cross these lines. There are two highly influential combinations of predictor variables associated with ARDS patients who are 65 and 67 years old, with relatively large outlying probabilities. Their residuals are 0.760 and -0.929, respectively.

3.5 Conclusion

This paper has presented a flexible Bayesian approach to a generalized linear model to describe the dependence of binary data on explanatory variables. The inverse EP cdf is used as the link of the binary regression model. The probit and an approximation to the logit models are particular cases of the proposed approach.

The approach uses the idea of data augmentation and a mixture representation for the EP distribution. This allows one to derive efficient Gibbs sampling algorithms for both informative and noninformative settings. Generation from all full conditional distributions is straightforward.

The proposed approach allows one to find the necessary degree of robustness by letting the kurtosis parameter θ vary over its range. Moreover, this range can be modified to provide models based on platykurtic and leptokurtic inverse link cdfs.

An empirical study over a wide range of datasets showed the proposed approach to be useful and to have good performance when compared with competing models. Therefore, it may be interesting to use the inverse of distributions of the EP family routinely in any Bayesian procedure where probit or logit binary regression models might conventionally be used.

Acknowledgements

We thank two anonymous referees for comments and suggestions which have highly improved this paper. This research has been partially supported by *Ministerio de Ciencia e Innovación*, Spain (Projects TIN2008-06796-C04-03 and MTM2011-28983-C03-02) and *Junta de Extremadura* (Project GRU10110).

Chapter 4

Bayesian analysis of some models that use the asymmetric exponential power distribution

Naranjo, L., Pérez, C. J., and Martín, J. (2014). Bayesian analysis of some models that use the asymmetric exponential power distribution. *Statistics and Computing*. In Press.

Abstract

The asymmetric exponential power (AEP) family includes the symmetric exponential power distribution as a particular case. It provides flexible distributions with lighter and heavier tails compared to the normal one. The distributions of this family can successfully handle both symmetry/asymmetry and light/heavy tails simultaneously. Even more, the distributions can fit each tail separately. This provides a great flexibility when fitting experimental data. The idea of using a scale mixture of uniform representation of the AEP distribution is exploited to derive efficient Gibbs sampling algorithms in three different Bayesian contexts. Firstly, a posterior exploration is performed, where the AEP distribution is considered for the likelihood model. Secondly, a linear regression model, that uses the AEP distribution for the error variable, is developed. And finally, a binary regression model is analyzed, by using the inverse of the AEP cumulative distribution function as the link function. These three models have been built in such a way that they share some full conditional distributions to sample from their respective posterior distributions. The theoretical results are illustrated by comparing with other competing models using some previously published datasets.

Keywords: Binary regression; Exponential power distribution; Gibbs sampling; Linear regression; Scale mixture of uniform distributions; Asymmetric exponential power distribution.

4.1 Introduction

The first formulation of the exponential power (EP) distribution (also known as generalized normal distribution or generalized error distribution) could be attributed to Subbotin (1923). Since then, several parameterizations have appeared in the literature (see e.g., Vianelli (1963), Box and Tiao (1973) and Gómez et al. (1998)). The EP family includes the normal distribution and incorporates additional shapes, including platykurtic (lighter tails compared to the normal) and leptokurtic (heavier tails compared to the normal), which is always an advantage when studying robustness.

However, the distributions of this family and their reparameterizations are symmetrical. Considering asymmetry can be useful when fitting experimental data. The skewed exponential power (SEP) or asymmetric exponential power (AEP) distributions consider asymmetry. These families include the EP distribution as a particular case and can provide distributions with lighter or heavier tails compared to the symmetric/asymmetric normal distribution.

In order to deal with both heavy tails and skewness, Azzalini (1986) presented a SEP distribution as an extension of the skew normal distribution introduced by him in Azzalini (1985). The SEP distribution proposed by Azzalini (1986) was further studied by DiCiccio and Monti (2004) and Monti (2003). In addition, Fernández et al. (1995) defined and studied a class of spliced-scale distributions suggested by independent sampling from a generalization of the EP distribution. The distributions of this class allow to successfully model skewness. Properties and estimation methods for the distributions of this class were presented by Theodossiou (2000), Ayebo and Kozubowski (2004), Komunjer (2007), and Delicado and Goria (2008). Based on the idea of spliced-scale distributions, several AEP distributions have been defined, see Arellano-Valle et al. (2005), Jones (2005), Zhu and Zinde-Walsh (2009) and Bottazzi and Secchi (2011). Zhu and Zinde-Walsh (2009) derived moments and moment-based measures for their AEP distribution; established consistency, asymptotic normality and efficiency of the maximum likelihood estimators over a large part of the parameter space by dealing with the problems created by a non-smooth likelihood function and derived an explicit analytical expression for the asymptotic covariance matrix.

The SEP and AEP families of distributions have been mainly studied from a frequentist approach. Up to the authors' knowledge, they have not been considered, except in Naranjo et al. (2012), from a Bayesian viewpoint. Possibly, the main reason has been the intractability of the posterior distribution. However, these distributions can play an interesting role in some contexts.

In this paper, Bayesian analyses of three models that use the rescaled AEP distribution of Zhu and Zinde-Walsh (2009) are presented and discussed. The implementations are based on the idea of using a modification of the scale mixture of uniform representation of the EP distribution suggested by Walker and Gutiérrez-Peña (1999). This idea is exploited to derive three efficient Gibbs sampling algorithms sharing some full conditional distributions. Damien et al. (1999) demonstrated the use of latent variables for sampling non-standard densities which arise in the context of the Bayesian analysis of non-conjugate and hierarchical models by using a Gibbs sampler. Specifically, they showed that all except one of the full conditional distributions are uniform and the remaining full conditional distribution is truncated. Although other

Markov chain Monte Carlo (MCMC) algorithms (see Gilks et al. (1996)), such as the Metropolis-Hastings one (see Chib and Greenberg (1995)), may be used to model the AEP distribution, our use of latent variables results in an easy-to-implement Gibbs sampler having standard full conditional distributions that are easy to sample from.

Besides the flexibility of the AEP distribution, some additional advantages due to the proposed Bayesian approaches. It is possible to introduce initial information by using prior distributions. This can be very useful in situations where expert or historical information can be obtained. When the prior distributions are not informative, the estimation results are similar to those obtained by using maximum likelihood methods. Moreover, these approaches can be applied even when the sample size is small.

The outline of the paper is as follows. In Section 4.2, the AEP distribution and the proposed scale mixture of uniform representation are presented. Section 4.3 explores the posterior distribution in a model where the AEP distribution is considered for the likelihood. Section 4.4 presents the linear regression model by using the AEP distribution for the error variable. Section 4.5 presents a binary regression model where the inverse of the AEP cumulative distribution function (cdf) is used as the link function. In Section 4.6 the conclusions are presented. Finally, some technical details are presented in two appendices.

4.2 The AEP distribution

In this paper the rescaled AEP distribution proposed by Zhu and Zinde-Walsh (2009) will be considered. This distribution has location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, skewness parameter $\alpha \in (0,1)$, and left and right tail parameters (shape parameters or power parameters controlling the kurtosis) $\theta_1 > 0$ and $\theta_2 > 0$. The probability density function (pdf) is given by

$$f_{AEP}(y|\mu,\sigma,\alpha,\theta_1,\theta_2) = \begin{cases} \frac{1}{\sigma}\exp\left(-\left|\frac{y-\mu}{\alpha\sigma/\Gamma(1+1/\theta_1)}\right|^{\theta_1}\right) & \text{if } y \leq \mu, \\ \frac{1}{\sigma}\exp\left(-\left|\frac{y-\mu}{(1-\alpha)\sigma/\Gamma(1+1/\theta_2)}\right|^{\theta_2}\right) & \text{if } y > \mu, \end{cases} \tag{4.1}$$

and it is denoted by $Y \sim \text{AEP}(\mu,\sigma,\alpha,\theta_1,\theta_2)$.

When $\alpha = 1/2$ and $\theta_1 = \theta_2$, the distribution is symmetric. The parameterization (4.1) allows to model each tail separately, and the effects of the shape parameters on the distribution are clearly observed. Figure 4.1 shows the pdfs and cdfs for some parameter values related to the skewness and left tail, with fixed values for the parameters related to location, scale and right tail. Analogous graphics can be obtained for the parameter related to the right tail. The distributions in this family are very flexible providing good fits to experimental data.

Scale mixtures of uniform distributions have been used to represent several distributions and statistical models. Qin (2000) developed a general family of statistical models using scale mixtures of uniform distributions. It enables the incorporation of difficult assumptions in a broad range of applications. In addition, it makes the

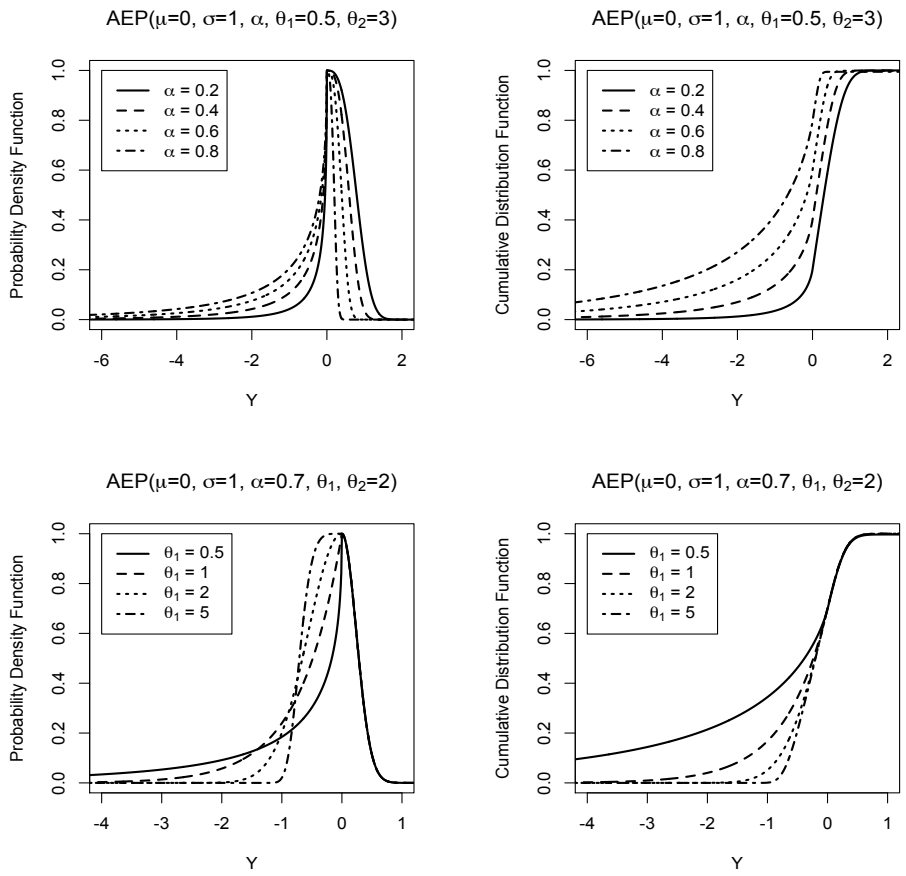

Figure 4.1: Pdfs and cdfs of the AEP$(\mu, \sigma, \alpha, \theta_1, \theta_2)$, for some parameter values related to the skewness (α) and left tail (θ_1), with fixed values $\mu = 0$, $\sigma = 1$ and θ_2.

analyses of real data straightforward by an appropriate use of latent variables in the resulting computational form of the model. The developments in Qin (2000) were applied to robust models, variance regression models, time series, econometric models and non-linear models. Other distributions have been defined as a scale mixture of uniforms, for instance, the uniform power distribution (Walker (1999)), the exponential power distribution (Walker and Gutiérrez-Peña (1999)), the generalized lognormal distribution (Martín and Pérez (2009)), and recently, the SEP family defined by Bottazzi and Secchi (2011) and applied by Naranjo et al. (2012) in a Bayesian context.

The mixture representation of the EP distribution defined by Walker and Gutiérrez-Peña (1999) is adapted here for the AEP distribution. The following result that has been proved in Appendix A, will be used in the next sections to derive the Gibbs sampling algorithms.

Proposition 4.1 *If* $Y|[U_1 = u_1] \sim \mathrm{U}\Big(\mu - \frac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_1^{1/\theta_1}\ ,\ \mu\Big)$, $U_1 \sim \mathrm{Ga}\big(1+1/\theta_1\ ,\ 1\big)$, *with probability* α *and* $Y|[U_2 = u_2] \sim \mathrm{U}\Big(\mu\ ,\ \mu + \frac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_2^{1/\theta_2}\Big)$, $U_2 \sim \mathrm{Ga}\big(1 + 1/\theta_2\ ,\ 1\big)$, *with probability* $(1-\alpha)$, *then* $Y \sim \mathrm{AEP}(\mu, \sigma, \alpha, \theta_1, \theta_2)$.

This proposition shows a scale mixture of uniform representation of a spliced-scale distribution that allows to model each tail separately from the property $\mathrm{p}(Y \leq \mu) = \alpha$. This proposition is also useful to derive the moments, the cumulative distribution function, and other properties of the AEP distribution.

The next sections present three Bayesian models that use AEP distributions.

4.3 Exploring the posterior distribution

This section explores the posterior distribution in a model where the AEP distribution is considered for the likelihood. The obtained posterior distribution is intractable for inferential purposes. The reason is that the posterior expectation of the parameters can not be analytically evaluated. The proposed solution involves expressing the AEP density as a mixture by using Proposition 4.1 and using numerical approximations based on Gibbs sampling algorithms.

The likelihood of a sample $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the latent vectors of mixing parameters $\mathbf{u}_1 = (u_{11}, \ldots, u_{1n})^T$ and $\mathbf{u}_2 = (u_{21}, \ldots, u_{2n})^T$ is

$$\begin{aligned} L(\mu, \sigma, \alpha, \theta_1, \theta_2 | \mathbf{y}, \mathbf{u}_1, \mathbf{u}_2) &= \prod_{i=1}^{n} \frac{1}{\sigma} \Bigg(\exp(-u_{1i}) I\Big[\mu - \tfrac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1} < y_i \leq \mu\Big] \\ &+ \exp(-u_{2i}) I\Big[\mu < y_i < \mu + \tfrac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\Big] \Bigg). \end{aligned}$$

Then the joint posterior distribution of the unobservables μ, σ, α, θ_1, θ_2, $\mathbf{u}_1$ and $\mathbf{u}_2$ is

$$\begin{aligned} &\pi(\mu, \sigma, \alpha, \theta_1, \theta_2, \mathbf{u}_1, \mathbf{u}_2 | \mathbf{y}) \\ &\quad \propto \pi(\mu)\pi(\sigma)\pi(\alpha)\pi(\theta_1)\pi(\theta_2) L(\mu, \sigma, \alpha, \theta_1, \theta_2 | \mathbf{y}, \mathbf{u}_1, \mathbf{u}_2). \end{aligned}$$

Jeffreys' noninformative prior distributions can be computed from the Fisher information matrix given by Zhu and Zinde-Walsh (2009). However, using these prior distributions does not provide full conditional distributions that are easy to sample from. Some alternative prior distributions are proposed in this section. They have the advantage of allowing the derivation of full conditional distributions that are easy to sample from. The full conditional distributions are presented in the following subsection.

4.3.1 Derivation of full conditional distributions

The full conditional distributions of u_{1i} and u_{2i}, for $i = 1, \ldots, n$, are

$$[u_{1i}, u_{2i}|\mathbf{y}, \mu, \sigma, \alpha, \theta_1, \theta_2] \tag{4.2}$$
$$\sim \begin{cases} \text{Exp}(1)\, I\left[u_{1i} > \left(\frac{\mu - y_i}{\alpha\sigma/\Gamma(1+1/\theta_1)}\right)^{\theta_1}\right] & \text{if } y_i \leq \mu, \\ \text{Exp}(1)\, I\left[u_{2i} > \left(\frac{y_i - \mu}{(1-\alpha)\sigma/\Gamma(1+1/\theta_2)}\right)^{\theta_2}\right] & \text{if } y_i > \mu, \end{cases}$$

where Exp(1) denotes the exponential distribution with parameter equal to 1. Note that if $y_i \leq \mu$ then $u_{1i} > 0$ and $u_{2i} = 0$, and if $y_i > \mu$ then $u_{1i} = 0$ and $u_{2i} > 0$.

The full conditional distribution of μ is

$$\pi(\mu|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \sigma, \alpha, \theta_1, \theta_2) \propto \pi(\mu) I\big[\underline{\mu} < \mu < \overline{\mu}\big], \tag{4.3}$$

where

$$\begin{aligned} \underline{\mu} &= \max\left\{ \max_{\{i:u_{2i}>0\}} \left\{ y_i - \tfrac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2} \right\}, \max_{\{i:u_{1i}>0,\ length(u_2[u_2>0])=0\}} \{y_i\} \right\}, \\ \overline{\mu} &= \min\left\{ \min_{\{i:u_{1i}>0\}} \left\{ y_i + \tfrac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1} \right\}, \min_{\{i:u_{2i}>0,\ length(u_1[u_1>0])=0\}} \{y_i\} \right\}. \end{aligned}$$

The full conditional distribution of σ is

$$\pi(\sigma|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \alpha, \theta_1, \theta_2) \propto \pi(\sigma) \frac{1}{\sigma^n} I\big[\sigma > \underline{\sigma}\big], \tag{4.4}$$

where

$$\underline{\sigma} = \max\left\{ \max_{\{i:u_{1i}>0\}} \left\{ \tfrac{(\mu - y_i)\Gamma(1+1/\theta_1)}{\alpha u_{1i}^{1/\theta_1}} \right\}, \max_{\{i:u_{2i}>0\}} \left\{ \tfrac{(y_i - \mu)\Gamma(1+1/\theta_2)}{(1-\alpha) u_{2i}^{1/\theta_2}} \right\} \right\}.$$

The full conditional distribution of α is

$$\pi(\alpha|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \sigma, \theta_1, \theta_2) \propto \pi(\alpha) I\big[\underline{\alpha} < \alpha < \overline{\alpha}\big], \tag{4.5}$$

where

$$\begin{aligned} \underline{\alpha} &= \max\left\{ 0, \max_{\{i:u_{1i}>0\}} \left\{ \tfrac{(\mu - y_i)\Gamma(1+1/\theta_1)}{\sigma u_{1i}^{1/\theta_1}} \right\} \right\}, \\ \overline{\alpha} &= \min\left\{ 0, \min_{\{i:u_{2i}>0\}} \left\{ 1 - \tfrac{(y_i - \mu)\Gamma(1+1/\theta_2)}{\sigma u_{2i}^{1/\theta_2}} \right\} \right\}. \end{aligned}$$

The full conditional distributions of θ_1 and θ_2 are

$$\pi(\theta_1|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \sigma, \alpha, \theta_2) \propto \pi(\theta_1) I\left[\theta_1 \in \bigcap_{\{i:u_{1i}>0\}} \Theta_{1i}\right], \tag{4.6}$$

$$\pi(\theta_2|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \sigma, \alpha, \theta_1) \propto \pi(\theta_2) I\left[\theta_2 \in \bigcap_{\{i:u_{2i}>0\}} \Theta_{2i}\right], \tag{4.7}$$

where

$$\Theta_{1i} = \left\{\theta_1 : \frac{\mu - y_i}{\alpha\sigma} < \frac{1}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1}\right\}, \quad \Theta_{2i} = \left\{\theta_2 : \frac{y_i - \mu}{(1-\alpha)\sigma} < \frac{1}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\right\}.$$

Some specific full conditional distributions for (4.3), (4.4) and (4.5) are obtained in Appendix B. Note that sampling from all the full conditional distributions is easy. Specifically, except (4.6) and (4.7), all the full conditional distributions are standard. The full conditional distributions of θ_1 and θ_2 can be easily sampled by using the acceptance-rejection method. The easiness of applying this method relies on the fact that these full conditional distributions are univariate and generally restricted to small domains.

The final algorithm consists of choosing initial values $\mu^{(0)}$, $\sigma^{(0)}$, $\alpha^{(0)}$, $\theta_1^{(0)}$ and $\theta_2^{(0)}$, and iteratively sampling $\mathbf{u}_1^{(j)}$ and $\mathbf{u}_2^{(j)}$, $\mu^{(j)}$, $\sigma^{(j)}$, $\alpha^{(j)}$, $\theta_1^{(j)}$ and $\theta_2^{(j)}$ from the full conditional distributions (4.2), (4.3), (4.4), (4.5), (4.6) and (4.7), respectively. The following initial points are proposed: $u_{1i} = 1$ and $u_{2i} = 0$ if $y_i \leq \mu$, $u_{1i} = 0$ and $u_{2i} = 1$ if $y_i > \mu$, $\mu = \text{mode}(\mathbf{y})$, $\alpha = \text{p}(y \leq \mu)$, $\theta_1 = 1$ and $\theta_2 = 1$. These initial points come from adapted results from Zhu and Zinde-Walsh (2009).

The next subsection presents an illustrative example.

4.3.2 Simulation example

In order to analyze the performance of the proposed procedures, simulation studies were conducted. An illustrative simulation experiment is presented here. Firstly, a convergence diagnostic is illustrated with one dataset, and secondly, the model has been implemented with various generated datasets.

Convergence diagnostic

A dataset is generated by $Y_i \sim \text{AEP}(\mu, \sigma, \alpha, \theta_1, \theta_2)$, $i = 1, \ldots, n$, by using the algorithm obtained from Proposition 4.1, where the sample size is $n = 300$ and the parameter values are: $\mu = 0$, $\sigma = 1$, $\alpha = 0.3$, $\theta_1 = 0.5$ and $\theta_2 = 2$. Two prior specifications are considered: non informative prior distributions given by $\pi(\mu) \propto 1$, $\pi(\sigma) \propto 1/\sigma^2$, $\pi(\alpha) \propto 1$, $\pi(\theta_1) \propto 1$, $\pi(\theta_2) \propto 1$, and informative prior distributions given by $\pi(\mu) = \text{N}(0, 1)$, $\pi(\sigma) \propto 1/\sigma^2$, $\pi(\alpha) = \text{Beta}(3, 7)$, $\pi(\theta_1) = \text{N}(0.5, 1)I[\theta_1 > 0]$, $\pi(\theta_2) = \text{N}(2, 1)I[\theta_2 > 0]$. The choice of these informative prior distributions is based on conjugacy, in the case of σ^2, and on the properties of the distributions related with the other parameters. Their hyperparameters have been chosen by using the relationship between the prior distributions and the true values of the parameters. For example, $\text{E}(\mu) = 0$, $\text{E}(\alpha) = 0.3$, $\text{mode}(\theta_1) = 0.5$ and $\text{mode}(\theta_2) = 2$. The initial values are $\mu = \text{mode}(\mathbf{y}) = 0.1$, $\sigma = \text{sd}(\mathbf{y}) = 0.8645$, $\alpha = \text{p}(Y \leq \mu) = 0.3733$, $\theta_1 = 1$ and $\theta_2 = 1$.

The algorithm has been implemented in `R`, and `BOA` package (see Smith (2007)) has been used to analyze the convergence. A total of 100,000 iterations have been performed by using the model proposed in this section.

In order to evaluate the efficiency of the algorithm, some summary statistics were calculated from the posterior sample of the parameters and are presented in Table

4.1. Note that estimates are close to the original values and standard deviations are small. The autocorrelations between batch means are close to one (Batch ACF), so batches (of default size 50) should be increased.

Table 4.1: Summary of MCMC.

Non informative prior

	True Value	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF	2.5%	50%	97.5% Precentiles
μ	0	0.040	3.09e-2	1.03e-4	2.23e-3	6.91e-4	0.817	-0.021	0.040	0.104
σ	1	1.022	9.49e-2	3.16e-4	1.30e-2	2.21e-3	0.958	0.832	1.020	1.205
α	0.3	0.299	3.10e-2	1.03e-4	2.23e-3	7.00e-4	0.844	0.240	0.299	0.364
θ_1	0.5	0.544	5.22e-2	1.74e-4	3.93e-3	1.14e-3	0.774	0.446	0.542	0.654
θ_2	2	2.247	4.30e-1	1.43e-3	5.16e-2	9.84e-3	0.912	1.565	2.193	3.227

Informative prior

	True Value	Mean	SD	Naive SE	MC Error	Batch SE	Batch ACF	2.5%	50%	97.5% Precentiles
μ	0	0.040	2.92e-2	9.75e-5	1.93e-3	6.52e-4	0.803	-0.017	0.040	0.100
σ	1	1.006	8.94e-2	2.98e-4	1.21e-2	2.08e-3	0.954	0.832	1.003	1.178
α	0.3	0.298	2.95e-2	9.83e-5	1.94e-3	6.62e-4	0.829	0.241	0.297	0.357
θ_1	0.5	0.539	5.02e-2	1.67e-4	3.65e-3	1.10e-3	0.766	0.444	0.537	0.644
θ_2	2	2.174	3.69e-1	1.23e-3	4.25e-2	8.40e-3	0.897	1.560	2.132	2.985

Table 4.2 gives the autocorrelation values before thinning for iterations 10,001-100,000 with lags 1, 10, 50, 100, 200, 500 and 1000. The lag1 autocorrelation values indicate that a batch procedure is necessary, however the lag1000 values indicate that the batch size 1000 is not enough for parameters σ and θ_2. Smith (2007) says that high autocorrelations suggest slow mixing of chains and, usually, slow convergence to the posterior distribution. However, Damien et al. (1999) exemplified that the introduction of auxiliary variables increases the autocorrelation, as it happens in this model.

Firstly, the convergence diagnostic method of Raftery and Lewis (1992) was considered. Sample size requirements were sought to ensure that posterior estimates of the $q = 0.025$ tail probabilities would be within an accuracy $r = \pm 0.02$ with probability equal to $s = 0.95$ with a precision of $\epsilon = 0.01$. The results suggest that 65,000 samples should be generated, the first 1,000 of which are discarded as a burn-in sequence and every 200 are saved as a thinning of the chain. The lower bound indicates that 235 independent samples are needed to estimate the posterior probability. The dependence factors are much larger than 5, indicating there are high autocorrelations.

The diagnostic of Heidelberger and Welch (1983) with level of confidence of 0.05 and accuracy of 0.1 indicates that, after a burn-in of 10,000, all the iterations are retained for posterior inference. Moreover, there is no significant evidence of non stationarity based on Cramer von Mises test statistics.

Finally, the convergence diagnostic of Gelman and Rubin (1992) and Brooks and Gelman (1998) is applied. For the purpose of assessing convergence, another parallel chain is sampled by using different starting values which are overdispersed with respect

Table 4.2: Autocorrelations.

	Non informative prior						
	lag1	lag10	lag50	lag100	lag200	lag500	lag1000
μ	0.993	0.934	0.728	0.549	0.342	0.089	0.017
σ	0.998	0.987	0.938	0.884	0.792	0.573	0.317
α	0.994	0.943	0.768	0.609	0.389	0.093	0.028
θ_1	0.990	0.912	0.660	0.486	0.348	0.118	0.030
θ_2	0.995	0.959	0.854	0.775	0.670	0.452	0.226
	Informative prior						
	lag1	lag10	lag50	lag100	lag200	lag500	lag1000
μ	0.992	0.930	0.709	0.526	0.303	0.044	0.015
σ	0.998	0.985	0.932	0.874	0.773	0.542	0.273
α	0.993	0.938	0.747	0.577	0.343	0.057	0.025
θ_1	0.989	0.908	0.648	0.469	0.326	0.097	0.023
θ_2	0.994	0.951	0.828	0.741	0.629	0.416	0.190

to the target distribution: $\mu = -1$, $\sigma = 2$, $\alpha = 0.7$, $\theta_1 = 2$ and $\theta_2 = 1$. The diagnostic provides the corrected scale reduction factors together with an upper limit (97.5%), the potential scale reduction factor, and the multivariate potential scale reduction factors. All of them are lower than 1.05, indicating convergence and that further simulations will not improve the values of the listed scalar estimators.

Although the introduction of auxiliary or latent variables increases the autocorrelation, as exemplified in Damien et al. (1999), the convergence here is satisfactory. Subsampling the output of a stationary Markov chain can result in poorer estimators (see MacEachern and Berliner (1994)), so that all the iterations after the burn-in have been used for estimation purposes both in non informative and informative scenarios.

Multiple datasets

Next, it is examined whether the parameter values are correctly recovered when simulating under various parameter values from AEP distributions. Some datasets were generated by $Y_i \sim \text{AEP}(\mu, \sigma, \alpha, \theta_1, \theta_2)$, $i = 1, \ldots, n$, by using the algorithm obtained from Proposition 1. The parameter values considered were $\mu = 0$, $\sigma = 1$, $\alpha = \{0.3, 0.5, 0.7\}$, $\theta_1, \theta_2 = \{0.5, 1, 2, 5\}$. Several sample sizes were considered, $n = \{100, 300, 500\}$. All the prior distributions were non informative, i.e. $\pi(\mu) \propto 1$, $\pi(\sigma) \propto 1/\sigma^2$, $\pi(\alpha) \propto 1$, $\pi(\theta_1) \propto 1$, $\pi(\theta_2) \propto 1$.

A total of 100,000 iterations were performed and 10,000 have been discarded as a burn-in sequence. The posterior means and standard deviations are reported in Table 4.3. Note that the posterior estimates with larger sample sizes provide means that are closer to the original values and smaller standard deviations. Some estimations are far from the original values. This usually happens when the sample size is not large enough and/or in extreme cases where θ_1 and/or θ_2 is large. It happens because it is difficult to distinguish between two pdfs or the pdf looks like a one-tail distribution

(where $\theta_1 \to \infty$ or $\theta_2 \to \infty$), and there is not enough information to distinguish the distribution shape. Table 4.3 also shows some NAs, what happens when the sample size is small, and the parameters lead to a one-tail distribution (for example when $(\theta_1, \theta_2) \in \{(0.5, 5), (1, 5)\}$ and $\alpha \in \{0.3, 0.5, 0.7\}$) or lead to an extremely platykurtic distribution (for example $(\theta_1, \theta_2) = (5, 5)$ and $\alpha \in \{0.3, 0.5, 0.7\}$). When the sample size is large this does not happen regardless of the values of θ_1 and θ_2.

Table 4.3: Estimations of the simulated data with distribution AEP($\mu = 0, \sigma = 1, \alpha, \theta_1, \theta_2$).

