KATHOLIEKE UNIVERSITEIT
LEUVEN

Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Science
Department of Mathematics

Pontificia Universidad Católica de Chile
Faculty of Mathematics
Department of Statistics

MULTIVARIATE MODELS FOR THE ANALYSIS OF CARIES EXPERIENCE DATA SUBJECT TO MISCLASSIFICATION

María José GARCÍA ZATTERA

Dissertation presented in partial fulfillment
of the requirements for the degree of

Doctor in Science

Doctor in Statistics

January 2011

MULTIVARIATE MODELS FOR THE ANALYSIS OF CARIES EXPERIENCE DATA SUBJECT TO MISCLASSIFICATION

María José GARCÍA ZATTERA

Supervisors:
Prof. Dr. E. Lesaffre
Prof. Dr. G. Marshall
Prof. Dr. A. Carbonez

Members of the Examination Committee:
Prof. Dr. D. Declerck
Prof. Dr. G. del Pino
Prof. Dr. G. Icaza
Prof. Dr. F. Quintana

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Science and Doctor in Statistics

January 2011

ISBN 978-90-8649-385-2
D/2011/10.705/2

Acknowledgements

The road has been rather long and not easy. There are many people who helped me through the different stages of this long journey and I would like to take this opportunity to thank those who made this thesis possible.

I would specially like to thank my supervisor in Leuven, Emmanuel Lesaffre, who always trusted me, even when I was about to give up. He gave me unconditional support and confidence in my work. This thesis would not have been possible without my supervisor in Santiago, Guillermo Marshall, who was willing to support my work and give me the opportunity to finish my PhD. Thanks to Emmanuel for introducing me to the challenging oral health research and to Guillermo for letting me continue working in the field I had started. Special thanks to Dominique Declerck, who played the role of a co-supervisor, for her continuous help and being always available for dental questions. Her discussions and suggestions were essential to a better understanding of the analysis oral health data. I would also like to thank An Carbonez and Geert Verbeke for their supervision along the progress of this thesis.

All the developments of this thesis were motivated by the data of the Signal Tandmobiel® study. I would like to thank to all who collected these interesting data and allowed me to use them in this thesis. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (L-BioStat, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

I would like to gratefully acknowledge the financial support during my PhD research from the Research Grants OT/00/35, OT/00/60 and OT/05/60, Katholieke Universiteit Leuven and from the National Scholarship for Doctoral Studies 2009, Conicyt (Chile). I would also like to show my gratitude to Emmanuel for giving me the opportunity to join the Biostatistical Centre, now L-Biostat at KUL. This experience not only gave me the chance to learn all about doing research, but to meet wonderful people who have helped me in different ways through all these years. Thanks to Kris, Samuel, Arnošt, Ann, Roula, Dimitris, Steffen and Timothy for your collaboration and friendship. Special gratitude goes to Silvia for being such a good friend, for her help, companionship and support during my visits to Leuven. I am also grateful for the administrative support provided by Jeannine and Kirsten.

I am indebted to the people with whom I shared my life in Belgium. I would specially like to thank Jessie, Peter, Jorge, Fanny, Katalin and István for helping me to feel that I was not alone and for staying by me when I needed most. Thanks

also to Ernesto, Gladys and their children for making me feel at home.

I wish to express sincere gratitude to my family. Thanks for your foolproof love and for holding me up. Special thanks go to my parents, Tachy and Toño, for all you have taught to me throughout my life, for your unconditional support and for always having a warm hug for me. Thanks to my sisters Consuelo and Constanza for your patience and confidence in me. Thanks to my grandmothers, Nona and Nena, my aunt, Ana María, my brother in law, Hernán and to Gloria for always being there to support me.

I could not forget the unconditional help of my friends Ale, Andrés, Ángela, Anita, Carola, Coté, Danitza and Marcia, to keep the horizon and being stand, even on the most difficult parts of this journey. Thanks also to Felipe for our daily chats about life in general and for always willing to listen and give an objective opinion. To Hannah for her dedication and help with English in the last stage of this manuscript and to Soledad for her administrative support.

Last but not least, I owe my deepest gratitude to Alejandro. There are no word nor enough space to say thanks to you. This long journey has not been easy to you either but, despite all what happened in the meanwhile, you were always willing to help. Your certainty that this was the right path and that I was able to do it, gave me the strength to continue till the end. Thanks for your patience, wisdom, peace, encouragement, support and unconditional confidence, but above all, thanks for your pure and everlasting love. This thesis is also dedicated to you.

María José García Zattera
Santiago, January 2011

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

Chapter 2:

García-Zattera, M. J., Jara, A., Lesaffre, E. & Declerck, D. (2007). Conditional independence of a multivariate binary data with an application in caries research. *Computational Statistics and Data Analysis* 51 3223–3234.

Chapter 3:

García-Zattera, M. J., Marshall, G. & Lesaffre, E. (2010). Effect of Misclassification on the Association Parameters of Multivariate Binary Data. *In preparation.*

Chapter 4:

García-Zattera, M. J., Mutsvari, T., Jara, A., Declerck, D. & Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine* 29(30) 3103–3117.

Chapter 5:

García-Zattera, M. J., Jara, A., Lesaffre, E. & Marshall, G. (2010). Multivariate modelling of a monotone disease process in the presence of misclassification. In A. Bowman, ed., *Proceedings of the 25th Workshop of Statistical Modelling.* Glasgow, UK: University of Glasgow, 221–226.

García-Zattera, M. J., Jara, A., Lesaffre, E. & Marshall, G. (2010). Modelling of multivariate monotone disease processes in the presence of misclassification. Invited re-submission to *Journal of the American Statistical Association.*

The author has also been co-author of the following publications:

2010

Fehlmann, E., Tapia, J. L., Fernández, R., Bancalari, A., Fabres, J., D'Apremont, I., García-Zattera, M. J., Grandi, C., Ceriani Cernadas, J. M. & Grupo Colaborativo Neocosur (2010). Impact of respiratory distress syndrome in very low birth weight infants: a multicenter South-American study. *Archivos Argentinos de Pediatría* 108(5) 393–400.

Mutsvari, T., García-Zattera, M. J., Declerck, D. & Lesaffre, E. (2010). Dealing with misclassification and missing data when estimating prevalence and incidence of caries experience. Submitted to *Community Dentistry and Epidemiology*.

Mutsvari, T., Lesaffre, E., García-Zattera, M. J., Diya, L. & Declerck, D. (2010). Factors that influence data quality in caries experience detection: a multi-level modeling approach. *Caries Research* 44(5) 438–444.

2009

Riquelme, A., Arrese, M., Soza, A., Morales, A., Baudrand, R., Pérez-Ayuso, R. M., González, R., Alvarez, M., Hernández, V., García-Zattera, M. J., Otarola, F., Medina, B., Rigotti, A., Miquel, J. F., Marshall, G. & Nervi, F. (2009). Non-alcoholic fatty liver disease and its association with obesity, insulin resistance and increased serum levels of C-reactive protein in Hispanics. *Liver International* 29(1) 82–88.

2008

Declerck, D., Leroy, R., Martens, L., Lesaffre, E., García-Zattera, M. J.,Vanden Broucke, S., Debyser, M. & Hoppenbrouwers, K. (2008). Factors associated with prevalence and severity of caries experience in pre-school children. *Community Dentistry and Oral Epidemiology* 36(2) 168–178.

2007

Jansen, F. H., Lesaffre, E., Penali, L. K., García-Zattera, M. J., Die-Kakou, H. & Bissagnene, E. (2007). Assessment of the relative advantage of various Artesunate-based combination therapies by a multi-treatment Bayesian random-effects meta-analysis. *American Journal of Tropical Medicine and Hygiene* 77(6) 1005–1009.

Jara, A., García-Zattera, M. J. & Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics and Data Analysis* 51(11) 5402–5415.

Lesaffre, E., García-Zattera, M. J., Redmond, C., Huber, H. & Needleman, I. (2007). Reported methodological quality of split-mouth studies. *Journal of Clinical Periodontology* 34(9) 756–761.

Martens, L., Leroy, R., Jara, A., García-Zattera, M. J., Lesaffre, E. & Declerck, D. (2007). Longitudinally registered plaque accumulation in a cohort of Flemish schoolchildren. *Quintessence international* 38(7) 555–564.

Vanobbergen, J., Lesaffre, E., García-Zattera, M. J., Jara, A., Martens, L. & Declerck, D. (2007). Caries patterns in primary dentition in 3-, 5- and 7-year-old children: spatial correlation and preventive consequences. *Caries Research* 41(1) 16–25.

2006

Nervi, F., Miquel, J. F., Alvarez, M., Ferreccio, C., García-Zattera, M. J., González, R., Pérez-Ayuso, R. M., Rigotti, A. & Villarroel, L. (2006). Gallbladder disease is associated with insulin resistance in a high risk Hispanic population. *Journal of Hepatology* 45(2) 299–305.

Contents

Summary

Oral diseases, such as dental caries, are a major health problem worldwide. Even though the prevalence of dental caries in children of Western countries has declined considerably in the last three decades, the disease has now become concentrated in a small group of children. Only a small proportion of the population experiences 50% of all caries lesions. The most likely explanation for the difference in oral health seems to be socio-economic environmental factors and it occurs during childhood. Therefore, to improve dental health, early identification of groups at a particular risk of developing caries becomes essential.

The identification of risk groups for dental caries is challenging because often oral health data show a complex structure. Caries experience (defined as past or present caries on teeth) data have a hierarchical structure (mouth, jaw, tooth, surface on tooth) with the lowest levels of highest interest to oral health researchers. This leads to the analysis of high dimensional correlated data, because events on tooth surfaces of the same child are dependent and, therefore, the conclusions arising from statistical methods ignoring such an association may be misleading. On top of that, the detection of dental caries might be difficult for a variety of reasons. Hence, misclassification of dental caries is likely to happen in practice. The fact that scoring caries is done with considerable error further complicates inference. The previous complexities of dental caries data sets, make necessary the development of adequate statistical methods that take into account all these aspects of the data at the same time in order to obtain valid inferences.

The understanding of the association structure of the caries process is important for the understanding of the etiology of the disease and can help the dentists in optimizing their clinical examination of the patient and direct preventive and restorative approaches. Motivated by dental data gathered from a longitudinal oral health study conducted in Flanders (Belgium), the Signal-Tandmobiel® study, we have studied the interpretation and the effect of the misclassification on the association parameters associated with two models for the analysis of multivariate binary data.

We have also proposed uni- and multi-variate Markov models for the analysis of longitudinal monotone binary data subject to misclassification. These models account for the effects of the covariates on the prevalences and incidences, and allow for the existence of different classifiers. Empirical and theoretical evidence are provided to show that the model parameters can be estimated from the main data without the need of external information on the misclassification parameters. In the multivariate Markov model, the joint distributions are defined through the specification of compatible full conditional distributions. The proposed multivariate hidden Markov model permits the study of the within- and across-time association parameters among the responses.

Samenvatting

Mondziekten, zoals tandcariës, zijn nog steeds een belangrijk gezondheidsprobleem en dit wereldwijd. Hoewel de prevalentie van cariës bij kinderen in de westerse landen aanzienlijk is gedaald in de laatste drie decennia, merken we dezer dagen een polarisatie op van het fenomeen. Namelijk cariës is nu vooral geconcentreerd in een vrij kleine groep van kinderen en dit reeds vanaf een zeer jonge leeftijd. Dit verschil in mondgezondheid is terug te brengen tot een verschil in sociaal-economische omgevingsfactoren. Om de tandheelkundige gezondheid in de populatie in zijn globaliteit te verbeteren, is daarom de vroege identificatie van de hoog risico groep (voor het ontwikkelen cariës) van essentiëel belang.

De identificatie van risicofactoren voor tandcariës (actieve of behandelde lesies) kan veeleisend zijn omwille van de hierarchische structuur van de gegevens. Tandcariësgegevens hebben namelijk een hiërarchische structuur (mond, tand, tandvlak) waarbij vooral het tandvlak de meeste interesse wegdraagt van de tandarts. De zoektocht naar risicofactoren voor tandcariës (nog actief of ontstaan in het verleden en verholpen) leidt snel tot statistische technieken voor hoog-dimensionale gecorreleerde data. Immers tandoppervlakken binnen éénzelfde mond zijn blootgesteld aan dezelfde omgevingsfactoren, bvb diëet en tandhygiëne. Het negeren van deze structuur in de statistische analyse kan leiden tot misleidende conclusies.

Daarenboven is de diagnose van tandcariës minder triviaal dan op het eerste gezicht de meesten vermoeden. De cariësstatus wordt immers vaak verkeerd ingeschat, men spreekt dan van misclassificatie van de cariësstatus. Uiteraard compliceert deze misclassificatie verder de besluittrekking. De statistische technieken zullen dan ook moeten rekening houden met dit scoringsproces om correcte besluiten te bekomen.

Het begrijpen van de spatiale spreiding van het cariësproces in de mond kan inzicht geven in het ontstaan van de ziekte en kan de tandartsen helpen om de klinische behandeling van de patiënt, zowel preventief als restauratief, te optimaliseren. Gemotiveerd door tandheelkundige gegevens uit een longitudinaal

mondgezondheidsstudie uitgevoerd in Vlaanderen (België), genaamd de Signal-Tandmobiel® studie, hebben we de spatiale structuur van tandcariës onderzocht alsook de impact van het fout scoren hierop en dit met behulp van twee multivariate statistische modellen voor binaire data. We hebben verder ook uni-en multivariate Markov modellen voorgesteld voor de analyse van longitudinale monotone binaire data onderworpen aan misclassificatie. Deze modellen meten het effect van covariaten op de prevalenties en incidenties en houden rekening met het feit dat meerdere tandheelkundige scoorders de gegevens noteerden. Empirische en theoretische details worden gegeven die moeten aantonen dat de model parameters kunnen worden geschat zonder gebruik te maken van externe validatie data. In het multivariate Markov-model zijn de gezamenlijke verdelingen bepaald aan de hand van compatibele conditionele verdelingen. Het voorgestelde latente Markov model laat toe om de verbanden tussen tandcariës (binair gescoord) te schatten zowel cross-sectioneel als longitudinaal.

Preface

The developments of this thesis were motivated by data gathered in a longitudinal oral health study, the Signal-Tandmobiel® study. The complex structure of dental data, makes unfeasible the use of standard statistical approaches and requires the development of adequate statistical methods that take into account all challenging aspects involved in order to obtain valid inferences. This thesis is devoted to the study and development of models for the analysis of multivariate binary data. Despite the fact that our motivation comes from dental data, the results equally apply to other research areas. Since the majority of the chapters of this manuscript correspond to either published or submitted papers, they are self-contained and every appendix corresponds to the supplementary material of the corresponding original paper.

Chapter 1 gives an introduction to oral health issues and describes the motivating data. An overview of the main topics involved in the analysis of dental data is also provided in this chapter. Specifically, Section 1.5 describes approaches to analyzing correlated data and Section 1.4 presents an outline of the effect of misclassification and different approaches to correct for it.

Even though, in general, the main interest is on the inferences on the mean structure and the association parameters can be considered as nuisance, in many situations the association structure is as crucial as the mean structure to answer the scientific questions. For instance, in oral health research it is of interest to assess the association of caries experience among different teeth. This knowledge can help the dentists in their clinical examination of the patient and in the understanding of the etiology of the disease. In Chapter 2 we study the relationship between the association structure induced by two different statistical models for the analysis of correlated binary data. We present a numerical example and a theoretical proof showing that the results and conclusions can be markedly different depending on the model considered.

Binary variables in epidemiological studies are often subject to misclassification. The diagnosis of caries experience is not an easy task, therefore misclassification is

likely to happen in oral health data. The effect of misclassification on the statistical inference has been widely investigated in the literature and, as a consequence, several approaches have been proposed to correct for it. However, most of the literature has been focused on the effect of misclassification on the inferences of regression coefficients and relatively less attention has been paid on its effect on the association parameters. Chapter 3 evaluates the effect of misclassification on the inferences about the association parameters induced by two models for the analysis of multivariate binary data.

The approaches proposed to correct for misclassification in cross-sectional studies rely on the availability of extra information about the misclassification process. Since this information is difficult to obtain, in Chapter 4 we investigate whether a misclassified monotone process contains all the necessary information to identify the model parameters without this extra information. In this chapter we study the identifiability of the parameters of a simple hidden Markov Model and extend it to allow the inclusion of covariates and different classifiers.

It is well known that the presence of misclassification implies a loss of power, which might be strengthened in case of univariate analyzes. This lead us to extend the simple hidden Markov model of Chapter 4 and to propose a multivariate hidden Markov model for monotone data. In Chapter 5 we propose and evaluate the small sample properties of a multivariate binary inhomogeneous Markov model in which unobserved monotone response variables are subject to misclassification. A Bayesian version of the model is described in detail where the multivariate baseline distributions and Markov transition matrices are defined as a function of covariates, throughout the specification of compatible full conditional distributions. In this proposal, the association structure is studied trough within- and across-time odds ratio parameters.

Finally, general conclusions and topics for future research are given in Chapter 6.

Nomenclature

Here, we give a list of the most often used notations and symbols in the thesis. The meaning of them is usually indicated once, when they first occur in each chapter.

$P(A)$	:	probability of the event A.
$X \perp\!\!\!\perp Y$	:	the random variables X and Y are independent.
$X \perp\!\!\!\perp Y \mid Z$	:	the random variables X and Y are conditionally independent given the random variable Z.
iid	:	independent and identically distributed.
ind	:	independent.
$N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	:	k-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\phi(\cdot)$	:	density of a standard normal distribution, $N(0, 1)$.
$\phi_k(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	:	density of a $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.
$\text{Beta}(a, b)$	:	Beta distribution with parameters a and b.
$\Gamma(a_0, b_0)$	:	gamma distribution with shape parameter a_0 and rate parameter b_0.
$I(\cdot)_A$	:	generic indicator function such that $I(x)_A = 1$ if $x \in A$ and 0 otherwise.

Abbreviations

Here, we give a list of the most often used abbreviations in the thesis. The meaning of them is usually indicated once, when they first occur in each chapter.