$\boldsymbol{\alpha}$, $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$	n	μ mean (s.d.)	σ mean (s.d.)	α mean (s.d.)	θ_1 mean (s.d.)	θ_2 mean (s.d.)
$\alpha = 0.3$	100	-0.02 (0.07)	1.01 (0.31)	0.26 (0.06)	0.58 (0.14)	0.53 (0.09)
$\theta_1 = 0.5$	300	-0.01 (0.02)	1.22 (0.19)	0.24 (0.03)	0.47 (0.05)	0.55 (0.05)
$\theta_2 = 0.5$	500	0.04 (0.03)	0.88 (0.12)	0.28 (0.03)	0.53 (0.05)	0.48 (0.03)
$\alpha = 0.3$	100	0.04 (0.06)	0.77 (0.23)	0.32 (0.06)	0.51 (0.09)	0.79 (0.20)
$\theta_1 = 0.5$	300	0.04 (0.04)	0.97 (0.13)	0.30 (0.04)	0.62 (0.08)	0.94 (0.12)
$\theta_2 = 1$	500	0.08 (0.03)	1.09 (0.12)	0.33 (0.03)	0.57 (0.05)	1.05 (0.12)
$\alpha = 0.3$	100	0.21 (0.13)	1.02 (0.18)	0.46 (0.11)	1.18 (0.37)	1.53 (0.41)
$\theta_1 = 0.5$	300	0.11 (0.05)	0.82 (0.10)	0.37 (0.05)	0.57 (0.06)	1.48 (0.26)
$\theta_2 = 2$	500	0.05 (0.02)	0.80 (0.07)	0.32 (0.03)	0.50 (0.04)	1.55 (0.19)
$\alpha = 0.3$	100	0.69 (0.35)	1.06 (0.20)	0.78 (0.26)	1.06 (0.35)	NA
$\theta_1 = 0.5$	300	0.03 (0.02)	0.93 (0.06)	0.29 (0.03)	0.50 (0.04)	5.69 (1.29)
$\theta_2 = 5$	500	0.03 (0.02)	0.94 (0.05)	0.28 (0.02)	0.52 (0.04)	4.54 (0.73)
$\alpha = 0.3$	100	0.00 (0.17)	1.31 (0.24)	0.28 (0.11)	2.00 (1.41)	1.22 (0.26)
$\theta_1 = 1$	300	0.04 (0.05)	0.87 (0.16)	0.32 (0.05)	0.82 (0.12)	0.94 (0.17)
$\theta_2 = 1$	500	0.07 (0.04)	0.91 (0.09)	0.33 (0.04)	0.94 (0.12)	0.96 (0.10)
$\alpha = 0.3$	100	0.22 (0.19)	1.12 (0.16)	0.48 (0.15)	2.06 (0.88)	2.24 (0.95)
$\theta_1 = 1$	300	0.14 (0.07)	0.81 (0.07)	0.43 (0.08)	1.08 (0.18)	1.29 (0.21)
$\theta_2 = 2$	500	0.21 (0.09)	0.93 (0.07)	0.44 (0.08)	1.08 (0.15)	1.76 (0.34)
$\alpha = 0.3$	100	NA	NA	NA	NA	NA
$\theta_1 = 1$	300	0.23 (0.13)	1.02 (0.07)	0.49 (0.12)	1.35 (0.30)	5.59 (2.38)
$\theta_2 = 5$	500	0.24 (0.10)	0.95 (0.06)	0.48 (0.10)	1.50 (0.25)	3.73 (1.09)
$\alpha = 0.3$	100	0.19 (0.12)	0.96 (0.16)	0.47 (0.11)	2.06 (0.80)	1.17 (0.30)
$\theta_1 = 2$	300	0.20 (0.13)	1.11 (0.07)	0.46 (0.11)	4.13 (1.45)	2.19 (0.46)
$\theta_2 = 2$	500	0.23 (0.08)	0.97 (0.05)	0.49 (0.07)	2.95 (0.55)	1.50 (0.20)
$\alpha = 0.3$	100	0.15 (0.18)	0.97 (0.12)	0.43 (0.17)	2.30 (1.09)	3.62 (2.09)
$\theta_1 = 2$	300	0.34 (0.14)	0.94 (0.05)	0.63 (0.14)	4.11 (1.22)	2.78 (0.98)
$\theta_2 = 5$	500	0.13 (0.07)	0.93 (0.05)	0.41 (0.07)	2.29 (0.45)	3.36 (0.64)
$\alpha = 0.3$	100	NA	NA	NA	NA	NA
$\theta_1 = 5$	300	NA	NA	NA	NA	NA
$\theta_2 = 5$	500	0.09 (0.11)	1.03 (0.04)	0.34 (0.11)	4.13 (1.38)	6.72 (1.71)

$\boldsymbol{\alpha}$, $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$	$\boldsymbol{n}$	$\boldsymbol{\mu}$ **mean (s.d.)**	$\boldsymbol{\sigma}$ **mean (s.d.)**	$\boldsymbol{\alpha}$ **mean (s.d.)**	$\boldsymbol{\theta_1}$ **mean (s.d.)**	$\boldsymbol{\theta_2}$ **mean (s.d.)**
$\alpha = 0.5$	100	-0.01 (0.07)	0.99 (0.32)	0.50 (0.06)	0.59 (0.12)	0.49 (0.08)
$\theta_1 = 0.5$	300	-0.01 (0.04)	1.06 (0.18)	0.43 (0.03)	0.46 (0.05)	0.54 (0.05)
$\theta_2 = 0.5$	500	0.07 (0.05)	1.01 (0.15)	0.49 (0.03)	0.54 (0.05)	0.50 (0.03)
$\alpha = 0.5$	100	0.04 (0.06)	0.79 (0.26)	0.52 (0.06)	0.49 (0.08)	0.72 (0.18)
$\theta_1 = 0.5$	300	0.05 (0.04)	0.98 (0.17)	0.49 (0.04)	0.58 (0.06)	0.99 (0.18)
$\theta_2 = 1$	500	0.04 (0.03)	0.96 (0.13)	0.49 (0.03)	0.47 (0.03)	1.02 (0.13)
$\alpha = 0.5$	100	0.06 (0.08)	0.81 (0.19)	0.52 (0.08)	0.44 (0.05)	2.60 (1.74)
$\theta_1 = 0.5$	300	0.15 (0.07)	0.90 (0.13)	0.57 (0.06)	0.50 (0.04)	1.96 (0.59)
$\theta_2 = 2$	500	0.10 (0.04)	0.67 (0.13)	0.55 (0.04)	0.47 (0.03)	1.16 (0.28)
$\alpha = 0.5$	100	0.62 (0.01)	1.08 (0.21)	0.99 (0.01)	0.94 (0.14)	NA
$\theta_1 = 0.5$	300	0.11 (0.10)	0.93 (0.12)	0.53 (0.09)	0.53 (0.05)	6.92 (3.06)
$\theta_2 = 5$	500	0.03 (0.02)	0.82 (0.06)	0.47 (0.03)	0.48 (0.03)	3.37 (0.62)
$\alpha = 0.5$	100	-0.05 (0.19)	1.42 (0.24)	0.43 (0.12)	1.60 (0.80)	1.38 (0.34)
$\theta_1 = 1$	300	0.02 (0.08)	0.93 (0.15)	0.48 (0.07)	0.87 (0.11)	1.10 (0.26)
$\theta_2 = 1$	500	0.08 (0.06)	1.06 (0.09)	0.52 (0.05)	1.04 (0.12)	1.09 (0.15)
$\alpha = 0.5$	100	0.04 (0.05)	0.59 (0.16)	0.56 (0.08)	0.66 (0.13)	1.11 (0.40)
$\theta_1 = 1$	300	0.08 (0.10)	0.98 (0.08)	0.52 (0.09)	1.12 (0.16)	2.08 (0.52)
$\theta_2 = 2$	500	0.05 (0.06)	0.90 (0.06)	0.50 (0.06)	0.96 (0.11)	2.19 (0.40)
$\alpha = 0.5$	100	NA	NA	NA	NA	NA
$\theta_1 = 1$	300	0.07 (0.06)	0.96 (0.09)	0.53 (0.06)	0.92 (0.13)	6.27 (2.67)
$\theta_2 = 5$	500	0.06 (0.05)	0.90 (0.06)	0.51 (0.05)	0.88 (0.09)	4.48 (1.04)
$\alpha = 0.5$	100	0.13 (0.13)	1.00 (0.16)	0.64 (0.12)	2.07 (0.70)	1.12 (0.39)
$\theta_1 = 2$	300	0.24 (0.15)	1.11 (0.07)	0.68 (0.13)	3.33 (0.88)	1.94 (0.67)
$\theta_2 = 2$	500	0.13 (0.06)	0.94 (0.05)	0.58 (0.06)	2.00 (0.25)	1.67 (0.24)
$\alpha = 0.5$	100	0.60 (0.01)	1.18 (0.09)	0.99 (0.01)	4.34 (1.16)	NA
$\theta_1 = 2$	300	0.18 (0.15)	1.01 (0.07)	0.64 (0.13)	2.25 (0.50)	4.60 (2.18)
$\theta_2 = 5$	500	0.28 (0.09)	0.93 (0.04)	0.75 (0.10)	2.35 (0.39)	4.26 (1.75)
$\alpha = 0.5$	100	NA	NA	NA	NA	NA
$\theta_1 = 5$	300	0.10 (0.17)	1.01 (0.05)	0.57 (0.17)	6.25 (2.61)	4.20 (1.80)
$\theta_2 = 5$	500	0.07 (0.13)	1.02 (0.04)	0.53 (0.12)	3.99 (1.04)	6.95 (2.32)
$\alpha = 0.7$	100	0.06 (0.10)	1.35 (0.40)	0.70 (0.07)	0.56 (0.08)	1.18 (0.55)
$\theta_1 = 0.5$	300	0.03 (0.04)	0.95 (0.19)	0.66 (0.04)	0.49 (0.05)	1.13 (0.27)
$\theta_2 = 1$	500	0.05 (0.03)	0.88 (0.17)	0.69 (0.03)	0.45 (0.03)	0.94 (0.17)
$\alpha = 0.7$	100	0.48 (0.08)	1.10 (0.23)	0.98 (0.05)	0.73 (0.10)	11.43 (1.55)
$\theta_1 = 0.5$	300	0.13 (0.05)	0.65 (0.15)	0.77 (0.06)	0.50 (0.05)	1.57 (0.71)
$\theta_2 = 2$	500	0.08 (0.05)	0.76 (0.10)	0.72 (0.05)	0.45 (0.02)	1.41 (0.39)
$\alpha = 0.7$	100	0.37 (0.01)	0.79 (0.23)	0.99 (0.01)	0.65 (0.11)	NA

$\alpha, \theta_1, \theta_2$	n	μ mean (s.d.)	σ mean (s.d.)	α mean (s.d.)	θ_1 mean (s.d.)	θ_2 mean (s.d.)
$\theta_1 = 0.5$	300	0.18 (0.06)	0.57 (0.16)	0.84 (0.07)	0.47 (0.05)	1.88 (1.16)
$\theta_2 = 5$	500	0.03 (0.04)	0.85 (0.07)	0.67 (0.04)	0.48 (0.02)	6.65 (2.38)
$\alpha = 0.7$	100	0.52 (0.01)	1.39 (0.17)	0.99 (0.01)	1.97 (0.39)	NA
$\theta_1 = 1$	300	0.12 (0.09)	1.09 (0.11)	0.76 (0.07)	1.22 (0.14)	1.95 (0.72)
$\theta_2 = 2$	500	0.02 (0.08)	0.87 (0.08)	0.68 (0.08)	0.91 (0.09)	2.15 (0.69)
$\alpha = 0.7$	100	0.35 (0.01)	0.86 (0.15)	0.99 (0.01)	1.24 (0.21)	NA
$\theta_1 = 1$	300	0.07 (0.07)	0.86 (0.08)	0.72 (0.07)	0.94 (0.10)	3.78 (1.83)
$\theta_2 = 5$	500	0.12 (0.04)	0.77 (0.07)	0.78 (0.05)	0.88 (0.06)	2.71 (0.92)
$\alpha = 0.7$	100	0.29 (0.01)	0.93 (0.11)	0.99 (0.01)	2.65 (0.68)	NA
$\theta_1 = 2$	300	0.10 (0.14)	1.05 (0.07)	0.78 (0.12)	2.58 (0.55)	5.31 (3.79)
$\theta_2 = 5$	500	0.07 (0.07)	0.92 (0.07)	0.75 (0.07)	1.69 (0.20)	4.61 (2.13)

The next section shows an application of the AEP distribution to linear regression.

4.4 Linear regression model with AEP error

A linear regression model whose error variable has asymmetric exponential power distribution is proposed.

4.4.1 Background

In the statistical literature, inference on linear regression models has been studied from a Bayesian point of view under the assumption that the error terms are symmetrically or asymmetrically distributed. Most of the research has been developed under a multivariate spherical normal distribution for the error vector. However, many authors provide data that can not be properly fitted by using a linear regression with normal distribution for the errors. This mainly happens because the possible shapes of the tails and the symmetry make the model not flexible enough to handle some types of data. Box and Tiao (1962) commented that in many problems the particular physical set-up is such that the errors involved might behave like a linear aggregate of component errors and, consequently, a central limit effect would operate. In fact, of course, the central limit theorem does not imply that a linear aggregate of a finite number of component errors would be exactly normal. For example, Blattberg and Gonedes (1974) observed that empirical distributions for stock prices had more kurtosis (heavier tails) than those predicted by the normal distribution, so that they considered another family of symmetric distributions that can also account for heavy tail distributions. Several robust models have been defined, see, for instance, Huber (1981) and Tiku et al. (1986). In addition, statistical inference based on the normal distribution is known to be vulnerable to outliers, what has increased the need to develop procedures directed at detecting outliers. Lange et al. (1989) illustrated the

ability of models based on the t distribution to handle outliers in a wide range of settings.

Some multivariate models considered in the recent literature (see Azzalini and Dalla-Valle (1996) and Arnold and Beaver (2000), among others) were designed to handle asymmetry. For example, Adcock (2010) discussed how some of the returns on most financial assets exhibit kurtosis and many also have skew distributions. In that paper, a general multivariate model for the probability distribution of assets returns, which incorporates both kurtosis and skewness, is described. Arellano-Valle and Genton (2005) discussed the increasing interest in finding more flexible methods to represent features of the data as adequately as possible and to reduce unrealistic assumptions. They highlight the book of Genton (2004) which provided a collection of applications in several areas where skew distributions are considered.

In many practical regression problems, a suitable transformation for symmetry is often considered for skewed data. For instance, Box and Cox (1964) discussed procedures for finding transformations in linear models. They made the assumption that a normal distributed, homoscedastic, linear model is appropriate after some suitable transformation applied to the outcome variable. The application of these transformations is sometimes useful. However, such a transformation can be difficult to define, or sometimes it is important to work with the original variables or scales.

Thus, several linear models have been defined to deal with flexible distributions of datasets, with skewness as well as tails that are lighter or heavier than those of the normal distribution, where the attention is focused on the robustness of the inferential theory. Bayesian inference in regression models with elliptical non-normal errors was initially presented by Box and Tiao (1973), who considered exponential power distributions for the error terms. Later, Zellner (1976) considered the multivariate Student's t regression distributions with zero location vector and scalar dispersion matrix of linear multiple regression models. The multivariate Cauchy and normal distributions are special cases. Fernández and Steel (1998) considered a Bayesian analysis of linear regression models that can account for skewed error distributions with heavy tails. They proposed a general procedure for introducing skewness into symmetric distributions, where the tail behavior is not affected. The main focus is the linear regression model with skewed Student's t error term. Arellano-Valle et al. (2000) reviewed and extended some results related to Bayesian analysis for elliptical linear models presented in the literature. Sahu et al. (2003) implemented a posterior regression analysis by considering skewed distributions for the error terms. They developed a new class of distributions by introducing skewness in multivariate elliptically symmetric distributions. This class contains many standard families including the multivariate skew normal and Student's t distributions. Recently, Arellano-Valle et al. (2008) introduced a class of shape mixtures of skewed distributions and studied some of its main properties. Specifically, they developed a Bayesian analysis of the skew-normal, skew-generalized-normal, skew-normal-t and skew-t-normal linear regression models under some special prior specifications for the model parameters. Other regression models with more flexible distributions of the error variable have been defined by Fernández and Steel (1999, 2000); Ma et al. (2004); Arellano-Valle et al. (2007); Ferreira and Steel (2007), among others.

In the next subsection, a Bayesian approach to a linear regression model is proposed. The error terms are identically and independently distributed as an AEP distribution, being symmetric or asymmetric, with light and/or heavy tails. This flexibility leads to a robust model, where it is possible to consider the skewness of the data while avoiding transformations. The implementation of the approach is based on Proposition 4.1, which is used to derive an efficient Gibbs sampling algorithm, which shares some full conditional distributions with the model presented in the previous section.

4.4.2 The approach

Suppose that n independent random variables are observed $y_1, \ldots, y_n$, and they are related to a set of covariables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$, for $i = 1, \ldots, n$, through the following linear regression model

$$y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where

$$\epsilon_i \quad \sim \quad \text{AEP}(0, \sigma, \alpha, \theta_1, \theta_2).$$

Note that the linear regression model usually includes an intercept, i.e.: $x_{i1} = 1$. The prior distribution for the regression vector $\boldsymbol{\beta}$ is the usual choice in the linear regression context, i.e.: a multivariate normal distribution, $\text{N}_k(\mathbf{b}, \mathbf{B})$.

In this model the mixture representation of the AEP distribution presented in Proposition 4.1 is considered. By using the notation introduced in the previous section, the joint posterior density of the unobservables $\mathbf{u}_1 = (u_{11}, \ldots, u_{1n})^T$, $\mathbf{u}_2 = (u_{21}, \ldots, u_{2n})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, σ, α, θ_1 and θ_2 given the data $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is given by

$$\begin{aligned}
&\text{p}(\mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}, \sigma, \alpha, \theta_1, \theta_2 | \mathbf{y}, \mathbf{x}) \propto \pi(\boldsymbol{\beta})\pi(\sigma)\pi(\alpha)\pi(\theta_1)\pi(\theta_2)\sigma^{-n} \\
&\quad \times \quad \prod_{i=1}^{n} \Bigg(\exp(-u_{1i}) I\Big[\mathbf{x}_i^T\boldsymbol{\beta} - \tfrac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1} < y_i \leq \mathbf{x}_i^T\boldsymbol{\beta}\Big] \\
&\quad + \quad \exp(-u_{2i}) I\Big[\mathbf{x}_i^T\boldsymbol{\beta} < y_i < \mathbf{x}_i^T\boldsymbol{\beta} + \tfrac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\Big] \Bigg).
\end{aligned}$$

Since the prior distribution of $\boldsymbol{\beta}$ is multivariate normal, then the conditional distribution of β_j given $\boldsymbol{\beta}_{(-j)}$ is normal, where $\boldsymbol{\beta}_{(-j)}^T = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)$. The subscript $(-j)$ denotes that the jth element has been removed. Then the posterior distribution of β_j given $\mathbf{y}$, $\mathbf{x}$, $\mathbf{u}_1$, $\mathbf{u}_2$, $\boldsymbol{\beta}_{(-j)}$, σ, α, θ_1 and θ_2 is given by

$$\big[\beta_j | \mathbf{y}, \mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}_{(-j)}, \sigma, \alpha, \theta_1, \theta_2\big] \sim \text{N}(\text{b}_j^* \ , \ \text{B}_j^*) I\Big[\beta_j \in \Big(\underline{\beta}_j \ , \ \overline{\beta}_j\Big)\Big], \tag{4.8}$$

for $j = 1, \ldots, k$, where

$$\begin{aligned}
\text{b}_j^* &= \mathbf{b}_j - \mathbf{B}_{j(-j)}\mathbf{B}_{(-j)(-j)}^{-1}\big(\boldsymbol{\beta}_{(-j)} - \mathbf{b}_{(-j)}\big), \\
\text{B}_j^* &= \mathbf{B}_{jj} - \text{B}_{j(-j)}\mathbf{B}_{(-j)(-j)}^{-1}\mathbf{B}_{(-j)j},
\end{aligned}$$

and

$$\begin{aligned}
\underline{\beta}_j &= \max\left\{ \max_{\{i:u_{1i}>0, x_{ij}<0\}} \left\{ \frac{y_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}}{x_{ij}} + \left(\frac{\frac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1}}{x_{ij}} \right) \right\}, \right. \\
&\quad \left. \max_{\{i:u_{2i}>0, x_{ij}>0\}} \left\{ \frac{y_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}}{x_{ij}} - \left(\frac{\frac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}}{x_{ij}} \right) \right\} \right\}, \\
\overline{\beta}_j &= \min\left\{ \min_{\{i:u_{1i}>0, x_{ij}>0\}} \left\{ \frac{y_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}}{x_{ij}} + \left(\frac{\frac{\alpha\sigma}{\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1}}{x_{ij}} \right) \right\}, \right. \\
&\quad \left. \min_{\{i:u_{2i}>0, x_{ij}<0\}} \left\{ \frac{y_i - \mathbf{x}_{i(-j)}^T \boldsymbol{\beta}_{(-j)}}{x_{ij}} - \left(\frac{\frac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}}{x_{ij}} \right) \right\} \right\}.
\end{aligned}$$

By considering $\pi(\boldsymbol{\beta}) \propto 1$, then

$$\left[\beta_j | \mathbf{y}, \mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}_{(-j)}, \sigma, \alpha, \theta_1, \theta_2\right] \sim \mathrm{U}\left(\underline{\beta}_j \, , \, \overline{\beta}_j\right).$$

Sampling from these two full conditional distributions for the regression parameters is straightforward. The final algorithm for this linear regression approach consists of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\sigma^{(0)}$, $\alpha^{(0)}$, $\theta_1^{(0)}$ and $\theta_2^{(0)}$, and iteratively sampling $\mathbf{u}_1^{(j)}$ and $\mathbf{u}_2^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\sigma^{(j)}$, $\alpha^{(j)}$, $\theta_1^{(j)}$ and $\theta_2^{(j)}$ from the full conditional distributions (4.2), (4.8), (4.4), (4.5), (4.6) and (4.7), respectively, replacing μ by $\mathbf{x}_i^T\boldsymbol{\beta}$ in (4.2), (4.4), (4.5), (4.6) and (4.7). Note that 5 out of the 6 full conditional distributions are shared with the model presented in Section 4.3.

4.4.3 Australian athletes dataset

A dataset from Cook and Weisberg (1994) on characteristics of Australian athletes, available from the Australian Institute of Sport (AIS), is considered in order to illustrate the AEP linear regression model performance. Specifically, the variables lean body mass (Lbm), height (Ht) and weight (Wt), associated with $n = 102$ Australian male athletes, are considered. The variables are highly correlated. Figure 4.2 displays the data. It shows the skewed dispersion of the data having asymmetric and heavy tails.

Arellano-Valle et al. (2008) developed a skew normal regression model without intercept to study the relationship between the athletes' lean body mass and their height and weight. Now, a model comparison framework is considered by using

$$\mathrm{Lbm}_i = \beta_1 \mathrm{Ht_i} + \beta_2 \mathrm{Wt}_i + \epsilon_i, \quad i = 1, \ldots, 102,$$

where both symmetrical and asymmetrical distributions for the error variable are considered. Specifically, the model with normal distribution (Normal), the skew normal regression model (SN) proposed by Arellano-Valle et al. (2008), and the AEP model proposed in this paper (AEP) are fitted. The criteria for model comparison will be the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), the Bayesian information criterion (BIC) proposed by Schwarz (1978), and the Akaike

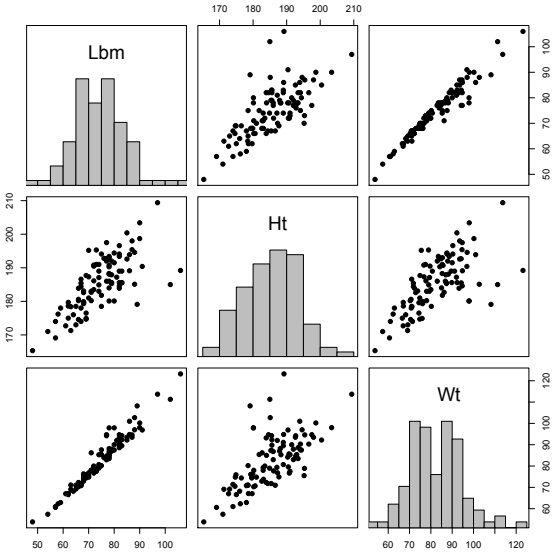

Figure 4.2: Australian athletes dataset.

information criterion (AIC) proposed by Akaike (1973). These criteria are evaluated as follows: $BIC = \overline{D(\eta)} + p\log(n)$, $AIC = \overline{D(\eta)} + 2p$, and $DIC = \overline{D(\eta)} + \widehat{p_D}$, where $D(\eta) = -2\log L(\eta)$ is the deviance of the model, $L(\eta)$ is the likelihood, $\overline{D(\eta)} = \mathrm{E}(D(\eta)|\text{data})$ is the posterior mean of the deviance, p is the number of parameters in the model, n is the sample size, $\widehat{p_D} = \overline{D(\eta)} - D(\overline{\eta})$ is the effective number of parameters, and $D(\overline{\eta})$ is the deviance at the posterior means of the parameters of interest $\overline{\eta} = \mathrm{E}(\eta|\text{data})$. Models with smaller values should be preferred over models with larger values.

For all the considered regression models, the prior distribution for $\boldsymbol{\beta}$ is a multivariate normal $\mathrm{N}_2(\mathbf{b}, \mathbf{B})$ with parameters $\mathbf{b}^T = (0, 0)$ and $\mathbf{B} = \mathrm{diag}(100, 100)$, and the prior distribution of the scale parameter is the improper $\pi(\sigma^2) \propto 1/\sigma^2$. The prior distributions for the other parameters are specified as follows: for the skew normal model, the prior distribution of the skew parameter is $\alpha \sim \mathrm{N}(0, 10)$; for the AEP linear regression model, the priors of the skew parameter and the shape parameters are improper (as presented in this section).

A total of $400,000$ iterations have been performed with a burn-in of $50,000$. With this specification, a convergence analysis showed that the chain converged. However, high autocorrelations have been obtained. This fact has made a long sample size necessary. By thinning the iterations with lag 10 and 100, very similar estimations have been obtained. Subsampling the output of a stationary Markov chain can result in poorer estimators, as MacEachern and Berliner (1994) have shown. Besides, this produces needless higher computational costs, as Kacperczyk et al. (2013) illustrated. Therefore, all iterations after the burn-in have been used.

Table 4.4 displays the summary of posterior estimations for the parameters and

the values for the three model comparison criteria (BIC, AIC and DIC). The estimated measures show that the model with AEP error gives the best performance with the three criteria, whereas the normal model is the poorest one. The data are better supported by asymmetrical errors. Besides, Figure 4.3 shows the histograms of the residuals, which confirm the skewness. The AEP model provides a very flexible alternative to fit these asymmetrical data. As a counterpart, the high autocorrelations make the sampling of a long chain necessary, as stated previously.

Table 4.4: Summary of posterior estimations and criteria values for the model parameters fitted to the Australian athletes dataset.

Model	Parameter	Mean	Standard deviation	95% HPD interval	Criteria
Normal	β_1	0.077	0.010	(0.057,0.096)	BIC=469.87
	β_2	0.731	0.022	(0.688,0.774)	AIC=461.92
	σ^2	5.328	0.774	(3.945,6.906)	DIC=461.96
SN	β_1	0.055	0.012	(0.033,0.079)	BIC=454.68
	β_2	0.811	0.028	(0.754,0.861)	AIC=444.18
	σ	3.517	0.319	(2.896,4.158)	DIC=443.68
	α	-4.227	1.377	(-7.160,-1.830)	
AEP	β_1	0.037	0.009	(0.019,0.056)	BIC=441.54
	β_2	0.833	0.022	(0.793,0.881)	AIC=428.41
	σ	2.836	0.562	(1.773,3.957)	DIC=424.11
	α	0.601	0.156	(0.345,0.886)	
	θ_1	0.764	0.159	(0.472,1.085)	
	θ_2	1.738	0.880	(0.471,3.455)	

Predictions have also been studied. Figure 4.4 displays the mean values and the 2.5% and 97.5% quantiles of 10 predicted data selected in a stratified way (the line represents the identity). The asymmetry is clearly showed through the quantiles. With the normal model, although the interval ranges are smaller, the upper bounds of the intervals are larger than those from the asymmetric models, and they are further away from the observed values. It is also reported that the estimated values from the SN and AEP models are closer to the observed values than those from the Normal model. Nevertheless, the AEP intervals have larger left limits than the SN ones, showing more flexibility for the AEP distribution. This matches the results from the residuals in Figure 4.3. The residuals show how the three models present asymmetry. The normal model shows a lack of fit. The residuals are leptokurtic and asymmetric, providing a heavy left tailed distribution. In the case of the SN model, the fit is better than in the normal case, but its residuals show a more leptokurtic shape. Finally, the residuals of the AEP model show a much better fit.

The possibility of including the intercept parameter in the models has been considered. In this case, the models with intercept do not provide improvements in terms of better fittings and the space parameter is needlessly increased. Besides, value zero is included in the 95% HPD intervals of the intercept parameter for the three models. The inclusion of the intercept term increases the consequences of the multicollinearity

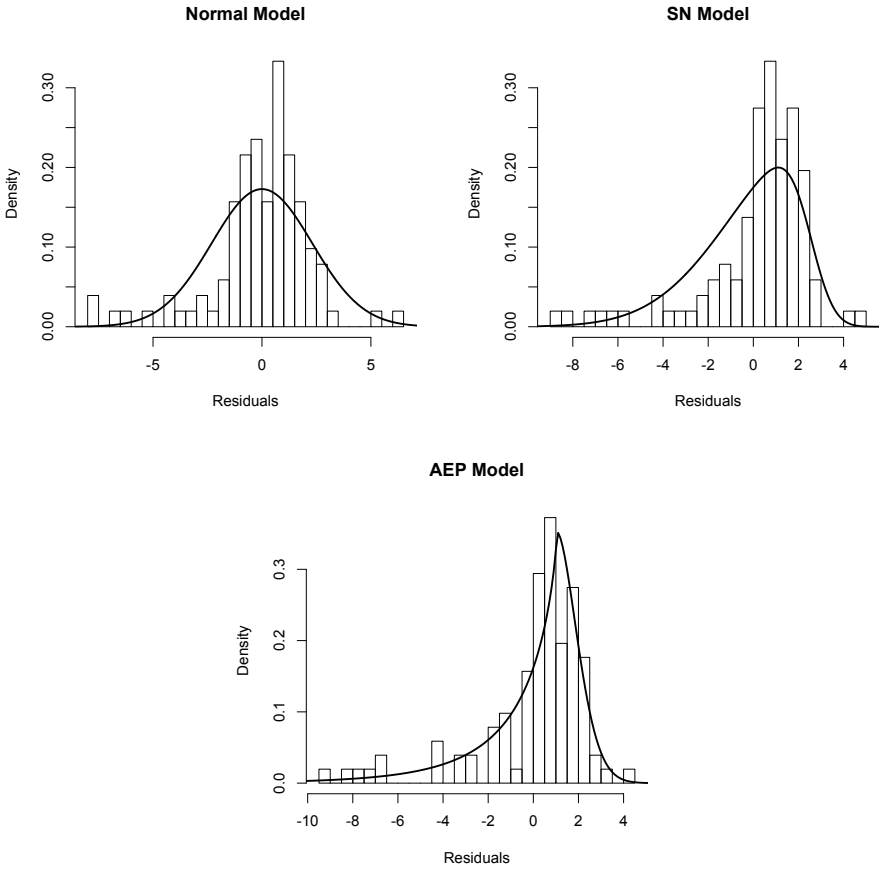

Figure 4.3: Residuals of the models fitted to the Australian athletes dataset.

problem.

The multicollinearity is an important issue that must be considered in linear regression. It can result in large standard errors for the regression parameters or regression parameters with signs that are the opposite of those expected. Curtis and Ghosh (2011) said that multicollinearity is a problem with the data and not necessarily a problem with statistical methodology. The linear regression model assumes each predictor has an independent effect on the response that can be encapsulated in the regression parameter. When predictors are highly correlated, the data do not contain much information on the independent effects of each variable. The data are deficient for determining the independent effects of a covariate on the response because the covariates themselves are not independent. Therefore one solution is to obtain better data, however this is not usually feasible. Sometimes a more realistic approach is to initially impose linear restrictions on the coefficients, such as specifying that one parameter is zero or to force two covariates into a group. Another solution is using

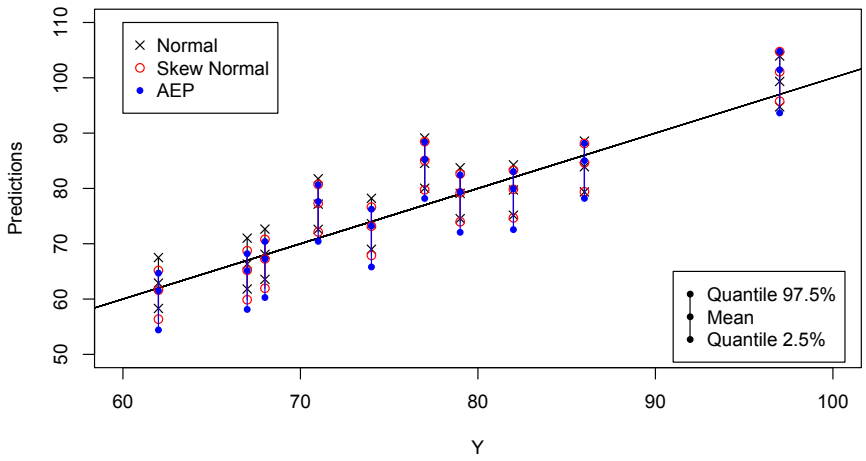

Figure 4.4: Predictions of the models fitted to the Australian athletes dataset.

an algorithm, as the Bayesian model proposed by Curtis and Ghosh (2011), that accounts for correlation among the predictors by simultaneously performing predictor selection and clustering.

In this example, the effect of the multicollinearity with the AEP error distribution is similar to the effect provided by the model based on the normal distribution. In general, the problem of multicollinearity in the AEP linear regression model should be reviewed as in the normal case, because it can similarly result in large standard errors for the regression parameters or in imprecise estimates. Moreover, depending on the data, the multicollinearity can highly affect the convergence in the same way as it happens with the normal model.

In conclusion, linear regression models based on error terms with AEP distribution turn out to be robust, allowing to deal with problems with asymmetry and heavy tails. Simulation studies were conducted showing the good performance of the proposed approach.

4.5 Binary regression model with AEP-based link function

A binary regression model, using the inverse of the asymmetric exponential power cumulative distribution function (cdf) as the link function, is proposed in this section.

4.5.1 Background

Suppose that n independent binary random variables $y_1, \ldots, y_n$ are observed, where y_i is Bernoulli distributed with success probability $\mathrm{p}(y_i = 1|\boldsymbol{\beta}, \mathbf{x}_i) = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$. $\boldsymbol{\beta}$ is a $k\times 1$ vector of unknown parameters, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ is a vector of known covariates, and Ψ is a known nonnegative function ranging between 0 and 1. The standard approach to modeling the dependence of binary data on explanatory variables under the generalized linear model setting is performed through a cdf Ψ. For instance, the probit model is obtained when Ψ is the standard normal cdf and the logit model when Ψ is the logistic cdf. See, for example, Cox (1971), McCullagh and Nelder (1989), and Collett (1991).

A wide choice of link functions is available. The most popular model for binary response data is the logistic regression based on the logit link. Other frequently used links include the probit and complementary log-log links. However, these do not always provide the best fit for a given dataset. Asymmetric links are sometimes good alternatives. For example, to describe a link, Chen et al. (1999a) considered the rates at which the probabilities of a given binary response approach 1 or 0. Under this notion, a link is symmetric if the rates are the same, otherwise the link is skewed or asymmetric. A skewed link can be characterized as positively skewed if the rate approaching 1 is faster than the rate approaching 0, otherwise it is negatively skewed. Chen et al. (1999a) suggested that an asymmetric link-based model may be more appropriate than a symmetric link model when the number of 1's and 0's are much different. Czado and Santner (1992) showed that a misspecification of the link function can result in a substantial increase in bias and mean square error of the success probability and regression parameter estimates.

Several link functions, more flexible than the logit and probit ones, have been defined. A seminal paper on binary regression was written by Albert and Chib (1993), who proposed a data augmentation framework for the Bayesian probit model by using Gibbs sampling. They also proposed extensions of the probit link by using mixtures of normal or Student's t distributions. The logit regression is studied as a particular case. Chen and Dey (1998) considered using a scale mixture of multivariate normal links to model binary responses when binary observations are taken from the same individuals or are taken over time in a longitudinal fashion. These include multivariate probit, Student's t link, logit, symmetric stable link, and exponential power link. Chen et al. (1999a) proposed a class of skewed link models, where the underlying latent variable has a mixed-effects model structure. However, the model has the limitation that the intercept term is confounded with the skewness parameters. Recently, Kim et al. (2008) introduced a link based on a generalized t distribution, which overcomes the problem of Chen et al. (1999a). Bazán et al. (2010) reviewed several asymmetrical links for binary regression models and presented a unified approach for two skew-probit links proposed in the literature. Other link functions have been defined by Aranda-Ordaz (1981); Stukel (1988); Czado (1994); Basu and Mukhopadhyay (2000a), among others. Besides, there are many other distribution functions that can be used to define asymmetrical link functions (see e.g. Zhu and Galbraith (2010)).

In the next subsection, a Bayesian approach to a binary regression model is proposed. The inverse of an AEP cdf is used as the link function. The implementation

of the approach is based on two main ideas. The first one is to develop a data augmentation framework by introducing latent variables in a similar way to Albert and Chib (1993). The second one is to use the scale mixture of uniform representation in Proposition 4.1. These two ideas are exploited to derive an efficient Gibbs sampling algorithm, sharing some full conditional distributions with the models presented in the two previous sections.

4.5.2 The approach

An AEP-based link model is assumed, $p_i = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$, where Ψ is the cdf of the distribution $\text{AEP}(0, \sigma, \alpha, \theta_1, \theta_2)$. Independent latent variables $z_1, \ldots, z_n$ are introduced, where z_i given $\boldsymbol{\beta}$, σ, α, θ_1 and θ_2 is distributed as $\text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma, \alpha, \theta_1, \theta_2)$, and one defines $y_i = 1$ if $z_i > 0$, and $y_i = 0$ if $z_i \leq 0$. With the random variables u_{1i} and u_{2i}, the distribution of z_i is written as a mixture

$$[z_i|\mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}, \sigma, \alpha, \theta_1, \theta_2] \sim \begin{cases} \text{U}\left(\mathbf{x}_i^T\boldsymbol{\beta} - \frac{\alpha\sigma}{\Gamma(1+1/\theta_1)}u_{1i}^{1/\theta_1}\ ,\ \mathbf{x}_i^T\boldsymbol{\beta}\right) & \text{with probability } \alpha \\ \text{U}\left(\mathbf{x}_i^T\boldsymbol{\beta}\ ,\ \mathbf{x}_i^T\boldsymbol{\beta} + \frac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)}u_{2i}^{1/\theta_2}\right) & \text{with probability } 1-\alpha \end{cases},$$

and

$$[u_{1i}|\theta_1] \sim \text{Ga}\left(1 + 1/\theta_1, 1\right), \quad [u_{2i}|\theta_2] \sim \text{Ga}\left(1 + 1/\theta_2, 1\right).$$

Note that if $z_i \sim \text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma, \alpha, \theta_1, \theta_2)$, where σ is unknown, the following equality holds

$$\begin{aligned} \text{p}(y_i = 1) &= \text{p}(z_i > 0) = \text{p}\left(\frac{z_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma} > -\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\right) \\ &= \text{p}(z_i^* - \mathbf{x}_i^T\boldsymbol{\beta}^* > -\mathbf{x}_i^T\boldsymbol{\beta}^*) = \text{p}(z_i^* > 0), \end{aligned}$$

where $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sigma$ and $z_i^* \sim \text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}^*, 1, \alpha, \theta_1, \theta_2)$. It is apparent that $\boldsymbol{\beta}$ and σ cannot be separately identified. Then, without loss of generality, from now on, $\sigma = 1$. Moreover, if $z_i \sim \text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}, 1, \alpha, \theta_1, \theta_2)$, then the following equality holds: if $0 < \mathbf{x}_i^T\boldsymbol{\beta}$, then $\text{p}(z_i < 0) = \text{p}(z_i^* < 0)$, where $\boldsymbol{\beta}^* = \boldsymbol{\beta}/(\alpha/2)$ and $z_i^* \sim \text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}^*, 1, 1/2, \theta_1, \theta_2)$, and if $\mathbf{x}_i^T\boldsymbol{\beta} < 0$, then $\text{p}(z_i > 0) = \text{p}(z_i^* > 0)$, where $\boldsymbol{\beta}^* = \boldsymbol{\beta}/((1-\alpha)/2)$ and $z_i^* \sim \text{AEP}(\mathbf{x}_i^T\boldsymbol{\beta}^*, 1, 1/2, \theta_1, \theta_2)$. Then set, without loss of generality, from now on, $\alpha = 1/2$. This means that the shape of the tail related with the observations equal to zero is modeled with the parameter θ_1, and the shape of the tail related with the observations equal to one is independently modeled with the parameter θ_2. If observed data are grouped, the model can be used by ungrouping the data, and the parameters θ_1 and θ_2 are related with proportions lower than 0.5 and greater than 0.5, respectively.

Then, the joint posterior density of the unobservable variables $\mathbf{z} = (z_1, \ldots, z_n)^T$, $\mathbf{u}_1 = (u_{11}, \ldots, u_{1n})^T$, $\mathbf{u}_2 = (u_{21}, \ldots, u_{2n})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, θ_1 and θ_2 given the

data $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is

$$\begin{aligned}
\mathrm{p}(\mathbf{z}, \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}, \theta_1, \theta_2 | \mathbf{y}, \mathbf{x}) &\propto \pi(\boldsymbol{\beta})\pi(\theta_1)\pi(\theta_2) \\
&\times \prod_{i=1}^{n} \Bigg(\Big(\exp(-u_{1i}) I\Big[\mathbf{x}_i^T\boldsymbol{\beta} - \tfrac{1}{2\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1} < z_i \leq \mathbf{x}_i^T\boldsymbol{\beta}\Big] \\
&+ \exp(-u_{2i}) I\Big[\mathbf{x}_i^T\boldsymbol{\beta} < z_i < \mathbf{x}_i^T\boldsymbol{\beta} + \tfrac{1}{2\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\Big] \Big) \\
&\times \big(I[z_i > 0] I[y_i = 1] + I[z_i \leq 0] I[y_i = 0] \big) \Bigg).
\end{aligned}$$

Note that this joint posterior distribution is complicated in the sense that it is difficult to normalize and sample from directly. But the posterior distribution can be estimated by using Gibbs sampling.

The full conditional distributions of the u_{1i} and u_{2i}, for $i = 1, \ldots, n$, are represented as

$$[u_{1i}|\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \theta_1, \theta_2] \sim \mathrm{Exp}(1) I\left[u_{1i} > \left(\frac{\max\{0\ ,\ \mathbf{x}_i^T\boldsymbol{\beta} - z_i\}}{1/2\Gamma(1+1/\theta_1)}\right)^{\theta_1}\right], \tag{4.9}$$

$$[u_{2i}|\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \theta_1, \theta_2] \sim \mathrm{Exp}(1) I\left[u_{2i} > \left(\frac{\max\{0\ ,\ z_i - \mathbf{x}_i^T\boldsymbol{\beta}\}}{1/2\Gamma(1+1/\theta_2)}\right)^{\theta_2}\right], \tag{4.10}$$

where Exp(1) denotes the exponential distribution with parameter equal to 1.

The full conditional distributions of the z_i, for $i = 1, \ldots, n$, are

$$[z_i|\mathbf{y}, \mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \boldsymbol{\beta}, \theta_1, \theta_2] \tag{4.11}$$

$$\sim \begin{cases} \mathrm{U}\Big(\max\Big\{0, \mathbf{x}_i^T\boldsymbol{\beta} - \frac{1}{2\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1}\Big\}\ ,\ \max\Big\{0, \mathbf{x}_i^T\boldsymbol{\beta} + \frac{1}{2\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\Big\} \Big) & \text{if } y_i = 1 \\ \mathrm{U}\Big(\min\Big\{0, \mathbf{x}_i^T\boldsymbol{\beta} - \frac{1}{2\Gamma(1+1/\theta_1)} u_{1i}^{1/\theta_1}\Big\}\ ,\ \min\Big\{0, \mathbf{x}_i^T\boldsymbol{\beta} + \frac{1}{2\Gamma(1+1/\theta_2)} u_{2i}^{1/\theta_2}\Big\} \Big) & \text{if } y_i = 0 \end{cases}.$$

Note that the distributions given in (4.9), (4.10) and (4.11) are standard and their sampling is straightforward.

The final algorithm consists of choosing initial values $\mathbf{z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\theta_1^{(0)}$ and $\theta_2^{(0)}$, and iteratively sampling $\mathbf{u}_1^{(j)}$, $\mathbf{u}_2^{(j)}$, $\mathbf{z}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\theta_1^{(j)}$ and $\theta_2^{(j)}$ from the full conditional distributions (4.9), (4.10), (4.11), (4.8), (4.6) and (4.7), respectively, replacing μ by $\mathbf{x}_i^T\boldsymbol{\beta}$ and y_i by z_i in (4.6) and (4.7) and replacing y_i by z_i in (4.8). In this case, 3 out of 6 full conditional distributions are shared with the models presented in the two previous sections.

4.5.3 Beetle mortality dataset

Bliss (1935) reported the results of a toxicological experiment based on the number of beetles dead after hours of exposure to carbon disulphide at various concentrations (transformed to $\log_{10}$ dose). The model of interest is given by

$$\eta_i = \Psi^{-1}(p_i) = \beta_0 + \beta_1 \log_{10} \mathrm{dose}_i, \quad i = 1, \ldots, 8,$$

where the observed binary outcome y_i is the death ($y_i = 1$) or survival ($y_i = 0$) of the flour beetle under the $i-$th dose, so that p_i is the proportion of flour beetle deaths under the $i-$th dose.

A model comparison framework is considered by using different cdfs Ψ. Symmetric and asymmetric link models are fitted to the data. Specifically, the following models are considered: the probit and t-link models proposed by Albert and Chib (1993), the skewed probit model (SP) proposed by Chen et al. (1999a), the skewed generalized t-link model (SGT) proposed by Kim et al. (2008), and the AEP-based link model proposed in this paper. The criteria for model comparison will be the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), the Bayesian information criterion (BIC) proposed by Schwarz (1978), the Akaike information criterion (AIC) proposed by Akaike (1973), and the absolute errors (AE) defined by $\sum_i |p_i - \widehat{p}_i|$. Criteria BIC, AIC and DIC have been evaluated as in subsection 4.4.3. The smaller the values of the criteria, the better the model.

The prior distribution for $\boldsymbol{\beta}$ for all the fitted regression models is a multivariate normal with parameters $\mathbf{b}^T = (-32.4, 18.3)$ (the estimates of $\boldsymbol{\beta}$ from a non informative probit model) and $\mathbf{B} = \text{diag}(10, 10)$. The prior distributions for the other parameters are specified as follows. For the t-link model, the degrees of freedom parameter is fixed and equal to 8 (approximation to the logit model). For the skewed probit model, the skew parameter is normally distributed, $\delta \sim \text{N}(0, 1)$. For the skewed generalized t-link model, the degrees of freedom parameter $\nu_1 \sim \text{Ga}(1, 0.1)$ and the parameter $\delta \sim \text{U}(0, 1)$, where $\nu_2 = 1/\delta^2$ is the scale parameter. For the AEP-based link model, the skew parameter is a fixed value, $\alpha = 0.5$, and the shape parameters $\theta_1 \sim \text{Ga}(1, 1)$ and $\theta_2 \sim \text{Ga}(1, 1)$.

For the AEP model, a total of one million iterations of MCMC are performed and a burn-in of $500,000$ is considered. With this specification all the chains seem to have converged. High autocorrelations have been obtained due to the model specification and also because the data have been ungrouped; note that the model was defined for binary data, but the data were originally represented as binomial. This is the reason for a large sample size and for a long burn-in. By thinning the iterations, very similar estimates have been obtained. However, all iterations after the burn-in have been used.

Table 4.5 displays the posterior mean, standard deviation and HPD interval of the model parameters and the values for the model comparison criteria (BIC, AIC, DIC, and AE). In case of SGT model, a reparameterization $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\delta$ is needed to properly compare the regression parameters by using the same measure scale, so that $\boldsymbol{\beta}^* = (-37.94, 21.46)^T$. The estimated measures show that the AEP-based link provides the best performance, whereas the T(8) model gives the worst one. The data are better supported by skewed models, of which the AEP distribution is the most flexible.

Figure 4.5 shows the data and the posterior estimates of the death proportions. Note how the proposed model is able to properly fit the proportions. Special attention should be paid to the discrepancy of the estimates in the central part of the logDose scale, i.e. around 1.75-1.80.

Simulation studies were conducted showing the good performance of the proposed

Table 4.5: Summary of posterior estimations and criteria values for the model parameters fitted to the beetle mortality dataset.

Model	Parameter	Mean	Standard deviation	95% HPD interval	Criteria	
Probit	β_0	-33.88	1.91	$(-37.55, -30.25)$	BIC=40.49	DIC=39.26
	β_1	19.14	1.04	$(17.10, 21.21)$	AIC=40.33	AE=0.348
T(8)	β_0	-35.17	2.03	$(-39.15, -31.17)$	BIC=42.50	DIC=41.22
	β_1	19.86	1.14	$(17.61, 22.10)$	AIC=42.34	AE=0.370
SP	β_0	-34.45	2.06	$(-38.52, -30.43)$	BIC=42.48	DIC=40.22
	β_1	19.43	1.14	$(17.19, 21.69)$	AIC=42.24	AE=0.352
	δ	0.08	0.45	$(-0.83, 0.95)$		
SGT	β_0	-8.30	1.73	$(-11.67, -4.97)$	BIC=38.73	DIC=36.72
	β_1	4.69	0.98	$(2.81, 6.61)$	AIC=38.41	AE=0.245
	ν_1	39.40	14.80	$(13.89, 69.65)$		
	δ	0.22	0.10	$(0.03, 0.40)$		
AEP	β_0	-35.16	0.36	$(-35.93, -34.70)$	BIC=38.15	DIC=36.43
	β_1	19.69	0.21	$(19.28, 20.13)$	AIC=37.83	AE=0.153
	θ_1	1.04	0.15	$(0.76, 1.33)$		
	θ_2	0.47	0.07	$(0.35, 0.61)$		

approach. The binary regression model with AEP-based link function has the flexibility to properly fit the data. It is clearer when data are binomial (grouped data), as in this example. However, in case of binary data, according to the description of links given by Chen et al. (1999a), the AEP distribution provides a flexible link function having symmetric/asymmetric and lighter/heavier tails.

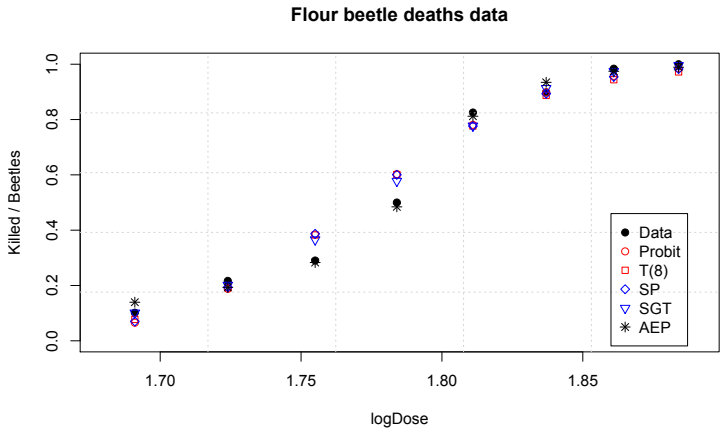

Figure 4.5: Observed and posterior proportions.

4.6 Conclusion

The AEP distribution family has been used in three different Bayesian contexts. This family includes the symmetric exponential power distribution as a particular case and provides flexible distributions with both lighter and/or heavier tails compared to the normal. These distributions are able to manage both symmetry/asymmetry and light/heavy tails simultaneously.

A scale mixture of uniform representation of the AEP distribution has been proposed to allow the development of efficient Gibbs sampling algorithms. Although the problems to solve are very different both in purpose and methodology (posterior exploration, linear regression, and binary regression), this representation has allowed them to share computational developments, i.e., most of the full conditional distributions are shared by the three models. Besides, these distributions are standard and easy to sample from.

The proposed models provide a great flexibility. The goodness of fit and the predictions obtained by using the AEP distribution are usually better than the ones obtained with standard methods when data are not symmetric. The main disadvantage is that there can exist high autocorrelations in the estimated samples, so that large MCMC chains are required to achieve convergence. Some examples that illustrate the performances of the proposed approaches and comparisons with competing models have been presented and discussed.

A very interesting research topic related to this work could be the use of a mixture of AEP distributions. This may be useful to provide more flexibility than the normal mixture-based approach in order to identify subpopulations and finding a better separation. This could be particularly suited to multimodal problems. Another interesting adaptation of the AEP-based proposed approach could be developed in the time series model context, where the asymmetrical data and the heaviness of the tails require a different distribution from the normal one. Specifically, many financial time series should be modeled by using more flexible distributions than the normal ones to accommodate for heavier/lighter tails and skewness. This flexibility is very important for the GARCH models, where the EP family has been widely used. The use of AEP and SEP distributions in this context is currently a challenge. Finally, a third research issue could be to generalize Proposition 4.1 to address a class of spliced-scale distributions of Bayesian semi-parametric scale mixture of Beta models. This might lead to a possibly more flexible approach.

Acknowledgements

The authors thank the editor and two anonymous referees for comments and suggestions which have improved the content and readability of the paper. This research has been partially funded by *Ministerio de Economía y Competitividad, Spain* (Project MTM2011-28983-C03-02), *Gobierno de Extremadura, Spain* (Project GRU10110), and *European Union* (European Regional Development Funds).

4.7 Appendix

4.7.1 Proofs

Proposition 4.1.
Proof. It is enough to show that

$$\begin{aligned}
&\int f(y|u_1)f(u_1)\mathrm{d}u_1 \\
&= \int \tfrac{1}{\alpha\sigma}\exp(-u_1)I\Big[\mu - \tfrac{\alpha\sigma}{\Gamma(1+1/\theta_1)}u_1^{1/\theta_1} < y \le \mu\Big]I[u_1 > 0]\mathrm{d}u_1 \\
&= \int \tfrac{1}{\alpha\sigma}\exp(-u_1)I\Big[u_1 > \big|\tfrac{y-\mu}{\alpha\sigma/\Gamma(1+1/\theta_1)}\big|^{\theta_1}\Big]I[y \le \mu]\mathrm{d}u_1 \\
&= \tfrac{1}{\alpha\sigma}\exp\Big(-\big|\tfrac{y-\mu}{\alpha\sigma/\Gamma(1+1/\theta_1)}\big|^{\theta_1}\Big)I[y \le \mu],
\end{aligned}$$

$$\begin{aligned}
&\int f(y|u_2)f(u_2)\mathrm{d}u_2 \\
&= \int \tfrac{1}{(1-\alpha)\sigma}\exp(-u_2)I\Big[\mu < y < \mu + \tfrac{(1-\alpha)\sigma}{\Gamma(1+1/\theta_2)}u_2^{1/\theta_2}\Big]I[u_2 > 0]\mathrm{d}u_2 \\
&= \int \tfrac{1}{(1-\alpha)\sigma}\exp(-u_2)I\Big[u_2 > \big|\tfrac{y-\mu}{(1-\alpha)\sigma/\Gamma(1+1/\theta_2)}\big|^{\theta_2}\Big]I[y > \mu]\mathrm{d}u_2 \\
&= \tfrac{1}{(1-\alpha)\sigma}\exp\Big(-\big|\tfrac{y-\mu}{(1-\alpha)\sigma/\Gamma(1+1/\theta_2)}\big|^{\theta_2}\Big)I[y > \mu],
\end{aligned}$$

and so

$$f(y) = \alpha\int f(y|u_1)f(u_1)\mathrm{d}u_1 + (1-\alpha)\int f(y|u_2)f(u_2)\mathrm{d}u_2$$

is the pdf of the distribution $\mathrm{AEP}(\mu,\sigma,\alpha,\theta_1,\theta_2)$. ∎

4.7.2 Specific full conditional distributions

Equation (4.3):
If $\pi(\mu) \propto 1$ this distribution becomes an uniform

$$[\mu|\mathbf{y},\mathbf{u}_1,\mathbf{u}_2,\sigma,\alpha,\theta_1,\theta_2] \sim \mathrm{U}\big(\underline{\mu}\ ,\ \overline{\mu}\big).$$

If $\pi(\mu)$ is the pdf of a normal distribution, $\mathrm{N}(m_\mu, s_\mu^2)$, the full conditional distribution becomes a truncated normal

$$[\mu|\mathbf{y},\mathbf{u}_1,\mathbf{u}_2,\sigma,\alpha,\theta_1,\theta_2] \sim \mathrm{N}(m_\mu, s_\mu^2)I\big[\underline{\mu} < \mu < \overline{\mu}\big].$$

Equation (4.4):
If $\pi(\sigma) \propto 1$, the full conditional distribution of σ is

$$[\sigma|\mathbf{y},\mathbf{u}_1,\mathbf{u}_2,\mu,\alpha,\theta_1,\theta_2] \sim \mathrm{ParetoI}\big(\mathrm{scale} = \underline{\sigma}\ ,\ \mathrm{shape} = n-1\big).$$

If $\pi(\sigma) \propto 1/\sigma^{m_\sigma}$, the full conditional distribution of σ is

$$[\sigma|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \alpha, \theta_1, \theta_2] \sim \text{ParetoI}\big(\text{scale} = \underline{\sigma}\ ,\ \text{shape} = n - 1 + m_\sigma\big).$$

If $\pi(\sigma)$ is the pdf of an inverse-gamma distribution, InvGamma(a_σ, b_σ), the full conditional distribution of σ is

$$[\sigma|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \alpha, \theta_1, \theta_2] \sim \text{InvGamma}\big(\text{shape} = n - 1 + a_\sigma\ ,\ \text{scale} = b_\sigma\big) I\big[\sigma > \underline{\sigma}\big].$$

Equation (4.5):
If $\pi(\alpha) \propto 1$, the full conditional distribution of α is

$$[\alpha|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \sigma, \theta_1, \theta_2] \sim \text{U}\big(\underline{\alpha}\ ,\ \overline{\alpha}\big).$$

If $\pi(\alpha)$ is the pdf of a beta distribution, Beta(a_α, b_α), the full conditional distribution of α is

$$[\alpha|\mathbf{y}, \mathbf{u}_1, \mathbf{u}_2, \mu, \sigma, \theta_1, \theta_2] \sim \text{Beta}(a_\alpha\ ,\ b_\alpha) I\big[\underline{\alpha} < \alpha < \overline{\alpha}\big].$$

Part III

Misclassification

Chapter 5

Addressing misclassification for binary data: probit and t-link regressions

Naranjo, L., Martín, J., Pérez, C. J., and Rufo, M. J. (2014). Addressing misclassification for binary data: probit and t-link regressions. *Journal of Statistical Computation and Simulation*. In Press.

Abstract

Generalized linear models are addressed to describe the dependence of data on explanatory variables when the binary outcome is subject to misclassification. Both probit and t-link regressions for misclassified binary data under Bayesian methodology are proposed. The computational difficulties have been avoided by using data augmentation. The idea of using a data augmentation framework (with two types of latent variables) is exploited to derive efficient Gibbs sampling and Expectation-Maximization algorithms. Besides, this formulation has allowed to obtain the probit model as a particular case of the t-link model. Simulation examples are presented to illustrate the model performance when comparing with standard methods that do not consider misclassification. In order to show the potential of the proposed approaches, a real data problem arising when studying hearing loss caused by exposure to occupational noise is analyzed.

Keywords: Bayesian methods; Binary regression; Data augmentation; Expectation-Maximization algorithm; Generalized linear models; Markov chain Monte Carlo methods; Misclassification.

5.1 Introduction

Sometimes data-generating processes are not error-free when data are collected in the real world. This fact can happen due to several causes that are especially critical in biomedical contexts. Even a small proportion of misclassified data can produce an important impact on inferences, because the effective amount of information can be dramatically reduced. In these contexts, additional parameters are necessary to correct the bias yielded by the use of misclassified data. If the misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non optimal decision making. Therefore, statistical models should address misclassification.

Generalized linear models are considered to describe the dependence of data on explanatory variables when the binary outcome is subject to misclassification. Cowling et al. (2001) presented two methods for logistic regression where the outcome is determined by an imperfect diagnostic test of unknown sensitivity and specificity. They first considered a large sample-based Bayesian approach in conjunction with the Expectation-Maximization algorithm. This approach allows for the inclusion of initial information and expert opinion in order to estimate odds ratios, probabilities and the sensitivity and specificity of the diagnostic test. The second approach utilizes a Gibbs sampler with Metropolis-Hastings steps. Later, Achcar et al. (2004) proposed a Bayesian logistic regression model. They concentrated on the sensitivity and specificity of medical tests in the presence of misclassification and they used a Metropolis-Hastings algorithm in the generation process. Paulino et al. (2005) presented a Bayesian analysis of misclassified binary data under the framework of a logistic regression model with random effects. They analyzed misclassified binary response data from a study of Human Papillomavirus infection and the random samples were generated by using a Metropolis-Hastings-within-Gibbs sampling method with a Gaussian proposal distribution.

Much effort has been paid to the logistic model. This is because the logistic regression has been used extensively in the biomedical context (the main development area for misclassification) and the odds ratios are easily interpretable. However, probit and t-link regressions offer alternatives to logistic regression to model binary data subject to misclassification, specially, when the underlying distribution is normal or Student's t, respectively. Note that the logistic model can be approximated by using the t-link one. Albert and Chib (1993) empirically observed that a Student's $t(\nu)$ random variable is approximately b times a logistic random variable with appropriate choices of positive values ν and b. Albert and Chib (1993) obtained, for probabilities between 0.001 and 0.999, that the logistic quantiles are approximately a linear function of Student's $t(8)$ quantiles with $b = 0.636$. Liu (2004) considered a Student's $t(7)$ with $b = 0.645$, whereas Chen and Dey (1998) obtained a Student's $t(7.581)$ with $b = 0.643$.

This paper addresses both the probit and the t-link regressions for misclassified binary data under the Bayesian methodology. The proposed models are presented for two prior distribution schemes. The computational difficulties have been avoided by using data augmentation. Although the augmented models increase the dimensionality, the generation processes become easier. The proposed data augmentation scheme has allowed to obtain the probit model as a particular case of the t-link model and

derive efficient algorithms (Gibbs sampling and Expectation-Maximization) for both models. The proposed approaches are extensions of the error-free regression model proposed by Albert and Chib (1993). At the same time, the probit model with normal prior distribution for the regression parameters is an extension to two misclassification parameters of the model presented by Rekaya et al. (2001).

The potential applicability of these approaches to many fields of knowledge makes this proposal interesting. Nevertheless, it is specially useful for biomedical contexts where the decisions and the nature of the data have specific characteristics (see Gustafson 2003). The application of the proposed models is illustrated with a real data problem arisen when studying the Noise-Induced Hearing Loss (NIHL). Hearing loss caused by exposure to occupational noise results in devastating disability that is virtually preventable. NIHL is the second most common form of sensorineural hearing deficit, after presbycusis. Pardo (2010) conducted an observational study in order to predict the hearing loss risk by determination of some factors. The study involved 190 workers from the Câmara Municipal de Monção (Portugal). All the workers had an audiometer and a clinic survey. The objective was to determine when a hearing loss is supposed to be induced by noise. Both false positives and false negatives appear in the sample when trying to determine suspicions of NIHL. This fact motivates the application of the proposed methodology.

The outline of the paper is as follows. In Section 5.2, the way misclassification is addressed in the proposed binary regression models is presented. Section 5.3 presents the prior distributions, and the posterior distributions are explored through different algorithms. Section 5.4 shows the performance of the proposed approaches through two simulation examples. In Section 5.5 a real-data problem about NIHL is analyzed. Section 5.6 presents the conclusion. Finally, the paper is completed with two appendices. Technical details of the EM algorithm implementation are presented in Appendix 5.7.1, whereas the R code for both Gibbs sampling and EM algorithms are given in Appendix 5.7.2.

5.2 Addressing misclassification in binary regression models

Suppose that n independent binary random variables $Y_1, \ldots, Y_n$ are observed, where Y_i is Bernoulli distributed with success probability $\mathrm{P}(Y_i = 1) = \theta_i$. The parameters θ_i are related to a set of covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ through a binary regression model addressing misclassification. They are defined as $\theta_i = p_i(1-\lambda_{10}) + (1-p_i)\lambda_{01}$, where p_i is the true positive probability for an observation with covariate pattern i, λ_{10} is the false negative probability, and λ_{01} is the false positive probability.

A binary response model with $p_i = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$ is assumed, where Ψ is a cumulative distribution function (cdf) and $\boldsymbol{\beta}$ is a $k-$dimensional vector of unknown parameters. The most common models used in practice are probit (standard Gaussian cdf) and logit (logistic cdf). In this section, the way of addressing misclassification for the probit model ($\Psi = \Phi$, the cdf of a standard Gaussian distribution) and the t-link one ($\Psi = T_\nu$, the cdf of a Student's t distribution with ν degrees of freedom) is pre-

sented. In both cases, the posterior distributions are intractable for direct generation. However, it is possible to generate from them by augmenting the models with latent variables. Although the models increase the dimensionality, the generation process becomes easier.

The first type of latent variables to introduce is related to the misclassification, since each observation can be classified in four different groups leading to true positive, false negative, false positive and true negative. Binary latent variables c^i_{hm}, $h, m = 0, 1$, are defined, where h represents the index for the true value and m represents the index for the observed value. When the latent variable takes value one, it denotes the group where the observation i has been assigned: true positive ($c^i_{11} = 1$), false negative ($c^i_{10} = 1$), false positive ($c^i_{01} = 1$), or true negative ($c^i_{00} = 1$). Note that $c^i_{11} + c^i_{01} = 1$ (when $y_i = 1$) or $c^i_{10} + c^i_{00} = 1$ (when $y_i = 0$). Then, the latent vectors $\mathbf{c}^i = (c^i_{11}, c^i_{10}, c^i_{01}, c^i_{00})^T$ and the latent matrix $\mathbf{c} = (\mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^n)^T$ are defined.

The second type of latent variables are introduced based on the idea of data augmentation from Albert and Chib (1993), i.e.: n independent latent variables $z_1, \ldots, z_n$ are considered, where z_i is distributed $\mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \gamma_i^{-1})$, and it is defined $c^i_{11} + c^i_{10} = 1$ if $z_i > 0$ and $c^i_{01} + c^i_{00} = 1$ if $z_i \leq 0$. If the t-link model is assumed, then γ_i is distributed as $\mathrm{Gamma}(\nu/2, 2/\nu)$, with probability density function $\mathrm{P}(\gamma_i) = c(\nu)\gamma_i^{\nu/2-1}\exp(-\nu\gamma_i/2)$, $c(\nu) = [\Gamma(\nu/2)(2/\nu)^{\nu/2}]^{-1}$. Then, the likelihood function is

$$
\begin{aligned}
& L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu|\mathbf{D}) \qquad (5.1)\\
& \propto \prod_{i=1}^{n}\Big[\big\{p_i(1-\lambda_{10}) + (1-p_i)\lambda_{01}\big\}^{y_i}\big\{p_i\lambda_{10} + (1-p_i)(1-\lambda_{01})\big\}^{1-y_i}\Big]\\
& \propto \prod_{i=1}^{n}\Big[\int\int\int\Big\{\phi(z_i; \mathbf{x}_i^T\boldsymbol{\beta}, \gamma_i^{-1})\\
& \times \big(I[z_i > 0]I[c^i_{11} + c^i_{10} = 1] + I[z_i \leq 0]I[c^i_{01} + c^i_{00} = 1]\big)\\
& \times \big(I[y_i = 1]I[c^i_{11} + c^i_{01} = 1] + I[y_i = 0]I[c^i_{10} + c^i_{00} = 1]\big)\\
& \times (1-\lambda_{10})^{c^i_{11}}\lambda_{10}^{c^i_{10}}\lambda_{01}^{c^i_{01}}(1-\lambda_{01})^{c^i_{00}}\mathrm{P}(\gamma_i)\Big\}\mathrm{d}\gamma_i\mathrm{d}z_i\mathrm{d}\mathbf{c}^i\Big],
\end{aligned}
$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{D} = \{\mathbf{y}, \mathbf{x}\}$.

This data augmentation scheme allows to obtain the likelihood function of the probit model as a particular case when $\gamma_i = 1$ and $\mathrm{P}(\gamma_i) = 1$ in (5.1). Then, the parameter ν is omitted and the likelihood function reduces to $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{D})$.

The following step is to define the prior distributions. There is some literature addressing informative prior elicitation for the generalized linear models. See, for example, Ibrahim and Laud (1991), Bedrick et al. (1996), Bedrick et al. (1997), and Chen et al. (1999). All of them assume binary regression models without errors. The literature about informative prior elicitation to binary regression models with misclassification is mainly focused on a normal distribution for the regression vector.

Two types of prior distributions for the regression vector $\boldsymbol{\beta}$ are considered in this paper: a multivariate normal one and a distribution based on eliciting prior

information. The latter distribution is based on the one proposed for generalized linear models by Bedrick et al. (1996), illustrated in Bedrick et al. (1997) with binomial data, and applied to misclassified binary data by Paulino et al. (2003) and McInturff et al. (2004). The prior distributions for λ_{10} and λ_{01} are beta ones. Finally, for the t-link model, the prior distribution for ν is a bounded discrete distribution.

The posterior distributions for the proposed models will be explored in the next section. The proposed algorithms have been implemented by using R software. This software gives a great flexibility to program data augmentation-based models. Other useful programming environments are WinBUGS and OpenBUGS. Congdon (2006) and McInturff et al. (2004) presented WinBUGS code for the logistic regression model considering misclassification. The first one used normal prior distribution for the regression parameters, whereas the second ones elicitated the prior distribution following the approach of Bedrick et al. (1996).

5.3 Exploring the posterior distributions

The previous formulation allows to derive Gibbs sampling algorithms (see, e.g. Gelfand and Smith (1990) and Gilks et al. (1996)) to generate from the posterior distributions. Besides, Expectation-Maximization (EM) algorithms are also derived (see Dempster et al. 1977).

5.3.1 Normal prior distribution

Given the data $\mathbf{D}$, the joint posterior distribution of the unobservable variables $\mathbf{c}$ and the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and ν is

$$\begin{aligned}
&\pi(\mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \nu|\mathbf{D}) \qquad (5.2)\\
&\quad\propto\quad \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})\pi(\nu)\\
&\quad\times\quad \prod_{i=1}^{n}\Big[\{p_i(1-\lambda_{10})\}^{c_{11}^i}\{p_i\lambda_{10}\}^{c_{10}^i}\{(1-p_i)\lambda_{01}\}^{c_{01}^i}\{(1-p_i)(1-\lambda_{01})\}^{c_{00}^i}\\
&\quad\times\quad \big(I[y_i=1]I[c_{11}^i+c_{01}^i=1]+I[y_i=0]I[c_{10}^i+c_{00}^i=1]\big)\Big].
\end{aligned}$$

The prior distribution for the regression parameter vector is a multivariate normal one $\mathrm{N}_k(\mathbf{b}_0, \mathbf{B}_0)$, i.e.: $\pi(\boldsymbol{\beta}) \propto \exp\big\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^T\mathbf{B}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)\big\}$. A prior distribution can be characterized as weakly informative if it is proper, but it is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available. This is met by using independent normal prior distributions with mean zero and large variances. Then, informative and weakly informative settings can be addressed.

Beta distributions are assumed for the misclassification parameters. This is the natural choice to model uncertainty about probabilities, i.e.: $\lambda_{10} \sim \mathrm{Be}(a_{10}, b_{10})$ and $\lambda_{01} \sim \mathrm{Be}(a_{01}, b_{01})$. Therefore, $\pi(\boldsymbol{\lambda}) \propto \lambda_{10}^{a_{10}-1}(1-\lambda_{10})^{b_{10}-1}\lambda_{01}^{a_{01}-1}(1-\lambda_{01})^{b_{01}-1}$. The non informative case has less sense than the informative one, since the strength of the

proposed approaches is based on using the prior information on the misclassification parameters. However, when $a_{10} = b_{10} = a_{01} = b_{01} = 1$, no information is provided on the misclassification parameters.

With respect to the number of degrees of freedom, a bounded discrete distribution $\pi(\nu)$ is considered. This allows for a great flexibility because this discrete distribution can be uniform or can give more probability mass to a particular value. When ν is considered as a fixed value, the t-link model is simplified.

In order to apply a Gibbs sampling algorithm for the joint posterior distribution (5.2), the full conditional distributions must be derived. The full conditional distributions of $\mathbf{c}$ and $\boldsymbol{\lambda}$ are easy to obtain, they are

$$\mathbf{c}^i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D} \sim \text{Multinomial}\left(1, \pi_{c^i}(c^i_{11}, c^i_{10}, c^i_{01}, c^i_{00})\right), \tag{5.3}$$

and

$$\begin{aligned}
\lambda_{10}|\mathbf{c}, \boldsymbol{\beta}, \nu, \mathbf{D} &\sim \text{Be}\left(a_{10} + \sum_{i=1}^{n} c^i_{10},\ b_{10} + \sum_{i=1}^{n} c^i_{11}\right), \\
\lambda_{01}|\mathbf{c}, \boldsymbol{\beta}, \nu, \mathbf{D} &\sim \text{Be}\left(a_{01} + \sum_{i=1}^{n} c^i_{01},\ b_{01} + \sum_{i=1}^{n} c^i_{00}\right),
\end{aligned} \tag{5.4}$$

where

$$\begin{aligned}
\pi_{c^i}(1,0,0,0) &= p_i(1-\lambda_{10})I[y_i = 1]/\theta_i, \\
\pi_{c^i}(0,1,0,0) &= p_i\lambda_{10}I[y_i = 0]/(1-\theta_i), \\
\pi_{c^i}(0,0,1,0) &= (1-p_i)\lambda_{01}I[y_i = 1]/\theta_i, \\
\pi_{c^i}(0,0,0,1) &= (1-p_i)(1-\lambda_{01})I[y_i = 0]/(1-\theta_i).
\end{aligned}$$

However, the full conditional distributions $\pi(\boldsymbol{\beta}|\mathbf{c}, \boldsymbol{\lambda}, \nu, \mathbf{D})$ and $\pi(\nu|\mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{D})$ do not have closed expressions to easily generate from. Generating from $\pi(\boldsymbol{\beta}, \nu|\mathbf{c}, \boldsymbol{\lambda}, \mathbf{D})$ could be addressed by using a Metropolis-Hastings step, leading to a Metropolis-Hastings-within-Gibbs algorithm. In this subsection, latent variables are introduced to allow a Gibbs sampling step, leading to a Gibbs-within-Gibbs algorithm. This avoids the difficulty of finding good proposal distributions that provide high acceptance rates (see, for example, Gilks et al. (1995)). The full conditional distributions are easy to generate from.