AIC	:	Akaike's information criteria.
BIC	:	Bayesian information criteria.
CE	:	Caries experience.
CI	:	Confidence interval.
CSLRM	:	Conditionally specified logistic regression model.
CPO	:	Conditional predictive ordinates.
FDI	:	Federation dentaire internationale.
GEE	:	Generalized estimating equation.
GLM	:	Generalized linear model.
GLMM	:	Generalized linear mixed model.
HMM	:	Hidden Markov model.
HPD	:	Highest posterior density.
LLM	:	Log-linear model.
LPML	:	Log pseudo marginal likelihood.
MCMC	:	Markov chain Monte Carlo.
MH	:	Metropolis-Hastings.
MLE	:	Maximum likelihood estimator.
MLM	:	Multvariate logistic model.
MPM	:	Multivariate probit model.
MSE	:	Mean squared error.
PsCP	:	Pseudo contour probability.
PML	:	Pseudo marginal likelihood.
PsBF	:	Pseudo Bayes factor.

ST : Signal Tandmobiel®.

List of Figures

List of Tables

Part I

General Introduction

Chapter 1

Introduction and Background Material

1.1 Why Oral Health?

Oral health means more than just an attractive smile. Poor oral health and untreated oral diseases and conditions can have a significant impact on quality of life. The link between oral infections and stroke, heart disease, and pre-term low-birth-weight babies, is becoming well documented and accepted within the health care community (see, e.g. Desvarieux et al., 2010). Likewise, more than 90% of all systemic diseases have oral manifestations, meaning the dentist may be the first health care provider to diagnose a health problem. Further, the early detection of oral diseases, contributes to the early diagnosis, prevention and treatment of major diseases like HIV/AIDS, which usually show up first in the form of oral fungal infections and injuries, bacterial or viral infections, and cardiovascular diseases.

According to the World Health Organization, oral diseases such as dental caries, periodontitis, and cancers of the mouth and pharynx are a major health problem worldwide, affecting developed countries and, with increasing frequency, developing countries, especially among poorer communities. The effects of oral diseases on pain, suffering and reduced quality of life are extensive and expensive. Treatment is estimated to represent between 5% and 10% of health costs in industrialized countries, and is beyond the resources of many developing countries. Dental caries and periodontal diseases have historically been considered the most important global oral health burdens (World Health Organization, 2003). Dental caries is the most prevalent oral disease in several Asian and Latin American countries. Although it seems that the problem is less severe in most African countries, the report states that, with the change in living conditions, is likely to increase dental caries in many developing countries of that continent, especially

due to growing consumption of sugars and inadequate exposure to fluorides.

Despite the fact that the past three decades have witnessed a dramatic decline in the prevalence of dental caries in children in countries of the Western World (see, e.g. Glass, 1982; Petersson & Bratthall, 1996), dental caries remains an important childhood disease affecting a considerable proportion of young children. About 10 to 15% of the children experience 50% of all caries lesions and 25 to 30% suffer 75% of the lesions (see, e.g. Marthaler et al., 1996). The most likely explanation for the difference in oral health seems to be socio-economic environmental factors, indicating that a considerable proportion of the target group does not benefit from traditional preventive approaches (Hausen et al., 2000). Therefore, the identification of groups at a particular risk of developing caries becomes essential to improve dental health. In particular, the estimation of the prevalence and incidence of dental caries, and the assessment of risk factors and of the association of caries among different teeth, is of major importance to direct preventive programs since childhood. This prevention is overriding in the deciduous dentition, given that the presence of caries in the primary dentition implies a higher risk of developing caries on the permanent teeth and accelerates the emergence of the successors (see, e.g. Leroy et al., 2003)

1.2 Dental Caries and Diagnostic

Dental caries, also known as tooth decay or cavity, is a disease wherein bacterial processes dissolve tooth enamel (outer layer of a tooth). This tissue progressively breaks down, producing dental caries (cavities, holes in the teeth), followed by the spread into the dentine and eventually the pulp. If left untreated, the disease can lead to pain, tooth loss, infection, and, in severe cases, death.

All sugars or carbohydrates present in the food (e.g. sucrose, fructose, and glucose) can easily remain in the mouth, sticking to the teeth if they are not cleaned regularly after every meal. These sugars are defined fermentable because they can be easily metabolized by the bacteria present in dental plaque to produce organic acid compounds, very aggressive for teeth enamel. On the other hand, exposure to alkali, such as sodium bicarbonate in saliva, reverses this process and aids in remineralization. Therefore, a tooth (which is primarily mineral in content) is in a constant state of back-and-forth de- and re-mineralization between the tooth and surrounding saliva. When demineralization proceeds faster than remineralization, dental caries occurs (see, e.g. Moynihan, 2000).

The presentation of caries is highly variable. Initially, it may appear as a small chalky area that may eventually develop into a large cavitation. Sometimes caries may be directly visible, however other methods of detection such as radiographs are used for less visible areas of teeth and to judge the extent of destruction.

Caries lesions are commonly scored in four levels of severity: d_4 (dentine caries with pulpal involvement), d_3 (dentine caries with obvious cavitation), d_2 (hidden dentine caries) and d_1 (white or brown-spot initial lesions in enamel without cavitation) (see, e.g. Fyffe et al., 2000; Pitts, 2004). Depending on the extent of tooth destruction, various treatments can be used to restore teeth to proper form, function, and aesthetics, but so far, there is no known method to regenerate large amounts of tooth structure. Instead, dental health organizations advocate preventive and prophylactic measures, such as regular oral hygiene and dietary modifications, to avoid dental caries.

1.3 Motivation

The motivation for the developments of this thesis, comes from dental data gathered in a longitudinal study. In the next sections, we introduce the motivating data set and explain the main difficulties found in this kind of data.

1.3.1 Motivating Data Set: the Signal-Tandmobiel® study

The Signal-Tandmobiel® (ST) study is a longitudinal prospective oral health screening study conducted in Flanders (the north part of Belgium) between 1996 and 2001. This study involved a sample of Flemish children born in 1989, which was obtained using a technique of stratified cluster (i.e. school) sampling without replacement. The fifteen considered strata were obtained combining the three types of educational systems (public, municipal and private schools) and the five Flemish provinces (West Flanders, East Flanders, Brabant, Antwerp and Limburg).

The selection was performed in such a way that each child had the same probability of being selected. Whenever a school was selected, all children in the first class of the selected school were included. Selecting individual children instead of schools would not have been feasible for ethical, practical and economical reasons. The schools were selected with a probability proportional to their size, i.e. the number of children in the first year.

The sample represents 7.3% of the corresponding Flemish population of the same age and consists of 4 468 schoolchildren, 2 153 (48.2%) girls and 2 315 (51.8%) boys. Detailed oral health data at tooth and tooth-surface level (caries experience, gingivitis, etc.) were annually collected on pre-scheduled visits (from the age of 7 to the age of 12) by a team of 16 dental examiners whose examination method was calibrated every six months. In addition, data on oral hygiene and dietary habits were collected using a questionnaire completed by the parents. Every survey year 117 to 1 177 children were not available for examination. The major reasons for

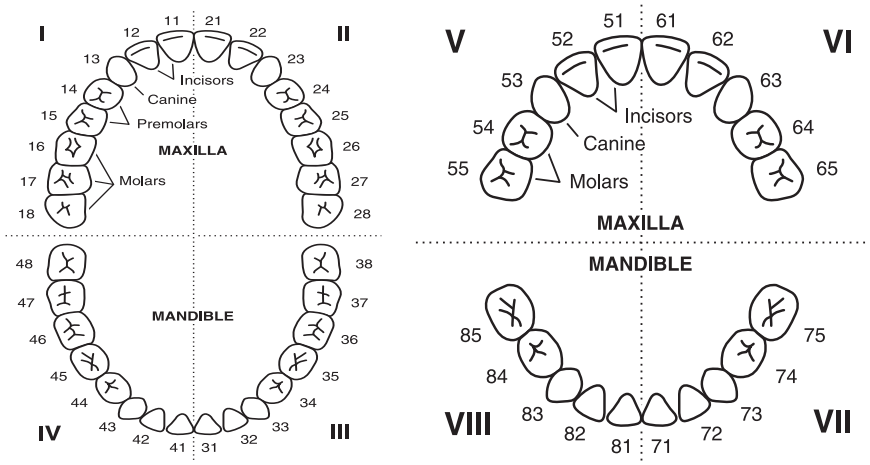
non-participation were not related to the objectives of the study: illness or absence on the day of the examination or change of school. Hence the data set consists of a series of at most 6 longitudinal dental observations and reported oral health habits.

In order to maintain a high level of intra- and inter-examiner reliability during the study period, three calibration exercises (involving 92, 32 and 24 children, respectively) were devoted to the scoring of caries experience at d_3 level, according to the guidelines of training and calibration published by the British Association for the Study of Community Dentistry (Pitts et al., 1997). At the end of each of the three calibration exercises the sensitivity and specificity of each dental examiner vis-a-vis a benchmark examiner was determined.

The information obtained in the calibration exercises were used as validation data set in previous work of the research team (see, e.g. Mwalili et al., 2005). However, these validation data were not taken at random from the main data. Rather a school was selected with a presumed high prevalence of caries experience. A pure random sample would be impractical, but also a validation data set sampled in a clustered manner (first sampling schools and then children within schools) would imply a too high investment in time and personnel. Further, both sampling approaches would likely involve too few children with caries experience implying that the sensitivity would be poorly estimated.

For a more detailed description of the study design and research methods we refer to Vanobbergen et al. (2000).

Here, we concentrate on caries experience on deciduous or permanent molars, which is defined as a binary variable indicating whether a tooth is decayed at d_3 level, missing or filled due to caries. The FDI (Federation Dentaire Internationale) notation to indicate the position of a tooth within the oral cavity is used throughout this thesis. In this two-digit notation, the first digit represents the quadrant number (1 to 4 for permanent teeth and 5 to 8 for deciduous teeth) starting with the maxillary right quadrant, moving around the maxillary arch to the left, then down and back to the right, and ending with the mandibular right quadrant. The second digit represents each tooth in the quadrant, numbered distally from the midline. The upper and lower jaws are also referred to as the maxilla and the mandible, respectively. Figure 1.1 (see page 7) shows the numbering of the permanent (panel a) and deciduous (panel b) teeth according to the FDI notation. Contralateral teeth (left-right), opponent teeth (upper-lower) and diagonal teeth (upper left-lower right or vice versa) are specific pairs of teeth that are analyzed within this manuscript.



(a) (b)

Figure 1.1: Federation Dentaire Internationale notation for the position of permanent (panel a) and deciduous (panel b) teeth. Maxilla = upper jaw, mandible = lower jaw. Quadrants 1 and 4, and quadrants 5 and 8 are at the right-hand side of the subject in panels (a) and (b), respectively. The left-hand side of the subject corresponds to quadrants 2 and 3, and quadrants 6 and 7 for permanent and deciduous teeth, respectively.

1.3.2 Challenging Statistical Problems

The analysis of oral health data and the identification of risk groups for dental caries are often challenging due to the complex data structure. The following three characteristics appear in most statistical problems related with dental data.

1. *High dimensionality of the problem:* there are more than 100 tooth surfaces in the permanent dentition and, in caries research, dentists are interested in lesions at tooth surfaces.

2. *Correlation among measurements:* events on tooth surfaces of the same child are dependent and, therefore, the conclusions arising from statistical methods ignoring such an association may be misleading. Furthermore, the understanding of the association structure of the caries process is important for the understanding of the etiology of the disease and can help the dentists in optimizing their clinical examination of the patient and direct preventive and restorative approaches.

3. *Misclassification:* the diagnosis of dental caries might be difficult for a variety of reasons. The process starts at microscopic level, passes initial stages of visible demineralization without loss of tooth substance and confinement to the enamel, which is difficult to detect. Additionally, nowadays composite materials can imitate the natural enamel so well that it is sometimes difficult to spot a restored lesion. Another reason is that the location of the cavity, e.g. far back in the mouth, hampers the view of dental examiners. Hence, overlooking dental caries is likely to happen in practice, but the dental examiner could also classify discolorations as dental caries.

The previous characteristics of dental caries data sets arise also in other areas of scientific research but usually independently, and have motivated the development of statistical methods for each problem separately. However, in oral health research these characteristics meet each other, requiring the development of adequate statistical methods that take into account all these aspects of the data at the same time in order to obtain valid inferences. Therefore, although this research focuses on the development of statistical methodology for the analysis of dental caries experience data, it equally applies to other research areas.

In Sections 1.4 and 1.5, we review the main issues of correlated data and misclassification, respectively.

1.4 Approaches to Analyzing Correlated Data

Correlated data arise from many epidemiological studies. This term can be used in a generic sense and understand it to encompass such structures as clustered data, multivariate observations, repeated measurements, longitudinal data, and spatially correlated data (Verbeke & Molenberghs, 2000).

The term *clustered data* is used whenever groups or clusters of individuals are randomized to an intervention or when naturally occurring groups in the population are randomly sampled. For example, families, households, hospitals wards, neighborhoods, and schools are instances of naturally occurring clusters in the population that might be the primary sampling units in a study. When more than one characteristic is measured on the same unit, we are dealing with *multivariate observations*. For instance, recording the presence or absence of caries experience on all the teeth of the same child of the ST study, corresponds to multivariate responses and individuals can be thought of as clusters. We refer to *repeated measurements* when the same measure is collected multiple times for the same subject but under different conditions. When repeated measurements are collected to study change in a response variable over time as well as to relate these changes in explanatory variables over time, we refer to *longitudinal data*. The prime advantage of longitudinal studies, over cross-sectional ones, is their

effectiveness for studying change, i.e. they can distinguish changes over time within individuals from differences among people in their baseline levels, something that a cross-sectional study cannot. *Spatial data* arise in a similar setting, but then the interest is to asses the effect of one or more spatial dimensions, instead of time.

Whatever the nature of the correlation, the lack of independence among observations must be accounted for when analyzing data from these studies in order to make valid inferences. There are several models for analyzing correlated data. In the following sections we give a brief overview of these models.

1.4.1 The Summary Statistic Approach

The simplest strategy to deal with correlated data is to avoid the multiple responses of each experimental unit, generating a summary statistic (i.e. mean, median, maximum, etc.) over all observations in the cluster. This produces independent observations and therefore, standard statistical methods can be applied. The latter is one of the main advantages of this approach, along with the fact that it is simple and intuitive.

The possible loss of power and precision due to the decreased sample size and the lack of a way to control for confounding at the site level, are some the disadvantages of this approach.

1.4.2 Generalized Linear Models for Correlated Data

Generalized linear models (GLMs) (Nelder & Wedderburn, 1972) can deal with continuous and discrete outcomes, and can be generalized to deal with unequal cluster sizes and general correlation structures. GLMs assume that a suitable transformation of the mean response is a linear function of the coefficients. Suppose data contain I independent clusters. Let Y_{ij} denote the outcome for observation j in cluster i ($i = 1, \ldots, I$, $j = 1, \ldots, J_i$). Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ_i})'$ be the vector of responses for cluster i and $\boldsymbol{x}_{ij}$ be a design vector for the jth response of cluster i. If observations are independent of each other, the data can be fitted using a GLM given by

$$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta},$$

where $\mu_{ij} = E(Y_{ij}|\boldsymbol{x}_{ij})$, $\boldsymbol{\beta}$ is the vector of regression coefficients quantifying the covariate effect, and $g(\cdot)$ is a monotone and differentiable function, known as the link function, which provides the relationship between the linear predictor, $\boldsymbol{x}'_{ij}\boldsymbol{\beta}$, and the mean of the distribution function. For continuous responses, the link function is usually the identity function $g(u) = u$. For binary responses, commonly used link functions include the logit link $g(u) = \log\{u/(1-u)\}$, the complementary

log-log link $g(u) = \log\{-\log(1-u)\}$, and the probit link $g(u) = \Phi^{-1}(u)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable.

In the presence of within-cluster associations, GLMs are no longer appropriate and need to be extended to deal with correlated data. Depending on the target of inference, there are three major extensions of GLMs for clustered data: *marginal*, *random effects*, and *conditional models*. The selection of the class of models depends on the research questions to be answered and on the assumptions the investigator is willing to make. Below we provide an overview of these three general approaches. For a deeper insight on GLMs to correlated data see, e.g. Diggle et al. (2002), Molenberghs & Verbeke (2005) and Fitzmaurice et al. (2009).

Marginal Models

In marginal models, the regression of the response, $\boldsymbol{Y}_i$, on explanatory variables is modelled separately from within-unit association. In a marginal model, a link function is specified to connect the marginal expectation of a response, $E(Y_{ij}|\boldsymbol{x}_{ij})$, to the linear predictor without conditioning on unobserved random components or on other outcomes, as opposed to random effects models and conditional models, respectively. The term *marginal* is used to emphasize that the average response over the sub-population that shares common values of the covariates, $E(Y_{ij}|\boldsymbol{x}_{ij})$, is being modelled. This implies that the regression coefficients have a *population-averaged* interpretation. In other words, marginal models are natural analogues for correlated data of GLMs for independent data.

It is important to note that these models assume that the conditional mean of the jth response, given $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iJ_i}$, depends only on $\boldsymbol{x}_{ij}$, that is $E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iJ_i}) = E(Y_{ij}|\boldsymbol{x}_{ij})$. This assumption necessarily holds with time-invariant covariates and time-varying covariates that are fixed by design of the study. However, the assumption may no longer hold when a time-varying covariate changes randomly over time. As a consequence, some precaution is required when fitting marginal models with time-varying covariates that are not fixed by design of the study. This problem has been studied by econometricians (see, e.g. Engle et al., 1983) and statisticians (see, e.g. Robins et al., 1999), producing an extensive statistical literature on this topic.

Parametric and semi-parametric marginal models have been discussed in the literature. Examples of parametric models include the multivariate probit model (Ashford & Sowden, 1970; Lesaffre & Molenberghs, 1991; Chib & Greenberg, 1998), the bivariate log-normal and t-student models (Albert, 1992), the scale mixture of normals (Chen & Dey, 1998), the multivariate logit model (O'Brien & Dunson, 2004), the multivariate skew-normal model (Chen, 2004), the Dale model (see, e.g. Dale, 1986; Molenberghs & Lesaffre, 1994), the Bahadur model (Bahadur, 1961)

and the marginal models of Heagerty (see, e.g. Heagerty & Zeger, 2000), among many others.