The new distribution of interest is

$$\begin{aligned}
&\pi(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \nu|\mathbf{c}, \boldsymbol{\lambda}, \mathbf{D}) \\
&\quad\propto\ \pi(\boldsymbol{\beta})\pi(\nu)\prod_{i=1}^{n}\Big\{\phi(z_i; \mathbf{x}_i^T\boldsymbol{\beta}, \gamma_i^{-1})\mathrm{P}(\gamma_i) \\
&\quad\times\ \left(I[z_i > 0]I[c^i_{11} + c^i_{10} = 1] + I[z_i \le 0]I[c^i_{01} + c^i_{00} = 1]\right)\Big\}.
\end{aligned}$$

Now, the four full conditional distributions are easily derived. The full conditional distributions of $z_1, \ldots, z_n$ are conditionally independent, then

$$z_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \nu, \mathbf{c}, \boldsymbol{\lambda}, \mathbf{D} \sim \begin{cases} \mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \gamma_i^{-1})I[z_i > 0] & \text{if } c^i_{11} + c^i_{10} = 1 \\ \mathrm{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \gamma_i^{-1})I[z_i \le 0] & \text{if } c^i_{01} + c^i_{00} = 1 \end{cases}. \tag{5.5}$$

For $\boldsymbol{\beta}$, it is obtained

$$\boldsymbol{\beta}|\mathbf{z},\boldsymbol{\gamma},\nu,\mathbf{c},\boldsymbol{\lambda},\mathbf{D} \sim \mathrm{N}_k\left(\mathbf{b}_k,\mathbf{B}_k\right), \tag{5.6}$$

where $\mathbf{b}_k = \mathbf{B}_k(\mathbf{x}^T\mathrm{W}\mathbf{z} + \mathbf{B}_0^{-1}\mathbf{b}_0)$, $\mathbf{B}_k = (\mathbf{x}^T\mathrm{W}\mathbf{x} + \mathbf{B}_0^{-1})^{-1}$, and $\mathrm{W} = \mathrm{diag}(\gamma_i)$.

The full conditional distributions of $\gamma_1,\ldots,\gamma_n$ are conditionally independent with

$$\gamma_i|\mathbf{z},\boldsymbol{\beta},\nu,\mathbf{c},\boldsymbol{\lambda},\mathbf{D} \sim \mathrm{Ga}\left(\frac{\nu+1}{2}, \frac{2}{\nu+(z_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}\right). \tag{5.7}$$

Finally, $\nu|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{D}$ is distributed according to a probability mass function proportional to

$$\pi(\nu)\prod_{i=1}^{n}\left\{c(\nu)\gamma_i^{\nu/2-1}\exp(-\nu\gamma_i/2)\right\}. \tag{5.8}$$

The final algorithm consists of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\mathbf{c}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $\nu^{(0)}$ and generating iteratively from the full conditional distributions. The initial points are proposed to be set: $\boldsymbol{\beta}^{(0)} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$, $\mathbf{c}^{i(0)} = (1,0,0,0)$ if $y_i = 1$ and $\mathbf{c}^{i(0)} = (0,0,0,1)$ if $y_i = 0$, $\lambda_{10}^{(0)} = 0$, $\lambda_{01}^{(0)} = 0$, $\gamma_i^{(0)} = 1$ and $\nu^{(0)} = 8$. Note that generation from the full conditional distributions is easy. Specifically, the full conditional distributions (5.3), (5.4), (5.5), (5.6), and (5.7) are standard, and the distribution given in (5.8) is a bounded discrete one, therefore, generation from it is trivial and efficient. The following order is proposed: $\mathbf{z}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\mathbf{c}^{(j)}$, $\boldsymbol{\lambda}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ and $\nu^{(j)}$ using (5.5), (5.6), (5.3), (5.4), (5.7) and (5.8), respectively.

When the probit model is considered, i.e. $\gamma_i = 1$ and $\mathrm{P}(\gamma_i) = 1$, the generation process is simplified since γ_i is a fixed value and ν is omitted. This means that generating from (5.5) and (5.6) constitutes the Gibbs-within-Gibbs algorithm, and the final Gibbs sampling algorithm only requires generating from (5.5), (5.6), (5.3) and (5.4). All of them are standard distributions.

By using the proposed formulation, an EM algorithm can also be derived. The EM algorithm was originally introduced by Dempster et al. (1977) and implemented for posterior distributions by Tanner and Wong (1987). It provides an iterative procedure to compute maximum likelihood estimations. Formulas and the way to obtain them are presented in the Appendix 5.7.1.

Note that the Gibbs sampling algorithm provides simulated posterior distributions for the parameters of interest (and, therefore, summary statistics can be easily calculated), whereas the EM algorithm only provides estimated modes. Appendix 5.7.2 presents the R code for the probit and t-link models considering misclassification for both Gibbs sampling and EM algorithms.

5.3.2 Eliciting prior information

In this subsection, the interest is focused on building a prior distribution for $\boldsymbol{\beta}$ based on the expert prior elicitation. Note that an empirical Bayes approach could be implemented by using the results in the previous subsection. However, the idea proposed

in Bedrick et al. (1996) and illustrated in Bedrick et al. (1997) is adapted to address misclassification.

Bedrick et al. (1996) proposed a method to induce a prior probability distribution on the regression vector $\boldsymbol{\beta}$ by using the so called conditional means prior (CMP) on $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_k)^T$, where, in binomial regression, $\tilde{p}_l = \mathrm{E}(\tilde{y}_l|\tilde{\mathbf{x}}_l)$ is the success probability for a potentially observable response $\tilde{y}_l$ at covariate vector $\tilde{\mathbf{x}}_l$. Prior knowledge from experts or previous studies is used to specify uncertainty about probabilities of the present condition, given various specified covariate configurations.

Assuming k regression coefficients (including the intercept), prior probabilities $\tilde{p}_l$ are elicited in the predictor space, $l = 1, \ldots, k$, for selected locations $\tilde{\mathbf{x}}_l$. The covariate vectors $\tilde{\mathbf{x}}_l$ are chosen subjected to expert opinion in the predictor variable range in order to make it reasonable to assume prior independence among the quantities $\tilde{p}_l$. With k linearly independent sets of covariate values, it is obtained a 1-1 transformation between $\boldsymbol{\beta}$ and $\tilde{\mathbf{p}}$, namely $\boldsymbol{\beta} = \tilde{\mathbf{x}}^{-1}\Psi^{-1}(\tilde{\mathbf{p}})$, where $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_1^T, \ldots, \tilde{\mathbf{x}}_k^T)^T$. Uncertainty about $\tilde{p}_l$ is modeled with independent distributions $\mathrm{Be}(a_l, b_l)$. The hyperparameters a_l and b_l are determined (indirectly in general) from expert prior judgements.

The independence CMP

$$\pi(\tilde{\mathbf{p}}) \propto \prod_{l=1}^{k} \tilde{p}_l^{a_l - 1}(1 - \tilde{p}_l)^{b_l - 1},$$

induces a prior on $\boldsymbol{\beta}$ given by

$$\pi(\boldsymbol{\beta}) \propto \prod_{l=1}^{k} \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})^{a_l - 1}\Big[1 - \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})\Big]^{b_l - 1}\psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta}),$$

where $\Psi = \Phi$ for the probit model and $\Psi = T_\nu$ for the t-link model, and ψ denotes the probability density function (pdf). Then, the posterior distribution is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu|\mathbf{D}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})\pi(\nu)L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu).$$

In order to build an algorithm to sample from $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and ν, the latent variables $\mathbf{c} = (\mathbf{c}^1, \ldots, \mathbf{c}^n)^T$ are introduced. The full conditional distributions for $\mathbf{c}$ and $\boldsymbol{\lambda}$ are the same as in (5.3) and (5.4). Now, the full conditional distributions for $\boldsymbol{\beta}$ and ν are given by

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{c}, \boldsymbol{\lambda}, \nu, \mathbf{D}) \quad &\propto \quad \prod_{l=1}^{k} \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})^{a_l - 1}\Big[1 - \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})\Big]^{b_l - 1}\psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta}) \\ &\times \quad \prod_{i=1}^{n} \Psi(\mathbf{x}_i^T\boldsymbol{\beta})^{c_{11}^i + c_{10}^i}\Big[1 - \Psi(\mathbf{x}_i^T\boldsymbol{\beta})\Big]^{c_{01}^i + c_{00}^i}, \qquad (5.9)\end{aligned}$$

$$\begin{aligned}\pi(\nu|\mathbf{c}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{D}) \quad &\propto \quad \pi(\nu)\prod_{l=1}^{k} \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})^{a_l - 1}\Big[1 - \Psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta})\Big]^{b_l - 1}\psi(\tilde{\mathbf{x}}_l^T\boldsymbol{\beta}) \\ &\times \quad \prod_{i=1}^{n} \Psi(\mathbf{x}_i^T\boldsymbol{\beta})^{c_{11}^i + c_{10}^i}\Big[1 - \Psi(\mathbf{x}_i^T\boldsymbol{\beta})\Big]^{c_{01}^i + c_{00}^i}. \qquad (5.10)\end{aligned}$$

In this case, by including latent variables, the generation process is not simplified. Therefore, a Metropolis-Hastings step is implemented, leading to a Metropolis-Hastings-within-Gibbs algorithm. In order to generate $\boldsymbol{\beta}^{(j)}$ from (5.9), a Metropolis-Hastings algorithm is used with the following normal proposal distribution, $\mathrm{N}_k(\boldsymbol{\beta}^{(j-1)}, (\mathbf{x}^T\mathbf{x})^{-1})$ (see, e.g. Gilks et al. 1996). On the other hand, the distribution given in (5.10) is a bounded discrete one, therefore, generating from it is trivial.

The final algorithm consists of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$ and $\nu^{(0)}$, and iteratively generating $\mathbf{c}^{(j)}$, $\boldsymbol{\lambda}^{(j)}$, $\boldsymbol{\beta}^{(j)}$ and $\nu^{(j)}$ from the full conditional distributions (5.3), (5.4), (5.9) and (5.10), respectively. When the probit model is considered, the generation process is simplified because there is no degrees of freedom to generate from. Appendix 5.7.2 presents the R code for these probit and t-link models.

Next section shows two simulation examples.

5.4 Simulation-based examples

An empirical study using a wide range of datasets showed us that the proposed approach is useful to address misclassification, having good performances. Two simulation examples are considered to illustrate the model performance.

In order to compare several competing models, the Deviance Information Criterion (DIC) is used. This criterion was proposed by Spiegelhalter et al. (2002) and is useful to assess the performance of models with different amounts of partial information. DIC is designed for complex hierarchical models with possibly improper prior distributions. It overcomes the problem of having to identify the number of parameters in the model, which is required for the calculation of the Akaike Information Criterion (AIC) (see Akaike 1973). The calculation of DIC is straightforward

$$DIC = \overline{D(\theta)} + \widehat{\rho_D},$$

where $D(\theta) = -2\log L(\theta)$ is the deviance of the model (that includes the likelihood $L(\theta)$) and $\widehat{\rho_D}$ is the effective number of parameters. The effective number of parameters is calculated as

$$\widehat{\rho_D} = \overline{D(\theta)} - D(\overline{\theta}),$$

where $\overline{D(\theta)} = \mathrm{E}(D(\theta)|\text{data})$ is the posterior mean of the deviance and $D(\overline{\theta})$ is the deviance at the posterior means, $\overline{\theta} = \mathrm{E}(\theta|\text{data})$, of the parameters of interest.

Models with smaller DIC should be preferred over models with larger DIC. Models are penalized both by $\overline{D(\theta)}$, which favours a good fit, and by the effective number of parameters.

DIC criterion is not applicable to the results obtained by using the EM algorithm. For this reason, another criterion should be used to compare the obtained results with this algorithm. The total variation distance (TVD) is proposed to measure the discrepancy between the real probabilities and the estimated ones. It is defined as $\text{TVD} = \sum_{i=1}^{n} |p_i - \widehat{p_i}|$, where $p_i = \Psi(\mathbf{x}_i^T\boldsymbol{\beta})$ and $\widehat{p_i} = \Psi(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}})$. So it is only applicable when data are simulated or a gold standard is available, and therefore, the true classification is known. Observed that DIC criterion is applied to observed data while TVD is applied to real data.

5.4.1 Model comparison

Firstly, a covariate set is generated independently by $x_{i1}, x_{i2} \sim \mathrm{N}(2, 0.25)$, for $i = 1, 2, \ldots, 100$, and the probabilities for the error-free model are obtained by

$$p_i = \Psi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}),$$

where $\boldsymbol{\beta} = (2, -3, 2)^T$ and $\Psi = \Phi$ is the cdf of a standard normal distribution. The true binary dependent variable y^{true} is randomly generated by using the following process: (i) generate $u_i \sim \mathrm{U}(0, 1)$, (ii) if $p_i \geq u_i$, then $y_i^{true} = 1$, else $y_i^{true} = 0$. Now, about 10% of the outcomes are randomly misclassified according to the following process: (iii) generate $v_i \sim \mathrm{U}(0, 1)$, (iv) if $v_i < 0.1$, then $y_i = 1 - y_i^{true}$, else $y_i = y_i^{true}$. The new response variable y remains equal to y^{true} for the non misclassified outcomes (about 90% of the outcomes). Thus the misclassification parameters are $\lambda_{10} \approx 0.1$ and $\lambda_{01} \approx 0.1$. Figure 5.1 graphically shows a randomly chosen covariate dataset.

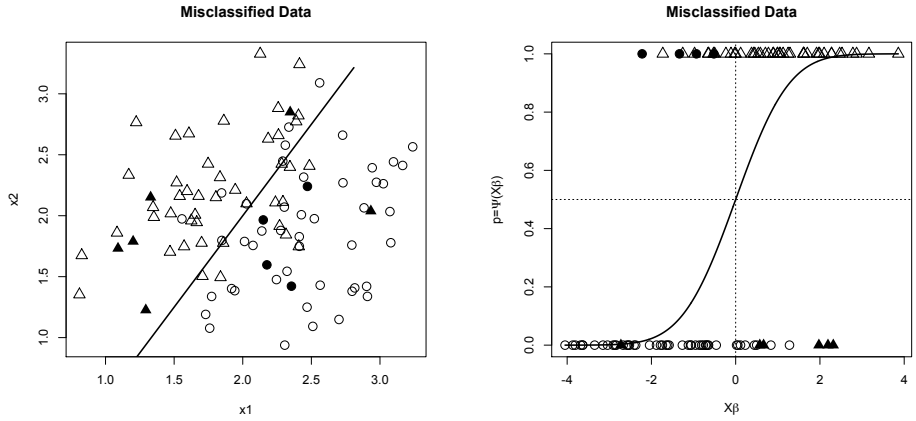

Figure 5.1: Datasets with misclassification: true negative (○), true positive (△), false positive (●), and false negative (▲).

The main objective is to compare the performance of the proposed models to the standard ones. This simulation-based scenario allows to compare the predictive outcomes with the real ones instead of the observed ones and, therefore, to know which model performs better.

Three prior specifications are considered for the regression parameters $\boldsymbol{\beta}$. Firstly, an informative prior distribution is used, i.e.: a multivariate normal distribution with parameters $\mathbf{b}_0^T = (2, -3, 2)$ and $\mathbf{B}_0 = \mathrm{diag}(1, 1, 1)$. The second prior specification is based on a weakly informative prior, i.e.: a multivariate normal distribution with parameters $\mathbf{b}_0^T = (0, 0, 0)$ and $\mathbf{B}_0 = \mathrm{diag}(10, 10, 10)$. Finally, the third prior specification is based on the proposal presented in Subsection 5.3.2. The following configurations and hyperparameters are considered:

Configurations	Hyperparameters	
$\tilde{\mathbf{x}}_1^T = (1, 1.7, 2.3)$	$a_{11} = 50$	$a_{21} = 5$
$\tilde{\mathbf{x}}_2^T = (1, 2.2, 2.0)$	$a_{12} = 25$	$a_{22} = 25$
$\tilde{\mathbf{x}}_3^T = (1, 2.5, 1.8)$	$a_{13} = 5$	$a_{23} = 50$

For the misclassification parameters, initial information is introduced according to the misclassification proportions that affect the data, i.e.: $\lambda_{10} \sim \text{Be}(5, 45)$ and $\lambda_{01} \sim \text{Be}(5, 45)$. In addition, for the t-link model, the degrees of freedom ν are random variates with finite discrete distribution. Its support is $\{4, 5, \ldots, 10\}$ and the probabilities are $\{0.05, 0.10, 0.20, 0.30, 0.20, 0.10, 0.05\}$.

The proposed models have been implemented in R software. Appendix 5.7.2 presents the R codes for the probit and t-link model considering misclassification for both Gibbs sampling and EM algorithms. Moreover, the logistic models have been considered with and without misclassification. When the prior distribution is normal, the logistic model considering misclassification has been implemented by considering the results in Congdon (2006). When the prior distribution is obtained by eliciting prior information following Bedrick et al. (1996), the results from McInturff et al. (2004) have been followed. The EM algorithms for the logistic models with and without misclassification have been implemented by using the results in Cowling et al. (2001) and Jaakkola and Jordan (2000), respectively. The EM algorithm considered by Cowling et al. (2001) uses the Newton-Raphson algorithm in the M-step to maximize the posterior distribution of $\boldsymbol{\beta}$. In contrast, the proposed EM algorithm for the probit model provides closed form expressions. In case of the t-link model, some expectations in the E-step and the maximum of ν in the M-step do not have closed form expressions and they must be computed by numerical methods (see Appendix 5.7.1).

A total of $40,000$ iterations are generated for each model by using Gibbs sampling. A burn-in of $5,000$ and saving one out of 20 values is considered. With these values the chains seem to have converged for all the models based on visual and statistical inspections obtained with BOA Package (see Smith (2007)). The EM algorithm has also been applied for all the models and the convergence has been attained after $40 - 115$ iterations at a tolerance level of 0.001.

By using a 2.8 GHz Intel Core i7 PC with 4 GB 1333 MHz DDR3 RAM, the run-times to fit each dataset with the models considering misclassification were computed. By using a normal prior distribution for the regression parameters, the run-times for the probit and logistic regression models were similar, and approximately about 8 minutes. The t-link model was slightly slower with an approximate time of 9 minutes. By eliciting the prior distribution for the regression parameter as in Subsection 5.3.2, the Metropolis-Hastings-within-Gibbs sampling algorithm exhibited acceptance rates between 50% and 60%, and the run-times were approximately 10 minutes for the probit and logistic regression models, and 18 minutes for the t-link regression model. In both prior distribution scenarios, probit and logistic models have a similar run-time, which is lower than the t-link one. Note that the t-link model has an additional parameter to sample, i.e., the degrees of freedom. Besides, when the prior distribution is elicited, the run-time is higher than in the other case. This

is because the Metropolis-Hastings-within-Gibbs algorithm is more costly than the Gibbs-within-Gibbs algorithm. The run-times of the EM algorithms for the probit and logistic models were under one second, and for the t-link model under 2 minutes. Note that the EM algorithm for the t-link model contains non standard expectations in contrast to its version for the probit model, what makes the former much slower. Appendix 5.7.1 provides the theoretical results for the EM algorithm.

In order to give more reliability to the results based on random misclassification, the experiment is replicated 100 times. The same covariate set and specifications are used, but data are generated and misclassified randomly at each time (steps (i)-(iv)). Table 5.1 shows the means and the standard deviations of the DIC values obtained for each different model by considering three different prior specifications. For the fitted models, the capital letter M indicates that the model considering misclassification has been used. Note that by applying the EM algorithm, the estimations are the modes of the parameters, so it is not possible to obtain the DIC values.

Table 5.1 shows that the proposed models considering misclassification are better than the ones that do not consider it. Besides, since data have been generated by using a probit model and they have been misclassified, the probit model that considers misclassification is the best one for the three prior specifications, as it was expected. The additional criterion to evaluate the discrepancy between real probabilities and estimations is considered in order to show the performance of the EM algorithm. The mean of the 100 TVD values obtained for each model is displayed in Table 5.2.

Table 5.1: DIC means (standard deviations) for probit datasets with misclassification.

Model	Informative	Weakly Informative	Elicited Information
Probit	100.664 (11.295)	100.763 (11.009)	101.517 (12.222)
Probit M	97.198 (9.823)	97.973 (9.780)	98.423 (9.368)
t	99.753 (11.339)	100.100 (11.014)	100.498 (11.764)
t M	97.559 (9.663)	98.802 (9.850)	98.581 (9.226)
Logit	99.006 (11.079)	100.014 (11.003)	100.731 (11.833)
Logit M	98.314 (9.234)	98.508 (9.674)	98.566 (9.282)

Table 5.2: TVD means for probit datasets with misclassification.

Model	Informative		Weakly Informative		Elicited Information
	Gibbs	EM	Gibbs	EM	Gibbs
Probit	8.390	8.635	10.274	10.620	7.885
Probit M	4.152	4.009	5.628	5.744	6.684
t	7.191	7.770	9.314	10.100	7.708
t M	4.099	4.117	5.541	5.833	7.170
Logit	8.692	10.321	9.829	10.456	7.793
Logit M	5.745	6.183	5.415	6.498	7.175

Again the models considering misclassification outperform the standard ones. It is remarkable that the TVD values are lower when using Gibbs sampling algorithms

than when using EM ones. In this case, by generating from the posterior distribution, better fits than by considering only the mode are provided. The EM algorithm is typically the fastest method.

5.4.2 Estimating misclassification

In this subsection, the interest is focused on analyzing if the misclassification parameters are recovered through estimations. The same covariate set as in the previous subsection is used, i.e., $x_{i1}, x_{i2} \sim \mathrm{N}(2, 0.25)$, for $i = 1, \ldots, 100$. However, a different simulation scheme is developed. The probabilities for the error-free model are defined by

$$p_i = \Psi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}),$$

where $\boldsymbol{\beta} = (2, -3, 2)^T$ and $\Psi = \Phi$ is the cdf of a standard normal distribution. The true binary dependent variable y^{true} is randomly generated by using the following process: (i) generate $u_i \sim \mathrm{U}(0, 1)$, (ii) if $p_i \geq u_i$, then $y_i^{true} = 1$, else $y_i^{true} = 0$. Now, different values of misclassification parameters have been applied, $\lambda_{10} = \{0.10, 0.15, 0.20\}$ and $\lambda_{01} = \{0.10, 0.15, 0.20\}$, to misclassify the outcomes according to the following process: (iii) generate $v_i \sim \mathrm{U}(0, 1)$, (iv) if $y_i^{true} = 1$ and $\mathrm{rank}_{\{v;\ y^{true}=1\}}(v_i) \leq \lambda_{10} n_1$, then $y_i = 1 - y_i^{true}$, if $y_i^{true} = 0$ and $\mathrm{rank}_{\{v;\ y^{true}=0\}}(v_i) \leq \lambda_{01} n_0$, then $y_i = 1 - y_i^{true}$, else $y_i = y_i^{true}$, where n_1 and n_0 are the number of y_i^{true} that are equal to 1 and 0, respectively. The new response variable y remains equal to y^{true} for the non misclassified outcomes, about $n_1(1 - \lambda_{10}) + n_0(1 - \lambda_{01})$ of the outcomes.

Two prior specifications are considered for the misclassification parameters λ_{10} and λ_{01}. Firstly, informative prior distributions are used, $\lambda_{10} \sim \mathrm{Be}(a_{10}, b_{10})$ and $\lambda_{01} \sim \mathrm{Be}(a_{01}, b_{01})$, where $a_{10} = \lambda_{10} n_1$, $b_{10} = (1 - \lambda_{10}) n_1$, $a_{01} = \lambda_{01} n_0$, and $b_{01} = (1 - \lambda_{01}) n_0$. The second prior specification is based on non informative prior distributions, $\lambda_{10} \sim \mathrm{Be}(1, 1)$ and $\lambda_{01} \sim \mathrm{Be}(1, 1)$. For the regression parameters, the informative prior distribution $\boldsymbol{\beta} \sim \mathrm{N}_3(\mathbf{b}_0, \mathbf{B}_0)$ is used, where $\mathbf{b}_0^T = (2, -3, 2)$ and $\mathbf{B}_0 = \mathrm{diag}(1, 1, 1)$. Finally, for the t-link model ν is a random variable with finite discrete distribution. Its support is $\{4, 5, \ldots, 10\}$ and the probabilities are $\{0.05, 0.10, 0.20, 0.30, 0.20, 0.10, 0.05\}$.

Following the same MCMC specifications as in the previous subsections, samples are generated for all the models. Table 5.3 displays the misclassification parameter estimates, $\hat{\lambda}_{10}$ and $\hat{\lambda}_{01}$, for several scenarios.

Table 5.3 shows that the estimations obtained by using informative prior distributions are generally closer to the real parameter value than those that use non informative prior distributions. The correct information on the misclassification parameters allows the proposed models to obtain better estimations. These estimations are translated into better predictions. In addition, the estimations obtained by using the Gibbs sampling algorithms are generally better than the ones obtained by using the EM algorithms. This also happens for the Gaussian mixture models presented by Dias and Wedel (2004), who compared EM, stochastic EM, and MCMC algorithms. Table 5.3 shows how some EM estimates that use non informative prior distributions are very far from the original values. This is because the EM algorithm is not be-

Table 5.3: Misclassification parameter estimations, $\widehat{\lambda_{10}}$ and $\widehat{\lambda_{01}}$.

λ_{10}	λ_{01}	Model	**Gibbs sampling**		**EM algorithm**	
			Infor.	Non Infor.	Infor.	Non Infor.
0.10	0.10	Probit M	0.078/0.073	0.092/0.060	0.062/0.059	0.049/5.421e-8
		t M	0.076/0.070	0.084/0.052	0.061/0.058	0.041/3.873e-9
		Logit M	0.072/0.067	0.074/0.046	0.057/0.054	0.019/2.976e-6
0.10	0.15	Probit M	0.084/0.165	0.107/0.208	0.071/0.155	0.074/0.190
		t M	0.080/0.156	0.097/0.190	0.067/0.152	0.061/0.185
		Logit M	0.075/0.151	0.081/0.181	0.059/0.142	0.004/0.161
0.10	0.20	Probit M	0.083/0.207	0.095/0.241	0.066/0.200	0.027/0.228
		t M	0.081/0.198	0.093/0.214	0.064/0.197	0.016/0.221
		Logit M	0.076/0.192	0.082/0.204	0.059/0.184	0.001/0.190
0.15	0.10	Probit M	0.113/0.090	0.099/0.110	0.098/0.079	0.054/0.091
		t M	0.111/0.081	0.094/0.087	0.097/0.076	0.044/0.076
		Logit M	0.108/0.075	0.082/0.070	0.093/0.063	0.007/0.001
0.15	0.15	Probit M	0.124/0.155	0.126/0.188	0.108/0.143	0.067/0.169
		t M	0.121/0.142	0.119/0.159	0.105/0.140	0.053/0.162
		Logit M	0.115/0.139	0.108/0.136	0.099/0.127	0.004/0.121
0.15	0.20	Probit M	0.112/0.201	0.091/0.229	0.095/0.187	1.937e-09/0.198
		t M	0.111/0.187	0.147/0.241	0.095/0.185	4.574e-10/0.192
		Logit M	0.107/0.187	0.074/0.172	0.091/0.176	5.521e-05/0.153
0.20	0.10	Probit M	0.186/0.112	0.175/0.163	0.171/0.100	0.129/0.150
		t M	0.182/0.098	0.163/0.135	0.169/0.095	0.120/0.142
		Logit M	0.174/0.095	0.138/0.120	0.161/0.080	0.058/0.097
0.20	0.15	Probit M	0.190/0.182	0.218/0.266	0.175/0.175	0.185/0.251
		t M	0.186/0.163	0.211/0.237	0.172/0.170	0.179/0.248
		Logit M	0.181/0.159	0.193/0.232	0.166/0.149	0.146/0.200
0.20	0.20	Probit M	0.200/0.190	0.216/0.185	0.189/0.180	0.200/0.165
		t M	0.192/0.184	0.190/0.164	0.186/0.177	0.189/0.155
		Logit M	0.186/0.180	0.176/0.146	0.173/0.170	0.129/0.123

ing able to properly identify the misclassification. Therefore, when no information is available it is advisable to use the Gibbs sampling approach.

Next section shows the applicability of the proposed approach to a real-data problem.

5.5 Noise-Induced Hearing Loss

Noise Induced Hearing Loss (NIHL) is an increasingly prevalent disorder caused by a one-time exposure to a very high-intensity sound, such as an explosion (acoustic trauma) or by continuous exposure to loud sound over a long period of time (gradually developing hearing loss). When the ear is exposed to excessive sound levels or loud

sounds over the time, the overstimulation of hair cells leads to a heavy production of reactive oxygen species, leading to oxidative cell death. Structural damage to hair cells will result in hearing loss that is characterized by the attenuation and distortion of incoming auditory stimuli.

Noise can cause permanent hearing loss at chronic exposures equal to an average sound pressure level of 85 dB. Based on a logarithmic scale, a 3 dB increase in sound pressure level represents a doubling of sound intensity. Therefore, 4 hours of noise exposure at 88 dB is considered to provide the same noise dose as 8 hours at 85 dB, and a single gunshot, which is approximately 140 to 170 dB, has the same sound energy as 40 hours at 90 dB. Sounds of less than 75 dB, even after long exposure, are unlikely to cause hearing loss. Avoiding continuous noise exposure stops further progression of the damage. A detailed discussion on NIHL can be found, for example, in Rabinowitz (2000b).

NIHL affects people of all ages and demographics, and it is the second most common form of sensorineural hearing deficit, after presbycusis. Exposure to occupational noise is the main cause of NIHL (see ACOM, Americam College of Occupational Medicine, Noise and Hearing Conservation Committee 1989). The U.S. Department of Labor's Occupational Safety and Health Administration states that exposure to 85 dB of noise for more than eight hours per day can result in permanent hearing loss (see OSHA, Occupational Safety and Health Administration 2002). NIHL can be prevented (except in cases of accidental exposure) by avoiding exposition to excessive noise and using hearing protection such as earplugs and earmuffs (see, e.g., Brookhouser 1994, Dobie 1995, and Rabinowitz 2000a).

NIHL generally affects person's hearing sensitivity in the higher frequencies, especially at 4000 Hz. Noise-induced impairments are usually associated with a notch-shaped high-frequency sensorineural loss that is worst at 4000 Hz, although the notch often occurs at 3000 or 6000 Hz as well (see Gelfand 2001). NIHL usually occurs initially at high frequencies (3000, 4000, or 6000 Hz), and then spreads to the low frequencies (500, 1000, or 2000 Hz) (see Chen and Tsai 2003).

Pardo (2010) conducted an observational study in order to predict NIHL risk by determination of some factors. The study involved 190 workers from the Câmara Municipal de Monção (Portugal). Dr. Pardo was the physician providing occupational health services to this company and he supervised the hearing conservation program. All the workers had a complete history, a physical examination and an audiometry. The audiometer was properly calibrated and used in a quiet room.

The objective of the study is to indicate when a hearing loss is supposed to be induced by noise. A probable diagnosis of NIHL is easy for many cases. However, the diagnosis is much less certain in others. Sometimes, NIHL is usually accompanied, and often, obscured, by age-associated hearing loss and sometimes by other additional forms of hearing impairment. The diagnostic task should be reduced to defining the likelihood of the presence of a NIHL component in the overall hearing impairment (see Coles et al. 2000). Therefore, both false positive and false negative data can appear (and in fact they appear in this sample). This fact motivates the need of applying the proposed methodology.

Bayesian binary regression models considering misclassification can help to pro-

vide better predictions than standard models. The proposed models are specially interesting in this case, because more precise predictions for the probability of having NIHL can be obtained. The proposed methodology uses the initial information provided by the physician and relates it to the data from the study to arrive at a posterior predictive distribution that is used to estimate the NIHL probability.

The observed binary outcome y is the suspicion of NIHL. The covariates considered in this study are: x_1 ="presence of risk" (0="no", 1="yes"), x_2 = "antecedents of risk" (0="no", 1="yes"), x_3 ="risk exposure time" (0="less than 5 years or no exposure", 1="5 to 15 years", 2="more than 15 years"), x_4 ="right ear audiometry at a frequency of 4000 Hz", and x_5 ="left ear audiometry at a frequency of 4000 Hz". Figure 5.2 and Table 5.4 show the data.

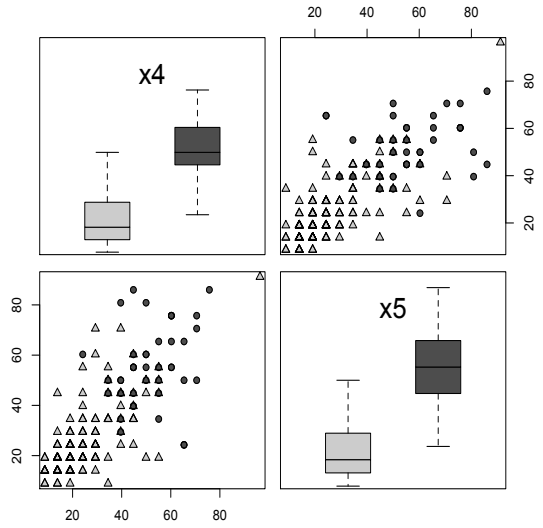

Figure 5.2: NIHL data: The dark (light) gray dots correspond to the patients with observed presence (no observed presence) of NIHL, i.e., $y_i = 1$ ($y_i = 0$).

Table 5.4: NIHL data: contingency table of the discrete variables.

		$x_1 = 0$		$x_1 = 1$	
		$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
$x_3 = 0$	$y = 0$	115	7	0	0
	$y = 1$	5	0	0	0
$x_3 = 1$	$y = 0$	0	13	2	1
	$y = 1$	0	4	0	0
$x_3 = 2$	$y = 0$	0	3	3	5
	$y = 1$	0	8	5	19

The models of interest are given by

$$\eta_i = \Psi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5},$$

where $\Psi = \Phi$ (probit model) and $\Psi = T_\nu$ (t-link model). Both the error-free models and the proposed ones are considered.

Following the method of constructing an informative prior distribution proposed in Subsection 5.3.2, the prior elicitation was carried out with the kind collaboration of Dr. Pardo. In order to induce the prior distribution on $\boldsymbol{\beta}$, six covariate configurations are chosen. These are the most common configurations in the data. For each configuration, the 25%, 50%, and 75% quantiles for the NIHL probabilities are chosen. This is an overspecification of the beta distributions, because only two quantiles are necessary. Specifically, two quantiles are used to calculate the prior hyperparameters and the third is used to check the choice. The elicitation of the prior hyperparameters for the misclassification parameters followed the same description. The complete set of prior hyperparameters elicited from quantiles is:

Configurations	Hyperparameters	
$\tilde{\mathbf{x}}_1^T = (1,0,0,0,20,20)$	$a_{11} = 1.35$	$a_{21} = 134.5$
$\tilde{\mathbf{x}}_2^T = (1,0,1,0,20,20)$	$a_{12} = 0.995$	$a_{22} = 106$
$\tilde{\mathbf{x}}_3^T = (1,0,1,1,25,25)$	$a_{13} = 1.2$	$a_{23} = 88$
$\tilde{\mathbf{x}}_4^T = (1,0,1,2,52.5,57.5)$	$a_{14} = 24$	$a_{24} = 3.3$
$\tilde{\mathbf{x}}_5^T = (1,1,0,2,60,50)$	$a_{15} = 22.95$	$a_{25} = 2.75$
$\tilde{\mathbf{x}}_6^T = (1,1,1,2,50,55)$	$a_{16} = 27.1$	$a_{26} = 3.48$
λ_{10}	$a_{10} = 2.55$	$b_{10} = 22.25$
λ_{01}	$a_{01} = 1.63$	$b_{01} = 150$

In addition, for the t-link model, the degrees of freedom are considered a r.v. with finite uniform discrete distribution. The support is $\{1, \ldots, 20\}$.

The Gibbs sampling algorithm defined in Subsection 5.3.2 is applied to four models: probit, probit considering misclassification, t-link, and t-link considering misclassification. The following specifications are used: $110,000$ iterations, a burn-in of $10,000$ and saving one out of 10 values. With these specifications, the chains seem to have converged for all the models.

The estimated DICs obtained for the four models are, respectively, $72.521 = \overline{D} + \widehat{\rho_D} = 69.154 + 3.367$, $67.546 = 64.214 + 3.332$, $70.340 = 66.747 + 3.592$, and $68.470 = 63.983 + 4.486$. This shows that the proposed models with misclassification are the best ones, having both the lowest posterior means of the deviance and the lowest effective numbers of parameters. Both probit and t-link models considering misclassification perform basically the same. The DICs are 67.546 and 68.470, respectively, what indicates that there is little difference between model performances by using this criterion. Specifically, the predictions are similar. In the subsequent analysis, the t-link model with misclassification is considered.

The estimated parameters obtained with the t-link model considering misclassification are summarized in Table 5.5. Figure 5.3 displays the plots of the prior (dashed lines) and posterior densities (solid lines) for the parameter of initial probabilities given the covariate configurations, $\tilde{\mathbf{p}}$, of the t-link model with misclassification.

These prior distributions induce the prior distributions for the regression parameters. The 95% Highest Posterior Density intervals (HPD) of the posterior distribution for the regression parameters show that the most important covariates to deduce NIHL are the right and left ear audiometries at 4000 Hz and the exposure time. The regression coefficients are positive (except the intercept), indicating that high values of the corresponding covariates are associated with a high probability of having NIHL. The posterior distributions for the regression parameters are presented in Figure 5.4.

Table 5.5: Summary of the posterior estimates for the parameters of the t-link model considering misclassification for NIHL data.

	Mean	Median	S.D.	95% HPD interval
β_0	-5.805	-5.678	0.957	$(-7.581, -4.108)$
β_1	0.486	0.485	0.422	$(-0.371, 1.303)$
β_2	0.287	0.265	0.450	$(-0.564, 1.113)$
β_3	0.562	0.565	0.319	$(-0.054, 1.181)$
β_4	0.059	0.059	0.019	$(0.023, 0.096)$
β_5	0.045	0.045	0.014	$(0.016, 0.075)$
ν	11.860	12.000	4.965	$(3.685, 20)$
λ_{10}	0.070	0.065	0.035	$(0.008, 0.136)$
λ_{01}	0.006	0.005	0.004	$(0.000, 0.018)$

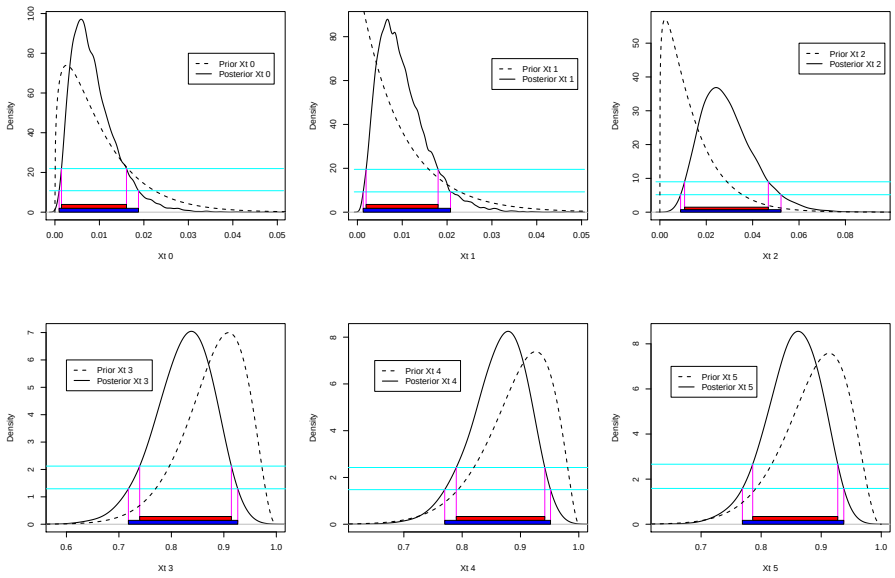

Figure 5.3: Prior (dashed lines) and posterior (solid lines) distributions with 90% and 95% HPD intervals for $\tilde{\mathbf{p}}$ of the t-link model considering misclassification for NIHL data.

The prior and posterior distributions for the misclassification parameters are displayed in Figure 5.5. The initial information elicited about the misclassification probabilities was quite precise. Finally, the posterior distribution is clearly platykurtic, since it has been very influenced by its prior uniform distribution. The posterior distribution for the degrees of freedom is presented in Figure 5.6.

The predictions obtained have been compared with the observed values. With this information, Dr. Pardo re-studied the conflictive cases. This led to find that some individuals were not correctly classified. For individuals with observed absence of NIHL, the predictions show that cases 24 and 89 should have been classified with presence of NIHL (the predictive probabilities are 0.75 and 0.92, respectively). They have been exposed to noise (more than 15 years and between 5 and 15 years, respectively) and have very high values of audiometry. So, they should have been classified as having NIHL. The model also detects case 95 as having NIHL (probability 0.55). This case belongs to the decision border. After reanalyzing the case, a moderate hearing loss attributable to the noise exposure is diagnosed. For individuals with observed presence of NIHL, cases 56 and 97 should have been classified with absence of NIHL (predictive probabilities 0.12 and 0.21, respectively). Probably, these two errors are attributable to miscoding, because almost any NIHL component appeared. Cases 1 and 120 have been detected as individuals with absence of NIHL (predictive probabilities 0.46 and 0.40, respectively). These cases are close to the decision border. However, they should not have been diagnosed as having NIHL. Cases 113 and 135 are special, since the hearing loss is not attributable to occupational noise (predictive probabilities 0.43 and 0.46). Finally, case 131 is correctly classified, although the predictive probability is 0.41.

A Bayesian residual analysis can also help to detect some misclassified data. A residual analysis as in Albert and Chib (1995) is performed. The boxplots of posterior distributions for the residuals $r_i = y_i - \theta_i$ against the fitted probabilities $\mathrm{E}(\theta_i|y_i)$ are presented in Figure 5.7. The middle sections of the boxplot correspond to the quartiles, and the extreme values correspond to the 5th and 95th percentiles of the distribution. The boxplots drawn with light gray color correspond to patients with observed absence of NIHL ($y_i = 0$), whose support is $(-1, 0)$, and the boxplots drawn with dark gray color correspond to the patients with observed presence of NIHL ($y_i = 1$), whose support is $(0, 1)$. Outlier observations correspond to densities of residuals that have locations away from zero, concentrated at extremes of the possible value range for the residuals and showing strong asymmetry. One way of gauging the relative sizes of these residuals is to compute the posterior probabilities that r_i exceed in absolute value some positive constant ε. Albert and Chib (1995) used $\varepsilon = 0.75$. Parallel horizontal lines are drawn at residual values -0.75 and 0.75. If the boxplot for a particular residual does not cross these lines, then the outlying probability is under 0.05. Residuals with large outlying probabilities correspond to boxplots that significantly cross these lines. There are four highly influential configurations of predictor variables associated with NIHL patients whose cases are 24 and 89 with residuals -0.752 and -0.927, and cases 56 and 97 with residuals 0.879 and 0.786, respectively. These are some of the misclassified observations identified by the proposed model.

The potential of the proposed model is not only to classify the individuals better,

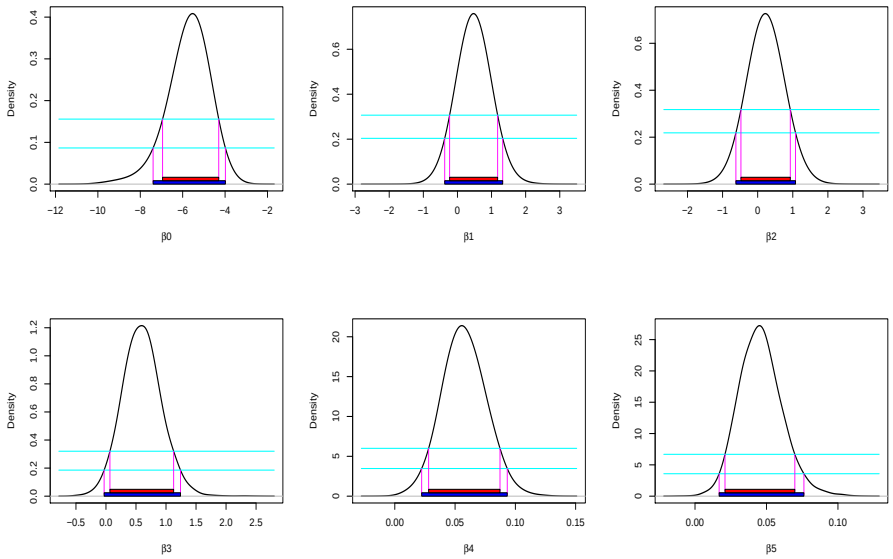

Figure 5.4: Estimated posterior distributions with 90% and 95% HPD intervals for the regression parameters $\boldsymbol{\beta}$ of the t-link model considering misclassification for NIHL data.

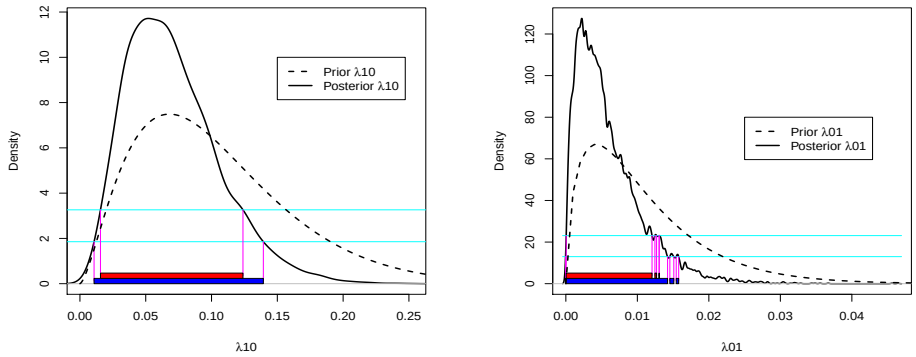

Figure 5.5: Prior (dashed lines) and posterior (solid lines) distributions with 90% and 95% HPD intervals for the misclassification parameters λ_{10} and λ_{01} of the t-link model considering misclassification for NIHL data.

but also to give probabilities of NIHL presence. Predictive probabilities between 0.35 and 0.65 should be carefully considered by conducting a re-study of the case, and making a follow up during the next years.

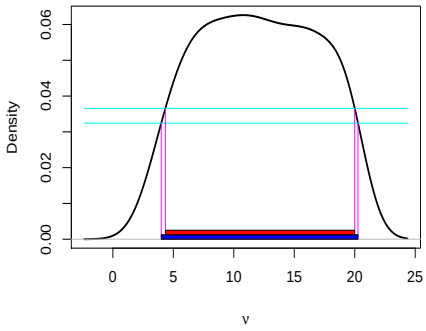

Figure 5.6: Estimated posterior distribution with 90% and 95% HPD intervals for the degrees of freedom ν of the t-link model considering misclassification for NIHL data.

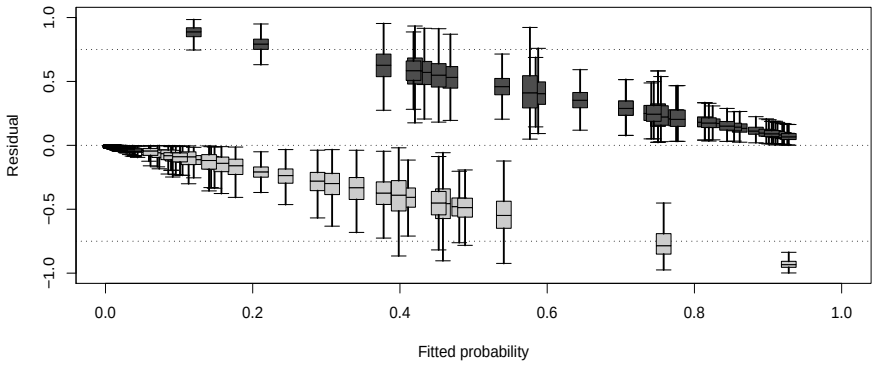

Figure 5.7: Boxplots of posterior distributions for residuals against the fitted probabilities of the t-link model with misclassification for NIHL data.

5.6 Conclusion

Misclassified data are usually found in many studies due to several causes that are especially critical in biomedical contexts. The impact that misclassified data may produce on inferences can be considerable, so it is recommended to build statistical models addressing misclassification. For this reason, some authors have considered the inclusion of new parameters in logistic models, allowing to correct the bias yielded by the misclassified data. Probit and t-link regressions are plausible alternatives to model binary data subject to misclassification, specially when the underlying distribution is normal or Student's t, respectively.

A Bayesian analysis of probit and t-link regression models when the binary outcome is subject to misclassification is described. The use of (two types of) latent

variables enables us to avoid computational difficulties, even by increasing the problem dimension. The proposed data augmentation scheme has allowed to obtain the probit model as a particular case of the t-link model and to derive efficient algorithms (Gibbs sampling and Expectation-Maximization) for both models.

Two simulation examples and a real data problem show the advantages of addressing misclassification. The proposed models are better than the standard ones when there are misclassified data and they can substantially increase the number of correct predictions.

The potential applicability of these approaches to many fields of knowledge makes this proposal interesting. However, more developments are of interest. The application of this framework to other generalized linear models should be straightforward and its extension to other settings should not pose great technical difficulties. The extension to polychotomous response data is an open problem that we will addressed in the near future.

Acknowledgements

The authors thank Dr. Pardo for providing the data and for helpful suggestions related to the medical context. The authors also thank an anonymous referee for comments and suggestions which have improved the content and the readability of the paper. This research has been partially funded by *Ministerio de Economía y Competitividad, Spain* (Projects TIN2008-06796-C04-03 and MTM2011-28983-C03-02), *Junta de Extremadura, Spain* (Project GRU10110), and *European Union* (European Regional Development Funds).

5.7 Appendix

5.7.1 EM algorithm

Given $\mathbf{D}$, The joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, ν, $\mathbf{z}$, $\mathbf{c}$, and γ is

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{z}, \mathbf{c}, \boldsymbol{\gamma}|\mathbf{D}) \\
&\quad \propto \quad \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})\pi(\nu) \prod_{i=1}^{n} \Big\{ \phi(z_i; \mathbf{x}_i^T \boldsymbol{\beta}, \gamma_i^{-1}) \mathrm{P}(\gamma_i) \\
&\quad \times \quad \big(I[z_i > 0] I[c_{11}^i + c_{10}^i = 1] + I[z_i \leq 0] I[c_{01}^i + c_{00}^i = 1] \big) \\
&\quad \times \quad \big(I[y_i = 1] I[c_{11}^i + c_{01}^i = 1] + I[y_i = 0] I[c_{10}^i + c_{00}^i = 1] \big) \\
&\quad \times \quad (1 - \lambda_{10})^{c_{11}^i} \lambda_{10}^{c_{10}^i} \lambda_{01}^{c_{01}^i} (1 - \lambda_{01})^{c_{00}^i} \Big\}.
\end{aligned}$$

In order to apply the EM algorithm, the expected value of the complete log-posterior given the current estimate of the parameter and the observations is computed

as

$$\begin{aligned}
&\mathrm{E}\big[\log \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu | \mathbf{z}, \mathbf{c}, \boldsymbol{\gamma}, \mathbf{D})\big] \\
&\quad \propto \quad -\boldsymbol{\beta}^T \big(\mathbf{x}^T \mathrm{W}^* \mathbf{x} + \mathbf{B}_0^{-1}\big)\boldsymbol{\beta} + 2\boldsymbol{\beta}^T \big(\mathbf{x}^T \big(\mathrm{W}\mathbf{z}\big)^* + \mathbf{B}_0^{-1}\mathbf{b}_0\big) \\
&\quad + \quad \left(b_{10} + \sum_{i=1}^{n} c_{11}^{i*} - 1\right) \log(1-\lambda_{10}) + \left(a_{10} + \sum_{i=1}^{n} c_{10}^{i*} - 1\right) \log(\lambda_{10}) \\
&\quad + \quad \left(a_{01} + \sum_{i=1}^{n} c_{01}^{i*} - 1\right) \log(\lambda_{01}) + \left(b_{01} + \sum_{i=1}^{n} c_{00}^{i*} - 1\right) \log(1-\lambda_{01}) \\
&\quad + \quad \log(\pi(\nu)) + n \log(c(\nu)) + \frac{\nu}{2} \left(\sum_{i=1}^{n} (\log(\gamma_i))^* - \sum_{i=1}^{n} \gamma_i^*\right),
\end{aligned}$$

where $\mathrm{W}^* = \mathrm{E}[\mathrm{W}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D}]$, $(\mathrm{W}\mathbf{z})^* = \mathrm{E}[\mathrm{W}\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D}]$, $\mathbf{c}^{i*} = \mathrm{E}[\mathbf{c}^i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D}]$, and $(\log(\gamma_i))^* = \mathrm{E}[\log(\gamma_i)|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D}]$. When the probit model is considered, W is the identity matrix and ν is omitted, so that $\mathrm{W}^* = \mathrm{W}$ and only $\mathbf{z}^* = \mathrm{E}[\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{D}]$ and $\mathbf{c}^{i*} = \mathrm{E}[\mathbf{c}^i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{D}]$ must be computed.

By maximizing $\mathrm{E}[\log \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu | \mathbf{z}, \mathbf{c}, \boldsymbol{\gamma}, \mathbf{D})]$, it is obtained that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{x}^T \mathrm{W}^* \mathbf{x} + \mathbf{B}_0^{-1})^{-1} (\mathbf{x}^T (\mathrm{W}\mathbf{z})^* + \mathbf{B}_0^{-1} \mathbf{b}_0), \tag{5.11}$$

$$\begin{aligned}
\widehat{\lambda_{10}} &= \frac{a_{10} + \sum_{i=1}^{n} c_{10}^{i*} - 1}{a_{10} + b_{10} + \sum_{i=1}^{n} c_{10}^{i*} + \sum_{i=1}^{n} c_{11}^{i*} - 2}, \qquad (5.12) \\
\widehat{\lambda_{01}} &= \frac{a_{01} + \sum_{i=1}^{n} c_{01}^{i*} - 1}{a_{01} + b_{01} + \sum_{i=1}^{n} c_{01}^{i*} + \sum_{i=1}^{n} c_{00}^{i*} - 2},
\end{aligned}$$

and

$$\widehat{\nu} = \arg\max_{\nu}(g(\nu)), \tag{5.13}$$

where

$$g(\nu) = \log(\pi(\nu)) + n \log(c(\nu)) + \frac{\nu}{2} \left(\sum_{i=1}^{n} (\log(\gamma_i))^* - \sum_{i=1}^{n} \gamma_i^*\right).$$

Given $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, ν and $\mathbf{D}$, the expected value of γ_i, $\gamma_i z_i$, c^i and $\log(\gamma_i)$ are

$$\begin{aligned}
\gamma_i^* &= \mathrm{E}[\gamma_i|\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu, \mathbf{D}] \qquad (5.14) \\
&= \frac{T_{\nu+2}\left(\mathbf{x}_i^T \boldsymbol{\beta} \sqrt{(\nu+2)/\nu}\right)(1-\lambda_{10})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10}) + T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{01}} I[y_i = 1] \\
&+ \frac{T_{\nu+2}\left(-\mathbf{x}_i^T \boldsymbol{\beta} \sqrt{(\nu+2)/\nu}\right)\lambda_{01}}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10}) + T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{01}} I[y_i = 1] \\
&+ \frac{T_{\nu+2}\left(\mathbf{x}_i^T \boldsymbol{\beta} \sqrt{(\nu+2)/\nu}\right)\lambda_{10}}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{10} + T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{01})} I[y_i = 0] \\
&+ \frac{T_{\nu+2}\left(-\mathbf{x}_i^T \boldsymbol{\beta} \sqrt{(\nu+2)/\nu}\right)(1-\lambda_{01})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{10} + T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{01})} I[y_i = 0],
\end{aligned}$$

$$
\begin{aligned}
(\gamma_i z_i)^* &= \mathrm{E}[\gamma_i z_i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}] \qquad (5.15)\\
&= \frac{\mathbf{x}_i^T\boldsymbol{\beta}T_{\nu+2}\left(\mathbf{x}_i^T\boldsymbol{\beta}\sqrt{(\nu+2)/\nu}\right)+t_\nu(\mathbf{x}_i^T\boldsymbol{\beta})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10})+T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{01}}(1-\lambda_{10})I[y_i=1]\\
&+ \frac{\mathbf{x}_i^T\boldsymbol{\beta}T_{\nu+2}\left(\mathbf{x}_i^T\boldsymbol{\beta}\sqrt{(\nu+2)/\nu}\right)+t_\nu(\mathbf{x}_i^T\boldsymbol{\beta})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{10}+T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{01})}\lambda_{10}I[y_i=0]\\
&+ \frac{\mathbf{x}_i^T\boldsymbol{\beta}T_{\nu+2}\left(-\mathbf{x}_i^T\boldsymbol{\beta}\sqrt{(\nu+2)/\nu}\right)-t_\nu(\mathbf{x}_i^T\boldsymbol{\beta})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10})+T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{01}}\lambda_{01}I[y_i=1]\\
&+ \frac{\mathbf{x}_i^T\boldsymbol{\beta}T_{\nu+2}\left(-\mathbf{x}_i^T\boldsymbol{\beta}\sqrt{(\nu+2)/\nu}\right)-t_\nu(\mathbf{x}_i^T\boldsymbol{\beta})}{T_\nu(\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{10}+T_\nu(-\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{01})}(1-\lambda_{01})I[y_i=0],
\end{aligned}
$$

$$
\begin{aligned}
c_{11}^{i*} &= \mathrm{E}[c_{11}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}] \qquad (5.16)\\
&= p_i(1-\lambda_{10})I[y_i=1]/\{p_i(1-\lambda_{10})+(1-p_i)\lambda_{01}\},\\
c_{10}^{i*} &= \mathrm{E}[c_{10}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}]\\
&= p_i\lambda_{10}I[y_i=0]/\{p_i\lambda_{10}+(1-p_i)(1-\lambda_{01})\},\\
c_{01}^{i*} &= \mathrm{E}[c_{01}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}]\\
&= (1-p_i)\lambda_{01}I[y_i=1]/\{p_i(1-\lambda_{10})+(1-p_i)\lambda_{01}\},\\
c_{00}^{i*} &= \mathrm{E}[c_{00}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}]\\
&= (1-p_i)(1-\lambda_{01})I[y_i=0]/\{p_i\lambda_{10}+(1-p_i)(1-\lambda_{01})\},
\end{aligned}
$$

and, finally,

$$
(\log(\gamma_i))^* = \mathrm{E}[\log(\gamma_i)|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}] \qquad (5.17)
$$

do not have closed form expressions, but they are computed by numerical integration.

When the probit model is considered, $\mathbf{c}^{i*} = \mathrm{E}[\mathbf{c}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{D}] = \mathrm{E}[\mathbf{c}^i|\boldsymbol{\beta},\boldsymbol{\lambda},\nu,\mathbf{D}]$ given in equation (5.16), and $\mathbf{z}^* = \mathrm{E}[\mathbf{z}|\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{D}]$ is

$$
\begin{aligned}
z_i^* &= \mathrm{E}[z_i|\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{D}] \qquad (5.18)\\
&= \mathbf{x}_i^T\boldsymbol{\beta}+\frac{\phi(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10}-\lambda_{01})}{\Phi(\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{10})+\Phi(-\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{01}}I[y_i=1]\\
&+ \frac{\phi(\mathbf{x}_i^T\boldsymbol{\beta})(\lambda_{10}-1+\lambda_{01})}{\Phi(\mathbf{x}_i^T\boldsymbol{\beta})\lambda_{10}+\Phi(-\mathbf{x}_i^T\boldsymbol{\beta})(1-\lambda_{01})}I[y_i=0].
\end{aligned}
$$

Therefore, the EM algorithm consists of

- E-step: Compute W^*, $(\mathrm{W}\mathbf{z})^*$, $\mathbf{c}^*$ and $(\log(\gamma_i))^*$ using equations (5.14), (5.15), (5.16) and (5.17), respectively. If the probit model is considered, we compute $\mathbf{z}^*$ and $\mathbf{c}^*$ using equations (5.18) and (5.16).

- M-step: Obtain new estimates $\widehat{\boldsymbol{\beta}}^{(t+1)}$, $\widehat{\boldsymbol{\lambda}}^{(t+1)}$ and $\widehat{\nu}^{(t+1)}$ by replacing the new values in equations (5.11), (5.12) and (5.13), respectively.

5.7.2 R code

Probit model considering misclassification. Gibbs sampling

```
### Prior distributions
a10 = 5; b10 = 45; a01 = 5; b01 = 45;   ### lambda prior
be0m = matrix(0,K); be0v = diag(1,K);   ### beta prior: normal

### Elicited prior information
Xt = rbind(c(1,1.7,2.3), c(1,2.2,2), c(1,2.5,1.8)) ;
a1 = c(50,25,5) ;   a2 = c(5,25,50) ;

BinaryMis_Probit <- function(Y,X, N,K, SIM) {
### Sample
   BETA = matrix(0,SIM,K) ;   LAM = matrix(0,SIM,2) ;
### Initial values
   Z = matrix(0,N) ;   betas = solve(t(X)%*%X) %*% (t(X)%*%Y) ;
   L10 = rbeta(1,a10,b10) ;   L01 = rbeta(1,a01,b01) ;
   CC = matrix(0,N,4) ;   CC[,1] = Y ;   CC[,4] = 1-Y ;
   p = the = matrix(0,N) ;   pC = matrix(0,N,4) ;

for(sim in 1:SIM) {

### Normal prior distribution
### Z
#   for(i in 1:N) {   if(CC[i,1]+CC[i,2]==1) {
#      Z[i] = normalleft(0,X[i,]%*%betas,1) } else {
#      Z[i] = normalright(0,X[i,]%*%betas,1) }
#   }
#   Z[Z==Inf | Z==-Inf] = 0
### betas: normal prior distribution
#   bv = solve(t(X)%*%X + solve(be0v))
#   bm = bv %*% (t(X)%*%Z + solve(be0v)%*%be0m)
#   betas <- t(rmnorm(n=1, mean=bm, varcov=bv))

### Eliciting prior information
### betas: elicited prior information
   betas1 <- t(rmnorm(n=1, mean=betas, varcov=solve(t(X)%*%X)))
   u <- runif(1,min=0,max=1)
   if( u<= alpha_probit_mis(betas,betas1,CC) ) { betas = betas1 }

### CC
   for(i in 1:N) {
      p[i] = pnorm(X[i,]%*%betas, mean=0,sd=1)
      the[i] = (p[i]*(1-L10) + (1-p[i])*L01)
      pC[i,1] = p[i]*(1-L10) * ifelse(Y[i]==1,1,0) / the[i]
      pC[i,2] = p[i]*L10 * ifelse(Y[i]==0,1,0) / (1-the[i])
      pC[i,3] = (1-p[i])*L01 * ifelse(Y[i]==1,1,0) / the[i]
      pC[i,4] = (1-p[i])*(1-L01) * ifelse(Y[i]==0,1,0) / (1-the[i])
      CC[i,] <- rmultinom(1,size=1,prob=pC[i,])
   }
```

```
### lambda
   L10 <- rbeta(1, a10+sum(CC[,2]), b10+sum(CC[,1]))
   L01 <- rbeta(1, a01+sum(CC[,3]), b01+sum(CC[,4]))
### Sample
   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;
}
list(beta=BETA, lam=LAM)
}

### Truncated normal distributions
normalright <- function(trb,mu,sig) { rp <- pnorm(trb, mean=mu, sd=sig) ;
   u <- rp*runif(1) ;   qnorm(u, mean=mu, sd=sig) } ### I[Z < trb]
normalleft <- function(tra,mu,sig) { rp <- pnorm(tra, mean=mu, sd=sig) ;
   u <- rp+(1-rp)*runif(1) ;   qnorm(u, mean=mu, sd=sig) } ### I[tra < Z]

### Step Metropolis-Hastings (betas: elicited prior information)
alpha_probit_mis <- function(B0,B1,CC) {
   fB0 = fB1 = matrix(0,N) ;   fI0 = fI1 = matrix(0,K) ;
   for(i in 1:N){
      fB0[i] = ((pnorm(X[i,]%*%B0))^(CC[i,1]+CC[i,2])) *
         ((1-pnorm(X[i,]%*%B0))^(CC[i,3]+CC[i,4]))
      fB1[i] = ((pnorm(X[i,]%*%B1))^(CC[i,1]+CC[i,2])) *
         ((1-pnorm(X[i,]%*%B1))^(CC[i,3]+CC[i,4]))
   }
   for(j in 1:K){
      fI0[j] = (dbeta(pnorm(Xt[j,]%*%B0), shape1=a1[j], shape2=a2[j])) *
         (dnorm(Xt[j,]%*%B0))
      fI1[j] = (dbeta(pnorm(Xt[j,]%*%B1), shape1=a1[j], shape2=a2[j])) *
         (dnorm(Xt[j,]%*%B1))
   }
(prod(fB1/fB0)) * (prod(fI1/fI0))
}
```

Probit model considering misclassification. EM algorithm

```
BinaryMis_Probit_EM <- function(Y,X, N,K, MAX,COT) {
### Sample
   BETA = matrix(0,MAX,K) ;   LAM = matrix(0,MAX,2) ;
### Initial values
   Z = matrix(0,N) ;   betas = solve(t(X)%*%X) %*% (t(X)%*%Y) ;
   L10 = a10/(a10+b10) ;   L01 = a01/(a01+b01) ;   CC = matrix(0,N,4) ;
   sim = 1 ;   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;

while( sim<MAX & max(abs(betas-BETA[sim,]))<COT ) {
   dZ = dnorm(X%*%betas,mean=0,sd=1)
   pZ1 = pnorm(X%*%betas,mean=0,sd=1)
   pZ0 = pnorm(-X%*%betas,mean=0,sd=1)
### Z = E[z | beta, lambda]
   Z = X%*%betas +
   dZ*(1-L10-L01) * ifelse(Y==1,1,0) / (pZ1*(1-L10) + pZ0*L01) +
```

```
   dZ*(-1+L10+L01) * ifelse(Y==0,1,0) / (pZ1*L10 + pZ0*(1-L01))
### CC = E[c | beta, lambda]
   CC[,1] = pZ1*(1-L10) * ifelse(Y==1,1,0) / (pZ1*(1-L10) + pZ0*L01)
   CC[,2] = pZ1*L10 * ifelse(Y==0,1,0) / (pZ1*L10 + pZ0*(1-L01))
   CC[,3] = pZ0*L01 * ifelse(Y==1,1,0) / (pZ1*(1-L10) + pZ0*L01)
   CC[,4] = pZ0*(1-L01) * ifelse(Y==0,1,0) / (pZ1*L10 + pZ0*(1-L01))
### betas: normal prior distribution
   betas = solve(t(X)%*%X+solve(be0v)) %*% (t(X)%*%Z+solve(be0v)%*%be0m)
### lambda
   L10 = (a10+sum(CC[,2])-1) / (a10+sum(CC[,2])-1 + b10+sum(CC[,1])-1)
   L01 = (a01+sum(CC[,3])-1) / (a01+sum(CC[,3])-1 + b01+sum(CC[,4])-1)
### Sample
   sim = sim+1 ;   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;
}
list(beta=BETA, lam=LAM)
}
```

t-link model considering misclassification. Gibbs sampling

```
### Prior distributions
a10 = 5; b10 = 45; a01 = 5; b01 = 45;   ### lambda prior
be0m = matrix(0,K); be0v = diag(1,K);   ### beta prior: normal
nu1 = 1; nu2 = 20; pnu = matrix(0.05,20);   ### nu prior

### Elicited prior information
Xt = rbind(c(1,1.7,2.3), c(1,2.2,2), c(1,2.5,1.8)) ;
a1 = c(50,25,5) ;   a2 = c(5,25,50) ;

BinaryMis_Trv <- function(Y,X, N,K, nu1,nu2,pnu, SIM){
### Sample
   BETA = matrix(0,SIM,K) ;   LAM = matrix(0,SIM,2) ;   NU = matrix(0,SIM) ;
### Initial values
   Z = matrix(0,N) ;   betas = solve(t(X)%*%X) %*% (t(X)%*%Y) ;   nu = 8 ;
   gam <- as.matrix(rgamma(N,shape=nu/2,scale=2/nu)) ;
   W = matrix(0,N,N) ;   diag(W) = gam ;
   L10 = rbeta(1,a10,b10) ;   L01 = rbeta(1,a01,b01) ;
   CC = matrix(0,N,4) ;   CC[,1] = Y ;   CC[,4] = 1-Y ;
   p = the = matrix(0,N) ;   pC = matrix(0,N,4) ;

for(sim in 1:SIM) {

### Normal prior distribution
### Z
#   for(i in 1:N) {   if(CC[i,1]+CC[i,2]==1) {
#      Z[i] = normalleft(0,X[i,]%*%betas, sqrt(1/gam[i])) } else {
#      Z[i] = normalright(0,X[i,]%*%betas, sqrt(1/gam[i])) }
#   }
#   Z[Z==Inf | Z==-Inf] = 0
### betas: normal prior distribution
#   bv = solve(t(X)%*%W%*%X + solve(be0v))
```

```
#   bm = bv %*% (t(X)%*%W%*%Z + solve(be0v)%*%be0m)
#   betas <- t(rmnorm(n=1, mean=bm, varcov=bv))
### gam
#   for(i in 1:N){
#   gam[i] <- rgamma(1,shape=(nu+1)/2,scale=2/(nu+(Z[i]-X[i,]%*%betas)^2))
#   }
#   diag(W) = gam
### nu
#   Fnu = F_nu(nu1,nu2,pnu, N,gam)
#   nu = Q_nu(runif(1),Fnu)

### Eliciting prior information
### betas: elicited prior information
   betas1 <- t(rmnorm(n=1, mean=betas, varcov=solve(t(X)%*%X)))
   u <- runif(1,min=0,max=1)
   if( u<=alpha_t_mis(betas,betas1, nu,CC) ) { betas = betas1 }
### nu: elicited prior information
   Fnu = FnuBed_mis(betas,nu1,nu2,pnu,CC)
   nu = Q_nu(runif(1),Fnu)

### CC
   for(i in 1:N){
      p[i] = pt(X[i,]%*%betas, df=nu)
      the[i] = (p[i]*(1-L10) + (1-p[i])*L01)
      pC[i,1] = p[i]*(1-L10) * ifelse(Y[i]==1,1,0) / the[i]
      pC[i,2] = p[i]*L10 * ifelse(Y[i]==0,1,0) / (1-the[i])
      pC[i,3] = (1-p[i])*L01 * ifelse(Y[i]==1,1,0) / the[i]
      pC[i,4] = (1-p[i])*(1-L01) * ifelse(Y[i]==0,1,0) / (1-the[i])
      CC[i,] <- rmultinom(1,size=1,prob=pC[i,])
   }
### lambda
   L10 <- rbeta(1, a10+sum(CC[,2]), b10+sum(CC[,1]))
   L01 <- rbeta(1, a01+sum(CC[,3]), b01+sum(CC[,4]))
### Sample
   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;   NU[sim] = nu ;
}
list(beta=BETA, lam=LAM, nu=NU)
}

### Truncated normal distributions
normalright <- function(trb,mu,sig) { rp <- pnorm(trb, mean=mu, sd=sig) ;
   u <- rp*runif(1) ;   qnorm(u, mean=mu, sd=sig) } ### I[Z < trb]
normalleft <- function(tra,mu,sig) { rp <- pnorm(tra, mean=mu, sd=sig) ;
   u <- rp+(1-rp)*runif(1) ;   qnorm(u, mean=mu, sd=sig) } ### I[tra < Z]

### Distribution function of nu, I[m1 <= nu <= m2], prior=pnu
F_nu <- function(m1,m2,pnu, N,gam) {
   ni = (m1:m2)
   Fni = fni = matrix(0,m2-m1+1)
   for(i in 1:(m2-m1+1)) {
```

```
      fni[i] = ((prod(gam))^(ni[i]/2-1)) * exp(-ni[i]*sum(gam)/2) /
         (((gamma(ni[i]/2))*((2/ni[i])^(ni[i]/2)))^N)
      fni[i] = fni[i]*pnu[i]
   }
   fni[fni=="NaN"] = 0
   Fni = cumsum(fni) / sum(fni,na.rm=TRUE)
   as.matrix(cbind(ni,fni,Fni))
}
### Quantil of nu
Q_nu <- function(u,Fnu) {
   m = dim(Fnu)[1]
   for(i in 1:m){
      if(u<=Fnu[i,3]){break}
   }
   Fnu[i,1]
}

### Step Metropolis-Hastings (betas, nu: elicited prior information)

alpha_t_mis <- function(betas0,betas1, nu,CC){
   fB0 = fB1 = matrix(0,N) ;   fI0 = fI1 = matrix(0,K) ;
   for(i in 1:N) {
      fB0[i] = ((pt(X[i,]%*%betas0,df=nu))^(CC[i,1]+CC[i,2])) *
         ((1-pt(X[i,]%*%betas0,df=nu))^(CC[i,3]+CC[i,4]))
      fB1[i] = ((pt(X[i,]%*%betas1,df=nu))^(CC[i,1]+CC[i,2])) *
         ((1-pt(X[i,]%*%betas1,df=nu))^(CC[i,3]+CC[i,4]))
   }
   for(j in 1:K){
      fI0[j] = (dbeta(pt(Xt[j,]%*%betas0,df=nu), shape1=a1[j],
         shape2=a2[j])) * (dt(Xt[j,]%*%betas0,df=nu))
      fI1[j] = (dbeta(pt(Xt[j,]%*%betas1,df=nu), shape1=a1[j],
         shape2=a2[j])) * (dt(Xt[j,]%*%betas1,df=nu))
   }
(prod(fB1/fB0)) * (prod(fI1/fI0))
}

### Distribution function of nu, I[m1 <= nu <= m2], prior=pnu
FnuBed_mis <- function(betas,nu1,nu2,pnu,CC){
   nui = (nu1:nu2) ;   Fni = fni = matrix(0,nu2-nu1+1) ;
   for(i in 1:(nu2-nu1+1)){
      fB = matrix(0,N) ;   fI = matrix(0,K) ;
      for(j1 in 1:N){
         fB[j1] = ((pt(X[j1,]%*%betas,df=nui[i]))^(CC[j1,1]+CC[j1,2])) *
            ((1-pt(X[j1,]%*%betas,df=nui[i]))^(CC[j1,3]+CC[j1,4]))
      }
      for(j2 in 1:K){
         fI[j2] = dbeta(pt(Xt[j2,]%*%betas,df=nui[i]), shape1=a1[j2],
            shape2=a2[j2]) * dt(Xt[j2,]%*%betas,df=nui[i])
      }
      fni[i] = pnu[i] * (prod(fB)) * (prod(fI))
```

```
   }
   Fni = cumsum(fni) / sum(fni,na.rm=TRUE)
   as.matrix(cbind(nui,fni,Fni))
}
```