Semi-parametrically specified models for multivariate categorical data have gained popularity because their parameters can most often be conveniently estimated using generalized estimating equations (GEE) (Liang & Zeger, 1986), which is an extension of the quasi-likelihood approach. In semi-parametric models, a full parametric assumption for the joint distribution of the multivariate responses is not imposed, but a regression model for the mean responses (i.e. making assumptions on the first and second moments of the responses). In particular, the association parameters are considered as nuisance characteristics of the joint distribution and left unspecified. By adopting a "working" assumption about the correlation structure, the GEE approach yields consistent estimators of the regression coefficients, even when the "working" association structure is not close to the true correlation structure. The GEE estimate of $\boldsymbol{\beta}$ is essentially a multivariate analog of the quasi-score function estimate based on quasi-likelihood method. Estimation can be carried out using the iterative Fisher scoring algorithm. Semi-parametric models where some association parameters are parametrically specified have been discussed in Prentice (1988), Zhao & Prentice (1990), Carey et al. (1993), among others.

Random Effects Models

Although marginal models account for correlated data, they do not provide any explanation for the potential source of correlation among the responses. The random effects approach provides a source for the within-unit association introducing random effects in the model of the mean response. These models are known as *generalized linear mixed models* (GLMMs), but in other areas, they are also known as *hierarchical*, *multilevel*, or *random coefficient* models.

In GLMMs, the model for the mean response is conditional on measured exogenous covariates and on unobserved random effects. The inclusion of the random effects induces the marginal correlation among the responses, when averaged over their distribution. In a GLMM, given a vector of random effects $\boldsymbol{b}_i$, the elements of the response vector $\boldsymbol{Y}_i$ are assumed to be conditionally independent following a distribution from the exponential family, such that

$$g\left\{E(Y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)\right\} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$$

where $\boldsymbol{\beta}$ is a vector of fixed effects associated to the design vector $\boldsymbol{x}_{ij}$ and $\boldsymbol{b}_i$ is a vector of random effects associated with the design vector $\boldsymbol{z}_{ij}$. The random effects are assumed to be independent of the covariates and to have a common multivariate distribution with mean zero, $f(\boldsymbol{b}_i|\boldsymbol{D})$. Usual implementations of the model, assume that $f(\boldsymbol{b}_i|\boldsymbol{D})$ is a multivariate normal distribution, where $\boldsymbol{D}$ is the

covariance matrix. The implied marginal distribution of $\boldsymbol{Y}_i$ is given by

$$f_i\left(\boldsymbol{Y}_i|\boldsymbol{\beta},\boldsymbol{D}\right)=\int\prod_{j=i}^{J_i}f\left(Y_{ij}|\boldsymbol{b}_i,\boldsymbol{\beta},\boldsymbol{D}\right)\ f\left(\boldsymbol{b}_i|\boldsymbol{D}\right)d\boldsymbol{b}_i, \tag{1.1}$$

and the covariance between the observations within a cluster is given by

$$\begin{aligned}\operatorname{cov}\left\{Y_{ij},Y_{ik}\right\} &= \operatorname{cov}\left\{E(Y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i),E(Y_{ik}|\boldsymbol{\beta},\boldsymbol{b}_i)\right\}+E\left\{\operatorname{cov}(Y_{ij},Y_{ik}|\boldsymbol{\beta},\boldsymbol{b}_i)\right\}\\ &= \operatorname{cov}\left\{\mu_{ij},\mu_{ik}\right\}+E(0)\\ &= \operatorname{cov}\left\{g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i),g^{-1}(\boldsymbol{x}'_{ik}\boldsymbol{\beta}+\boldsymbol{z}'_{ik}\boldsymbol{b}_i)\right\}.\end{aligned}$$

It is important to note that in general, the marginal distribution in expression (1.1) no longer follows a GLM, due to the non-linearity of the link function typically adopted in regression models for discrete responses. Furthermore, in these cases the marginal mean of $\boldsymbol{Y}_i$ is given by

$$\begin{aligned}E\left(Y_{ij}\mid\boldsymbol{\beta}\right) &= E\left\{E(Y_{ij}\mid\boldsymbol{\beta},\boldsymbol{b}_i)\right\}\\ &= E\left\{g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\boldsymbol{z}'_{ij}\boldsymbol{b}_i)\right\},\end{aligned}$$

which in general cannot be simplified. This shows that the interpretation of the fixed effects parameters $\boldsymbol{\beta}$ is at the level of the conditional mean, given the random effects, and not at the population level. In other words, the regression parameters have *cluster-specific* interpretation because they represent the effects of covariates on changes in an individual's possible transformed mean response per unit change in the covariate, while controlling for all other covariates and the random effects.

Since marginal models yield straightforward interpretation of the regression coefficients, but the random effects approach offers a basis for the interpretation of the association structure, some authors have sought for models that combine the features of both approaches. For example, Heagerty (1999) and Heagerty & Zeger (2000) have developed models that combine the versatility of GLMMs for modelling the within-cluster association with a marginal regression model for the marginal expectation of the responses. They refer to their general class of models as marginalized random effects models. Unlike the standard GLMMs, the marginalized random effects models of Heagerty (1999) have no closed form expression for the conditional probability of response (conditional on the random effects).

Conditional Models

In a conditional model, the parameters describe a feature of responses, for instance, expectation, probability, odds or logit, conditioning on values of the other outcomes (Cox, 1972). A popular class of conditional models is given by the so-called transition models, used for the analysis of longitudinal data. Members of this class are extensions of GLMs for describing the conditional distribution of each response Y_{ij} as an explicit function of the past responses $Y_{i1}, \ldots, Y_{ij-1}$ and covariates $\boldsymbol{x}_{ij}$. Popular members of this class are Markov chains for which the conditional distribution $Y_{ij} \mid Y_{i1}, \ldots, Y_{ij-1}$ depends only on the q prior observations $Y_{ij-q}, \ldots, Y_{ij-1}$. The integer q is referred to as the order of the model. Examples of applications of binary Markov chains can be found in Korn & Whittemore (1979), Stern & Coe (1984) and Zeger et al. (1985), among others. Applications of Markov chains for count data can be found, for instance, in Wong (1986) and Zeger & Qaqish (1988).

Transition models have been criticized because the interpretation of the regression coefficient parameters of one response is conditional on the past responses of the same subject, which is not necessarily of interest if the focus is on marginal inferences. To overcome this problem, Azzalini (1994) proposed a non-homogeneous first-order Markov chain, parameterized such that the regression coefficients have a marginal interpretation. This approach has been extended by Heagerty (2002) and Chen et al. (2009), to allow for more general dependence structures.

Transition models are applicable for longitudinal data but not for other classes of correlated data types. Conditionally specified models for correlated data can be considered to account for more general association structures. An example of this approach is the conditionally specified logistic regression model of Joe & Liu (1996). These authors define a multivariate distribution for binary data by specifying compatible Bernoulli conditional distributions with the conditional probabilities expressed as logistic regression models. This model covers a wide range of dependence structure, allows for the use of known diagnostic methods for logistic regression in the model assessment, and belongs to the exponential family, making the estimation process computationally feasible. Other examples of this approach can be found, for instance, in Rosner (1984) and Connolly & Liang (1988).

1.5 Measurement and Misclassification Errors

Measurement error occurs whenever we cannot exactly observe one or more of the variables that enter into a model of interest and are present in nearly

every discipline. When the true and observed values are both categorical, then measurement error is more specifically referred to as *misclassification*. For instance, in epidemiology, the outcome variable is often presence or absence of a disease, such as AIDS, breast cancer, caries experience, etc. This is often assessed through an imperfect diagnostic procedure, which can lead to either false positives or false negatives.

There are two main issues in a misclassification problem: (i) the evaluation of the consequences of the naive analyses which ignore the misclassification, and (ii) the development of methods to correct for misclassification. With some exceptions, correcting for measurement error requires information or data from external sources. These aspects are reviewed in the following sections.

1.5.1 Effects of Misclassification

The effect of measurement errors has been studied at the response and covariate levels. Much of the research interest in this area has been focused on measurement error in the covariates, particularly in continuous covariates. We refer the reader to Fuller (1987) and Carroll et al. (1995) for general overviews of measurement error problems at the covariates level in multiple linear regression models and GLMs, respectively. Errors in the response, however, has received relatively less attention in the literature. In the remaining of the section we restrict ourselves to misclassification on the response variables.

The effect of misclassification on the responses depends on whether the misclassification generating mechanism is *non-differential* or *differential*. Suppose the true categorical response is denoted by Y and the possible error-corrupted response by Y^*. Consider a regression of the response Y on covariates X. Non-differential misclassification of the response means Y^* is conditionally independent of Y given X, i.e. $f(Y^*|Y,X) = f(Y^*|Y)$. On the other hand, if $f(Y^*|Y,X) \neq f(Y^*|Y)$, differential misclassification of the response has occurred. Reviews on the effects of misclassification include Dalenius (1977), Chen (1989) and Kuha & Skinner (1997).

In an early reference, Bross (1954) showed that non-differential misclassification on a binary response does not affect the validity of the significance test used to compare samples from two populations but the power may be drastically reduced. He also showed that severely biased estimates can be obtained when misclassification is ignored. Tests about the difference between proportions are further discussed by Rubin et al. (1956), Katz & McSweeney (1979), and Zelen & Haitovsky (1991) for the binary case, and Mote & Anderson (1965) and Tenenbein (1970) for the multinomial case. The bias associated to the estimator of relative risk has been studied by Copeland et al. (1977) and Hofler (2005). Several authors have extended these analyzes to the regression context (see, e.g. Buonaccorsi, 2010). In

general, the results suggest that under non-differential misclassification the bias in the regression coefficients is predictable in direction, namely toward the null value. Contrary to popular misconceptions, however non-differential misclassification can sometimes produce bias away from the null, if the response variable has more than two levels (see, e.g. Dosemeci et al., 1990) or if the classification errors depend on errors made in other variables (see, e.g. Chavance et al., 1992; Kristensen, 1992).

The effects of differential misclassification is unpredictable and the induced bias can be in any direction. Because of that, some investigators go through elaborated sampling designs to ensure that the misclassification will be non-differential. Despite that, data manipulations can bring back the problem. For instance, changes in the categorization of a misclassified variable may turn a non-differential misclassification into a differential one (see, e.g. Wacholder et al., 1991), and also if a non-differentially mismeasured continuous variable is dichotomized, differential error may be induced (see, e.g. Flegal et al., 1991).

1.5.2 Approaches to Correcting for Misclassification

Different approaches to correcting for misclassification processes have been developed to study the effect of misclassification errors and to protect the validity of data analyzes. Such approaches have been applied in a number of biostatistical contexts and analytic epidemiology. Specific scientific questions in these contexts may require inferences about (i) an underlying biological process, reflected in statistical associations that might be obscured or distorted by nuisance misclassification, (ii) the misclassification process itself, or (iii) both the underlying and misclassification processes.

The approaches to correcting for misclassification can be classified in two major groups: (i) the approaches that correct the estimators obtained without considering the misclassification process (naive estimators), and (ii) the approaches that estimate the parameters of interest based on the proposal of a full probability model for the observed and unobserved variables. Examples of the former includes the matrix method (Barron, 1977; Morrissey & Spiegelman, 1999), the inverse matrix method (Marshall, 1990), and the MC-SIMEX method (Küchenhoff et al., 2006).

The proposal of a model for correcting for misclassification includes three ingredients: (i) a model for the true (unobserved) values, which can be essentially any statistical model, (ii) a misclassification model, which involves specification of the relationship between the true and the observed values, and (iii) extra data, information or assumptions that may be needed to correct for measurement error. Using these ingredients, several authors have proposed model-based approaches for correcting for misclassification in regression settings for uncorrelated or correlated data under both differential and non-differential misclassification. We refer the

reader to Geng & Asano (1989), Evans et al. (1996), Magder & Hughes (1997), Neuhaus (1999), Paulino et al. (2003), Mwalili et al. (2005), Küchenhoff et al. (2006), McGlothlin et al. (2008), and references therein, for different approaches for the correction for misclassification in uncorrelated data contexts. Methods for correcting for misclassified correlated data have been proposed by Espeland et al. (1988), Espeland et al. (1989), Nagelkerke et al. (1990), Schmid et al. (1994), Singh & Rao (1995), Albert et al. (1997), Cook et al. (2000), Rekaya et al. (2001), Rosychuk & Thompson (2001), Neuhaus (2002), Rosychuk & Thompson (2003), Paulino et al. (2005), Rosychuk & Islam (2009) and Roy & Banerjee (2009).

In all of the previously described approaches a misclassification model needs to be assumed. The simplest misclassification model for misclassified binary data can be completely described through the misclassification probabilities

$$P\left(Y^* = j | Y = k\right) = \tau_{jk}, \qquad j, k \in \{0, 1\},$$

which may be located in a 2×2 matrix as follows:

$$\Pi = \begin{pmatrix} \tau_{00} & 1 - \tau_{11} \\ 1 - \tau_{00} & \tau_{11} \end{pmatrix},$$

where $\tau_{11} = P(Y^* = 1 | Y = 1)$ is called the *sensitivity* of the measuring instrument or examiner, and $\tau_{00} = P(Y^* = 0 | Y = 0)$ is the *specificity*. In other words, the sensitivity is the probability of testing positive when the disease is present, and the specificity corresponds to the probability of testing negative when the disease is absent.

In many practical situations, typically in cross-sectional studies, the available data (main data) contain no information about the misclassification parameters. Thus, to correct for misclassification, external information about these parameters is needed. The auxiliary data sources can be grouped into two main categories: *internal study*, i.e. a random subset of the primary data, and *external study*, which corresponds to an independent study. An internal validation data set is the ideal, because it can be used with all known analytical techniques, permits direct examination of the error structure and tests of critical error model assumptions, typically leads to much greater precision of estimation and inference, and has strong links to the well developed theory of missing data analysis (see, e.g. Little & Rubin, 1987). With external validation data, one must assume that the error structure in those data also applies to the primary data set. An external validation study is useful when there are a priori reasons to believe that misclassification is non-differential. For both validity and efficiency considerations, internal validation studies are preferred over external ones.

Within each of the previously described broad categories, there are three types of data: (i) *validation data*, in which the true (or latent) variable is observable together with its possible corrupted version, (ii) *replication data*, in which

replicates of the misclassified variables are available, and (iii) *instrumental data*, in which another variable, correlated with the true variable, is measured without error, jointly with the misclassified variable. Validation data are typically obtain by comparing the results from the classifier with the ones from a *gold standard*. A gold standard refers to a measuring instrument or examiner that is error-free. However, in practice many practical situations an infallible classifier may not exist or may be prohibitively expensive. Thus, the measurements are often made by what is called a *benchmark scorer*, an experienced examiner or a tested measuring instrument which is assumed to be error-free or is nearly so.

1.6 Aims of the Thesis

Motivated by the ST study, we propose and evaluate models to tackle the main challenges mentioned in Section 1.3.2. From a methodological point of view, the main objectives of this thesis are:

1. To evaluate two regression models for the analysis of correlated binary data. Specifically, we considered the multivariate probit model (Ashford & Sowden, 1970) and the conditionally specified logistic regression model (Joe & Liu, 1996) for the evaluation of risk factors and of the association structure of caries experience in the primary dentition. Special attention is given to the different interpretation of association structures arising from these models. These analyses, assuming error-free responses, are presented in Chapter 2.

2. To evaluate the effect of misclassification on the inferences about the association parameters for multivariate binary data. Although there is a rich literature on methods for correcting for misclassification in regression models for categorical data, the impact about the inference on model parameters has received relatively less attention and almost exclusively focused on the effect of the inferences on the mean structure. Because in the understanding of the etiology of caries experience it is of relevance to understand the association structure, we study the impact of different misclassification processes on the multivariate probit model (Ashford & Sowden, 1970) and the conditionally specified logistic regression model (Joe & Liu, 1996) in Chapter 3.

3. To propose and study the properties of models for univariate longitudinal binary data. In Chapter 4, we study whether the parameters associated to binary Markov models in which the response variable is subject to an unconstrained misclassification process and follows a progressive behavior, can be estimated without the need of external information on the misclassification parameters. We propose an extension of the simple version of the binary Markov model to describe the relationship between covariates and prevalence and incidence allowing for different classifiers.

4. To propose and study the properties of models for multivariate longitudinal binary data. In Chapter 5, we propose and evaluate the small sample properties of a multivariate binary inhomogeneous Markov model in which unobserved correlated response variables are subject to an unconstrained misclassification process and have a monotone behavior. The multivariate baseline distributions and Markov transition matrices of the unobserved processes are defined as a function of covariates, throughout the specification of compatible full conditional distributions. Distinct misclassification models are discussed, where the existence of different classifiers for each subject across time is taken into account.

References

Albert, J. H. (1992). Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation* 44 47–61.

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

Ashford, J. R. & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* 26 535–546.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* 81 767–775.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In H. Solomon, ed., *Studies in Item Analysis and Prediction*. Stanford, USA: Stanford University Press, 158–176.

Barron, B. A. (1977). Effects of misclassification on estimation of relative risk. *Biometrics* 33 414–418.

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

Buonaccorsi, J. P. (2010). *Measurement Error*. New York, USA: Chapman & Hall/CRC.

Carey, V., Zeger, S. L. & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 80 517–526.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London, UK: Chapman & Hall/CRC.

Chavance, M., Dellatolas, G. & Lellouch, J. (1992). Correlated nondifferential misclassifications of disease and exposure. *International Journal of Epidemiology* 21 537–546.

Chen, M. H. (2004). Skewed link models for categorical response data. In M. G. Genton, ed., *Skew-elliptical Distributions and their Applications: A Journey Beyond Normality*. New York, USA: Chapman & Hall/CRC, 131–151.

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine* 8 1095–1106.