t-link model considering misclassification. EM algorithm

```
BinaryMis_Trv_EM <- function(Y,X, N,K, nu1,nu2,pnu, MAX,COT){
### Sample
BETA = matrix(0,MAX,K) ;   LAM = matrix(0,MAX,2) ;   NU = matrix(0,MAX) ;
### Initial values
   Z = matrix(0,N) ;   betas = solve(t(X)%*%X) %*% (t(X)%*%Y) ;
   L10 = a10/(a10+b10) ;   L01 = a01/(a01+b01) ;   CC = matrix(0,N,4) ;
   W = matrix(0,N,N) ;   gam = matrix(1,N) ;   diag(W) <- gam ;
   sim = 1 ;   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;

while( sim<MAX & max(abs(betas-BETA[sim,]))<COT ) {
   dZ = dt(X%*%betas,df=nu)
   pZ1 = pt(X%*%betas,df=nu)
   pZ0 = pt(-X%*%betas,df=nu)
   pZ3 = pt(X%*%betas*sqrt((nu+2)/nu),df=nu+2)
   pZ2 = pt(-X%*%betas*sqrt((nu+2)/nu),df=nu+2)
   the = (pZ1*(1-L10) + pZ0*L01)   ### 1-the = (pZ1*L10 + pZ0*(1-L01))
### CC = E[c | beta, lambda, nu]
   CC[,1] = pZ1*(1-L10) * ifelse(Y==1,1,0) / the
   CC[,2] = pZ1*L10 * ifelse(Y==0,1,0) / (1-the)
   CC[,3] = pZ0*L01 * ifelse(Y==1,1,0) / (the)
   CC[,4] = pZ0*(1-L01) * ifelse(Y==0,1,0) / (1-the)
### gam = E[gam | beta, lambda, nu]
   gam = (pZ3*(1-L10) + pZ2*L01) * ifelse(Y==1,1,0) / the  +
         (pZ3*L10 + pZ2*(1-L01)) * ifelse(Y==0,1,0) / (1-the)
   diag(W) = gam
### WZ = E[gam*z | beta, lambda, nu]
   WZ = (X%*%betas*pZ3 + dZ)*(1-L10) * ifelse(Y==1,1,0) / the +
        (X%*%betas*pZ3 + dZ)*L10 * ifelse(Y==0,1,0) / (1-the) +
        (X%*%betas*pZ2 - dZ)*L01 * ifelse(Y==1,1,0) / the +
        (X%*%betas*pZ2 - dZ)*(1-L01) * ifelse(Y==0,1,0) / (1-the)
### logW = E[log(gam) | beta, lambda, nu]
   for(i in 1:N) { logW[i] = Eloglam(10,i, X,betas,nu,L10,L01) }
### betas : normal prior distribution
   betas = solve(t(X)%*%W%*%X + solve(be0v)) %*%
      (t(X)%*%WZ + solve(be0v)%*%be0m)
### nu
   nu = argmax_gnu(nu1,nu2,pnu, N,logW,W)
### lambda
   L10 = (a10+sum(CC[,2])-1) / (a10+sum(CC[,2])-1 + b10+sum(CC[,1])-1)
   L01 = (a01+sum(CC[,3])-1) / (a01+sum(CC[,3])-1 + b01+sum(CC[,4])-1)
### Sample
   sim = sim+1 ;   BETA[sim,] = betas ;   LAM[sim,] = c(L10,L01) ;
   NU[sim] = nu ;
```

```
}
list(beta=BETA, lam=LAM, nu=NU)
}

###  E[log(gam_i) | beta, lambda, nu]
Eloglam <- function(Ulam,i, X,betas,nu,L10,L01) {
   lami = (1:(Ulam*100))/100
   dN1 = pnorm(sqrt(lami)*X[i,]%*%betas)
   dN0 = pnorm(sqrt(lami)*X[i,]%*%betas)
   dT1 = pt(X[i,]%*%betas,df=nu)
   dT0 = pt(-X[i,]%*%betas,df=nu)
   if(Y[i]==1) { A = (dN1*(1-L10) + dN0*L01) / (dT1*(1-L10) + dT0*L01)) }
   if(Y[i]==0) { A = (dN1*L10 + dN0*(1-L01)) / (dT1*L10 + dT0*(1-L01)) }
   logf = log(lami) * dgamma(lami, shape=nu/2, scale=2/nu) * A ;
   0.01*sum(logf)
}

###
argmax_gnu <- function(m1,m2,pnu, N,logW,W) {
   ni = (m1:m2)
   gnu = matrix(0,m2-m1+1)
   for(j in 1:(m2-m1+1)) {
      gnu[j] = log(pnu[j]) + (ni[j]/2)*sum(logW) - (ni[j]/2)*sum(W) +
               N*log((gamma(ni[j]/2)*((2/ni[j])^(ni[j]/2)))^(-1))
   }
   ni[gnu==max(gnu)]
}
```

Chapter 6

A Bayesian approach for misclassified ordinal response data

Naranjo, L., Pérez, C. J., Martín, J., and Lesaffre, E. (2014). A Bayesian approach for misclassified ordinal response data. Preprint 155, Universidad de Extremadura, Badajoz, Spain.

Abstract

Motivated by a longitudinal oral health study, the Signal-Tandmobiel® study, a Bayesian approach has been addressed to model misclassified ordinal response data. Two regression models have been considered to incorporate misclassification in the categorical response. Specifically, probit and logit models have been developed. The computational difficulties have been avoided by using data augmentation frameworks. This idea is exploited to derive efficient Markov chain Monte Carlo methods. Although the method is proposed for ordered categories, it can also be implemented for unordered ones in a simple way. The model perfomance is shown through a simulation-based example, and the analysis of the motivating study is presented.

Keywords: Bayesian analysis; Data augmentation; Latent variables; Markov chain Monte Carlo methods; Misclassification; Ordinal regression model.

6.1 Introduction

Dental caries is one of the most prevalent chronic diseases worldwide, affecting persons in all age groups. Many epidemiological surveys and clinical studies are carried out to obtain a further understanding of this disease entity. However, the process of detecting caries experience (CE) is not an obvious issue. CE scoring may not perfectly reflect the tooth's true condition, and therefore, the presence of CE can be misdiagnosed, leading to misclassified outcomes. In order to standardize data collection techniques in epidemiological surveys and clinical trials, CE assessment guidelines have been developed by the International Caries Detection and Assessment System (ICDAS (2005)). These guidelines highlight the need for training the examiners and measuring the reliability of the obtained scores. However, despite these criteria, the process of CE detection is subject to misclassification.

In situations where misclassification may happen, additional parameters are necessary to correct the bias yielded by the use of non free-error data. If misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non-optimal decision-making. Therefore, a correction for misclassification is needed to obtain unbiased estimates for the regression coefficients. Statistical models addressing misclassification should be available in these contexts.

Several models to address misclassification on binary regression have been proposed in the scientific literature, see, e.g., Rekaya et al. (2001), Paulino et al. (2003), McInturff et al. (2004), Paulino et al. (2005), McGlothlin et al. (2008), and Naranjo et al. (2014a). Gustafson (2003) and Buonaccorsi (2010) presented reviews on the effects of misclassification on model estimates. However, the number of models considering measurement errors for polychotomous response data is dramatically reduced. Computations in multidimentional settings are more difficult, and this case is not an exception. In fact, approaches to polychotomous response data regression models addressing misclassification have not been yet exhaustively developed in the literature. Albert et al. (1997) proposed a class of models to analyze repeated monotonic ordinal responses with diagnostic misclassification. They separately modeled the underlying monotonic response and the misclassification process, by developing an EM algorithm for maximum likelihood estimation that incorporates covariates and randomly missing data. Mwalili et al. (2005) presented a Bayesian approach for correcting interobserver measurement error in an ordinal logistic regression model taking into account the variability of the estimated correction terms. Roy and Banerjee (2009) considered a multivariate probit model for correlated binary responses. Some of the responses were subject to classification errors and hence they were not directly observable. Besides, measurements on some of the predictors were not available, instead measurements on their surrogate were available. However, the conditional distribution of the unobservable predictors given the surrogate was completely specified. They proposed models based on likelihood methodologies that take into account either or both of these sources of errors.

Motivated by a longitudinal oral health study, the Signal-Tandmobiel® study (ST), a Bayesian approach to address misclassified ordinal response data is proposed and discussed in this paper. A regression model is developed to incorporate mis-

classification in categorical response. A data augmentation framework is proposed to derive efficient Markov chain Monte Carlo (MCMC) algorithms to polychotomous response data that are subject to misclassification. Although only the ordered case is explored, the approach can be extended for unordered categories. This approach generalizes the binary probit regression model addressing misclassification proposed by Naranjo et al. (2014a) and the data augmentation scheme for ordinal regression models proposed Albert and Chib (1993). The model performance is illustrated with a simulated-based example, and the analysis of the motivating ST data is presented.

The outline of the paper is as follows. The ST study is introduced in Section 6.2, illustrating the need of addressing misclassification. The way misclassification is addressed in polychotomous response data models is presented in Section 6.3. In Section 6.4, the prior distributions are described and the posterior distributions are explored. Ordered categories for both probit and logit models are considered. Section 6.5 shows the model performance for a simulation-based example, whereas the analysis of the ST data is presented in Section 6.6. Finally, Section 6.7 presents the conclusion and some future research lines.

6.2 The Signal-Tandmobiel® study

The Signal-Tandmobiel® study is a longitudinal prospective oral health intervention project conducted in Flanders (North of Belgium), between 1996 and 2001. For this project, 4468 children (2315 boys and 2153 girls) were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen trained dentists (examiners) based on visual and tactile observations. The clinical examinations took place in a mobile dental clinic, with a standard chair and artificial dental light. No radiographs were taken. Data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents. For a more detailed description of the study design and research methods see Vanobbergen et al. (2000).

In this work, caries lesions were scored in four ordinal levels of lesion severity. Caries experience (CE) is considered as an ordinal variable indicating whether the tooth is at the level defining a progressive disease. The statistical findings were applied to the scoring of the four permanent first molars, i.e., teeth 16 and 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). The numbering of teeth follows the notation of the Federation Dentaire Internationale, which indicates the position of the tooth in the mouth. Diagnosing CE is difficult for a variety of reasons. For instance, composite materials can imitate the natural enamel so well that it is difficult to spot a restored lesion; or the location of the cavity, far back in the mouth, hampers the view of the dental examiner. Hence, overlooking CE is likely to happen in practice, but the dental examiner could also classify discolorations as CE.

In the ST study, 16 dental examiners were calibrated for scoring CE. The calibration exercises were performed according to the guidelines of training and calibration published by the British Association for the Study of Community Dentistry (Pitts et al. (1997)). The calibration of the dental examiners was performed by comparing

their scores on the tooth surfaces of a group of children to those of a benchmark examiner. Note that there exists no infallible scorer for CE. The best one can do is to take a very experienced dental examiner, called benchmark. In order to maintain a high level of intra- and inter-examiner reliability, calibration exercises were carried out twice a year for all examiners involved. During the study period (1996-2001), three calibration exercises were devoted to the scoring of CE (1996, 1998, 2000), involving 92, 32 and 224 children, respectively. A contingency table of dental examiners and the benchmark examiner was determined, yielding a table with misclassified scores. Data of the three calibration exercises were combined into one validation dataset, and also examiners' data were combined into one. All examiners were lumped together, but the approach can be generalized to take into account multiple examiners. The results suggested that examiners overscore or underscore the true CE status.

In the main dataset the dental examiners scored the children, but their scores are likely to be prone to error. Ignoring in the statistical analysis that the levels of CE lesion severity are prone to misclassification might lead to wrong estimates, and so, to wrong conclusions. Bayesian ordinal regression models considering misclassification can help to provide better predictions than standard models.

6.3 Addressing misclassification in polychotomous response data models

Suppose that n independent random variables $Y_1, \ldots, Y_n$ are observed, where Y_i takes one of J categories, $i = 1, \ldots, n$. Suppose that $Y_1, \ldots, Y_n$ are prone to error. Let $\theta_{is} = \mathrm{p}(Y_i = s|\mathbf{x}_i)$ denote the probability that the i-th observation with covariate pattern $\mathbf{x}_i$ is classified in the s-th category (it is possibly misclassified), $s = 1, \ldots, J$. The parameters θ_{is} are related to a set of covariates $\mathbf{x}_i$ through a regression model that considers misclassification. They are defined as

$$\theta_{is} = \mathrm{p}(Y_i = s|\mathbf{x}_i) = \sum_{r=1}^{J} \mathrm{p}(Y_i = s|\xi_i = r)\mathrm{p}(\xi_i = r|\mathbf{x}_i) = \sum_{r=1}^{J} \lambda_{rs} p_{ir},$$

where $\lambda_{rs} = \mathrm{p}(Y_i = s|\xi_i = r)$ is the probability that an observation y_i is classified in the s-th category when the true category ξ_i is the r-th one, and $p_{ir} = \mathrm{p}(\xi_i = r|\mathbf{x}_i)$ denotes the probability that the true category for an observation with covariate pattern $\mathbf{x}_i$ is the r-th.

Note that $\boldsymbol{\xi}$ is an unknown random vector of the true classifications, and ξ_i has a categorical distribution $\xi_i \sim \mathrm{Cat}(p_{i1}, \ldots, p_{iJ})$ whose vector of probabilities is $(p_{i1}, \ldots, p_{iJ})$, where $p_{ir} = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}, r)$, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown regression parameters, and g is the link function that usually depends on a cumulative distribution function (cdf). Various link functions can be used for g, but the most common are the logit and probit links, where g depends on the cdf of a logistic and a normal distribution, respectively (see, for example, Johnson and Albert (1999) and Congdon (2006)). These two link functions will be considered to develop the proposed regression models.

The likelihood function for a model considering misclassification can be expressed as

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda}|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \sum_{s=1}^{J} \sum_{r=1}^{J} \lambda_{rs} p_{ir} I[y_i = s],$$

where $I[\cdot]$ denotes the indicator function, i.e., $I[A] = 1$ if A is true, and $I[A] = 0$ otherwise, and $\boldsymbol{\lambda}$ is a matrix

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_J \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1J} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{J1} & \lambda_{J2} & \cdots & \lambda_{JJ} \end{pmatrix},$$

where $\sum_{s=1}^{J} \lambda_{rs} = 1$, and the elements of the diagonal, λ_{rr} for $r = 1, \ldots, J$, denote the probabilities of correct classification.

Latent variables related with the misclassification are introduced to simplify the generation process. Binary latent variables c_{rs}^i are defined, $r, s = 1, \ldots, J$, where r represents the index for the true value and s represents the index for the observed value. When the latent variable takes value one, it denotes the group where the ith observation has been assigned: the true category is r and the observed category is s, i.e. $c_{rs}^i = 1$ if $\xi_i = r$ and $y_i = s$. Note that $c_{+s}^i = \sum_{r=1}^{J} c_{rs}^i = 1$ when the observed category is s, i.e. $y_i = s$, and $c_{r+}^i = \sum_{s=1}^{J} c_{rs}^i = 1$ means that the true category is r, i.e. $\xi_i = r$. For each $i = 1, \ldots, n$, a latent matrix $\mathbf{c}^i$ is defined

$$\mathbf{c}^i = \begin{pmatrix} c_{11}^i & c_{12}^i & \cdots & c_{1J}^i \\ c_{21}^i & c_{22}^i & \cdots & c_{2J}^i \\ \vdots & \vdots & \ddots & \vdots \\ c_{J1}^i & c_{J2}^i & \cdots & c_{JJ}^i \end{pmatrix}.$$

Then, an augmented likelihood function is considered

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda}|\mathbf{c}, \mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \left[\prod_{r=1}^{J} \prod_{s=1}^{J} \lambda_{rs}^{c_{rs}^i} \right] \left[\prod_{r=1}^{J} p_{ir}^{c_{r+}^i} \right] \left[\sum_{s=1}^{J} I[y_i = s] I[c_{+s}^i = 1] \right].$$

This data augmentation scheme allows to derive easy-to-implement MCMC algorithms in the context of polychotomous regression models considering misclassification.

6.4 Exploring the posterior distributions in ordered categories

In this section the prior distributions are presented, which together with the specifications of the previous section allow to derive MCMC sampling algorithms (see Gilks et al. (1996)) to sample from the posterior distributions.

For ordered response categories the ordinal regression model is defined by cutpoints $\gamma_0, \gamma_1, \ldots, \gamma_{J-1}, \gamma_J$ considering that $p_{ir} = \Psi\left(\gamma_r - \mathbf{x}_i^T\boldsymbol{\beta}\right) - \Psi\left(\gamma_{r-1} - \mathbf{x}_i^T\boldsymbol{\beta}\right)$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})^T$ is the vector of unknown cutpoints, $\gamma_0 = -\infty$, $\gamma_J = \infty$, and Ψ is a cdf. Note that if a constant term is included in $\mathbf{x}_i$ and $\boldsymbol{\beta}$ includes an intercept, then there are only $J-2$ unknown cutpoints $\gamma_2, \ldots, \gamma_{J-1}$, with $\gamma_1 = 0$ (see Johnson and Albert (1999)). This option is often less complex for numerical stability in sampling (see Congdon (2005)).

6.4.1 Prior distributions

The prior distribution for the regression parameter vector is usually a multivariate normal $\mathrm{N}_k(\mathbf{b}_0, \mathbf{B}_0)$, that is, $\pi(\boldsymbol{\beta}) \propto \exp\left\{-(\boldsymbol{\beta} - \mathbf{b}_0)^T\mathbf{B}_0^{-1}(\boldsymbol{\beta} - \mathbf{b}_0)/2\right\}$. There is some literature addressing informative prior elicitation for generalized linear models. Bedrick et al. (1996) proposed the conditional means prior approach to introduce a prior distribution on the regression parameters, and Chen et al. (2000a) proposed power prior distributions for the regression parameters based on the notion of availability of historical data. All of them assume models without errors. The literature about informative prior elicitation to binomial regression models with misclassification is mainly focused on the regression parameter vector, see some applications in McInturff et al. (2004), Paulino et al. (2005) and Naranjo et al. (2014a).

The prior distribution for the cutpoints $\boldsymbol{\gamma}$, $\pi(\boldsymbol{\gamma})$, is usually multivariate normal, $\mathrm{N}_{J-1}(\mathbf{l}_0, \mathbf{L}_0)$, or proportional to 1, $\pi(\boldsymbol{\gamma}) \propto 1$, see Albert and Chib (1993).

For the misclassification parameters, since $\lambda_{rs} \in (0,1)$ and $\sum_{s=1}^{J}\lambda_{rs} = 1$, for $r, s = 1, \ldots, J$, the natural prior distributions for $\boldsymbol{\lambda}_r$ are Dirichlet, $\boldsymbol{\lambda}_r \sim \mathrm{Dirichlet}(a_{r1}, \ldots, a_{rJ})$, where $a_{rs} > 0$, whose probability density function is

$$\pi(\boldsymbol{\lambda}_r) = \frac{\Gamma\left(\sum_{s=1}^{J} a_{rs}\right)}{\prod_{s=1}^{J}\Gamma(a_{rs})}\prod_{s=1}^{J}\lambda_{rs}^{a_{rs}-1} \propto \prod_{s=1}^{J}\lambda_{rs}^{a_{rs}-1}.$$

When the response is presented in an ordinal scale, adjacent categories have bigger risk to be misclassified, so that the natural constraints are $\lambda_{r1} < \cdots < \lambda_{r,r-1} < \lambda_{rr}$ and $\lambda_{rr} > \lambda_{r,r+1} > \cdots > \lambda_{rJ}$. In case of nominal response data, there is no correlation between categories, however, a natural constraint is to assume that the correct classification probability is greater than the misclassification probabilities, i.e., $\lambda_{rr} > \lambda_{rs}$. Alternative sets of constraints on the parameters are presented by Swartz et al. (2004).

6.4.2 Posterior distributions

The joint posterior distribution of the unobservables $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\mathbf{c}$, and $\boldsymbol{\lambda}$ is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{c}, \boldsymbol{\lambda}|\mathbf{y}, \mathbf{x}) \quad \propto \quad \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\lambda})\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda}|\mathbf{c}, \mathbf{y}, \mathbf{x}).$$

In order to derive Gibbs sampling algorithms, the full conditional distributions must be obtained. The full conditional distributions for $\boldsymbol{\lambda}$ and $\mathbf{c}$ are easy to obtain. Specifically, the full conditional distributions for $\mathbf{c}$ given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, the data $\mathbf{y}$, and the

covariates $\mathbf{x}$ is

$$\pi(\mathbf{c}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}) \propto \prod_{i=1}^{n}\left[\prod_{r=1}^{J}\prod_{s=1}^{J}\lambda_{rs}^{c_{rs}^i}\right]\left[\prod_{r=1}^{J}p_{ir}^{c_{r+}^i}\right]\left[\sum_{s=1}^{J}I[y_i=s]I[c_{+s}^i=1]\right],$$

where $\sum_{r=1}^{J}\sum_{s=1}^{J}c_{rs}^i=1$, so that for $i=1,\ldots,n$,

$$\left[c_{+s}^i|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}\right] \sim \text{Multinomial}\left(1,\pi_{c_{+s}^i}\right)I[y_i=s], \tag{6.1}$$

and $c_{+j}^i=(0,\ldots,0)$ for $j\neq s$, where

$$\pi_{c_{+s}^i}=\frac{\lambda_{rs}p_{ir}}{\sum_{j=1}^{J}\lambda_{js}p_{ij}}I[y_i=s].$$

The full conditional distribution for $\boldsymbol{\lambda}$ given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\mathbf{c}$, the data $\mathbf{y}$, and the covariates $\mathbf{x}$ is

$$\pi(\boldsymbol{\lambda}|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{c},\mathbf{y},\mathbf{x}) \propto \prod_{r=1}^{J}\prod_{s=1}^{J}\lambda_{rs}^{\sum_{i=1}^{n}c_{rs}^i+a_{rs}-1},$$

where $\sum_{s=1}^{J}\lambda_{rs}=1$, so that for $r=1,\ldots,J$,

$$[\boldsymbol{\lambda}_r|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{c},\mathbf{y},\mathbf{x}] \sim \text{Dirichlet}\left(\sum_{i=1}^{n}c_{r1}^i+a_{r1},\ldots,\sum_{i=1}^{n}c_{rJ}^i+a_{rJ}\right). \tag{6.2}$$

If a constraint on $\boldsymbol{\lambda}$ is considered, $\lambda_{r1}<\cdots<\lambda_{r,r-1}<\lambda_{rr}$ and $\lambda_{rr}>\lambda_{r,r+1}>\cdots>\lambda_{rJ}$ in case of ordinal data, or $\lambda_{rs}<\lambda_{rr}$ in case of nominal data, the distribution is a truncated Dirichlet according to the chosen constraint.

However, the full conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x})$ and $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x})$ do not have closed expressions to easily generate from, because these are given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}) \propto \pi(\boldsymbol{\beta})\prod_{i=1}^{n}\prod_{r=1}^{J}p_{ir}^{c_{r+}^i} \quad \text{and} \quad \pi(\boldsymbol{\gamma}|\boldsymbol{\beta},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}) \propto \pi(\boldsymbol{\gamma})\prod_{i=1}^{n}\prod_{r=1}^{J}p_{ir}^{c_{r+}^i}.$$

Our proposal considers the introduction of latent variables in order to allow other easy-to-sample steps within this Gibbs sampling. These latent variables are based on the data augmentation framework of the ordinal regression model proposed by Albert and Chib (1993). Independent latent continuous random variables $Z_1,\ldots,Z_n$ are assumed, whose cdf is given by Ψ, $c_{r+}^i=1$ if $\gamma_{r-1}<z_i<\gamma_r$, and $c_{r+}^i=0$ otherwise. The new joint posterior distribution of interest is given by

$$\pi(\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\gamma}|\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma})\prod_{i=1}^{n}\psi\left(z_i-\mathbf{x}_i^T\boldsymbol{\beta}\right)\sum_{r=1}^{J}I\left[\gamma_{r-1}<z_i<\gamma_r\right]I\left[c_{r+}^i=1\right],$$

where ψ is the probability density function (pdf) of Z_i.

In order to generate from this distribution, again Gibbs sampling is considered, so a Gibbs-within-Gibbs sampling algorithm is developed. In case of a logit link function, a Metropolis-Hastings step is required.

Ordinal probit addressing misclassification

Let $\Psi = \Phi$ be the cdf of a standard normal distribution, $\mathrm{N}(0,1)$, then the conditional posterior distributions of the Gibbs-within-Gibbs sampling are the following:

$$[Z_i|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}] \sim \mathrm{N}\left(\mathbf{x}_i^T\boldsymbol{\beta},1\right) \times \sum_{r=1}^{J} I\left[\gamma_{r-1} < z_i < \gamma_r\right] I\left[c_{r+}^i = 1\right] \quad (6.3)$$

$$[\boldsymbol{\beta}|\mathbf{z},\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}] \sim \mathrm{N}_k\left(\widehat{\mathbf{b}},\widehat{\mathbf{B}}\right), \quad (6.4)$$

$$\left[\gamma_r|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\gamma}_{(-r)},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}\right] \sim \mathrm{N}\left(\mathbf{l}_{0r},\mathbf{L}_{0rr}\right) I\left[\underline{\gamma_r},\overline{\gamma_r}\right], \quad (6.5)$$

where $\boldsymbol{\gamma}_{(-r)} = \{\gamma_l : l \neq r\}$,

$$\widehat{\mathbf{b}} = \left(\mathbf{B}_0^{-1} + \mathbf{x}^T\mathbf{x}\right)^{-1}\left(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{x}^T\mathbf{z}\right), \quad \underline{\gamma_r} = \max\left\{\max\left\{z_i : c_{r+}^i = 1\right\}, \gamma_{r-1}\right\},$$
$$\widehat{\mathbf{B}} = \left(\mathbf{B}_0^{-1} + \mathbf{x}^T\mathbf{x}\right)^{-1}, \quad \overline{\gamma_r} = \min\left\{\min\left\{z_i : c_{r+1,+}^i = 1\right\}, \gamma_{r+1}\right\}.$$

If the prior distribution $\pi(\boldsymbol{\gamma}) \propto 1$ is used for the cutpoints, then the conditional posterior distribution for γ_r is $\left[\gamma_r|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\gamma}_{(-r)},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}\right] \sim \mathrm{U}\left(\underline{\gamma_r},\overline{\gamma_r}\right)$. Note that sampling from all the full conditional distributions is easy. Specifically, all the full conditional distributions are standard.

The final algorithm consists of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $\mathbf{c}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, and iteratively sampling $\boldsymbol{z}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$, $\mathbf{c}^{(j)}$ and $\boldsymbol{\lambda}^{(j)}$ from the full conditional distributions (6.3), (6.4), (6.5), (6.1) and (6.2) respectively. The following initial values are proposed to be set: $\boldsymbol{\beta}^{(0)} = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{y}$, $\gamma_j^{(0)} = q_z\left(\frac{1}{2} + \frac{j-1}{2(J-1)}\right)$, for $j = 2,\ldots,J-1$, where $q_z(t)$ is the t-th quantile of a distribution $\mathrm{N}(0,1)$, $c_{y_iy_i}^{i(0)} = 1$ and $c_{rs}^{i(0)} = 0$ if $r \neq y_i$ and $s \neq y_i$ (i.e. $\xi_i = y_i$), and $\boldsymbol{\lambda}^{(0)} = \mathrm{diag}_J(1)$.

Ordinal logit addressing misclassification

Let Ψ be the cdf of a standard logistic distribution, $\mathrm{L}(0,1)$, then it is possible to use a similar algorithm as in case of probit link by replacing the link function and including Metropolis-Hastings updates for $\boldsymbol{\beta}$.

Specifically, the following changes are needed. In equation (6.3) the latent variables z_i are sampled from a truncated normal distribution, but now they are sampled from a truncated logistic distribution

$$[Z_i|\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}] \sim \mathrm{L}\left(\mathbf{x}_i^T\boldsymbol{\beta},1\right) \times \sum_{r=1}^{J} I\left[\gamma_{r-1} < z_i < \gamma_r\right] I\left[c_{r+}^i = 1\right]. \quad (6.6)$$

Now, the conditional posterior distribution of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|\mathbf{z},\boldsymbol{\gamma},\mathbf{c},\boldsymbol{\lambda},\mathbf{y},\mathbf{x}] \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} l\left(z_i - \mathbf{x}_i^T\boldsymbol{\beta}\right) \sum_{r=1}^{J} I\left[\gamma_{r-1} < z_i < \gamma_r\right] I\left[c_{r+}^i = 1\right], \quad (6.7)$$

where l is the pdf of a standard logistic distribution. This full conditional distribution is not standard, Metropolis-Hastings updates can be performed by using a multivariate normal proposal density centered on the previously sampled value of $\boldsymbol{\beta}$ and having covariance matrix proportional to $(\mathbf{x}^T\mathbf{x})^{-1}$.

The final algorithm consists of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $\mathbf{c}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, and iteratively sampling $\boldsymbol{z}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$, $\mathbf{c}^{(j)}$ and $\boldsymbol{\lambda}^{(j)}$ from the full conditional distributions (6.6), (6.7), (6.5), (6.1) and (6.2), respectively. In order to sample from (6.7), Metropolis-Hastings updates are required, where the proposal distribution is $\mathrm{N}_k\left(\boldsymbol{\beta}^{(j-1)}, (\mathbf{x}^T\mathbf{x})^{-1}\right)$. The initial values are proposed in a similar way as those from the probit link.

6.5 Simulation-based example

A simulation-based study has been carried out to analyze the model performance of the proposed approach. In this section, an example is presented to illustrate the advantages of using this approach.

Several criteria have been considered for model performance. The deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) is evaluated as $DIC = 2\overline{D(\eta)} - D(\overline{\eta})$, where $D(\eta) = -2\log L(\eta)$ is the deviance of the model, $L(\eta)$ is the likelihood, $\overline{D(\eta)} = \mathrm{E}(D(\eta)|\text{data})$ is the posterior mean of the deviance, and $D(\overline{\eta})$ is the deviance at the posterior means of the parameters of interest $\overline{\eta} = \mathrm{E}(\eta|\text{data})$. Another criterion is the total variation distance (TVD) between the true and estimated probabilities. It is defined as $TVD = \sum_{i=1}^{n}\sum_{r=1}^{J}|p_{ir} - \widehat{p_{ir}}|$, and it has been proposed to measure the discrepancy between the true probabilities and the estimated ones in simulation-based scenarios (see Naranjo et al. (2014a)). Finally, the third criterion that will be considered in this section is the pseudo-predictors method (see Czado et al. (2011)). The variables p_{ir}^{obs}, where $p_{ir}^{obs} = 1$ if $y_i = r$ and $p_{ir}^{obs} = 0$ otherwise, correspond to the observed probabilities for category r at the ith observation in contrast to the predicted probabilities $\widehat{p_{ir}}$. When category r is observed at the observation i, it is clear that a good model fit leads to a high probability $\widehat{p_{ir}}$, and to small probabilities $\widehat{p_{ij}}$ for other categories $j \neq r$. Large differences should be penalized more than small differences. Then, the verification score is defined as $S = \frac{1}{n}\sum_{r=1}^{J}\sum_{i=1}^{n}\left(p_{ir}^{obs} - \widehat{p_{ir}}\right)^2$, providing an idea of the model fit. For all the criteria, models with smaller criteria values are preferred over models with large values.

Multiple misclassified ordinal response data are generated. The main objective is to compare the performance of the proposed ordinal regression model addressing misclassification with the standard ordinal regression model. This simulation-based scenario allows to compare the predictive outcomes with the true ones instead of comparing them with the observed ones (which are subject to misclassification). This allows to know which model performs better.