Chen, M. H. & Dey, D. (1998). Bayesian modelling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhya* 60 322–343.

Chen, B., Yi, G. Y. & Cook, R. J. (2009). Likelihood analysis of joint marginal and conditional models for longitudinal categorical data. *The Canadian Journal of Statistics* 37 182–205.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85 347–361.

Connolly, M. A. & Liang, K. -Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* 75 501–506.

Cook, R. J., Ng, E. T. M. & Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* 56 1109–1117.

Copeland, K. T., Checkoway, H., McMichael, A. J. & Holbrook, R. H. (1977). Bias due to misclassification in estimation of relative risk. *American Journal of Epidemiology* 105 488–495.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* 21 113–120.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* 42 721–727.

Dalenius, T. (1977). Bibliography of non-sampling errors in surveys. *International Statistical Review* 45 71–89, 181–197, 303–317.

Desvarieux, M., Demmer, R. T., Jabocbs, D. R. Jr., Rundek, T., Boden-Albala, B., Sacco, R. L. & Papapanou, P. N. (2010). Periodontal bacteria and hypertension: the oral infections and vascular disease epidemiology study (INVEST) *Journal of Hypertension* 28(7) 1413–1421.

Diggle, P. J., Heagerty, P. J., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data (Second Edition)*. Oxford, UK: Oxford University Press.

Dosemeci, M., Wacholder, S. & Lubin, J. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology* 132 746–749.

Engle, R. F., Hendry, D. F. & Richard, J. F. (1983). Exogeneity. *Econometrica* 51 277–304.

Espeland, M. A., Murphy, W. C. & Leverett, D. H. (1988). Assessing diagnostic reliability and estimating incidence rates associated with a strictly progressive disease: dental caries. *Statistics in Medicine* 7 403–416.

Espeland, M. A., Platt, O. S. & Gallagher, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. *Journal of the American Statistical Association* 84(408) 972–979.

Evans, M., Guttman, I., Haitovsky, Y. & Swartz, T. (1996). Bayesian analysis of binary data subject to misclassification. In D. Berry, K. Chaloner & J. Geweke, eds., *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. New York, USA: North Holland, 67–771.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2009). *Longitudinal Data Analysis*. New York, USA: Chapman & Hall/CRC.

Flegal, K. M., Keyl, P. M. & Nieto, F. J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology* 134 1233–1244.

Fuller, W. A. (1987). *Measurement Error Models*. New York, USA: Wiley.

Fyffe, H. E., Deery, C., Nugent, Z. J., Nuttall, N. M. & Pitts, N. B. (2000). Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). *Community Dentistry and Oral Epidemiology* 28(1) 42–51.

Glass, R L. (1982). The first international conference on the declining prevalence of dental caries. *Journal of Dental Research* 61(Special Issue) 1304.

Geng, Z. & Asano, C. (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics* 8 2935–2954.

Hausen, H., Karkkainen, S. & Seppa, L. (2000). Application of the high-risk strategy to control dental caries. *Community Dentistry and Oral Epidemiology* 28 26–34.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 55(3) 688–698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for categorical longitudinal data. *Biometrics* 58 342–351.

Heagerty, P. J. & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science* 15 1–26.

Hofler, M. (2005). The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research* 14 92–101.

Joe, H. & Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.

Katz, B. M. & McSweeney, M. (1979). Misclassification errors and categorical data analysis. *Journal of Experimental Education* 47 331–338.

Korn E. L. & Whittemore, A. S. (1979). Methods for analyzing pane studies of acute health effects of air pollution. *Biometrics* 35 795–802.

Kristensen, P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 3 210–215.

Küchenhoff, H., Mwalili, S. M. & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 2 85–96.

Kuha, J. & Skinner, C. (1997). Categorical data analysis and misclassification. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin, eds., *Survey Measurement and Process Quality*. New York, USA: Wiley, 633–670.

Leroy, R., Bogaerts, K., Lesaffre, E. & Declerck, D. (2003). Impact of caries experience in the deciduous molars on the emergence of the successors. *European Journal of Oral Sciences* 111 1066–110.

Lesaffre, E. & Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine* 10 1391–1403.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73 13–22.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York, USA: Wiley.

Magder, L. S. & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 146 195–203.

Marshall, R. J. (1990). Validation-study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology* 43 941–947.

Marthaler, T. M., O'Mullane, D. M. & Vrbic, V. (1996). The prevalence of dental caries in Europe 1990-1995. *Caries Research* 30 237–255.

McGlothlin, A., Stamey, J. D. & Seaman, J. W. (2008). Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach. *Biometrical Journal* 50 123–134.

Molenberghs, G. & Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* 89 633–644.

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York, USA: Springer-Verlag.

Morrissey, M. J. & Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* 55 338–344.

Mote, V. L. & Anderson, R. L. (1965). An investigation of effect of misclassification on properties of χ^2-tests in analysis of categorical data. *Biometrika* 52 95–109.

Moynihan, P. (2000). Foods and factors that protect against dental caries. *Nutrition Bulletin* 25 281–286.

Mwalili, S. M., Lesaffre, E. & 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* 54(1) 77–93.

Nagelkerke, N. J. D., Chunge, R. N. & Kinot, S. N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 9 1211–1219.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* 135(3) 370–384.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86(4) 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* 58 675–683.

O'Brien, S. & Dunson, D. (2004). Bayesian multivariate logistic regression. *Biometrics* 60 739–746.

Paulino, C. D., Silva, G. & Achcar, J. A. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics and Data Analysis* 49 1120–1131.

Paulino, C. D., Soares, P. & Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics* 59 670–675.

Petersson, G. H. & Bratthall, D. (1996). The caries decline: a review of reviews. *European Journal of Oral Sciences* 104 436–443.

Pitts, N. B. (2004). Modern concepts of caries measurement. *Journal of Dental Research* 83(Suppl 1) C43–C47.

Pitts, N. B., Evans, D. J. & Pine, C. M. (1997). British association for the study of community dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. *Community Dental Health* 14(Suppl 1) 6–9.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44 1033–1048.

Rekaya, R., Weigel, K. A. & Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics* 57 1123–1129.

Robins, J. M., Greenland, S. & Hu, F. C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association* 94 687–712.

Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired data situations. *Biometrics* 40 1025–1035.

Rosychuk, R. J. & Islam, M. S. (2009). Parameter estimation in a model for misclassified Markov data - a Bayesian approach. *Computational Statistics and Data Analysis* 53 3805–3816.

Rosychuk, R. J. & Thompson, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 19 394–404.

Rosychuk, R. J. & Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine* 22 2035–2055.

Roy, S. & Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal* 51 420–432.

Rubin, T., Rosenbaum, J. & Cobb, S. (1956). The use of interview data for the detection of associations in field studies. *Journal of Chronic Diseases* 4 253–266.

Schmid, C. H., Segal, M. R. & Rosner, B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference* 42(1–2) 1–18.

Singh, A. C. & Rao, J. N. K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey. *Journal of the American Statistical Association* 90(430) 478–488.

Stern R. D. & Coe R. (1984). A model fitting analysis of daily rainfall data (with discussion). *Journal of the Royal Statistical Society, Series A* 147 1–34.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* 65(331) 1350–1361.

Vanobbergen, J., Martens, L., Lesaffre, E. & 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.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York USA: Springer-Verlag.

Wacholder, S., Dosemeci, M. & Lubin, J. H. (1991). Blind assignment of exposure does not prevent differential misclassification. *American Journal of Epidemiology* 134 433–437.

Wong, W. H. (1986). Theory of partial likelihood. *Annals of Statistics* 14 88–123.

World Health Organization (2003). *The World Oral Health Report 2003. Continuous improvement of oral health in the 21st century - the approach of the WHO Global Oral Health Programme*. Geneva, Switzerland: World Health Organization.

Zeger S. L., Liang K. -Y. & Self S. G. (1985). The analysis of binary longitudinal data with time-dependent covariates. *Biometrika* 72 31–38.

Zeger S. L. & Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* 44 1019–1031.

Zelen, M. & Haitovsky, Y. (1991). Testing hypotheses with binary data subject to misclassification errors: analysis and experimental design. *Biometrika* 78 857–865.

Zhao, L. P. & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* 77 642–648.

Part II

Multivariate Models for Possibly Misclassified Binary Data

Chapter 2

Conditional independence of multivariate binary data with an application in caries research

This chapter has been published as:

García-Zattera, M. J., Jara, A., Lesaffre, E. & Declerck, D. (2007). Conditional independence of multivariate binary data with an application in caries research. *Computational Statistics and Data Analysis* 51 3223–3234.

Abstract

For the analysis of caries experience in seven year old children the association between the presence or absence of caries experience among deciduous molars within each child is explored. Some of the high associations have an etiological basis (e.g., between symmetrically opponent molars), while others (diagonally opponent molars) are assumed to be the result of the transitivity of association and to disappear once conditioned on the caries experience status of the other deciduous molars, covariates and random effects. However, using discrete models for multivariate binary data, conditioning does not remove the diagonal association. When the association is explored on a latent scale, e.g. by a multivariate probit model, then conditional independence can be concluded. This contrast is confirmed when using other models on the (observed) binary scale and on the latent scale. Depending on the point of view, the differences in conditional independence might be seen as a consequence of different types of measurements or as a consequence of different models. An example shows that the results and conclusions can be markedly different with important consequences on model building. The explanation for this result is exemplified mathematically and illustrated using dental data from the Signal Tandmobiel® study.

Key Words: Conditional independence, Multivariate binary data, Latent variable representation, Multivariate probit model.

2.1 Introduction

In oral health research it is of interest to assess the association of caries experience (CE) among different teeth. The knowledge that caries development on one tooth is related to caries development on another tooth can help the dentists in optimizing their clinical examination of the patient and directs preventive and restorative approaches. Further, the exploration of CE patterns in the mouth can also help in further refining the understanding of the etiology of the disease. Indeed, it is still not established whether caries is a spatially local disease or not and the answer to that question might be related to a variety of factors determining caries activity (see, e.g. Hujoel et al., 1994, and references therein).

Based on data obtained in seven-year old children recruited in the Signal Tandmobiel® (ST) study, we examined the association between the presence/absence of CE on the eight deciduous molars and found a high association between symmetrically opponent molars, vertically opponent molars (maxilla versus mandible) and diagonally opponent molars. The first association is known and relatively easy to explain (Psoter et al., 2003). The second association is somewhat more difficult to understand. However, the high association between

diagonally opponent molars is believed to be the result of the (assumed) transitivity of the associations, i.e. due to the high association between symmetrically opponent molars and vertically opponent molars. This was verified by fitting a classical random effects logistic regression model (with subject as random effect) explaining the occurrence of CE on a deciduous molar by the CE on the other molars and subject specific characteristics. However, this model was not able to remove this high association, and the same was true for all other considered discrete models for multivariate binary vectors. In contrast, when the association was explored on a latent scale, say by a multivariate probit model (MPM), then the partial correlation matrix indicated conditional independence.

While we acknowledge that conclusions can change when different statistical models are used, we were initially surprised to see such a major difference when switching from the observed binary scale (one class of models) to the latent continuous scale (another class of models). A similar behavior would be observed if measurement error is added to the latent variable (see Section 2.2). In this paper we will highlight a possible reason why conditional independence is not invariant to the scale used for the analysis.

To illustrate the markedly different conclusions that can be obtained from different statistical models for multivariate binary responses, we analyzed the CE data with (a) the conditionally specified logistic regression model (CSLRM) as suggested by Joe & Liu (1996), and (b) the MPM (see, e.g., Ashford & Sowden 1970, Lesaffre & Molenberghs 1991 or Chib & Greenberg 1998). As for the log-linear model (LLM), the CSLRM acts on the observed binary scale, but the CSLRM allows the inclusion of covariates. Further, the CSLRM is intimately related to logistic regression. Namely, in the CSLRM, the association is measured by their odds ratio of a pair of binary responses conditional on the remaining binary responses and covariates. Consequently, the estimated odds ratios automatically express conditional (in)dependence. The MPM expresses the association between the binary responses via the correlation matrix of a multivariate normal latent random vector. Conditional (in)dependence can then be evaluated by the partial correlation matrix.

In Section 2.2, independence and conditional independence are reviewed. In Section 2.3, we briefly review the CSLRM and the MPM. An application to oral health data from the ST study is shown in Section 2.4. Finally, Section 2.5 gives some concluding remarks.

2.2 Independence and Conditional Independence

Suppose that $\boldsymbol{V}$ is a m-dimensional normally distributed random vector and that a random sample of n individuals is available yielding vectors $\boldsymbol{V}_i$ $(i = 1, \ldots, n)$.

However, we assume that $\boldsymbol{V}$ is not observed but is latent and that either $\boldsymbol{Y}$ or $\boldsymbol{Z}$ is observed. The random vector $\boldsymbol{Z}_i$ is continuous, namely $\boldsymbol{Z}_i = \boldsymbol{V}_i + \boldsymbol{\varepsilon}_i$ $(i = 1, \ldots, n)$, where $\boldsymbol{\varepsilon}_i$ is normally distributed and independent of $\boldsymbol{V}_i$. On the other hand, $\boldsymbol{Y}_i$ is a random multivariate binary response vector defined as $Y_{ij} = I(V_{ij} > c_j)$, where c_j $(j = 1, \ldots, m)$ are specific cut off points.

The correlation matrix $\boldsymbol{R} \equiv (\rho_{jk})_{jk}$ corresponding to $\boldsymbol{V}$ describes the association structure of the latent vector and conditional independence is seen from the elements of the partial correlation matrix $\boldsymbol{C} \equiv (c_{jk})_{jk}$, obtained from appropriately standardizing $\boldsymbol{R}^{-1}$. Namely, V_j is conditionally independent of V_k given the other V_m for $m \neq k, j$ when $c_{jk} = 0$. This property does not hold for other multivariate distributions and hence in these cases a partial correlation equal to zero does not automatically imply conditional independence.

In this paper we are interested in the relationship between the association structure on the latent scale (of $\boldsymbol{V}$) and that on the observed scale (of $\boldsymbol{Y}$ and $\boldsymbol{Z}$), especially with respect to conditional independence. Clearly, if $\boldsymbol{R}$ is the identity matrix, then also the components of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are statistically independent. Further, the association structure of $\boldsymbol{Z}$ depends on the magnitude of the measurement error component defined by $\boldsymbol{\varepsilon}$. For instance, even when the components of $\boldsymbol{V}$ are perfectly related, the components of $\boldsymbol{Z}$ could show a poor correlation if the variability of $\boldsymbol{\varepsilon}$ is quite high. Furthermore, conditional independence can not be expected for $\boldsymbol{Z}$ even when it holds for $\boldsymbol{V}$. For the binary case, $\rho_{jk} = 0$, $j \neq k$ implies independence of Y_j and Y_k. But again, conditional independence for $\boldsymbol{V}$ does not imply conditional independence for $\boldsymbol{Y}$ and this will be illustrated now.

Consider the random vector $\boldsymbol{V} \sim N_3(\boldsymbol{\mu}, \boldsymbol{R})$, with,

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} 1.00 & 0.64 & 0.80 \\ 0.64 & 1.00 & 0.80 \\ 0.80 & 0.80 & 1.00 \end{pmatrix},$$

and the categorical variables Y_j, $(j = 1, 2, 3)$ defined as above. The partial correlation matrix then becomes

$$\boldsymbol{C} = \begin{pmatrix} 1.00 & 0 & 0.62 \\ 0 & 1.00 & 0.62 \\ 0.62 & 0.62 & 1.00 \end{pmatrix}.$$

Since $c_{12} = 0$ the partial correlation coefficient $\rho_{V_1, V_2 . V_3} = 0$ and thus $V_1 \perp\!\!\!\perp V_2 | V_3$. However, the probability of Y_1 and Y_2 given Y_3 is,

$$P(Y_1 = 1, Y_2 = 1 | Y_3 = 1) = 0.6557,$$

while $P(Y_1 = 1 | Y_3 = 1) = P(Y_2 = 1 | Y_3 = 1) = 0.7952$, and,

$$P(Y_1 = 1 | Y_3 = 1) P(Y_2 = 1 | Y_3 = 1) = 0.6323.$$

Hence, we have shown numerically that conditional independence of variables V_1 and V_2 given V_3 does not imply $Y_1 \perp\!\!\!\perp Y_2|Y_3$. The evaluation of these expressions involves the computation of multivariate normal probabilities which was carried out using the methodology described in Genz (1992) and Genz (1993). A theoretical proof of this result is shown in Section A.1 of Appendix A.

The concept of conditional independence is a key notion in graphical models (see, e.g. Whittaker, 1990; Cox & Wermuth, 1996). Indeed, the key idea is to utilize the correspondence between separation in graphs and conditional independence in probability. Therefore, the results from this paper could have been derived from the theory of graphical models. Namely, our paper in fact deals with the result that $X \perp\!\!\!\perp Y|Z$ does not necessarily imply that $h(X) \perp\!\!\!\perp h(Y)|\ h(Z)$.

2.3 Two models for the analysis of multivariate binary responses

In this section we will describe two models that were used to illustrate the difference between analyzing the multivariate binary response on the observed binary scale and on the latent continuous scale. But, the contrast remains when these models are replaced by other similar models, as indicated below.