The generating process is as follows. A covariate set is generated by $x_{ij} \sim \mathrm{N}(1,1)$, for $i = 1,\ldots,300$ and $j = 1,\ldots,3$ ($n = 300$ and $J = 3$). The vector of regression parameters is $\boldsymbol{\beta} = (-2,2,2)^T$ ($k = 3$) and the vector of cutpoints is $\boldsymbol{\gamma} = (0,5)^T$ ($\gamma_0 = -\infty$ and $\gamma_3 = \infty$). Two link functions are considered and $\mathcal{D}_\Psi$ denotes the

distribution related with the cdf Ψ of the standard normal distribution, $\mathrm{N}(0,1)$, or the standard logistic distribution, $\mathrm{L}(0,1)$. The true ordinal dependent variable ξ is randomly generated by using the following process: (i) generate $\varepsilon_i \sim \mathcal{D}_\Psi$ and define $z_i = \mathbf{x}_i^T\boldsymbol{\beta}+\varepsilon_i$, (ii) define $\xi_i = r$ if $\gamma_{r-1} < z_i < \gamma_r$. Now, about 30% of the outcomes are randomly misclassified according to the following process: (iii) generate $u_i \sim \mathrm{U}(0,1)$, (iv) define $y_i = s$ if $\xi_i = r$ and $\sum_{j=1}^{s-1}\lambda_{rj} < u_i < \sum_{j=1}^{s}\lambda_{rj}$, where λ_{rs} are the elements of the matrix of misclassification probabilities

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.75 & 0.20 & 0.05 \\ 0.20 & 0.60 & 0.20 \\ 0.05 & 0.20 & 0.75 \end{pmatrix}.$$

Note that adjacent categories are more likely to be misclassified.

Figure 6.1 shows a randomly chosen data set. In both graphics, the black lines represent the true probabilities. The first graphic shows the three shaded areas of the stacked bar chart that represent the empirical probabilities. It is evident that there exists misclassification because the empirical probabilities and the true probabilities are different. The second graphic shows the misclassified data. In this graphic, there are data whose highest probability are, for example, the category 1 (drawn as empty light gray dots), but these are classified as category 2 (drawn as filled dark gray dots) or category 3 (drawn as filled black dots).

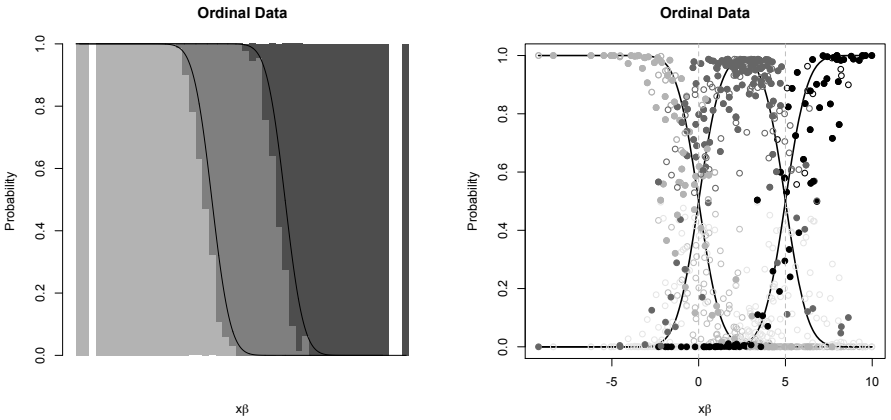

Figure 6.1: Dataset with ordinal misclassified data. The black lines represent the theoretical probabilities. The first graphic shows the three shaded areas of the stacked bar chart that represent the empirical probabilities. The second graphic shows the misclassified data: $\circ$ (empty dots) denote the probabilities p_{ir}, and $\bullet$ (filled dots) denote the category of each observation y_i.

Informative prior distributions have been used. For the regression parameters, a multivariate normal distribution $\mathrm{N}_3(\mathbf{b}_0, \mathbf{B}_0)$ is used, with hyperparameters $\mathbf{b_0} = (-2,2,2)^T$ and $\mathbf{B}_0 = \mathrm{diag}(1,1,1)$. For the cutpoints, the improper prior distributions $\pi(\boldsymbol{\gamma}) \propto 1$ have been used. For the misclassification parameters, initial information is

introduced according to the misclassification proportions that have been considered for the data, that is, $\boldsymbol{\lambda}_r \sim \text{Dirichlet}(\boldsymbol{a}_r)$ where $\boldsymbol{a}_r$ is the rth row of the matrix of misclassification probabilities multiplied by 10. Therefore $\boldsymbol{\lambda}_1 \sim \text{Dirichlet}(7.5, 2, 0.5)$, $\boldsymbol{\lambda}_2 \sim \text{Dirichlet}(2, 6, 2)$ and $\boldsymbol{\lambda}_3 \sim \text{Dirichlet}(0.5, 2, 7.5)$.

The algorithms have been implemented in `R` software. The standard ordinal models that have been used are the probit and logit models defined by Albert and Chib (1993) and Johnson and Albert (1999), respectively. A total of 110,000 iterations of MCMC have been generated for each model. Then, 10,000 iterations were taken as burn-in and one out of 100 values have been saved (thinning equal to 100). With these specifications the chains seem to have converged.

In order to avoid that the results depend on a single simulation, the experiment has been replicated 200 times. The same covariate set, parameters and specifications are used, but data are generated and randomly misclassified at each time (steps (i)-(iv)).

Table 6.1 shows the posterior estimations for the regression parameters. The estimations from the models addressing misclassification are better than the ones from the models that do not consider misclassification. Although their standard deviations are larger, the means are much closer to the original values of the parameters. Therefore, the estimations from the proposed models are less biased than those from models do not consider misclassification. Table 6.2 shows the posterior estimations for the misclassification parameters. The correct information provided for the misclassification parameters allows the proposed model to obtain good estimations. Table 6.3 shows the means and the standard deviations of the considered criteria that have been calculated with the 200 generated datasets. The three criterion values for the models considering misclassification are smaller than the ones for the standard models. Results show that the proposed models considering misclassification are better than the models that do not consider it.

Table 6.1: Estimated means (standard deviations) for the regression parameters for ordinal datasets with misclassification.

Dataset	Parameter	Probit	Logit	Probit Mis	Logit Mis
Probit Mis	$\beta_1 = -2$	−0.440 (0.076)	−0.793 (0.129)	−2.121 (0.328)	−2.207 (0.248)
	$\beta_2 = 2$	0.497 (0.074)	0.874 (0.129)	2.165 (0.358)	2.251 (0.281)
	$\beta_3 = 2$	0.477 (0.069)	0.843 (0.119)	2.058 (0.317)	2.147 (0.239)
	$\gamma_2 = 5$	1.364 (0.108)	2.367 (0.201)	5.477 (0.884)	5.844 (0.710)
Logit Mis	$\beta_1 = -2$	−0.391 (0.069)	−0.710 (0.122)	−2.495 (0.234)	−2.500 (0.233)
	$\beta_2 = 2$	0.423 (0.070)	0.744 (0.124)	2.433 (0.237)	2.438 (0.239)
	$\beta_3 = 2$	0.410 (0.066)	0.725 (0.116)	2.369 (0.209)	2.367 (0.198)
	$\gamma_2 = 5$	1.128 (0.092)	1.942 (0.172)	5.987 (0.603)	6.181 (0.629)

6.6 The analysis of the Signal-Tandmobiel® data

The proposed methodology uses the initial information provided by the validation dataset and relates it to the data from the study to arrive at a posterior predictive distribution that is used to estimate probabilities of the levels of CE lesion severity.

Table 6.2: Estimated means (standard deviations) for the misclassification parameters for ordinal datasets with misclassification.

Parameter	Dataset Probit Mis		Dataset Logit Mis	
	Probit Mis	Logit Mis	Probit Mis	Logit Mis
$\lambda_{11} = 0.75$	0.756 (0.051)	0.782 (0.050)	0.755 (0.048)	0.775 (0.050)
$\lambda_{12} = 0.20$	0.196 (0.048)	0.174 (0.047)	0.130 (0.031)	0.119 (0.032)
$\lambda_{13} = 0.05$	0.049 (0.027)	0.044 (0.027)	0.115 (0.037)	0.106 (0.041)
$\lambda_{21} = 0.20$	0.192 (0.035)	0.169 (0.037)	0.203 (0.034)	0.188 (0.038)
$\lambda_{22} = 0.60$	0.608 (0.041)	0.646 (0.041)	0.588 (0.046)	0.615 (0.050)
$\lambda_{23} = 0.20$	0.200 (0.037)	0.185 (0.039)	0.208 (0.036)	0.198 (0.041)
$\lambda_{31} = 0.05$	0.044 (0.027)	0.039 (0.025)	0.092 (0.038)	0.080 (0.039)
$\lambda_{32} = 0.20$	0.193 (0.048)	0.169 (0.045)	0.147 (0.040)	0.135 (0.038)
$\lambda_{33} = 0.75$	0.763 (0.054)	0.791 (0.051)	0.760 (0.054)	0.784 (0.054)

Table 6.3: Estimated criterion means (standard deviations) for ordinal datasets with misclassification.

Dataset	Model	DIC	TVD	S
Probit Mis	Probit	548.464 (21.383)	199.121 (13.161)	0.374 (0.025)
	Logit	545.171 (21.325)	190.720 (13.037)	0.355 (0.025)
	Probit Mis	537.549 (20.335)	177.286 (11.653)	0.309 (0.026)
	Logit Mis	539.681 (20.515)	180.207 (11.184)	0.319 (0.024)
Logit Mis	Probit	578.341 (18.554)	182.919 (11.616)	0.450 (0.025)
	Logit	574.224 (19.141)	176.314 (11.929)	0.434 (0.026)
	Probit Mis	550.442 (19.907)	163.247 (10.017)	0.383 (0.026)
	Logit Mis	556.740 (19.760)	164.004 (9.820)	0.390 (0.024)

The interest of the present analysis is to evaluate the misclassification probabilities of the levels of CE lesion severity, and to address the influence of oral hygiene and geographical information on the levels of CE lesion severity.

The ordinal outcome y is the level of CE lesion severity. The covariates considered in the model were the following: gender, age, frequency of brushing, plaque index proscimal surfaces, plaque index occlusal surfaces, and geographical location (represented by the standardized (x, y) coordinate of the municipality of the school to which the child belongs).

In order to compare the proposed methodologies, three different models for both logit and probit link functions have been considered for the main dataset. The first models are the ordinal logistic and probit regression ones (Logistic-Standard and Probit-Standard), i.e. the standard models without considering misclassification. In the second models (Logistic-Validation and Probit-Validation), the validation dataset has been used to estimate the misclassification probabilities $\boldsymbol{\lambda}$, and afterwards, the regression parameters $\boldsymbol{\beta}$ and the cutpoints $\boldsymbol{\gamma}$ of the ordinal regression models have been estimated for the main dataset. Note that the main dataset has not been used to

estimate the misclassification probabilities. The regression parameters and cutpoints are estimated by using some full conditional distributions of the algorithm proposed in Section 6.4.2. Specifically, the algorithms consist of choosing initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $\mathbf{c}^{(0)}$, and iteratively sampling $\boldsymbol{z}^{(j)}$, $\boldsymbol{\beta}^{(j)}$, $\boldsymbol{\gamma}^{(j)}$ and $\mathbf{c}^{(j)}$ from the following full conditional distributions. For the logit link they are sampled from (6.6), (6.7), (6.5) and (6.1), respectively, where $\boldsymbol{\beta}^{(j)}$ are sampled by using Metropolis-Hastings updates, and for the probit link they are sampled from (6.3), (6.4), (6.5) and (6.1), respectively. Finally, the third models (Logistic-Misclassification and Probit-Misclassification) are the algorithms proposed in Sections 6.4.2 and 6.4.2. In these cases, the validation dataset has been used to elicite the prior distribution for the misclassification parameters $\boldsymbol{\lambda}$, and then the algorithms are applied to the main dataset.

Note that the validation dataset has been used in two different ways. In the second models, the validation dataset is used to compute the misclassification probabilities. This is the common way when there exists a validation dataset, because the scores from the examiners and from the benchmark are available. Therefore, the misclassification probabilities can be estimated. In the third models, the validation dataset has been used to construct a prior distribution, by eliciting the initial information. If a validation dataset is available, there is no need to build a prior distribution. However, frequently a validation dataset is not available in the real life. So this way is useful to exemplify that if a prior distribution is correctly elicited from the initial information, then the conclusions are better.

The way how the validation dataset has been used is as follows. Let y^{exa} and y^{ben} be the scores of the examiners and the benchmark in the validation dataset, respectively. The hierarchical model $y^{exa}|y^{ben} = r \sim \text{Multinomial}(1, \boldsymbol{\lambda}_r)$, $\boldsymbol{\lambda}_r \sim \text{Dirichlet}(\boldsymbol{a}_r)$, and $a_{rs} \sim \text{Gamma}(0.01, 0.01)$, for $r, s = 1, \ldots, 4$, allows to estimate the posterior distributions of the misclassification probabilities in the validation dataset. Then, the posterior estimations are given by (mean $\pm$ standard deviation)

$$\widehat{\boldsymbol{\lambda}} = \begin{pmatrix} 0.880 \pm 0.019 & 0.120 \pm 0.019 & 0 & 0 \\ 0.199 \pm 0.039 & 0.701 \pm 0.046 & 0.100 \pm 0.030 & 0 \\ 0.061 \pm 0.039 & 0.252 \pm 0.072 & 0.687 \pm 0.077 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.8}$$

and the marginal posterior distributions are given by

$$\begin{aligned} \boldsymbol{\lambda}_1 &\sim \text{Dirichlet}(22.312,\ 3.164,\ 0,\ 0), \\ \boldsymbol{\lambda}_2 &\sim \text{Dirichlet}(7.082,\ 24.011,\ 3.551,\ 0), \\ \boldsymbol{\lambda}_3 &\sim \text{Dirichlet}(1.892,\ 7.254,\ 19.278,\ 0), \\ \boldsymbol{\lambda}_4 &\sim \text{Dirichlet}(0,\ 0,\ 0,\ 1). \end{aligned} \tag{6.9}$$

From the estimations of the misclassification probabilities is evident that adjacent categories are correlated, in the meaning that the probability of to be misclassified in an adjacent category is higher than the one of to be misclassified in a not adjacent one. This misclassification probabilities obtained from the validation dataset can be used to correct for misclassification. For the Logistic-Validation and Probit-Validation models, the estimated misclassification probabilities $\widehat{\boldsymbol{\lambda}}$ are given in (6.8).

For the Logistic-Misclassification and Probit-Misclassification models, the distributions of (6.9) are used as the prior distributions of $\boldsymbol{\lambda}$.

For the three models, noninformative prior distributions have been used for the regression parameters, specifically, a multivariate normal distribution $\mathrm{N}_7(\mathbf{0}_7, \mathbf{I}_7 \times 100)$. For the cutpoints, the improper prior distributions $\pi(\boldsymbol{\gamma}) \propto 1$ has been used.

The estimated parameters obtained with the logistic and probit models are summarized in Table 6.4. These show the posterior means, standard deviations (SD), and the 95% credible intervals. The standard deviations estimated with the models considering misclassification are larger than those obtained with the standard models due to the inclusion of other parameters. Therefore, by using the models considering misclassification the 95% credible intervals are wider. Moreover, the estimations from the proposed model considering misclassification (Logistic-Misclassification and Probit-Misclassification) are closer to the estimations by using the misclassification probabilities from the validation dataset (Logistic-Validation and Probit-Validation) than those from the standard models (Logistic-Standard and Probit-Standard). Models considering misclassification are less unbiased than models that do not consider it.

Positive regression coefficients reflect higher probabilities of CE lesion severity compared to the reference level for categorical covariates. For the variable gender, the category of boys was taken as the reference. The girls have higher probability of having CE than the boys. The reason is that the permanent teeth emerge earlier with girls than with boys, and hence teeth of girls are longer at risk at the same age as those of boys. The probability of CE lesion severity increases as the age of children increases, which is a biologically expected result due to the fact that CE is a progressive illness. The regression coefficient of brushing frequency is negative, indicating that the brushing frequency is a protection factor against CE. The regression coefficients of plaque index on both proscimal and occlusal surfaces are positive, indicating that high values of the corresponding covariates are associated with high probabilities of having high levels of CE lesion. Moreover, there was a significant effect of the x-coordinate, but not of the y-coordinate of the school geographical location. See the obtained results by Vanobbergen et al. (2000) and Mwalili et al. (2005).

The estimated posterior parameters of misclassification probabilities obtained with the proposed models are the following. The posterior estimations by using the logit link, $\widehat{\boldsymbol{\lambda}}_{logit}$, are closer to the ones obtained from the validation dataset in (6.8), than the estimations obtained by using the probit link, $\widehat{\boldsymbol{\lambda}}_{probit}$.

$$\widehat{\boldsymbol{\lambda}}_{logit} = \begin{pmatrix} 0.880 \pm 0.026 & 0.119 \pm 0.026 & 0 & 0 \\ 0.213 \pm 0.072 & 0.680 \pm 0.077 & 0.105 \pm 0.051 & 0 \\ 0.080 \pm 0.048 & 0.276 \pm 0.049 & 0.643 \pm 0.062 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\widehat{\boldsymbol{\lambda}}_{probit} = \begin{pmatrix} 0.918 \pm 0.027 & 0.081 \pm 0.027 & 0 & 0 \\ 0.227 \pm 0.070 & 0.655 \pm 0.074 & 0.117 \pm 0.054 & 0 \\ 0.061 \pm 0.040 & 0.224 \pm 0.054 & 0.713 \pm 0.061 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Table 6.4: Summary of the posterior estimates for the parameters of the ST data.

Parameter	**Mean±SD**	**95% credible interval**	**Mean±SD**	**95% credible interval**
	Logistic-Standard		**Probit-Standard**	
Covariates				
Gender (girl)	0.319±0.075	(0.171,0.463)	0.197±0.045	(0.108,0.285)
Age	0.073±0.035	(0.004,0.140)	0.032±0.020	(-0.008,0.073)
Brushing	-0.156±0.026	(-0.209,-0.105)	-0.096±0.016	(-0.128,-0.065)
Proscimal	0.455±0.078	(0.305,0.605)	0.262±0.045	(0.173,0.354)
Occlusal	0.573±0.195	(0.202,0.959)	0.353±0.114	(0.129,0.574)
x-coordinate	0.003±0.001	(0.001,0.004)	0.002±0.001	(0.001,0.002)
y-coordinate	-0.002±0.002	(-0.006,0.001)	-0.001±0.001	(-0.003,0.001)
Cutpoints				
γ_1	0.797±0.044	(0.718,0.901)	0.378±0.034	(0.292,0.430)
γ_2	1.948±0.047	(1.866,2.050)	1.052±0.031	(0.990,1.104)
γ_3	4.828±0.168	(4.529,5.176)	2.422±0.072	(2.279,2.559)
	Logistic-Validation		**Probit-Validation**	
Covariates				
Gender (girl)	0.431±0.096	(0.247,0.619)	0.254±0.058	(0.134,0.370)
Age	0.028±0.046	(-0.057,0.123)	0.001±0.026	(-0.050,0.054)
Brushing	-0.202±0.033	(-0.268,-0.134)	-0.120±0.019	(-0.156,-0.081)
Proscimal	0.537±0.100	(0.339,0.730)	0.303±0.059	(0.188,0.417)
Occlusal	0.760±0.233	(0.301,1.210)	0.440±0.130	(0.177,0.695)
x-coordinate	0.004±0.001	(0.002,0.006)	0.002±0.001	(0.001,0.003)
y-coordinate	-0.003±0.002	(-0.008,0.001)	-0.001±0.001	(-0.004,0.001)
Cutpoints				
γ_1	0.313±0.061	(0.160,0.441)	0.081±0.039	(0.030,0.173)
γ_2	0.961±0.062	(0.821,1.065)	0.456±0.017	(0.418,0.487)
γ_3	4.237±0.163	(3.927,4.568)	2.069±0.077	(1.918,2.222)
	Logistic-Misclassification		**Probit-Misclassification**	
Covariates				
Gender (girl)	0.438±0.107	(0.236,0.657)	0.240±0.055	(0.136,0.350)
Age	0.011±0.046	(-0.080,0.102)	0.004±0.025	(-0.045,0.054)
Brushing	-0.209±0.036	(-0.278,-0.139)	-0.116±0.019	(-0.155,-0.079)
Proscimal	0.536±0.109	(0.334,0.745)	0.301±0.056	(0.192,0.414)
Occlusal	0.769±0.252	(0.282,1.263)	0.412±0.129	(0.170,0.666)
x-coordinate	0.004±0.001	(0.002,0.006)	0.002±0.001	(0.001,0.003)
y-coordinate	-0.003±0.002	(-0.008,0.001)	-0.001±0.001	(-0.004,0.001)
Cutpoints				
γ_1	0.073±0.062	(-0.0.079,0.172)	-0.015±0.038	(-0.082,0.053)
γ_2	0.620±0.093	(0.484,0.848)	0.564±0.072	(0.420,0.680)
γ_3	4.039±0.220	(3.626,4.448)	2.084±0.100	(1.903,2.298)

6.7 Conclusion

A Bayesian approach to polychotomous response data that are subject to misclassification has been proposed and discussed in this paper. The idea of using a data augmentation framework has been exploited to derive efficient MCMC algorithms. This model has been explored for ordered categories in the response variable by using both probit and logit link functions.

The applicability of the proposed approach has been illustrated through a simulated example that shows their good performance when compared with models that do not consider misclassification. A longitudinal oral health study conducted in Flanders (North of Belgium), the Signal-Tandmobiel® study, has been analyzed. The main advantage of the proposed model is provided better estimations than the standard ones. Through the simulated example we have shown that, when data are misclassified, the estimates from models that do not consider misclassification are biased, and that the estimates from models considering misclassification are closer to the real ones. Moreover, by using latent variables and considering prior information it is possible to update misclassification probabilities. Therefore, when ordinal data are subjected to misclassification, it is highly recommended to consider model that take into account this fact.

Although the approach has been explored to ordered categories, it also can be extended for unordered categories as follows. The data augmentation scheme provided in Section 6.3 is firstly considered. The latent variables $\mathbf{c}$ and the misclassification probabilities $\boldsymbol{\lambda}$ are introduced in the nominal response data model. The Gibbs sampling described in Section 6.4.2 is used to sample $\mathbf{c}$ and $\boldsymbol{\lambda}$ from the full conditional posterior distributions (6.1) and (6.2), respectively. Then, the latent vector $\boldsymbol{\xi}$ is obtained from $\mathbf{c}$, which correspond to the true classifications, where $\xi_i = r$ if $c^i_{r+} = 1$. Finally, the outcomes $\mathbf{Y}$ are replaced by the latent vector $\boldsymbol{\xi}$ to obtain the probabilities $\boldsymbol{p}$. The regression parameters are estimated by using other algorithms (see, for example, McCulloch et al. (2000)). Moreover, the approach can also be extended to other link functions.

The potential applicability of this approach to many fields of knowledge makes this proposal interesting. A very interesting research topic related to this work is to generalize the models of misclassification to model multivariate ordinal response data, i.e., when the outcomes consist of several ordinal variables that are correlated.

Acknowledgements

The Signal-Tandmobiel® study comprises the following partners: D. Declerck (Dental School, Katholieke Universiteit Leuven), L. Martens (Dental School, University of Ghent), J. Vanobbergen (Dental School, University of Ghent), P. Bottenberg (Dental School, University of Brussels), E. Lesaffre (L-BioStat, Katholieke Universiteit Leuven), and K. Hoppenbrouwers (Youth Health Department, Katholieke Universiteit Leuven, and Flemish Association for Youth Health Care).

The first three authors have been partially supported by *Ministerio de Educación y Ciencia*, Spain (Project MTM2011-28983-C03-02) and *Gobierno de Extremadura*, Spain (Project GRU10110).

Chapter 7

A Bayesian multilevel model for misclassified ordinal response data: an application to caries experience lesion severity

Naranjo, L., Mutsvari, T., Pérez, C. J., Martín, J., and Lesaffre, E. (2014). A Bayesian multilevel model for misclassified ordinal response data: an application to caries experience lesion severity. Preprint 154, Universidad de Extremadura, Badajoz, Spain.

Abstract

Motivated by a longitudinal oral health study, the Signal-Tandmobiel®, a Bayesian approach has been addressed to correct for misclassification in a logistic multilevel model where ordinal response is subject to error. The caries experience lesion severity represents the ordinal variable of interest that is prone to misclassification. Caries experience data have a hierarchical structure since the data are recorded for the teeth nested within mouth. The proposed approach allows a bias correction when misclassification probabilities can be estimated from a validation dataset. Another way to address the correction is by including historical or expert information in the model through the elicitation of prior distributions. The proposed approach provides right estimations leading to realistic predictions. This approach can be used in many contexts different from the caries experience one.

Keywords: Bayesian analysis; Caries experience; Hierarchical data; Misclassification; Multilevel model; Ordinal response data.

7.1 Introduction

Dental caries is one of the most prevalent chronic diseases worldwide affecting persons in all age groups. Many epidemiological surveys and clinical studies are carried out to obtain a further understanding of this disease entity. However, the process of detecting caries experience (CE) is not an obvious issue. CE scoring may not perfectly reflect true condition of the tooth, and therefore, the presence of CE can be misdiagnosed, leading to misclassified outcomes. In order to standardize data collection techniques in epidemiological surveys and clinical trials, CE assessment guidelines have been developed by the World Health Organization (WHO (1997)), the British Association for the Study of Community Dentistry (BASCD, see Pine et al. (1997)), and the International Caries Detection and Assessment System (ICDAS (2005)). These guidelines highlight the need for training the examiners and measuring the reliability of the obtained scores. However, despite these criteria, the process of CE detection is subject to incorrect classification. In addition, CE data have a natural hierarchical structure since data are recorded for the teeth nested within mouth. Teeth from the same mouth tend to be more alike in their physical and hygienic characteristics than teeth chosen at random from the population at large.

Many kinds of data, including observational data collected in the human and biological sciences, have a hierarchical (nested or clustered) structure. When lower level units are nested within one or more higher level strata, conventional single level regression analysis is not appropriate since observations are no longer independent. Such dependency means standard errors are biased if the nesting is ignored, and incorrect inferences concerning prediction effects may be achieved. Multilevel models are methodologies for dealing appropriately with nested or clustered data (see, e.g., Gelman and Hill (2007), Congdon (2010), Hox (2010) and Goldstein (2011)). Multilevel analysis allows characteristics of different groups to be included in models of individual behavior. At each level of hierarchy, a different set of variables may be defined. Moreover, the inclusion of random effects changes the scope of inference, allows the variability analysis in parameter sets, accounts for internal structure in data, achieves a more honest accounting for the uncertainties in a modeled system, and may result in improved estimates of each parameter by borrowing strength from the ensemble. Recent developments in Markov chain Monte Carlo (MCMC) methods allow fully Bayesian analyses of sophisticated multilevel models for complex referenced data (Congdon (2010)).

Many measurements are made with substantial error components, especially in the social and biological sciences. Thus, if the measurement has to be repeated it is not expected to get always identical result. Measurement errors for categorical variables are called misclassifications. In situations where misclassification may happen, it is necessary to correct the bias yielded by the use of non free-error data. If misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non-optimal decision-making. Therefore, a correction for misclassification is needed to obtain unbiased estimates for the involved parameters. Statistical models addressing misclassification should be available in these contexts. This is specially well-known in generalized linear models (see, e.g., Gustafson (2003), Carroll et al. (2006) and

Buonaccorsi (2010)). However, the effect of measurement error and misclassification in multilevel models has not been enough explored. The behavior of bias associated with measurement error in covariates or in the response, for such hierarchically clusters data, is not well-known and can be complex (see Woodhouse et al. (1996), Goldstein et al. (2008) and Ferrão and Goldstein (2014)).

Motivated by a longitudinal oral health study, the Signal-Tandmobiel® study (ST), Bayesian approaches to address misclassified ordinal multilevel response data are proposed and discussed in this paper. Naranjo et al. (2014b) and Mutsvari et al. (2013) proposed two relevant approaches. Naranjo et al. (2014b) developed a Bayesian approach to model misclassified ordinal response data, where no hierarchy levels were taken into account. They used data augmentation frameworks and Markov chain Monte Carlo (MCMC) methods in probit and logit models. Mutsvari (2012) studied the misclassification process in detecting the presence or absence of CE by considering the hierarchical structure of data. Related to the same data, Mutsvari et al. (2010) investigated the factors affecting misclassification errors, and Mutsvari et al. (2013) explored and suggested possible ways of correcting for misclassification using validation datasets in a binary multilevel scheme. The purpose of this paper is to extend the methods proposed by Naranjo et al. (2014b) and Mutsvari et al. (2013) to model misclassified ordinal response data having a hierarchical structure.

The outline of this paper is as follows. The ST study is introduced in Section 7.2, illustrating the need of correction for misclassification. The way misclassification is addressed in multilevel ordinal models is presented in Section 7.3. In Section 7.4, the analysis of the ST data is presented. Finally, Section 7.5 presents the conclusion.

7.2 The Signal-Tandmobiel® study

The Signal-Tandmobiel® study is a longitudinal prospective oral health intervention project, conducted in Flanders (North of Belgium) between 1996 and 2001. For this project, 4468 children (2315 boys and 2153 girls) were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen trained dentists (examiners) based on visual and tactile observations. The clinical examinations took place in a mobile dental clinic, with a standard chair and artificial dental light. No radiographs were taken. Data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents. For a more detailed description of the study design and research methods see Vanobbergen et al. (2000).

In this work, caries lesions were scored in six severity levels. Caries experience (CE) can be represented as an ordinal variable indicating whether the tooth is at the level defining a progressive disease. The scoring of the four permanent first molars, i.e., teeth 16 and 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants) were considered. The numbering of the teeth follows the FDI (Federation Dentaire Internationale) notation, which indicates the position of the tooth in the mouth. The diagnosis of CE may be difficult for several reasons. For instance, composite materials can imitate the natural enamel so well that it is difficult to spot a restored lesion; or the location of the cavity, far back in the mouth,

hampers the view of the dental examiner. Hence, overlooking CE is likely to happen in practice, but the dental examiner could also classify discolorations as CE.

In the ST study, the 16 dental examiners were calibrated for scoring CE. The calibration exercises were performed according to the guidelines of training and calibration published by the British Association for the Study of Community Dentistry (Pitts et al. (1997)). The calibration of the dental examiners was performed by comparing their scores on the tooth surfaces of a group of children to those of a benchmark examiner. Note that there exists no infallible scorer for CE, but a very experienced dental examiner (called benchmark) is considered in this context. In order to maintain a high level of intra- and inter-examiner reliability, calibration exercises were carried out twice a year for all involved examiners. During the study period (1996-2001), three calibration exercises were devoted to the scoring of CE (1996, 1998, 2000), involving 92, 32 and 224 children, respectively. A contingency table of dental examiners and the benchmark examiner was determined, yielding a table with misclassified scores. Data from the three calibration exercises were combined into one validation dataset, and also data from the examiners were combined into one. The results suggested that there were some examiners overscoring or underscoring the true CE status.

In the main dataset the dental examiners scored the children, but their scores are likely to be prone to error. In the statistical analysis, ignoring that the levels of lesion severity of CE are prone to misclassification generally leads to wrong estimates, and so, to wrong conclusions. Bayesian ordinal multilevel models considering misclassification can help to provide better predictions than standard models. The proposed methodology uses the validation dataset in two possible ways: i) to calculate misclassification probabilities, and ii) to elicitate prior distributions of the misclassification probabilities, and relates it to the data from the study to arrive at a posterior predictive distribution that is used to estimate the levels of lesion severity of CE probabilities.

7.3 The approach

In this section the multilevel model for ordinal response is described for notation purposes. Later, the multilevel model for ordinal responses subject to misclassification is proposed as a generalization.

7.3.1 The ordinal logistic multilevel model

In order to set the notation, let m denote the level-mouth units (clusters, level 2), let t denote the level-tooth units (nested observations, level 1), and let e denote the level-examiner units. Assume that there are $m = 1, 2, \ldots, M$ subjects and $t = 1, 2, \ldots, T_m$ teeth nested within each subject. The total number of teeth observations across subjects is given by $n = \sum_{m=1}^{M} T_m$. Let Y_{mte} be the ordinal outcome variable denoting the level of CE lesion severity on tooth t in subject m examined by e. Let the J ordered response categories be coded as $r = 1, 2, \ldots, J$. Note that Y_{mte} is a random variable with a categorical distribution whose parameter vector is $\boldsymbol{p}_{mte} = (p_{mte,1}, \ldots, p_{mte,J})$, denoted by $Y_{mte} \sim \text{Cat}(p_{mte,1}, \ldots, p_{mte,J})$, where

$p_{mte,r} = \mathrm{p}(y_{mte} = r)$ is the probability that the variable Y_{mte} takes the rth category. Ordinal response models often utilize cumulative comparisons of the ordinal outcome. The cumulative probabilities for the J categories of the ordinal outcome Y_{mte} are defined as $\pi_{mte,r} = \mathrm{p}(y_{mte} \leq r) = \sum_{j=1}^{r} p_{mte,j}$. The model uses $\pi_{mte,r}$, which is the true conditional probability for the rth level of CE lesion severity on tooth t in mouth m by examiner e.

The logistic multilevel model for ordered categories is given in terms of the cumulative log odds:

$$\mathrm{logit}(\pi_{mte,r}) \quad = \quad \log\left[\frac{\pi_{mte,r}}{1-\pi_{mte,r}}\right] = \kappa_r - \boldsymbol{x}_{mte}^T\boldsymbol{\beta} + u_m + u_e, \tag{7.1}$$

for $r = 1, \ldots, J-1$, with $J-1$ strictly increasing model thresholds κ_r, i.e. $\kappa_1 < \kappa_2 < \cdots < \kappa_{J-1}$, also known as cutpoints. It is common to set a threshold to zero because it is often less complex for numerical stability in sampling, typically, this is done in terms of the first threshold, $\kappa_1 = 0$. Let $\boldsymbol{x}_{mte}$ denote the vector of covariates (design vector) of the tooth t of subject m examined by e and $\boldsymbol{\beta}$ is the vector of associate regression coefficients.

The quantities u_m and u_e are random intercepts at mouth and examiner levels, respectively. These random effects are independently distributed with mean zero and variances σ_m^2 and σ_e^2, respectively, that is, $(u_m, u_e)^T \sim \mathrm{N}_2\big(\mathbf{0}, \mathrm{diag}(\sigma_m^2, \sigma_e^2)\big)$, where $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the k-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. As the score observations are marked by a panel of examiners, it may be more appropriate to consider the random intercept at examiner level, with u_e representing the response bias for the examiner. If the examiners are considered a random sample of possible examiners, u_e can be treated as random, giving a two-way random effects model. Note that the random effects for subjects and examiners are not nested, but they are crossed if each person is assessed by each examiner.

7.3.2 Addressing misclassification

Suppose that Y_{mte} are prone to error. Let $\theta_{mte,s} = \mathrm{p}(y_{mte} = s|\boldsymbol{x}_{mte})$ denote the probability that the observation of the tooth t of subject m examined by e is classified in the sth category. The way parameters $\theta_{mte,s}$ are related to a set of covariates $\boldsymbol{x}_{mte}$ is defined as

$$\theta_{mte,s} = \sum_{r=1}^{J} \mathrm{p}(y_{mte} = s|\xi_{mte} = r)\mathrm{p}(\xi_{mte} = r|\boldsymbol{x}_{mte}) = \sum_{r=1}^{J} \lambda_{rs} p_{mte,r},$$

where $\lambda_{rs} = \mathrm{p}(y_{mte} = s|\xi_{mte} = r)$ is the probability that an observation from Y_{mte} is classified in the sth category when the true category ξ_{mte} is the rth one, and $p_{mte,r}$ denotes the probability that the true category for an observation with covariate pattern $\boldsymbol{x}_{mte}$ is the rth one. Note that $\boldsymbol{\xi}$ is an unknown random vector representing

the true classification, and $\boldsymbol{\lambda}$ is the matrix of misclassification probabilities

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_J \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1J} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{J1} & \lambda_{J2} & \ldots & \lambda_{JJ} \end{pmatrix},$$

where $\sum_{s=1}^{J} \lambda_{rs} = 1$, and the elements of the diagonal, λ_{rr}, for $r = 1, \ldots, J$, denote the probabilities of correct classification.