2.3.1 The Conditionally Specified Logistic Regression Model: a model on the observed binary scale

Let $\boldsymbol{Y}_i$ be defined as before and let $\boldsymbol{x}_{ij}$ be the corresponding covariate vector. Joe & Liu (1996), suggested a model for multivariate binary responses with covariates. The conditional distribution of each binary response Y_{ij} given the other binary responses $Y_{ik} = y_{ik},\ k \neq j$ and the covariates $\boldsymbol{x}_{ij}$ is equivalent to a logistic regression with parameter vector $\boldsymbol{\beta}_j$ and parameters $\gamma_{jk},\ k \neq j$. That is, for $j = 1, ..., m$,

$$\text{logit } P(Y_{ij} = 1|Y_{ik} = y_{ik}, k \neq j, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j + \sum_{k \neq j} \gamma_{jk} y_{ik}. \tag{2.1}$$

Joe & Liu (1996) showed that a necessary and sufficient condition for compatibility of conditional distributions is that $\gamma_{jk} = \gamma_{kj},\ j \neq k$, and that the joint distribution is given by,

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{n} \left[c(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})^{-1} \exp \left\{ \sum_{j=1}^{m} \left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j \right) y_{ij} + \sum_{1 \leq j < k \leq m} \gamma_{jk}\ y_{ij} y_{ik} \right\} \right], \tag{2.2}$$

with normalizing constant,

$$c(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{y_1=0}^{1} \cdots \sum_{y_m=0}^{1} \exp\left\{ \sum_{j=1}^{m} \left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j\right) y_j + \sum_{1 \le j < k \le m} \gamma_{jk}\, y_j y_k \right\}. \quad (2.3)$$

In (2.1) to (2.3), the parameters γ_{jk} are interpreted as conditional log-odds ratios, since,

$$\begin{aligned} \exp\{\gamma_{jk}\} \quad = \quad & \frac{P(Y_{ij}=1, Y_{ik}=1 | \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}{P(Y_{ij}=1, Y_{ik}=0 | \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)} \times \\ & \frac{P(Y_{ij}=0, Y_{ik}=0 | \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}{P(Y_{ij}=0, Y_{ik}=1 | \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}. \end{aligned}$$

Note that for $m = 2$, there are no Y_{il}'s so that γ_{12} is also the unconditional log-odds ratio and it is constant over the covariates. For $m \geq 3$, it is straightforward to show that the bivariate marginal distributions from (2.2), and the log-odds ratios depend on the covariates. Note also that the exponential family in (2.2) is not closed under marginalization and can be easily extended if interaction terms are needed.

In the absence of covariates, it is popular to analyze conditional independence on the observed binary scale with a log-linear model. For Y_1, Y_2 and Y_3 a LLM up to two-way interactions is given by

$$\log\left(\mu_{jkl}\right) = \lambda + \lambda_j^{Y_1} + \lambda_k^{Y_2} + \lambda_l^{Y_3} + \lambda_{jk}^{Y_1 Y_2} + \lambda_{jl}^{Y_1 Y_3} + \lambda_{kl}^{Y_2 Y_3},$$

where λ is the overall mean of the natural logarithm of the expected frequencies, $\lambda_j^{Y_1}$, $\lambda_k^{Y_2}$, $\lambda_l^{Y_3}$ represent the main effects for variables Y_1, Y_2 and Y_3, respectively; and $\lambda_{jk}^{Y_1 Y_2}$, $\lambda_{jl}^{Y_1 Y_3}$, and $\lambda_{kl}^{Y_2 Y_3}$ represent the respective interaction effects. In this case, the null hypothesis of conditional independence between two variables given the other one, for instance Y_1 and Y_2 given Y_3, is $H_0 : \lambda_{jk}^{Y_1 Y_2} = 0, \ \forall \ j,k$. We applied also the LLM to the dental example but with basically the same results.

An R-program (R Development Core Team, 2004), calling FORTRAN subroutines, was written for the analysis of the multivariate binary data with the CSLRM (`cslogistic`) using a likelihood and a Bayesian approach. The program `cslogistic` is available from the Comprehensive R Archive Network (CRAN) or upon request to the authors.

2.3.2 The Multivariate Probit Model: a model on the latent continuous scale

A commonly used alternative modelling strategy for multivariate binary (or ordinal) data involves the introduction of latent variables, i.e. by considering the binary variables as a discretized continuous variables. Indeed, the key idea is to introduce an m-dimensional latent variable vector $\boldsymbol{V}_i = (V_{i1}, ..., V_{im})$ such that

$$Y_{ij} = I(V_{ij} > 0),$$

with $j = 1, ..., m$. A common distributional assumption, leading to the MPM, is $\boldsymbol{V}_i \sim N_m(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{R})$, where $\boldsymbol{X}_i$ is a matrix of covariates associated to the regression parameters vector $\boldsymbol{\beta}$ and, for identifiability reasons the matrix $\boldsymbol{R}$ must be in correlation form (Chib & Greenberg, 1998). The correlations $\rho_{jk} = \text{corr}(V_j, V_k)$ are known as the *tetrachoric correlation coefficients*.

Likelihood and Bayesian analyzes were performed. Maximum likelihood estimates were obtained using the SAS procedure QLIM (version 9.1). For the Bayesian analysis, noninformative prior distributions were given for all parameters of the model. Posterior distributions of the parameters were estimated using Markov Chain Monte Carlo techniques and the Metropolized hit-and-run algorithm proposed by Chen & Schmeiser (1993) was used to generate correlation matrices. The Markov chain was initialized with all the regression coefficients, except the intercepts, equal to zero. The first 10 000 samples were discarded as burn-in and an additional 400 000 iterations were used to compute posterior summaries (posterior mean and 95% highest posterior density (95% HPD) credible intervals using the method of Chen & Shao, 1999). Convergence was checked using standard criteria (Cowles & Carlin, 1996) as implemented in the BOA package (Smith, 2005).

The Bayesian Multivariate Logistic Model (MLM) of O'Brien & Dunson (2004) was also fitted. Since the posterior distribution of regression coefficients, marginal and partial correlation coefficients were basically the same as for the MPM, they are not shown. However, it is important to note that in the MLM framework a zero partial correlation does not imply conditional independence.

2.4 Analysis of the Oral Health Example

2.4.1 The Oral Health Question

The ST study is a longitudinal prospective oral health screening study conducted in Flanders (Belgium) between 1996 and 2001. For this project, 4468 children were examined on a yearly basis during their primary school time by one of sixteen trained and calibrated dental examiners. Data on oral hygiene and dietary habits

were obtained through structured questionnaires, completed by the parents. For a more detailed description of the ST study we refer to Vanobbergen et al. (2000).

Based on the first year oral health data, we examined the association pattern of CE in the mouth. It is well known that a strong association between neighboring teeth exists. However, it is also of interest to know whether other relationships exist.

Here, CE of the 8 deciduous molars was analyzed using a CSLRM and a MPM. For ease of exposition, the European notation to indicate the location of a deciduous tooth in the mouth is shown in Figure 2.1. Covariates included in the models were age (in years; `Age`), gender (boys versus girls; `Gender`), age at start of brushing (in years; `Startbr`), regular use of fluoridated supplements (yes versus no; `Sysfl`), daily use of sugar containing drinks (no versus yes; `Drinks`), number of between-meal snacks (two or less than two a day versus more than two a day; `Meals`) and frequency of tooth brushing (once or more a day versus less than once a day; `Freqbrus`). Except for the intercept, it was assumed that the covariates have a common effect on the probabilities of CE for all teeth molars.

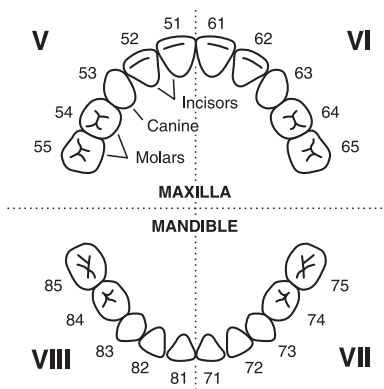


Figure 2.1: European notation to indicate the location of the deciduous teeth in the mouth.

Table 2.1 (see page 35) shows the unconditional odds ratios (95% confidence interval; 95% CI) expressing the association of CE in the eight deciduous molars. The table shows that adjacent molars (e.g., 54 and 55), homologous molars (e.g., 54 and 64) and vertically opponent molars (e.g., 54 and 84) have a high association. However, also the association between diagonally opponent molars (e.g., 54 and 74) seems to be high. Observe that in this analysis no correction for covariates was made nor did we take into account the CE pattern of other molars in the mouth. The dentists speculated that the high association between the diagonally opponent molars was due to the high association between the homologous molars and the

high association between the opponent molars. It was hoped that a conditional analysis could demonstrate this.

Table 2.1: Signal-Tandmobiel® Study: unconditional odds ratios (95% confidence interval) for caries experience in deciduous molars.

Tooth	Tooth						
	64	**74**	**84**	**55**	**65**	**75**	**85**
54	16.54 (13.40 ; 20.34)	8.59 (7.08 ; 10.43)	7.87 (6.49 ; 9.55)	11.00 (9.03 ; 13.41)	7.08 (5.85 ; 8.56)	5.68 (4.71 ; 6.85)	5.60 (4.65 ; 6.75)
64		8.33 (6.89 ; 10.07)	7.48 (6.20 ; 9.03)	7.05 (5.85 ; 8.50)	11.84 (9.74 ; 14.41)	5.29 (4.40 ; 6.35)	5.22 (4.35 ; 6.26)
74			24.18 (19.88;29.40)	6.64 (5.58 ; 7.91)	6.19 (5.20 ; 7.36)	9.46 (7.93 ; 11.29)	7.58 (6.38 ; 9.01)
84				6.48 (5.44 ; 7.71)	6.46 (5.43 ; 7.68)	8.27 (6.95 ; 9.84)	8.88 (7.46 ; 10.58)
55					14.69 (12.12 ; 17.79)	8.89 (7.42 ; 10.65)	8.61 (7.19 ; 10.31)
65						7.79 (6.52 ; 9.30)	8.13 (6.80 ; 9.72)
75							20.31 (16.70 ; 24.70)

Since the results obtained in the likelihood and Bayesian approaches were the same for both models, we have opted for the Bayesian solution.

2.4.2 Conditionally Specified Logistic Regression

Table 2.2 presents the posterior summaries of the regression coefficients of the CSLRM. The results indicate clear differences in CE with respect to age of the child, age at start of brushing, regular use of fluoridated supplements, daily use of sugar containing drinks and number of between-meal snacks. The posterior

Table 2.2: Signal-Tandmobiel® Study: posterior means and 95% highest posterior density (95% HPD) credible intervals of regression coefficients obtained from the conditionally specified logistic regression model for caries experience in eight deciduous molars.

Covariate	**Estimate**	**95% HPD**
Age (years)	0.075	(0.061 ; 0.089)
Gender (girls)	0.015	(-0.014 ; 0.042)
Startbr (years)	0.033	(0.020 ; 0.046)
Sysfl (no)	0.103	(0.071 ; 0.134)
Drinks (yes)	0.099	(0.065 ; 0.129)
Meals (> 2/day)	0.044	(0.016 ; 0.078)
Freqbrus (< 1/day)	0.015	(-0.033 ; 0.057)

summaries of the conditional odds ratios for CE in deciduous molars are shown

in Table 2.3. The table shows that adjacent, homologous and vertically opponent molars have a high association. However, all the associations between diagonally opponent molars remained highly positive and significant.

Table 2.3: Signal-Tandmobiel® Study: posterior means (95% highest posterior density credible intervals) of conditional odds ratios for caries experience in deciduous molars, obtained from the conditionally specified logistic regression model.

Tooth	Tooth						
	64	**74**	**84**	**55**	**65**	**75**	**85**
54	5.80 (4.06 ; 7.52)	1.90 (1.54 ; 2.37)	2.02 (1.75 ; 2.23)	5.42 (4.11 ; 6.41)	0.88 (0.67 ; 1.04)	1.05 (0.83 ; 1.26)	1.08 (0.94 ; 1.24)
64		2.07 (1.75 ; 2.54)	1.61 (1.27 ; 1.89)	0.79 (0.68 ; 0.95)	6.34 (5.45 ; 7.39)	1.12 (0.93 ; 1.37)	1.00 (0.84 ; 1.18)
74			10.68 (9.35 ; 12.28)	1.37 (1.13 ; 1.67)	1.05 (0.81 ; 1.24)	3.29 (2.80 ; 4.01)	1.23 (0.96 ; 1.41)
84				1.15 (0.99 ; 1.30)	1.46 (1.27 ; 1.71)	1.51 (1.33 ; 1.68)	2.49 (2.21 ; 2.79)
55					6.71 (5.28 ; 8.11)	1.84 (1.57 ; 2.19)	2.36 (2.04 ; 2.71)
65						1.97 (1.64 ; 2.33)	2.07 (1.61 ; 2.38)
75							9.61 (8.19 ; 11.07)

2.4.3 Multivariate Probit Model

In the MPM model, associations of CE in the mouth were high and significant for symmetrical and vertically opponent molars but also important for diagonally opponent molars, see Table 2.4 (page 37). The analysis revealed that all correlation coefficients were significant and considerably high. The posterior estimate of the correlation matrix indicates that the equicorrelation assumption on the correlation structure is not valid. For example, the 95% HPD intervals for the tetrachoric correlation coefficient between tooth 55 and 64, and tooth 54 and 75 are (0.74 ; 0.81) and (0.50 ; 0.59), respectively.

The posterior summaries of the regression coefficients in the model showed basically the same results as for the CSLRM and hence not shown.

From the estimated correlations one can calculate the partial correlation matrix. Here all partial correlations are smaller than the corresponding correlations, but the difference was the largest for the diagonally opponent molars. For instance, Figure 2.2 (see page 37) shows the posterior distributions of tetrachoric correlations and partial correlations for the homologous pair 54 and 64 (panel a), and the diagonal opponent pair 54 and 74 (panel b), respectively. Clearly, the largest difference between the two correlation coefficients is seen for molars 54 and 74.

Table 2.4: Signal-Tandmobiel® Study: posterior means (95% highest posterior density credible intervals) of latent marginal correlation matrix for caries experience, obtained from the multivariate probit model.

Tooth	Tooth						
	64	**74**	**84**	**55**	**65**	**75**	**85**
54	0.78 (0.74 ; 0.81)	0.65 (0.62 ; 0.69)	0.62 (0.58 ; 0.66)	0.70 (0.65 ; 0.74)	0.61 (0.56 ; 0.66)	0.55 (0.50 ; 0.59)	0.55 (0.49 ; 0.62)
64		0.64 (0.60 ; 0.68)	0.61 (0.56 ; 0.65)	0.61 (0.56 ; 0.66)	0.72 (0.68 ; 0.76)	0.53 (0.48 ; 0.58)	0.53 (0.46 ; 0.58)
74			0.85 (0.82 ; 0.87)	0.61 (0.56 ; 0.66)	0.60 (0.55 ; 0.64)	0.69 (0.64 ; 0.73)	0.64 (0.59 ; 0.69)
84				0.60 (0.55 ; 0.65)	0.60 (0.56 ; 0.64)	0.66 (0.61 ; 0.70)	0.67 (0.62 ; 0.71)
55					0.77 (0.73 ; 0.81)	0.67 (0.63 ; 0.71)	0.67 (0.62 ; 0.72)
65						0.64 (0.60 ; 0.68)	0.66 (0.60 ; 0.70)
75							0.82 (0.79 ; 0.85)

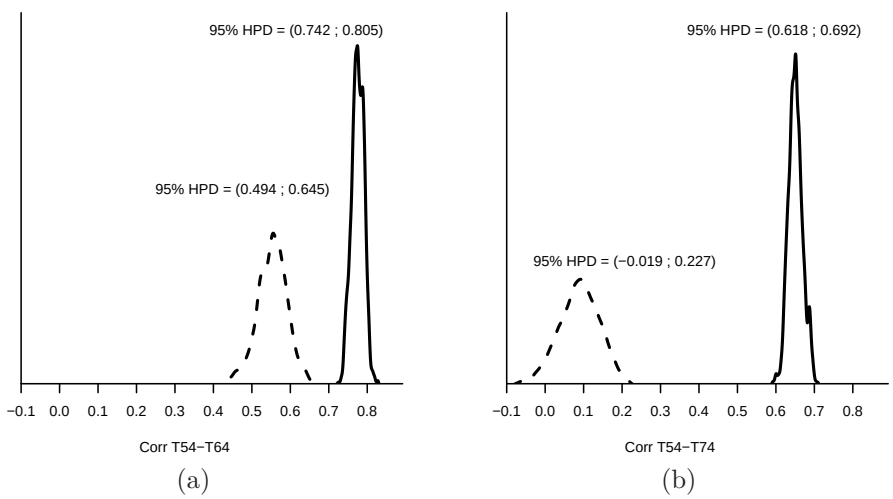


Figure 2.2: Signal-Tandmobiel® Study: tetrachoric correlation coefficients for caries experience in molars 54 and 64 (panel a) and in molars 54 and 74 (panel b). The marginal and the partial correlations are shown in solid and dashed lines, respectively

Table 2.5 presents posterior summaries of the whole partial correlation matrix. A zero entry in this matrix corresponds to conditional independence between corresponding latent variables. While all the associations between neighboring and symmetrical molars remained highly positive and significant, the association between opponent and diagonally opponent molars were in most of the cases not significant, suggesting that the highly observed marginal association could be explained to a large extent by the transitivity of the correlation structure. Further, compared to the results of the CSLRM, a total of ten discordant results were found. In eight of these cases, the conditional odds ratios are significant while the partial correlations are not.

Table 2.5: Signal-Tandmobiel® Study: posterior means (95% highest posterior density credible intervals) of latent partial correlation matrix for caries experience, obtained from the multivariate probit model.

Tooth	Tooth						
	64	**74**	**84**	**55**	**65**	**75**	**85**
54	0.57 (0.49 ; 0.65)	0.10 (-0.02 ; 0.23)	0.06 (-0.07 ; 0.18)	0.42 (0.33 ; 0.51)	-0.23 (-0.37 ; -0.13)	-0.03 (-0.16 ; 0.09)	0.02 (-0.09 ; 0.15)
64		0.14 (0.02 ; 0.26)	0.01 (-0.11 ; 0.14)	-0.21 (-0.32 ; -0.10)	0.49 (0.41 ; 0.58)	-0.03 (-0.16 ; 0.09)	-0.04 (-0.17 ; 0.08)
74			0.65 (0.60 ; 0.70)	0.03 (-0.08 ; 0.14)	-0.05 (-0.17 ; 0.06)	0.26 (0.16 ; 0.36)	-0.08 (-0.19 ; 0.02)
84				-0.01 (-0.13 ; 0.09)	0.06 (-0.05 ; 0.18)	-0.05 (-0.16 ; 0.06)	0.22 (0.11 ; 0.32)
55					0.50 (0.42 ; 0.58)	0.14 (0.03 ; 0.26)	0.10 (-0.03 ; 0.21)
65						0.07 (-0.06 ; 0.19)	0.13 (0.02 ; 0.26)
75							0.58 (0.52 ; 0.65)

2.4.4 Model Comparison

To exclude that our conclusion of conditional independence in the MLM is due to a badly fitted model, we compared the goodness-of-fit of both models. As measures we used Akaike's information criteria (AIC) for the frequentist approaches. For the Bayesian approaches we used the Bayesian information criteria (BIC), conditional predictive ordinates (CPO) and pseudo Bayes factors (PsBF). In Table 2.6 (see page 39), we see that MPM performed slightly better than the CSLRM for AIC and BIC. The cross validation model comparison criteria showed basically the same results as information criteria. In Figure 2.3 (see page 39) we present a scatter plot of the CPO for the MPM versus CSLRM. Clearly, the MPM was better than the CSLRM and the PsBF confirmed this, namely the value of $2 \times \log_{10} PsBF$ for MPM versus CSLRM was 115.50.

Table 2.6: Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) for the conditionally specified logistic regression model (CSLRM) and the multivariate probit model (MPM).

	AIC	**BIC**
CSLRM	20471.36	20772.09
MPM	20346.00	20606.73

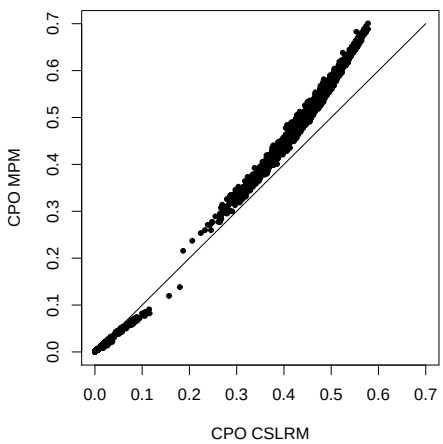


Figure 2.3: Scatter plot of the conditional predictive ordinates (CPO) for the multivariate probit model (MPM) and conditionally specified logistic regression model (CSLRM).

In order to compare the conditional independence structure of the binary random vector implied by the fitted MPM reporting conditional independence of the underlying latent vector, the conditional odds ratios were calculated. Table 2.7 (see page 40) shows the posterior summaries of the conditional odds ratios derived from the MPM.

The conditional odds ratios derived from the MPM gave the same conclusions of the results obtained with the CSLRM shown in Table 2.3 (see page 36). Again, the importance of this finding is that although we base our analysis on the better model, conditional independence on the latent continuous scale is not transferable to the observed binary scale.

Finally, based on the symmetry on the association structure observed in Table 2.4 (see page 37), we fitted the MPM under equality constraints in the correlation matrix. The conclusions regarding the association structure remained.

Table 2.7: Signal-Tandmobiel® Study: posterior means (95% highest posterior density credible intervals) of conditional odds ratios for caries experience in deciduous molars, based on the results of the multivariate probit model.

Tooth	Tooth						
	64	**74**	**84**	**55**	**65**	**75**	**85**
54	5.58 (4.23 ; 6.88)	1.92 (1.41 ; 2.45)	1.62 (1.19 ; 2.03)	3.37 (2.53 ; 4.20)	0.90 (0.64 ; 1.61)	1.15 (0.88 ; 1.44)	1.24 (0.92 ; 1.55)
64		1.92 (1.42 ; 2.42)	1.48 (1.06 ; 1.88)	0.95 (0.71 ; 1.22)	4.11 (3.22 ; 5.11)	1.07 (0.82 ; 1.34)	1.06 (0.80 ; 1.34)
74			9.64 (7.45 ; 12.05)	1.36 (1.02 ; 1.71)	1.14 (0.82 ; 1.45)	2.76 (2.04 ; 3.46)	1.33 (0.98 ; 1.68)
84				1.29 (0.94 ; 1.63)	1.46 (1.09 ; 1.85)	1.57 (1.14 ; 2.02)	2.36 (1.81 ; 2.97)
55					4.85 (3.63 ; 6.04)	2.08 (1.60 ; 2.64)	1.97 (1.46 ; 2.49)
65						1.71 (1.28 ; 2.16)	2.02 (1.52 ; 2.59)
75							6.97 (5.44 ; 8.44)

2.5 Concluding Remarks

Conditional independence is regarded as a fundamental concept not only in the theory of statistical inference (see, e.g. Dawid, 1979; Nogales et al., 2000), but also in structural modelling (Pearl, 1995). Model building most often deals with structural properties underlying a process generating latent as well as observed variables.

The MPM represents one of the strategies for the analysis of clustered multivariate binary data, which is described in terms of a correlated Gaussian distribution for underlying latent variables that are manifested as discrete variables through a threshold specification. Although latent variable modelling could be viewed as a dubious exercise fraught with unverifiable assumptions and naive inferences regarding causality, the MPM is a natural way of relating stimulus and response where such an interpretation for a threshold approach is readily available; examples include attitude measurement, assigning pass/fail gradings for examinations based on mark cut-off, and bioassay settings were the underlying continuous scale can be a lethal dose of a drug.

On the other hand, from a formal point of view a statistical model is defined as a family of probability distributions on a sample space, i.e., explains the observed data (see, e.g. McCullagh, 2002). In this context the latent variable representation is only a convenient stochastic representation of the statistical model. Unfortunately, whether this is a comparison of models or types of measurements, is an unverifiable hypothesis. An interesting discussion about the difference between true variables measured with error and the latent variables, can be found in Skrondal & Rabe-Hesketh (2004).

We showed that the association structure on the latent continuous scale is not transferable to the observed binary scale. In particular that conditional independence on the latent scale does not transfer to the observed scale. Which of the two scales provide us with the answer will depend on the problem and biological evidence there is. The important issue of this paper is to show that the two analyses can and often will yield two different interpretations. With regard to our oral health example on CE we conclude that, while there will always be an apparent relationship between diagonally opponent molars, they are indeed (conditionally) independent for CE. The basis for this conclusion is: (a) our findings with the MPM and (b) the absence of a biological explanation for a direct association of CE in diagonally opponent teeth. There is further dental evidence for our conclusion. Indeed, Veerkamp & Weerheijm (1995) pointed out that CE also very much depends on the eruption stage. Namely, that caries can only develop when the respective tooth has been exposed long enough. Now teeth in the maxilla emerge earlier than teeth in the mandible. Hence, symmetrically opponent molars have about the same emergence time while opponent and diagonally opponent teeth emerge at different ages providing extra evidence that these associations are not etiological.

Our findings are also of importance in model building exercises in general. In fact, the decision to increase the complexity of the model depends on whether the extra variate has a (significant) relationship with a particular response, conditional on the already included covariates and the remaining responses. In this context, Webb & Forster (2004) suggested a MPM, characterized by the structure of the inverse correlation matrix of the latent variables. Their model building exercise was based on tests for conditional dependence on the latent scale while the interpretations were done on the observed binary scale. Hence if their analyses were done on the observed scale, quite different models could have been obtained implying a quite different interpretation.

Finally, it is worth mentioning that a similar phenomenon will occur when the actual data are continuous but discretized for the sake of the analysis, a practice that is often seen in medical papers. For the same reason as pointed out above, markedly different conclusions might be drawn from the analysis on the continuous scale and the analysis on the discretized scale.

Acknowledgements

The first author is supported by the Research Grant OT/00/60, Catholic University Leuven, and by the Research Grant OT/00/35, Catholic University Leuven, which also supported the second author. The authors also acknowledge the partial support from the Interuniversity Attraction Poles Program P5/24 − Belgian State − Federal Office for Scientific, Technical and Cultural Affairs. Data

collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

Ashford, J. R. & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* 26 535–546.

Chen, M. H. & Schmeiser, B. W. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *Journal of Computational and Graphical Statistics* 2 251–272.

Chen, M. H. & Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 8 (1) 69–92.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85 347–361.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative study. *Journal of the American Statistical Association* 91 883–904.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies. Models, Analysis and Interpretations.* London, UK: Chapman & Hall/CRC

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B* 41 1–31.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* 1 141–150.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics* 25 400–405.

Hujoel, P. P., Lamont, R. J., DeRouen, T. A., Davis, S. & Lerouxi, B. G. (1994). Within-subject coronal caries distribution patterns: an evaluation of randomness with respect to the midline. *Journal of Dental Research* 73 (9) 1575–1580.

Joe, H. & Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.

Lesaffre, E. & Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine* 10 1391–1403.

McCullagh, P. (2002). What is a statistical model? (with discussion). *Annals of Statistics* 30 1225–1310.

Nogales, A. G., Oyola, J. A., & Pérez, P. (2000). On conditional independence and the relationship between sufficiency and invariance under the Bayesian point of view. *Statistics and Probability Letters* 46 75–84.

O'Brien, S. & Dunson, D. (2004). Bayesian multivariate logistic regression. *Biometrics* 60 739–746.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* 82 669–710.

Psoter, W. J., Zhang, H., Pendrys, D. G., Morse, D. E. & Mayne, S. T. (2003). Classification of dental caries patterns in the primary dentition: a multidimensional scaling analysis. *Community Dent Oral Epidemiol* 31 231–238.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal, an Structural Equation Models*. New York, USA: Chapman & Hall/CRC

Smith, B. J. (2005). *Bayesian Output Analysis Program (BOA) for MCMC*. College of Public Health, University of Iowa, Iowa, USA. URL `http://www.public-health.uiowa.edu/boa`.

Vanobbergen, J., Martens, L., Lesaffre, E. & 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.

Veerkamp, J. S. & Weerheijm, K. L. (1995). Nursing-bottle caries: the importance of a development perspective. *Journal of Dentistry for Children* 62(6) 381–386.

Webb, E. L. & Forster, J. J. (2004). Bayesian model determination for multivariate ordinal and binary data. Tech. rep., University of Southampton, School of Mathematics.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York, USA: Wiley.

Chapter 3

Effect of Misclassification on the Association Parameters of Multivariate Binary Data

This draft is under preparation for publication:

García-Zattera, M. J., Marshall, G. & Lesaffre, E. (2010). Effect of misclassification on the association parameters of multivariate binary data. *In preparation*.

Abstract

We evaluate the small sampling behavior of naive maximum likelihood estimators of the association parameters of multivariate models for correlated binary responses, when the responses are subject to an unconstraint misclassification process. The small sample properties of the estimators of the association parameters of multivariate probit and conditionally specified logistic regression models, are evaluated using Monte Carlo simulation. Differential and non-differential misclassification processes are considered in the simulation scenarios. The results indicate that naive estimators are strongly biased towards the null of no association, regardless the type of misclassification process.

Key Words: Misclassification; Multivariate Binary Data; Association Parameters.

3.1 Introduction

Standard approaches to analyze correlated binary data are often based on the assumption that there are no classification errors in the observations. However, in many research areas, the observed status may not perfectly reflect the true state of a subject due, for instance, to limitations of experience/knowledge of the examiners, or to imperfect diagnostic instruments or procedures. The problem is especially important in health sciences where misdiagnoses occur when sick individuals are diagnosed as healthy or vice versa, or when the severity of the disease is misjudged. The evaluation of caries experience (CE), typically defined as a binary variable indicating whether a tooth is decayed, missing or filled due to caries, is an oral health example where misclassification occurs. The diagnosis of CE is not an easy task for a variety of reasons, including the existence of high quality composite materials, location of the cavity, discolorations, among many others (see, e.g. García-Zattera et al., 2010).

The effect of misclassification on the statistical inference has been widely investigated in the literature. An early reference is Bross (1954) who discusses the biases caused by misclassification in binary contingency tables. He shows that when misclassification is ignored, the estimated difference between two proportions is biased toward the null of no difference, the significance level of an hypothesis test is correct if both populations have the same misclassification probabilities (non-differential or covariate independent), but its power is reduced. Tests about the difference between proportions are further discussed by Rubin et al. (1956), Katz & McSweeney (1979), and Zelen & Haitovsky (1991) for the binary case, and Mote & Anderson (1965) for the multinomial case. Gladen & Rogan (1979) show that the power of tests about relative risk is reduced when data are affected

by misclassification. Schwartz (1985) studies the bias of the naive estimator of a single probability and shows how the misclassification probabilities affect the coverage probability of conventional confidence intervals.

Several approaches have been proposed in the literature to correct for misclassification. These can be classified into the approaches that correct the naive estimators and the approaches that estimate the parameters of interest based on the proposal of a full probability model for the true and error-prone variables. Examples of the former, in the context of contingency tables, include the matrix method (Barron, 1977; Morrissey & Spiegelman, 1999) and the inverse matrix method (Marshall, 1990). By assuming that the misclassification probabilities are known, Barron (1977) proposes a correction method, known as the matrix method (Morrissey & Spiegelman, 1999) or indirect method (Marshall, 1990), to obtain unbiased estimators from misclassified 2×2 contingency table data. Marshall (1990) compares the relative efficiency of the direct method (also known as reclassification or inverse matrix method) with the indirect method and shows that the direct method is more efficient than the indirect method. Greenland (1988) derives the variance of corrected estimators when the misclassification probabilities are estimated from external or internal validation data.

With some exceptions, the development of model-based approaches for correcting for misclassification, requires information from external sources about the misclassification parameters. Tenenbein (1970, 1971) proposes double sampling (i.e. internal validation data) to obtain the maximum likelihood estimator (MLE) and its asymptotic variance for misclassified binomial data. Espeland & Hui (1987) demonstrate how to model misclassified data with validation data as an incomplete data problem using a log-linear model.

Several authors have extended the previous approaches to regression settings for uncorrelated or correlated data. We refer the reader to Geng & Asano (1989), Magder & Hughes (1997), Neuhaus (1999), Paulino et al. (2003), Mwalili et al. (2005), McGlothlin et al. (2008), and references therein, for different approaches for the correction for misclassification in uncorrelated data contexts. Methods for correcting for misclassified correlated data have been proposed by Espeland et al. (1988), Espeland et al. (1989), Nagelkerke et al. (1990), Schmid et al. (1994), Singh & Rao (1995), Albert et al. (1997), Cook et al. (2000), Rekaya et al. (2001), Rosychuk & Thompson (2001), Neuhaus (2002), Rosychuk & Thompson (2003), Paulino et al. (2005), Rosychuk & Islam (2009), Roy & Banerjee (2009) and García-Zattera et al. (2010).

Compared to the rich literature on methods for correcting for misclassification in regression models for categorical data, the impact about the inference on model parameters has received relatively less attention and almost exclusively focused on the effect of the inferences on the mean structure (i.e., regression coefficients). Neuhaus (1999) examines the magnitude of bias and efficiency loss

due to misclassification in binary regression with a single covariate and obtained some approximate bias-correction factor for regression parameter. Neuhaus (2002) studies the influence of response misclassification in generalized linear mixed models for the analysis of data from clustered and longitudinal studies. Although the association parameters can be considered as nuisance parameters in many correlated data problems, there are also many instances where understanding the association structure is just as central as understanding the mean structure to the proper solution of the scientific problems. For instance, in oral health research it is of interest to assess the association of caries experience among different teeth. The knowledge that caries development on one tooth is related to caries development on another tooth can help the dentists in optimizing their clinical examination of the patient and direct preventive and restorative approaches. Further, the exploration of caries experience patterns in the mouth can also help in further refining the understanding of the etiology of the disease.

In this paper, we study the impact of differential and non-differential misclassification in response variables on the association parameters of some common models for the analysis of multivariate binary data. Specifically, we investigate the small sample behavior, using Monte Carlo simulation, of naive MLE for the association parameters associated to the multivariate probit and conditionally specified logistic regression models, when the misclassification of the response is ignored. The paper is organized as follows. Section 3.2 discusses the models under consideration. Although most of the material in this section is not original, the discussion is necessary for the sake of completeness. The evaluation of the properties of the naive estimators under both misclassification types are presented in Section 3.3. A final discussion section concludes the article.

3.2 Two Regression Models for Multivariate Binary Data

Assume that for each of I experimental units and J variables of interest, the regression data $(Y_{ij}, \boldsymbol{x}'_{ij})$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ is recorded, where $Y_{ij} \in \{0, 1\}$ is the response variable of interest and $\boldsymbol{x}_{ij} \in \mathbb{R}^p$ is a p-dimensional design vector. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ and $\boldsymbol{X}_i = \text{diag}(\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{iJ})$ be the vector of binary responses and design matrix, respectively. In the remaining of this section we describe two regression models describing the conditional distribution of $P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{X}_i)$, $\boldsymbol{y}_i \in \{0, 1\}^J$, that are used to illustrate the effect of the misclassification on the association parameters.

3.2.1 The Multivariate Probit Model

The multivariate probit model (MPM) was introduced by Ashford & Sowden (1970), and further considered, for instance, by Amemiya (1972), Ochi & Prentice (1984), Lesaffre & Molenberghs (1991), Lesaffre & Kaufmann (1992), McCulloch (1994), Chan & Kuk (1997), Chib & Greenberg (1998), Edwards & Allenby (2003), and Jara et al. (2007). Under the MPM the joint distribution of the binary responses is given by

$$P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{X}_i\right) = \int_{A\left(y_{iJ}, \boldsymbol{x}_{iJ}, \boldsymbol{\beta}_J^P\right)} \cdots \int_{A\left(y_{i1}, \boldsymbol{x}_{i1}, \boldsymbol{\beta}_1^P\right)} \phi_J\left(\boldsymbol{v} \mid \boldsymbol{0}_J, \boldsymbol{R}\right) d\boldsymbol{v},$$

where $\phi_J\left(\boldsymbol{v} \mid \boldsymbol{0}_J, \boldsymbol{R}\right)$ is the density of a J-variate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{R}$, $\boldsymbol{\beta}_j^P \in \mathbb{R}^p$, $j = 1, \ldots, J$, are regression coefficients for the jth-response variable, and $A\left(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}_j^P\right)$ is the interval given by

$$A(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}_j^P) = \begin{cases} \left(-\infty\ , \boldsymbol{x}_{ij}'\boldsymbol{\beta}_j^P\right] & \text{if } y_{ij} = 1, \\ \\ \left[\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j^P,\ \infty\ \right) & \text{if } y_{ij} = 0, \end{cases}.$$

Due to identifiability issues, $\boldsymbol{R}$ must be in correlation form. In fact, a parameterization in terms of a unconstraint covariance matrix is not identified.