The logistic multilevel model for ordered categories addressing misclassification is

$$\begin{aligned} Y_{mte} &\sim \text{Cat}(\theta_{mte,1}, \ldots, \theta_{mte,J}), \qquad (7.2) \\ \theta_{mte,s} &= \sum_{r=1}^{J} \lambda_{rs} p_{mte,r}, \\ \pi_{mte,r} &= \sum_{j=1}^{r} p_{mte,j}, \\ \text{logit}(\pi_{mte,r}) &= \log\left[\frac{\pi_{mte,r}}{1-\pi_{mte,r}}\right] = \kappa_r - \boldsymbol{x}_{mte}^T \boldsymbol{\beta} + u_m + u_e. \end{aligned}$$

7.4 Application to Signal-Tandmobiel® data

The proposed methodology uses the initial information provided by the validation dataset and relates it to the data from the study to arrive at a posterior predictive distribution that is used to estimate probabilities of the levels of CE lesion severity. The interest of the present analysis is to use the validation dataset in two different ways: i) to directly calculate the misclassification probabilities of the levels of CE lesion severity and, ii) to elicitate the prior distribution of the misclassification probabilities. The proposed approach is applied to the main dataset to address the influence of oral hygiene and geographical information on the levels of CE lesion severity.

A total of 11232 teeth, corresponding to 2808 children, was involved in the analysis. The covariates included in the model were the following. At level of mouth: gender (girls *vs.* boys), age, frequency of brushing, and geographical location (represented by the standardized (x, y) coordinate of the municipally of the school to which the child belongs). At level of tooth: plaque index proscimal surfaces, plaque index occlusal surfaces, and jaw (mandible *vs.* maxilla).

In order to compare the proposed methodologies, three different models for logit link functions have been considered for the main dataset. The first model is the ordinal logistic multilevel model (Multilevel) presented in (7.1), i.e. the multilevel model without considering misclassification. In the second model (Multilevel-Mis) the validation dataset has been used to estimate the misclassification probabilities $\boldsymbol{\lambda}$, and afterwards, the regression parameters $\boldsymbol{\beta}$ and the thresholds $\boldsymbol{\kappa}$ of the ordinal multilevel model have been estimated for the main dataset by using the model presented in (7.2). The last model (Multilevel-Mis-Prior) is also the ordinal multilevel model addressing

misclassification presented in (7.2), but now the validation dataset has been used to elicit the prior distribution for the misclassification parameters $\boldsymbol{\lambda}$, and then the algorithms are applied to the main dataset.

Note that the validation dataset has been used in two different ways. In the Multilevel-Mis model, the validation dataset is directly used to compute the misclassification probabilities. This is the common way to proceed when there exists a validation dataset, because the scores from the examiners and from the benchmark are available. Therefore, the misclassification probabilities can be estimated. In the last model (Multilevel-Mis-Prior), the validation dataset has been used to build the prior distribution by eliciting the initial information. If a validation dataset is available, there is no need to build a prior distribution. However, this information can be used in a Bayesian framework. This model is extremely useful when there is no validation dataset (as it frequently happens), but historical or expert information is available. This allows to obtain less biased estimates and better predictions.

The way how the validation dataset has been used is as follows. Let Y^{exa} and Y^{ben} be the scores of the examiners and the benchmark in the validation dataset, respectively. The hierarchical model $Y^{exa}|y^{ben} = r \sim \text{Cat}(\boldsymbol{\lambda}_r)$, $\boldsymbol{\lambda}_r \sim \text{Dir}(\boldsymbol{a}_r)$, and $a_{rs} \sim \text{Gamma}(0.01, 0.01)$, for $r, s = 1 \ldots, 6$, allows to estimate the posterior distributions of the misclassification probabilities in the validation dataset, where $\text{Cat}(\boldsymbol{\lambda}_r)$ denotes the categorical distribution with vector of parameters $\boldsymbol{\lambda}_r = (\lambda_{r1}, \ldots, \lambda_{rJ})$, $\text{Dir}(\boldsymbol{a}_r)$ denotes the Dirichlet distribution with vector of hyperparameters $\boldsymbol{a}_r = (a_{r1}, \ldots, a_{rJ})$, and $\text{Gamma}(a, b)$ denotes the Gamma distribution with shape parameter a and rate parameter b.

From the validation dataset the following estimated misclassification probabilities are obtained for model Multilevel-Mis (mean±SD)

$$\widehat{\boldsymbol{\lambda}} = \begin{pmatrix} 0.924 \pm 0.009 & 0.069 \pm 0.009 & 0.007 \pm 0.003 & 0 & 0 & 0 \\ 0.259 \pm 0.040 & 0.656 \pm 0.043 & 0.067 \pm 0.022 & 0.018 \pm 0.012 & 0 & 0 \\ 0.057 \pm 0.037 & 0.079 \pm 0.042 & 0.659 \pm 0.074 & 0.205 \pm 0.063 & 0 & 0 \\ 0 & 0.502 \pm 0.274 & 0 & 0.498 \pm 0.274 & 0 & 0 \\ 0 & 0 & 0.490 \pm 0.184 & 0.179 \pm 0.138 & 0.331 \pm 0.171 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{7.3}$$

and the following Dirichlet distributions are used as prior distributions for the Multilevel-Mis-Prior model

$$\begin{aligned} \boldsymbol{\lambda}_1 &\sim \text{Dir}(46.736, 3.640, 0.647, 0, 0, 0), \\ \boldsymbol{\lambda}_2 &\sim \text{Dir}(11.787, 30.069, 3.111, 1.096, 0, 0), \\ \boldsymbol{\lambda}_3 &\sim \text{Dir}(2.879, 3.617, 30.722, 9.611, 0, 0), \\ \boldsymbol{\lambda}_4 &\sim \text{Dir}(0, 11.349, 0, 11.218, 0, 0), \\ \boldsymbol{\lambda}_5 &\sim \text{Dir}(0, 0, 16.401, 5.445, 10.143, 0), \\ \boldsymbol{\lambda}_6 &\sim \text{Dir}(0, 0, 0, 0, 0, 1). \end{aligned} \tag{7.4}$$

In order to estimate the model parameters, Bayesian approaches are considered (see Bernardo and Smith (1994) and Congdon (2010)). In Bayesian analysis, the

initial knowledge about the parameters (prior distribution) is combined with the observed data (likelihood) to get the posterior distribution. The posterior estimates are obtained by using Markov chain Monte Carlo (MCMC) methods (see Gilks et al. (1996)). JAGS software has been used to implement the MCMC simulations (Ntzoufras (2009)).

For all the models, non informative prior distributions have been used for the regression parameters, specifically, $\beta_k \sim \mathrm{N}(0, 0.01)$. For the thresholds, non informative normal distributions have been used $\kappa_j \sim \mathrm{N}(0, 0.01)$. The random intercepts at mouth and examiner levels are distributed as $u_m \sim \mathrm{N}(0, \sigma_m^2)$, $u_e \sim \mathrm{N}(0, \sigma_e^2)$, $\sigma_m^2 \sim \mathrm{InvGamma}(0.01, 0.01)$, and $\sigma_e^2 \sim \mathrm{InvGamma}(0.01, 0.01)$, where $\mathrm{InvGamma}(a, b)$ denotes the inverse Gamma distribution with shape parameter a and rate parameter b.

The estimated parameters obtained are summarized in Table 7.1 by posterior means, standard deviations (SD) and the 95% credible intervals. The standard deviations estimated with the models considering misclassification are larger than those obtained with the standard model due to the inclusion of additional parameters. Therefore, by using the models considering misclassification the 95% credible intervals are wider. Moreover, the estimates from the proposed models considering misclassification are closer to the estimates by using the misclassification probabilities from the validation dataset than those from the standard model. The variance estimates of the random effects show that there was more variability at mouth level than at examiner level.

Positive regression coefficients reflect higher probabilities of CE lesion severity compared to the reference level for categorical covariates. For the variable gender, the category of boys was taken as the reference. The girls have higher CE probability than the boys. The reason is that the permanent teeth emerge earlier in girls than in boys, and hence teeth of girls have been longer at risk at the same age than those of boys. For the variable jaw, the category of mandible was taken as the reference. The maxilla has a higher probability of having CE lesion severity than the mandible. The probability of CE lesion severity decreases as the age of children increases. The regression coefficient of brushing frequency is negative, indicating that the brushing frequency is a protection factor against CE. The regression coefficients of plaque index on both proscimal and occlusal surfaces are positive, indicating that high values of the corresponding covariates are associated with high probabilities of having high levels of CE lesion severity. Moreover, there were significant effects of the x-coordinate and y-coordinate of the school geographical location.

7.5 Conclusion

A Bayesian approach has been addressed to correct for misclassification in multilevel models where ordinal response is prone error. This is an extension of both, the Bayesian approach proposed by Naranjo et al. (2014b) to address multilevel data, and the analysis proposed by Mutsvari et al. (2013) to allow ordinal data. Based on the validation dataset, both misclassification probabilities and prior distribution of the misclassification probabilities have been estimated. Thus, these have been

Table 7.1: Summary of the posterior estimates for the parameters of the different models.

Parameter	Multilevel		Multilevel-Mis		Multilevel-Mis-Prior	
	Mean±SD	95% Interval	Mean±SD	95% Interval	Mean±SD	95% Interval
Mouth level						
Gender (girl)	0.558±0.126	(0.312,0.810)	1.288±0.290	(0.748,1.870)	0.921±0.221	(0.502,1.371)
Age	-0.124±0.039	(-0.196,-0.045)	-0.226±0.100	(-0.429,-0.056)	-0.128±0.094	(-0.315,0.060)
Brushing	-0.274±0.041	(-0.361,-0.196)	-0.622±0.101	(-0.813,-0.422)	-0.445±0.093	(-0.664,-0.271)
x-coordinate	0.066±0.015	(0.038,0.096)	0.154±0.034	(0.094,0.226)	0.109±0.027	(0.062,0.169)
y-coordinate	-0.063±0.020	(-0.104,-0.027)	-0.122±0.059	(-0.236,-0.005)	-0.099±0.054	(-0.219,0.021)
Tooth level						
Jaw (maxilla)	0.094±0.056	(-0.019,0.205)	0.343±0.116	(0.113,0.575)	0.190±0.097	(0.012,0.386)
Proscimal	0.659±0.128	(0.402,0.904)	1.287±0.289	(0.696,1.850)	1.012±0.231	(0.579,1.479)
Occlusal	0.968±0.302	(0.367,1.556)	2.251±0.661	(0.970,3.539)	1.695±0.534	(0.700,2.811)
Thresholds						
κ_2	2.124±0.052	(2.024,2.228)	4.173±0.275	(3.670,4.758)	3.343±0.325	(2.769,4.045)
κ_3	4.054±0.096	(3.870,4.251)	7.073±0.386	(6.334,7.875)	6.373±0.623	(5.281,7.837)
κ_4	5.193±0.140	(4.924,5.475)	7.368±0.397	(6.634,8.205)	7.066±0.707	(5.865,8.753)
κ_5	5.555±0.159	(5.252,5.871)	8.268±0.385	(7.539,9.072)	7.897±0.675	(6.821,9.570)
Random						
σ^2_m	5.934±0.378	(5.252,6.748)	23.851±2.932	(18.759,30.161)	14.796±3.469	(9.794,23.844)
σ^2_e	0.066±0.057	(0.008,0.212)	0.284±0.271	(0.014,1.022)	0.158±0.151	(0.010,0.580)

related to the multilevel model under these two different assumptions. Under these assumptions estimates are similar.

A very interesting research topic related to this work is to consider multilevel models when explanatory variables are prone to misclassification or measurement error. Moreover, the introduction of level-dependence in misclassification or measurement error in multilevel models can be pursued.

Acknowledgements

The Signal-Tandmobiel® study comprises the following partners: D. Declerck (Dental School, Katholieke Universiteit Leuven), L. Martens (Dental School, University of Ghent), J. Vanobbergen (Dental School, University of Ghent), P. Bottenberg (Dental School, University of Brussels), E. Lesaffre (L-BioStat, Katholieke Universiteit Leuven), and K. Hoppenbrouwers (Youth Health Department, Katholieke Universiteit Leuven, and Flemish Association for Youth Health Care).

The first, third and fourth authors have been partially funded by *Ministerio de Economía y Competitividad*, Spain (Project MTM2011-28983-C03-02), *Gobierno de Extremadura*, Spain (Project GRU10110), and *European Union* (European Regional Development Funds).

Part IV

Conclusion

Chapter 8

Conclusion and further research

This last chapter concludes with a summary of the main contributions presented in the preceding chapters. It also presents a discussion on possible future researches.

8.1 Conclusion

This thesis has focused on two important topics in Statistics. The first one is the development of robust regression models by using flexible link functions and, the second one is the extension of categorical data models to address misclassification. Several approaches have been developed in this thesis. Bayesian methodology has been considered.

Models that use asymmetric exponential power distributions

In Part II, models that use distributions belonging to the symmetric exponential power (EP) and the asymmetric exponential power (AEP) family have been studied in three different Bayesian contexts: posterior distribution exploration, linear regression, and binary regression. Scale mixture of uniform representations of the EP and AEP distributions have been used to allow the development of efficient Gibbs sampling algorithms.

Firstly, the AEP distribution proposed by Zhu and Zinde-Walsh (2009) has been considered for the likelihood model in Chapter 4. The AEP family includes the EP distribution as a particular case and provide flexible distributions with both lighter and/or heavier tails compared to the normal. These distributions are able to manage both symmetry/asymmetry and light/heavy tails simultaneously. The proposed approaches represent alternatives to analyze data that do not verify the normality assumption. These distributions provide flexible fits to many types of experimental or observational data.

Secondly, a linear regression model that uses the AEP distribution for the error variable has been developed in Chapter 4. The flexibility of the AEP distribution allows to consider the skewness of the data avoiding transformations and reducing

the quantity of outliers. This model turns out to be a robust model, allowing to deal with problems for which it is important to properly model errors with symmetry/asymmetry and light/heavy tails.

Finally, binary regression models based on the inverse of the EP and AEP cdf's as the link functions have been proposed and analyzed in Chapters 3 and 4, respectively. The implementations are based on the development of data augmentation frameworks. These frameworks introduce latent variables that follow linear regression structures with EP and AEP distributions. The EP link allows to find the necessary degree of robustness by letting the kurtosis parameter vary over its range. This range can be modified to provide models based on links with platykurtic or leptokurtic cdf's. Besides, according to the description of links given by Chen et al. (1999a), asymmetric links are good alternatives when the rates at which the probabilities of a given binary response approach 1 or 0 are not the same, and the number of 1's and 0's are much different. Therefore, the AEP distribution provides a great flexibility to fit binary regression-based models.

Models with misclassified categorical response data

Categorical response data models addressing misclassification have been studied in Part III. Specifically, binary, polychotomous and multilevel regression models have been proposed. All of them have considered that the outcomes are subject to misclassification.

Firstly, binary regression models when the binary outcome is misclassified have been presented in Chapter 5. The probit model has been obtained as a particular case of the t-link model. Data augmentation schemes have been developed to derive efficient Gibbs sampling and Expectation-Maximization algorithms. The results show that, when data are misclassified, the proposed approaches considering misclassification are better than the standard ones that do not consider it, and can substantially increase the number of correct predictions.

Secondly, an approach to polychotomous response data that are subject to misclassification has been discussed in Chapter 6. The idea of using a data augmentation framework has been exploited to derive efficient Markov chain Monte Carlo (MCMC) algorithms. This model has been mainly explored for ordered categories in the response variable by using both probit and logit link functions, but it can also be applied for unordered categories. Through simulation-based examples it is shown that, when data are misclassified, the estimates from models that do not consider misclassification are biased, and that the estimates from models considering misclassification are closer to the real ones. Therefore, when ordinal data are subject to misclassification, it is highly recommended to consider this type of approach.

Finally, motivated by a longitudinal oral health study, a different type of ordinal model considering misclassification is developed in Chapter 7. This model exploits the multilevel structure of the data. The hierarchical structure of the model is beneficial for a better understanding of the data. Moreover, the inclusion of random effects allows to study the variability in parameter sets.

8.2 Further research

The work presented in this thesis provides foundations for interesting future research avenues and extensions. In this section some possible future research topics are suggested.

Models that use asymmetric exponential power distributions

The EP and AEP distributions that have been used are univariate. An interesting research topic is multivariate extensions. These distributions can be used as the likelihood, the error distributions in linear regression and to provide a link function in a multinomial regression model. The main problem is the intractability of the posterior distributions. However, this can be overtaken by using a data augmentation scheme. A scale mixture of uniform representation of the multivariate EP distribution can be defined. However, this representation does not allow to estimate posterior distributions in an easy way. The development of another data augmentation scheme to model the multivariate EP distribution also could be of interest.

Another interesting research topic could be the use of a mixture of EP or AEP distributions. This may be useful to provide more flexibility than the normal mixture-based approach in order to identify subpopulations and to find a better separation. This could be particularly suited to multimodal problems. Although the computational cost could be higher than by using other standard models, it is expected that its flexibility is able to capture skewness and kurtosis features better than other competing models.

Another interesting use of the AEP-based proposed approach could be the adaptation to the time series model context, where the asymmetrical data and the heaviness of the tails require a different distribution from the normal one. Specifically, many financial time series should be modeled by using more flexible distributions than the normal ones to accommodate for heavier/lighter tails and skewness. This flexibility is very important for the GARCH models, where the EP family has been widely used. The use of AEP distributions in this context is currently a challenge.

Finally, a research issue could be the generalization of the scale mixture of uniform representation for the AEP distribution. For example, a class of spliced-scale distributions of Bayesian semi-parametric scale mixture of Beta models could be considered. This might lead to a possibly more flexible approach.

Models with misclassified categorical response data

A very interesting research topic related to this work is to generalize the misclassification approach to model multivariate ordinal response data, i.e., when the outcomes consist of several ordinal variables that are correlated. Moreover, the development of multivariate models when validation data are available, by using gold standard or benchmark samples, could be studied.

Another interesting research topic is the study of the relation between misclassification and asymmetric link functions in categorical regression models. In this thesis,

symmetric link functions have been considered for binary and ordinal regression models addressing misclassification. However, it could be interesting to study the effect of skewness due to asymmetric link functions on misclassification probabilities.

Covariates prone to misclassification (categorical covariates) or measurement error (continuous covariates) could be considered in the generalized linear model context. Some advances have been achieved, but more research is needed in this topic.

Finally, complex model structures can be taken into account to address misclassification. The extension of models considering misclassification for handling multinomial data or higher order of dependence could be considered. The inclusion of time-dependence association in longitudinal models to address misclassification or measurement error could be a challenge. The introduction of level-dependence in misclassification or measurement error in hierarchical models can also be pursued.

References

Achcar, J. A., Martínez, E. Z., and Louzada-Neto, F. (2004). Binary data in the presence of misclassifications. In Antoch, J., editor, *Proceedings of the Computational Statistics Conference*, pages 581–588, Prague, Czech Republic. Physica-Verlag/Springer.

ACOM (American College of Occupational Medicine) Noise and Hearing Conservation Committee (1989). Occupational noise-induced hearing loss. *Journal of Occupational Medicine*, 31(12):996–1001.

Adcock, C. J. (2010). Asset pricing and portfolio selection based on the multivariate extended skew-Student-t distribution. *Annals of Operations Research*, 176:221–234.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, New Jersey, second edition.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Proceedings of 2nd International Symposium on Information Theory*, pages 267–281, Budapest, Hungary. Académiai Kiado.

Albert, J. (2009). *Bayesian Computation with R*. Springer, New York, second edition.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82(4):747–759.

Albert, P. S., Hunsberger, S. A., and Biro, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association*, 92(440):1304–1311.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, 68(2):357–363.

Arellano-Valle, R. B. and Bolfarine, H. (1995). On some characterisations of the t-distribution. *Statistic and Probability Letters*, 25:79–85.

Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2007). Bayesian inference for skew-normal linear mixed model. *Journal of Applied Statistics*, 34:663–682.

Arellano-Valle, R. B., Castro, L. M., Genton, M. G., and Gómez, H. W. (2008). Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis. *Bayesian Analysis*, 3(3):513–540.

Arellano-Valle, R. B., Galea-Rojas, M., and Iglesias, P. (1999-2000). Bayesian analysis in elliptical linear regression models. *Journal of the Chilean Statistical Society*, 16-17:59–104.

Arellano-Valle, R. B. and Genton, M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, 96:93–116.

Arellano-Valle, R. B., Gómez, H. W., and Quintana, F. A. (2005). Statistical inference for a general class of asymmetric distributions. *Journal of Statistical Planning and Inference*, 128:427–443.

Arnold, B. C. and Beaver, R. J. (2000). Some skewed multivariate distributions. *American Journal of Mathematical and Management Sciences*, 20(1-2):27–38.

Ayebo, A. and Kozubowski, T. J. (2004). An asymmetric generalization of Gaussian and Laplace laws. *Journal of Probability and Statistical Science*, 1(2):187–210.

Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208.

Azzalini, A. and Dalla-Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Basu, S. and Mukhopadhyay, S. (2000a). Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhya: The Indian Journal of Statistics*, 62:372–387.

Basu, S. and Mukhopadhyay, S. (2000b). Binary response regression with normal scale mixture links. In Dey, D. K., Ghosh, S. K., and Mallick, B. K., editors, *Generalized Linear Models: A Bayesian Perspective*, pages 231–241. Marcel Dekker, New York.

Bazán, J. L., Bolfarine, H., and Branco, M. D. (2010). A framework for skew-probit links in binary regression. *Communications in Statistics - Theory and Methods*, 39(4):678–697.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91(436):1450–1460.

Bedrick, E. J., Christensen, R., and Johnson, W. (1997). Bayesian binomial regression: predicting survival at a Trauma Center. *The American Statistician*, 51(3):211–218.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, Chichester.

Blattberg, R. C. and Gonedes, N. J. (1974). A comparison of the Stable and Student distributions as statistical models for stock prices. *The Journal of Business*, 47(2):244–280.

Bliss, C. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1):134–167.

Bottazzi, G. and Secchi, A. (2011). A new class of asymmetric exponential power densities with applications to economics and finance. *Industrial and Corporate Change, Oxford University Press*, 20(4):991–1030.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Box, G. E. P. and Tiao, G. C. (1962). A further look at robustness via Bayes's theorem. *Biometrika*, 49(3/4):419–432.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Massachusetts.

Brookhouser, P. E. (1994). Prevention of noise-induced hearing loss. *Preventive Medicine*, 23(5):665–669.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Bross, I. (1954). Misclassification in 2×2 tables. *Biometrics*, 10:478–486.

Buonaccorsi, J. P. (2010). *Measurement Error*. Chapman & Hall/CRC, London.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, Florida, second edition.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674.

Chen, J. D. and Tsai, J. Y. (2003). Hearing loss among workers at an oil refinery in Taiwan. *Archives of Environmental Health*, 58(1):55–58.

Chen, M.-H. and Dey, D. K. (1998). Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyã: The Indian Journal of Statistics, Series A*, 60:322–343.

Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999a). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448):1172–1186.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000a). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84:121–137.

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999b). Prior elicitation, variable selection, and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society, Series B*, 61:223–242.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000b). *Monte Carlo Methods in Bayesian Computation*. Series in Statistics. Springer, New York.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49(4):327–335.

Coles, R. R. A., Lutman, M. E., and Buffin, J. T. (2000). Gudelines on the diagnosis of noise-induced hearing loss for medicolegal purposes. *Clinical Otolaryngology*, 25:264–273.

Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall, London.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. John Wiley & Sons, Chichester.

Congdon, P. (2006). *Bayesian Statistical Modelling*. John Wiley & Sons, Chichester, second edition.

Congdon, P. D. (2010). *Applied Bayesian Hierarchical Methods*. Chapman & Hall/CRC, Boca Raton, Florida.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons, New York.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Cowling, D., Johnson, W. O., and Gardner, I. A. (2001). Bayesian modelling of risk when binary outcomes are subject to error. Technical report, Department of Statistics, University of California, Davis.

Cox, D. R. (1971). *The Analysis of Binary Data*. Methuen, London.

Curtis, S. M. and Ghosh, S. K. (2011). A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression. *Journal of Statistical Theory and Practice*, 5(4):715–735.

Czado, C. (1994). Parametric link modification of both tails in binary regression. *Statistical Papers*, 35:189–201.

Czado, C., Heyn, A., and Müller, G. (2011). Modeling individual migraine severity with autoregressive ordered probit models. *Statistical Methods and Applications*, 20(1):101–121.

Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33(2):213–231.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B*, 61(2):331–344.

Delicado, P. and Goria, M. N. (2008). A small sample comparison of maximum likelihood, moments and l-moments methods for the asymmetric exponential power distribution. *Computational Statistics and Data Analysis*, 52:1661–1673.

Dellaportas, P. and Smith, A. M. F. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, 42(3):443–459.

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

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker, New York.

Diaconis, P. and Freedman, D. A. (1993). Nonparametric binary regression: A Bayesian approach. *The Annals of Statistics*, 21(4):2108–2137.

Dias, J. G. and Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic Gaussian mixture likelihoods. *Statistics and Computing*, 14:323–332.

DiCiccio, T. J. and Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association*, 99(466):439–450.

Dobie, R. A. (1995). Prevention of noise-induced hearing loss. *Archives of Otolaryngology Head & Neck Surgery*, 121(4):385–91.

Eyheramendy, S. and Madigan, D. (2007). A flexible Bayesian generalized linear model for dichotomous response data with an application to text categorization. *IMS Lecture Notes-Monograph Series*, 54:76–91.

Fernández, C., Osiewalski, J., and Steel, M. F. J. (1995). Modeling and inference with v-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340.

Fernández, C. and Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Fernández, C. and Steel, M. F. J. (1999). Multivariate Student-t regression models: Pitfalls and inference. *Biometrika*, 86:153–167.

Fernández, C. and Steel, M. F. J. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16:80–101.

Ferrão, M. E. and Goldstein, H. (2014). Adjusting for differential misclassification in multilevel models: the relationship between child exposure to smoke and cognitive development. *Quality & Quantity*, 48:251–258.

Ferreira, J. T. A. S. and Steel, M. F. J. (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica*, 17:505–529.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton, Florida, second edition.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelfand, S. (2001). *Auditory system and related disorders*. Essentials of Audiology. Thieme, New York: Thieme, second edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman & Hall/CRC, Boca Raton, Florida.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4):445–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348.

Goldstein, H. (2011). *Multilevel Statistical Models*. Wiley, Chichester, 4th edition.

Goldstein, H., Kounali, D., and Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling*, 8(3):243–261.

Gómez, E., Gómez-Villegas, M. A., and Marín, J. M. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3):589–600.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman and Hall, Boca Raton, Florida.

Haro-López, R. A., Mallick, B. K., and Smith, A. F. M. (2000). Binary regression using data adaptive robust link functions. In Dey, D. K., Ghosh, S. K., and Mallick, B. K., editors, *Generalized Linear Models: A Bayesian Perspective*, pages 243–253. Marcel Dekker, New York.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, second edition.

Huang, Y. (2008). Long-term HIV dynamic models incorporating drug adherence and resistance to treatment for prediction of virological responses. *Computational Statistics and Data Analysis*, 52:3765–3778.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86(416):981–986.

ICDAS (2005). *Criteria Manual. International Caries Detection and Assessment System (ICDAS II)*. International Caries Detection and Assessment System (ICDAS) Coordinating Committee.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer, New York.

Jones, M. C. (2005). In discussion of R. A. Rigby and D. M. Stasinopoulos (2005) Generalized additive models for location, scale and shape. *Applied Statistics*, 54(3):507–554.

Kacperczyk, M., Damien, P., and Walker, S. G. (2013). A new class of Bayesian semi-parametric models with applications to option pricing. *Quantitative Finance*, 13(6):967–980.

Kim, S., Chen, M.-H., and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika*, 95(1):93–106.

Komunjer, I. (2007). Asymmetric power distribution: Theory and applications to risk measurement. *Journal of Applied Econometrics*, 22(5):891–921.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896.

Liu, C. (2004). Robit regression: A simple robust alternative to logistic and probit regression. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 227–238. John Wiley and Sons, New York.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2012). *The BUGS Book - A Practical Introduction to Bayesian Analysis*. CRC Press / Chapman & Hall, Boca Raton, Florida.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Ma, Y., Genton, M., and Davidian, M. (2004). Linear mixed effects models with flexible generalized skew-elliptical random effects. In Genton, M. G., editor, *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, pages 339–358. Chapman and Hall/CRC, Boca Raton, Florida.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48(3):188–190.

Martín, J. and Pérez, C. J. (2009). Bayesian analysis of a generalized lognormal distribution. *Computational Statistics and Data Analysis*, 53:1377–1387.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton, Florida, second edition.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99:173–193.

McGlothlin, A., Stamey, J. D., and Seaman, J. W. J. (2008). Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach. *Biometrical Journal*, 50(1):123–134.

McGrory, C. and Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, 51:5352–5367.

McInturff, P., Johnson, W. O., Cowling, D., and Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in Medicine*, 23:1095–1109.

Meza, C., Jaffrézic, F., and Foulley, J.-L. (2009). Estimation in the probit normal model for binary outcomes using the SAEM algorithm. *Computational Statistics and Data Analysis*, 53:1350–1360.

Monti, A. C. (2003). A note on the estimation of the skew normal and the skew exponential power distributions. *METRON - International Journal of Statistics*, LXI(2):205–219.

Mudholkar, G. S. and George, E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika*, 65:667–668.

Mutsvari, T. (2012). *Misclassification in Multilevel Models with Application to Dental Caries Research*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Mutsvari, T., Bandyopadhyay, D., Declerck, D., and Lesaffre, E. (2013). A multilevel model for spatially correlated binary data in the presence of misclassification: an application in oral health research. *Statistics in Medicine*, 32(30):5241–5259.

Mutsvari, T., Lesaffre, E., García-Zattera, M. J., Diya, L., and Declerck, D. (2010). Factors that influence data quality in caries experience detection: a multilevel modeling approach. *Caries Research*, 44:438–444.

Mwalili, S. M., Lesaffre, E., and Declerck, D. (2005). A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, 54(1):77–93.

Naranjo, L., Martín, J., Pérez, C. J., and Rufo, M. J. (2014a). Addressing misclassification for binary data: probit and t-link regressions. *Journal of Statistical Computation and Simulation*. In Press.

Naranjo, L., Pérez, C. J., and Martín, J. (2012). Bayesian analysis of a skewed exponential power distribution. In *Proceedings of COMPSTAT 2012, 20th International Conference on Computational Statistics*, pages 641–652, Limassol, Cyprus.

Naranjo, L., Pérez, C. J., Martín, J. R., and Lesaffre, E. (2014b). A Bayesian approach for misclassified ordinal response data. Preprint 155, Universidad de Extremadura, Badajoz, Spain.

Newton, M. A., Czado, C., and Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91(433):142–153.

Nott, D. J. and Leng, C. (2010). Bayesian projection approaches to variable selection in generalized linear models. *Computational Statistics and Data Analysis*, 54:3227–3241.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, New Jersey.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester.

OSHA (Occupational Safety and Health Administration) (2002). *Hearing Conservation, Document OSHA 3074*. U.S. Department of Labor.

Pardo, E. (2010). Perda de acuidade auditiva en trabalhadores da Câmara Municipal de Monção. Technical report, Câmara Municipal de Monção, Portugal.

Paulino, C. D., Silva, G., and Achcar, J. A. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics and Data Analysis*, 49:1120–1131.

Paulino, C. D., Soares, P., and Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, 59:670–675.

Pérez, C. J., Girón, F. J., Martín, J., Ruiz, M., and Rojano, C. (2007). Misclassified multinomial data: a Bayesian approach. *RACSAM*, 101(1):71–80.

Pine, C. M., Pitts, N. B., and Nugent, Z. J. (1997). British association for the study of community dentistry (based) guidance on the statistical aspects of training and calibration of examiners for surveys of child dental health, a based coordinated dental epidemiology program me quality standard. *Community Dental Health*, 14 (Suppl 1):18–29.

Pitts, N. B., Evans, D. J., and Pine, C. M. (1997). British association for the study of community dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. *Community Dent Health*, 14(Suppl 1):6–9.

Qian, S. S., Lavine, M., and Stow, C. A. (2000). Univariate Bayesian nonparametric binary regression with application in environmental management. *Environmental and Ecological Statistics*, 7:77–91.

Qin, Z. (2000). *Uniform Scale Mixture Models with Applications to Bayesian Inference*. PhD thesis, University of Michigan, Michigan, USA.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabinowitz, P. M. (2000a). How to prevent noise-induced hearing loss. *American Family Physician*, 61(9):2759–2760.

Rabinowitz, P. M. (2000b). Noise-induced hearing loss. *American Family Physician*, 61(9):2749–2756.

Raftery, A. E. and Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In Bernardo, J. M., Smith, A. F. M., Dawid, A. P., and Berger, J. O., editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, New York.

Rekaya, R., Weigel, K. A., and Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics*, 57(4):1123–1129.

Robert, C. P. (1994). *The Bayesian Choice*. Springer-Verlag, New York.

Rocker, G. M., Pearson, D., Stephens, M., and Shale, D. J. (1988). An assessment of a double-isotope method for the detection of transferrin accumulation in the lungs of patients with widespread pulmonary infiltrates. *Clinical Science*, 75:47–52.

Roy, S. and Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when the covariates are subject to measurement error. *Biometrical Journal*, 51(3):420–432.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics*, 31(2):129–150.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Smith, B. J. (2007). BOA: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11):1–37.

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

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83:426–431.

Subbotin, M. (1923). On the law of frecuency errors. *Mathematicheskii Sbornik*, 31:296–301.

Swartz, T., Haitovsky, Y., Vexler, A., and Yang, T. (2004). Bayesian identifiability and misclassification in multinomial data. *The Canadian Journal of Statistics*, 32(3):285–302.

Tan, M. T., Tian, G.-L., and Ng, K. W. (2010). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation.* Chapman & Hall/CRC Biostatistics Series, Boca Raton, Florida.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540.

Theodossiou, P. (2000). Skewed generalized error distribution of financial assets and option pricing. http://ssrn.com/abstract=219679. SSRN Working Paper.

Tiku, M. L., Tan, W. Y., and Balakrishnan, N. (1986). *Robust Inference.* Marcel Dekker, New York.

Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. (2000). The Signal-Tandmobiel® project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, 2:87–96.

Vianelli, S. (1963). La misura della variabilità condizionata in uno schema generale delle curve normali di frequenza. *Statistica*, 23:447–474.

Walker, S. G. (1999). The uniform power distribution. *Journal of Applied Statistics*, 26(4):509–517.

Walker, S. G. and Gutiérrez-Peña, E. (1999). Robustifying Bayesian procedures. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 685–710, New York. Oxford University Press.

WHO (1997). *Oral Health Surveys: Basic Methods*. World Health Organization, Geneva, 4 edition. Public Health Report.

Wood, S. and Kohn, R. (1998). A Bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association*, 93(441):203–213.

Woodhouse, G., Yang, M., Goldstein, H., and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(2):201–212.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *Journal of the American Statistical Association*, 71:400–405.

Zellner, A. and Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, 25:365–393.

Zhu, D. and Galbraith, J. W. (2010). A generalized asymmetric Student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157(2):297–305.

Zhu, D. and Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, 148(1):86–99.

Zhu, L. and Carlin, B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19:2265–2278.