An alternative and useful formulation of the MPM is in terms of Gaussian latent variables. In this formulation, the binary responses are seen as indicators of the event that some unobserved latent variables exceed a threshold value of zero. The key idea is to introduce a J-dimensional latent variable vector $\boldsymbol{V}_i = (V_{i1}, ..., V_{iJ})$ following a multivariate linear model, such that

$$Y_{ij} = I(V_{ij})_{\{V_{ij} > 0\}},$$

where $I(\cdot)_A$ is an indicator function for the set A. The MPM arises when it is assumed that

$$\boldsymbol{V}_i \overset{ind.}{\sim} N_J\left(\boldsymbol{X}_i\boldsymbol{\beta}^P, \boldsymbol{R}\right),$$

where $\boldsymbol{\beta}^P = \left(\boldsymbol{\beta}_1^{P'}, \ldots, \boldsymbol{\beta}_J^{P'}\right)' \in \mathbb{R}^{Jp}$. Under this formulation, the joint distribution of the binary variables may be expressed as

$$P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{X}_i\right) = \int_{B(y_{iJ})} \cdots \int_{B(y_{i1})} \phi_J\left(\boldsymbol{V}_i \mid \boldsymbol{X}_i^P\boldsymbol{\beta}^P, \boldsymbol{R}\right) d\boldsymbol{V}_i, \tag{3.1}$$

where

$$B(y_{ij}) = \begin{cases} (-\infty,\ 0\ ] & \text{if } y_{ij} = 0, \\ \\ (\ \ 0\ \ , \infty) & \text{if } y_{ij} = 1, \end{cases}.$$

This latent variable representation shows how the model parameters determine the joint distribution of the observed variables. In particular, the correlation matrix $\boldsymbol{R} = \{\rho_{jk}\}$ completely captures the association among the observed variables and the correlations $\rho_{jk} = \text{corr}\,(V_{ij}, V_{ik})$ are known as the *tetrachoric correlation coefficients*. This modelling perspective is both flexible and general. In contrast, attempts to model the correlation of the binary responses directly may lead to difficulties.

The latent variable representation also forms the basis for likelihood inference based on the EM algorithm and posterior sampling. Classical inference in the MPM has been considered by Ashford & Sowden (1970), Amemiya (1972), Ochi & Prentice (1984), Lesaffre & Molenberghs (1991), McCulloch (1994), Chan & Kuk (1997) and Chib & Greenberg (1998). Bayesian inference has been discussed by Chib & Greenberg (1998), McCulloch et al. (2000), Liu (2001), Edwards & Allenby (2003), Liu & Daniels (2006) and Zhang et al. (2006).

In the analyzes presented in Section 3.3, maximum likelihood estimates are obtained using the R (R Development Core Team, 2010) library `mprobit`, which is available from CRAN. In this library, quasi-Newton minimization of negative log-likelihood is used with the approximation of Joe (1995) for rectangle multivariate normal probabilities in expression (3.1).

3.2.2 The Conditionally Specified Logistic Regression Model

The conditionally specified logistic regression model (CSLRM) for correlated binary data was introduced by Liu (1994) and Joe & Liu (1996). Under this model, the multivariate joint distributions $P\,(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{X}_i)$ are defined by the specification of their full conditional distributions. Specifically, Liu (1994) and Joe & Liu (1996) assume that, for $j = 1, \ldots, J$, the conditional distribution of the corresponding binary response Y_{ij}, given the other binary responses $Y_{ik} = y_{ik}$, $\forall k \neq j$, and the covariates $\boldsymbol{x}_{ij}$, is a Bernoulli distribution with probability following a logistic regression model with parameter vector $\boldsymbol{\beta}_j^L \in \mathbb{R}^p$ and parameters $\gamma_{jk} \in \mathbb{R}$, $k \neq j$, given by

$$\text{logit}\,\{P(Y_{ij} = 1 \mid Y_{ik} = y_{ik}, k \neq j, \boldsymbol{x}_{ij})\} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j^L + \sum_{k \neq j} \gamma_{jk} y_{ik}.$$

Joe & Liu (1996) show that a necessary and sufficient condition for compatibility of conditional distributions is that $\gamma_{jk} = \gamma_{kj}$, $j \neq k$, and that the joint distribution of the binary vector $\boldsymbol{Y}_i$ is given by

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{X}_i) = c(\boldsymbol{X}_i, \boldsymbol{\beta}^L, \boldsymbol{\gamma})^{-1} \exp\left\{ \sum_{j=1}^{J} (\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j^L) y_{ij} + \sum_{1 \leq j < k \leq J} \gamma_{jk}\ y_{ij} y_{ik} \right\}, \quad (3.2)$$

where $\boldsymbol{\beta}^L = \left(\boldsymbol{\beta}_1^{L'}, \ldots, \boldsymbol{\beta}_J^{L'}\right)' \in \mathbb{R}^{Jp}$, $\boldsymbol{\gamma} = \{\gamma_{jl} : 1 \leq j < k \leq J\}$, and $c(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is a normalizing constant given by

$$c(\boldsymbol{X}_i, \boldsymbol{\beta}^L, \boldsymbol{\gamma}) = \sum_{y_1=0}^{1} \cdots \sum_{y_J=0}^{1} \exp\left\{ \sum_{j=1}^{J} (\boldsymbol{x}'_{ij} \boldsymbol{\beta}_j^L) y_j + \sum_{1 \leq j < k \leq J} \gamma_{jk}\ y_j y_k \right\}.$$

In the CSLRM, the γ_{jk} parameters are interpreted as conditional log-odds ratios, since

$$\begin{aligned} \exp\{\gamma_{jk}\} \quad = \quad & \frac{P(Y_{ij}=1, Y_{ik}=1 \mid \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}{P(Y_{ij}=1, Y_{ik}=0 \mid \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)} \times \\ & \frac{P(Y_{ij}=0, Y_{ik}=0 \mid \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}{P(Y_{ij}=0, Y_{ik}=1 \mid \boldsymbol{x}_{ij}, \boldsymbol{x}_{ik}, Y_{il}=y_{il}, l \neq j,k)}. \end{aligned}$$

Liu (1994) and Joe & Liu (1996) show that for $J = 2$, there are no Y_{il}'s so that γ_{12} is also the unconditional log-odds ratio and it is constant over the covariates. For $J \geq 3$, it is straightforward to show that the bivariate marginal distributions from (3.2), and the log-odds ratios depend on the covariates. Note also that the exponential family in (3.2) is not closed under marginalization and can be easily extended if interaction terms are needed.

Liu (1994) discusses likelihood inferences based on the Newton-Raphson algorithm and evaluate the quality of approximations to point estimators of the regression parameters based on an approximated likelihood inference. The approximation is based on the likelihood arising from the conditional distributions of the model. An R-program (R Development Core Team, 2010), calling FORTRAN subroutines, was written for the analysis of the multivariate binary data with the CSLRM (`cslogistic`) using likelihood and Bayesian approaches. The program `cslogistic` is available from the Comprehensive R Archive Network (CRAN) and was used in Section 3.3.

3.3 The Empirical Evaluation of the Misclassification Effect

We evaluated the finite sample performance of naive MLE of the parameters associated to the CSLRM and MPM under response misclassification, by using Monte Carlo simulation. For each scenario under consideration, we generated 1000 data sets from the true model. The responses generated from the model, $\boldsymbol{Y}_i$, were misclassified using non-differential and differential misclassification processes, yielding $\boldsymbol{Y}_i^* = (Y_{i1}^*, \ldots, Y_{iJ}^*)'$. For each simulated data set, we fitted

the corresponding model to $\boldsymbol{Y}^*$. The maximum likelihood estimates of the association parameters were used to obtain Monte Carlo estimates of the bias and mean squared error (MSE) of the corresponding estimators. In order to quantify the effect of misclassification, the bias and MSE of the MLE under no misclassification were also estimated, for the same true models and simulation scenarios. The different true models and misclassification models are described in the next sections.

3.3.1 The True Models

We simulated data from the CSLRM and MPM for $J = 3$ response variables, and assumed the same design vector and regression effects for each response, i.e. $\boldsymbol{x}_i \equiv \boldsymbol{x}_{i1} = \cdots = \boldsymbol{x}_{iJ}$, $\boldsymbol{\beta}_C^L \equiv \boldsymbol{\beta}_1^L = \cdots = \boldsymbol{\beta}_J^L$ and $\boldsymbol{\beta}_C^P \equiv \boldsymbol{\beta}_1^P = \cdots = \boldsymbol{\beta}_J^P$. We considered design vectors containing a single continuous predictor $w_i \in \mathbb{R}$, i.e. $\boldsymbol{x}_i = (1, w_i)'$, where the w_i's were simulated independently from a standard normal distribution, $w_i \overset{iid.}{\sim} N(0,1)$, and took $\boldsymbol{\beta}_C^L = \boldsymbol{\beta}_C^P = (-1, 1)'$.

In order to reduce the number of possible simulation scenarios, we considered equal association parameters between the variables. Specifically, for the CSLRM we took $\gamma \equiv \gamma_{12} = \gamma_{13} = \gamma_{23}$. Equivalently, for the MPM we considered an exchangeable correlation matrix with parameter ρ. Different degrees of association between the binary response variables were considered. Specifically, we considered $\gamma \in \{0.4, 1.1, 1.8, 2.5\}$ and $\rho \in \{0.2, 0.4, 0.6, 0.8\}$ as values for the CSLRM and MPM, respectively. For each association scenario, three different sample sizes were considered ($I = 200$, 400 and 1000). Therefore, 12 different scenarios were considered for each model. Finally, for each sample size, the same realization of the continuous predictor w_i was used to simulate the data under the different models.

3.3.2 The Misclassification Models

We considered non-differential and differential misclassification processes for each of the 12 scenarios described in the previous section and for each model. For the former, the misclassified responses, Y_{ij}^*, were generated from a Bernoulli distribution, with parameter depending on the status of the realization of the corresponding true response,

$$Y_{ij}^* \mid Y_{ij} = y_{ij} \overset{ind.}{\sim} \begin{cases} \text{Bernoulli}\left(\tau^{11}\right), & \text{if } y_{ij} = 1, \\ \text{Bernoulli}\left(1 - \tau^{00}\right), & \text{if } y_{ij} = 0, \end{cases}$$

where $\tau^{11} \in [0,1]$ is the sensitivity and $\tau^{00} \in [0,1]$ is the specificity of the classification procedure. Four non-differential misclassification processes were considered by taking $\left(\tau^{11}, \tau^{00}\right) \in \{0.85, 0.95\} \times \{0.85, 0.95\}$.

For the differential misclassification process, we considered three different values for the sensitivity and specificity parameters, τ_1^{11}, τ_2^{11} and τ_3^{11}, and τ_1^{00}, τ_2^{00} and τ_3^{00}, respectively. In this case, the Y_{ij}^*'s were generated from a Bernoulli distribution, with parameter depending on the status of the realization of the corresponding true response and on the continuous predictor w_i. Specifically, we considered

$$Y_{ij}^* \mid Y_{ij} = 1, w_i \overset{ind.}{\sim} \begin{cases} \text{Bernoulli}\left(\tau_1^{11}\right), & \text{if } w_i < -0.44, \\ \text{Bernoulli}\left(\tau_2^{11}\right), & \text{if } w_i \in [-0.44, 0.44), \\ \text{Bernoulli}\left(\tau_3^{11}\right), & \text{if } w_i \geq 0.44, \end{cases}$$

and

$$Y_{ij}^* \mid Y_{ij} = 0, w_i \overset{ind.}{\sim} \begin{cases} \text{Bernoulli}\left(1-\tau_1^{00}\right), & \text{if } w_i < -0.44, \\ \text{Bernoulli}\left(1-\tau_2^{00}\right), & \text{if } w_i \in [-0.44, 0.44), \\ \text{Bernoulli}\left(1-\tau_3^{00}\right), & \text{if } w_i \geq 0.44. \end{cases}$$

Two differential misclassification processes were considered. In the first case, a positive association between the precision of the classification process and the value of the continuous predictor, by taking $\tau_1^{11} = \tau_1^{00} = 0.75$, $\tau_2^{11} = \tau_2^{00} = 0.85$, and $\tau_3^{11} = \tau_3^{00} = 0.95$. In the second case, a negative association between the precision of the classification process and the value of the continuous predictor, by taking $\tau_1^{11} = \tau_1^{00} = 0.95$, $\tau_2^{11} = \tau_2^{00} = 0.85$, and $\tau_3^{11} = \tau_3^{00} = 0.75$.

3.3.3 The Results

We fitted the MPM and CSRLM for each of the simulated misclassified data sets, assuming a common intercept and predictor effect, and an unstructured association structure. In order to illustrate the effect of the misclassification, the bias and MSE are expressed as the ratio with respect to the corresponding values under no misclassification. The results are shown for non-differential and differential misclassification separately in the next sections.

Non-Differential Misclassification

For the MPM, the results suggested that under non-differential misclassification, the MLE of the tetrachoric correlations and the associated partial correlations can be strongly and negatively biased. Under all the considered scenarios, the bigger the sample size the greater is the ratio between the absolute bias of the estimators of the tetrachoric correlations under misclassification and no misclasssifcation, which is explained by the reduction in the bias of the estimators as long as the sample size increases, under no misclassification. In general, a similar pattern was

observed for the partial tetrachoric correlations. For a given sample size and error structure, the stronger the association between the unobserved binary variables, the bigger the bias of the naive estimators with respect to the one observed without misclassification in the corresponding scenario. Similarly, the difference between the biases increases, with increasing misclassification errors. Under the worst scenario, the bias of the naive estimator was as big as 496 times the one observed under no misclassification. The results for the worst misclassification scenario are presented in Table 3.1. The results of the remaining scenarios follow a similar pattern and are given in Tables B.1, B.3 and B.5 of Section B.1 of Appendix B.

Table 3.1: Bias of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.85$. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	**Sample Size**	**Correlation**			**Partial Correlation**		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	7.5	6.1	15.6	9.3	5.7	55.0
	400	13.5	10.7	7.8	26.5	14.1	7.4
	1000	18.4	31.8	25.2	20.2	49.0	32.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	14.2	20.8	15.4	15.4	85.0	14.8
	400	19.1	29.2	19.1	24.0	82.0	18.7
	1000	65.0	43.3	43.5	163.0	54.3	41.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	19.1	29.1	14.1	25.6	107.0	8.0
	400	65.8	57.0	49.3	102.0	52.0	40.6
	1000	95.8	77.2	96.3	98.0	50.0	198.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	39.9	31.9	34.5	32.3	10.3	22.5
	400	56.3	83.8	126.3	11.6	106.5	72.3
	1000	123.8	496.0	164.3	23.4	70.7	209.0

The results regarding the MSE for the MPM under non-differential misclassification were similar to the ones described for the bias. Specifically, when the sample size, degree of association or misclassification errors increase, the difference between the MSE of the MLE of the tetrachoric correlations with and without misclassification increases. The results for the partial correlations followed a similar pattern but the ratios were lower in magnitude. The results suggest that the MSE of the naive estimator for the tetrachoric correlations and partial correlations can be as big as 248 and 6.2 times bigger than the corresponding values without misclassification, respectively. The MSE results for the worst non-differential misclassification scenario in the MPM are shown in Table 3.2 (see page 55). The remaining scenarios are given in Tables B.2, B.4 and B.6 of Section B.1 of Appendix B.

Table 3.2: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.85$. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	1.5	1.5	1.4	1.1	1.0	1.0
	400	2.1	2.1	2.0	1.5	1.4	1.4
	1000	4.0	4.0	3.8	2.8	2.6	2.6
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	4.2	4.0	4.5	1.5	1.5	1.6
	400	7.9	8.7	7.3	2.6	2.6	2.5
	1000	17.8	17.8	17.8	6.0	5.0	5.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	11.1	13.1	12.3	1.8	1.9	1.8
	400	27.3	27.8	23.4	3.2	3.3	3.2
	1000	75.0	50.7	50.3	7.0	7.2	7.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	40.6	39.3	40.3	1.5	1.3	1.6
	400	88.3	86.7	87.3	2.4	2.7	2.7
	1000	247.0	249.0	246.0	6.0	6.0	6.7

Under the CSLRM, the results followed a similar behavior regarding the direction of the bias and MSE than for the MPM. However, the difference between the bias of the estimators with and without misclassification were bigger than the ones observed for the MPM. These results can be explained by the smaller bias obtained under no misclassification for the MLE of the association parameters in the CSLRM than in the MPM, which can be due to the higher marginal prevalence considered in the simulation scenarios of the CSRLM than in the MPM and, to the fact that the MPM requires an approximation of the likelihood function in order to obtain the inferences. The magnitude of the MSE ratios of the parameters in the CSLRM were similar to the ones observed for the partial tetrachoric correlations of the MPM. However, in contrary to the observed in the MPM, the MSE of the estimator under misclassification can be as big as 57.6 times the one observed under no misclassification. Table 3.3 (see page 56) shows the bias and MSE results for the CSLRM in the worst misclassification scenarios. The results under the remaining scenarios showed a similar behavior and are given in Tables B.7, B.8 and B.9 of Section B.2 of Appendix B.

Table 3.3: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.85$. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification, B^* and MSE^* respectively, and the corresponding values of the MLE when there is no misclassification, B and MSE respectively.

True Values	**Sample Size**	$\|B^*/B\|$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	24.6	14.2	11.0	1.4	1.5	1.5
	400	35.7	31.3	220.8	2.1	2.2	2.1
	1000	24.0	603.9	48.0	3.5	3.8	3.5
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.1$	200	84.2	339.3	148.9	5.4	5.2	5.2
	400	455.4	265.2	662.9	10.5	8.9	9.6
	1000	11516.6	685.0	1210.4	23.1	24.7	22.1
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	34.1	39.1	50.2	8.9	9.4	8.4
	400	60.2	52.2	397.0	17.3	18.0	20.9
	1000	81161.3	620.2	142.9	48.5	51.5	50.2
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	21.5	26.1	19.2	9.4	9.0	9.5
	400	147.7	364.8	40.8	21.1	20.3	21.3
	1000	128.9	277.8	143.9	55.3	57.6	54.1

Differential Misclassification

Under differential misclassification, the MLE of the association parameters of the MPM showed a similar behavior, namely, they are strongly biased to the null of no association and the MSE, in comparison with the results obtained under no misclassification, increases with the sample size and the degree of association. The effect of the differential misclassification was bigger when there was a negative association between the continuous predictor and the precision of the classification than in the positive case. In general, the results under positive association were similar to the ones obtained under an intermediate non-differential misclassification process. The results under negative association were similar to the ones obtained under the worst non-differential misclassification scenario. Tables 3.4 and 3.5 (see pages 57 and 58, respectively) show the results for the bias and MSE, respectively, for the MPM in this case. The results obtained for the positive association between the covariate and the precision of the classification showed a similar pattern with a lower magnitude and shown in Tables B.10 and B.11 of Section B.3 of Appendix B.

Similarly to the observed in the non-differential misclassification processes, the results for the CSLRM under differential misclassification showed a bigger effect of the misclassification for this model than for the MPM. The bias and MSE for the

Table 3.4: Bias of the estimators of the association parameters of the multivariate probit model under differential misclassification with positive association between the precision of the classification and the continuous predictor. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	**Sample Size**	**Correlation**			**Partial Correlation**		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	4.5	4.0	10.0	5.3	3.6	33.5
	400	8.1	6.5	4.8	15.0	8.3	4.4
	1000	11.0	19.3	15.0	11.4	28.5	18.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	11.7	17.2	12.2	12.2	68.0	10.8
	400	14.7	23.0	15.1	17.4	61.5	14.1
	1000	49.5	32.8	33.3	117.0	38.7	30.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	15.9	24.3	11.8	19.9	84.5	6.4
	400	53.5	46.1	40.8	77.5	39.3	32.2
	1000	78.8	62.8	77.8	77.0	38.3	149.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	33.8	27.7	28.9	25.9	8.8	17.1
	400	47.4	70.5	105.5	9.3	84.5	55.7
	1000	101.3	406.0	136.0	17.8	54.0	164.0

estimators under positive and negative differential misclassification are presented in Tables 3.6 and 3.7 (see pages 59 and 60), respectively. The results showed that the negative differential misclassification has a stronger effect on the estimates of the association parameters. These results also showed that for the CSLRM, the effect of the misclassification on the estimation of the association parameters is bigger under differential than under non-differential misclassification.

3.4 Concluding Remarks

This paper attempts to shed light on the effect of response misclassification on the small sample behavior of naive estimators of the association parameters of regression models for multivariate binary data. The simulation results show that the MLE of the association parameters can be strongly biased if the misclassification process is ignored. Furthermore, regardless the type of misclassification, the naive MLE are strongly biased towards the null of no association, thus showing a different behavior than the one of the estimators of regression coefficients. Indeed, under a non-differential misclassification process, the estimators of the regression coefficients are attenuated towards to the null of

Table 3.5: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under differential misclassification with positive association between the precision of the classification and the continuous predictor. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	**Sample Size**	**Correlation**			**Partial Correlation**		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	0.8	0.9	0.8	0.7	0.7	0.7
	400	1.0	1.2	1.2	0.8	0.8	0.9
	1000	1.8	1.8	1.8	1.2	1.2	1.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	3.0	2.8	3.0	1.1	1.1	1.0
	400	4.9	5.6	4.6	1.6	1.6	1.5
	1000	10.5	10.5	10.8	3.4	2.8	2.8
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	7.8	9.2	8.7	1.2	1.3	1.3
	400	18.2	18.5	16.1	2.0	2.0	2.1
	1000	51.0	33.7	33.0	4.5	4.3	4.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	29.3	29.7	28.4	1.0	1.0	1.0
	400	62.7	61.3	61.3	1.6	1.8	1.7
	1000	167.0	167.0	169.0	3.6	3.6	4.3

no effect, while under differential misclassification, the bias of the estimators can be in both directions, leading to an apparent effect or an apparent lack of effect of the covariate when the reverse is true (see, e.g. Buonaccorsi, 2010).

Although the results reported here could be intuited from the analogy of a misclassification process with classical continuous measurement error models, we argue that the conclusions derived from the continuous cases cannot be directly extended to the discrete one because of the different assumptions of the models. In particular, when a mismeasured variable is binary (or more generally has a known finite support), the independence assumption between the measurement error and the true values of the variable invoked by the standard models for measurement error is particularly untenable.

Similar results to the ones obtained here are expected for other models for multivariate categorical data, such as generalized linear mixed models, multivariate logistic models and log-linear models. Therefore, the development and study of strategies for the correction for misclassification for multivariate binary data seems to be an important subject of research.

The natural next step on this research, would be to try to correct for

Table 3.6: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under differential misclassification with positive association between the precision of the classification and the continuous predictor. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification, B^* and MSE^* respectively, and the corresponding values of the MLE when there is no misclassification, B and MSE respectively.

True Values	**Sample Size**	$\|B^*/B\|$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	19.5	15.6	17.3	1.2	1.6	2.5
	400	29.7	42.0	181.1	1.6	3.1	1.6
	1000	24.0	713.9	74.8	3.6	5.0	7.5
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.1.1$	200	89.6	324.2	196.5	6.0	4.8	8.5
	400	551.5	370.0	520.3	14.9	16.7	6.1
	1000	12991.6	712.7	1730.3	29.2	26.7	44.4
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	41.0	35.0	61.3	12.6	7.6	12.1
	400	81.3	71.3	324.8	31.2	33.1	14.2
	1000	92781.6	614.7	188.8	63.2	50.5	87.1
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	25.9	24.8	20.8	13.4	8.1	11.1
	400	190.6	470.3	36.6	35.0	33.6	17.2
	1000	146.7	279.9	169.6	71.5	58.5	75.1

misclassification and to analyze whether this correction would decrease the attenuation of the estimators of the association parameters towards the null. However, we argue that it is important to evaluate the advantages of the usage of approaches to correct for misclassification. Although the use of this approaches can reduce the bias of a particular estimator, it may introduce more variability at the same time and thus, yield an estimator with a greater MSE than the naive one (see, e.g. Luan et al., 2005). This issue should be explored in the context of misclassified binary data because it might not always be beneficial to correct for misclassification.

Moreover, as the data could contain no information on the misclassification parameters, the identification study of proposals trying to estimate the misclassification parameters without using external information is also needed. These and other topics are the subject of ongoing research.

Table 3.7: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under differential misclassification with negative association between the precision of the classification and the continuous predictor. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification and the corresponding values of the MLE when there is no misclassification.

True Values	**Sample Size**	$\lvert B^*/B \rvert$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	31.7	13.5	5.2	1.8	1.5	1.0
	400	37.1	18.2	227.6	2.3	1.3	2.3
	1000	24.4	709.0	76.5	3.6	5.0	7.7
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.1$	200	76.1	376.7	89.7	4.6	6.2	2.4
	400	309.5	130.0	843.1	5.3	2.7	15.2
	1000	12966.5	706.2	1732.0	29.2	26.2	44.6
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	25.3	46.3	37.2	5.2	13.0	4.8
	400	29.8	24.5	507.0	4.8	4.5	33.8
	1000	92760.3	614.1	189.1	63.1	50.5	87.4
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	16.0	28.7	16.4	5.3	10.8	7.0
	400	76.3	195.1	48.0	5.9	6.1	29.4
	1000	146.3	278.9	169.4	71.1	58.1	74.9

Acknowledgements

The first author is supported by the National Scholarship for Doctoral Studies 2009, Conicyt (Chile) and by the Research Grant OT/05/60. She also acknowledges the partial support from the Interuniversity Attraction Poles Program P6/03, Belgian State, Federal Office for Scientific, Technical and Cultural Affairs.

References

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

Amemiya, T. (1972). Bivariate probit analysis: minimum chi-square methods. *Journal of the American Statistical Association* 69 940–944.

Ashford, J. R. & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* 26 535–546.

Barron, B. A. (1977). Effects of misclassification on estimation of relative risk. *Biometrics* 33 414–418.

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

Buonaccorsi, J. P. (2010). *Measurement Error*. New York, USA: Chapman & Hall/CRC.

Chan, J. S. K. & Kuk, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* 53 86–97.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85 347–361.

Cook, R. J., Ng, E. T. M. & Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* 56 1109–1117.

Edwards, Y. D. & Allenby, G. M. (2003). Multivariate analysis of multiple response data. *Journal of Marketing Research* 40 321–334.

Espeland, M. A. & Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* 43 1001–1012.

Espeland, M. A., Murphy, W. C. & Leverett, D. H. (1988). Assessing diagnostic reliability and estimating incidence rates associated with a strictly progressive disease: dental caries. *Statistics in Medicine* 7 403–416.

Espeland, M. A., Platt, O. S. & Gallagher, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. *Journal of the American Statistical Association* 84(408) 972–979.

García-Zattera, M. J., Mutsvari, T., Jara, A., Declerck, D. & Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine* (To appear).

Geng, Z. & Asano, C. (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics* 8 2935–2954.

Gladen, B. & Rogan, W. J. (1979). Misclassification and the design of environmental studies. *American Journal of Epidemiology* 109 607–616.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine* 7 745–757.

Jara, A., Garcia-Zattera, M. J. & Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics and Data Analysis* 51 5402–5415.

Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association* 90 957–964.

Joe, H. & Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.

Katz, B. M. & McSweeney, M. (1979). Misclassification errors and categorical data analysis. *Journal of Experimental Education* 47 331–338.

Lesaffre, E. & Kaufmann, H. (1992). Existence and uniqueness of the maximum likelihood estimator for a multivariate probit. *Journal of the American Statistical Association* 87 805–811.

Lesaffre, E. & Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine* 10 1391–1403.

Liu, Y. (1994). *A model for multivariate binary data with covariates based in compatible conditionally specified logistic regressions.* Unpublished doctoral thesis, Department of Statistics, University of British Columbia.

Liu, C. (2001). Bayesian analysis of multivariate probit models - Discussion on the art of data augmentation by van Dyk and Meng. *Journal of Computational and Graphical Statistics* 10 75–81.

Liu, X. & Daniels, M. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *Journal of Computational and Graphical Statistics* 15 897–914.

Luan, X., Pan, W., Gerberich, S. G. & Carlin, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine* 24 2221–2234.

Magder, L. S. & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 146 195–203.

Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology* 43 941–947.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* 89 330–335.

McCulloch, R., Polson, N. & Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* 99 173–193.

McGlothlin, A., Stamey, J. D. & Seaman, J. W. (2008). Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach. *Biometrical Journal* 50 123–134.

Morrissey, M. J. & Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* 55 338–344.

Mote, V. L. & Anderson, R. L. (1965). An investigation of effect of misclassification on properties of χ^2-tests in analysis of categorical data. *Biometrika* 52 95–109.

Mwalili, S. M., Lesaffre, E. & 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* 54(1) 77–93.

Nagelkerke, N. J. D., Chunge, R. N. & Kinot, S. N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 9 1211–1219.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86(4) 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* 58 675–683.

Ochi, Y. & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* 71 531–543.

Paulino, C. D., Silva, G. & Achcar, J. A. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics and Data Analysis* 49 1120–1131.

Paulino, C. D., Soares, P. & Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics* 59 670–675.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Rekaya, R., Weigel, K. A. & Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics* 57 1123–1129.

Rosychuk, R. J. & Islam, M. S. (2009). Parameter estimation in a model for misclassified Markov data - a Bayesian approach. *Computational Statistics and Data Analysis* 53 3805–3816.

Rosychuk, R. J. & Thompson, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 19 394–404.

Rosychuk, R. J. & Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification *Statistics in Medicine* 22 2035–2055.

Roy, S. & Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal* 51 420–432.

Rubin, T., Rosenbaum, J. & Cobb, S. (1956). The use of interview data for the detection of associations in field studies. *Journal of Chronic Diseases* 4 253–266.

Schmid, C. H., Segal, M. R. & Rosner, B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference* 42(1–2) 1–18.

Schwartz, J. E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods and Research* 13 435–466.

Singh, A. C. & Rao, J. N. K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey. *Journal of the American Statistical Association* 90(430) 478–488.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* 65(331) 1350–1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. *Biometrics* 27 935–944.

Zelen, M. & Haitovsky, Y. (1991). Testing hypotheses with binary data subject to misclassification errors: analysis and experimental design. *Biometrika* 78 857–865.

Zhang, X., Boscardin, W. J. & Belin, T. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics* 15 880–896.

Chapter 4

Correcting for Misclassification for a Monotone Disease Process with an Application in Dental Research

This chapter has been published as:

García-Zattera, M. J., Mutsvari, T., Jara, A., Declerck, D. & Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine* 29(30) 3103–3117.

Abstract

Motivated by a longitudinal oral health study, we evaluate the performance of binary Markov models in which the response variable is subject to an unconstrained misclassification process and follows a monotone or progressive behavior. Theoretical and empirical arguments show that the simple version of the model can be used to estimate the prevalence, incidences and misclassification parameters without the need of external information and that the incidence estimators associated to the model outperformed approaches previously proposed in the literature. We propose an extension of the simple version of the binary Markov model to describe the relationship between covariates and prevalence and incidence allowing for different classifiers. We implemented a Bayesian version of the extended model and show that, under the settings of our motivating example, the parameters can be estimated without any external information. Finally, the analyses of the motivating problem are presented.

Key Words: Misclassified binary data; Monotone processes; Incidence estimation; Hidden Markov model; Identifiability.

4.1 Introduction

Errors in the determination of disease outcome occur very often in epidemiological studies. Diagnostic tests or classifiers may not perfectly reflect the true individuals' condition leading to misclassified disease outcomes where sick individuals may be diagnosed as healthy, healthy individuals may be diagnosed as sick, or the severity of the case may be misjudged. The effect of misclassification on estimation and hypothesis testing has been widely investigated in the literature. Bross (1954) showed that misclassification on a binary response does not affect the validity of the significance test used to compare samples from two populations but the power may be drastically reduced. He also showed that, although the data contain no information regarding the misclassification probabilities, severely biased estimates can be obtained when misclassification is ignored. Tenenbein (1970, 1971) extended the analysis to multinomial data and proposed a double sampling scheme to obtain information concerning the misclassification probabilities.

Several authors have extended the analyses to regression settings for uncorrelated or clustered data gathered in cross-sectional studies under both differential (covariate dependent) and non-differential (covariate independent) misclassification, and proposed strategies for correcting the estimates. See, for instance, Geng & Asano (1989), Magder & Hughes (1997), Neuhaus (1999, 2002), Rekaya et al. (2001), Mwalili et al. (2005) and Küchenhoff et al. (2006), for different

approaches to misclassified categorical data under several sampling schemes. The findings are that under non-differential misclassification the regression coefficients are attenuated to the null. However, the attenuation effect of non-differential misclassification does not preclude that for an individual study the misclassification process had the opposite effect, (see, e.g. Jurek et al., 2005). Under differential misclassification the bias can lead either to an apparent (but non-existing) association or to an apparent lack of association when it does exist.

The basic difficulty in the proposed models for correcting for misclassification in cross-sectional studies is the identifiability of the model parameters by the available data. Without additional information beyond the main data, it is not possible to take into account the effect of misclassification. Therefore, the approaches rely on 1) the existence of a validation study, along the lines of Tenenbein's double sampling approach, or on 2) expert information, incorporated through out a prior distribution in a Bayesian analysis, to provide the required additional information on the misclassification parameters. A difficulty with 1) is the necessity of an infallible classifier, which may not exist or may be prohibitively expensive. Further, since the variability on the estimation of the misclassification parameters has an important effect on the variability of the regression parameters, the validation study should be taken as large as possible. Unfortunately, often an internal validation study cannot be taken large, if it can be taken at all in practice, as will be seen in the next section. A difficulty with 2) is the existence of expert knowledge on parameters that have not been estimated previously.

In the context of longitudinal studies, the impact of misclassification on transition probability estimates for an unobservable alternating binary Markov process, i.e. where a subject may alternate between two states over time (e.g. uninfected/infected, healthy/sick, etc) has been discussed by Cook et al. (2000), Nagelkerke et al. (1990), Rosychuk & Thompson (2001, 2003), Rosychuk & Islam (2009). Motivated by research questions associated to a longitudinal oral health study conducted in Flanders (Belgium), the Signal-Tandmobiel® (ST) study, this paper focuses on misclassified longitudinal binary data where the true response follows a progressive or monotone process, i.e. when the subjects cannot alternate between the two stages once the severe stage (e.g. infected, diseased, death) is reached over time. Some medical examples of progressive processes are rheumatoid arthritis, systemic lupus, osteoporosis, AIDS, chronic kidney disease and caries experience (CE).

Hidden Markov models (HMM) for longitudinal monotone data have been considered by Espeland et al. (1988, 1989), Schmid et al. (1994), Singh & Rao (1995) and Albert et al. (1997). In these proposals the authors did not make use of external information on the misclassification parameters and suggested that those parameters can be estimated from the main data. However, neither formal proofs nor empirical evidences have been provided establishing that the model parameters are identified. Although in a simple inhomogeneous HMM for

a monotone binary process, the existence of at least three observations over time and constant misclassification parameters ensure that the number of identified parameters is greater than the number of parameters of interest, these conditions do not ensure the identifiability of model parameters. For instance, restriction on the misclassification parameters of the kind $\eta + \theta > 1$ may be still needed, where $\eta \in [0, 1]$ and $\theta \in [0, 1]$ is the sensitivity and specificity, respectively. Further, even though those restrictions may be sufficient for the identification of the model parameters in the simple version of the model they do not ensure the identification of the parameters in more realistic extensions of the simple HMM.

In this paper we evaluate the identifiability properties of simple inhomogeneous HMM under the absence of external information on the misclassification parameters, and compare the performance of the associated estimators with early approaches proposed for incidence estimation. We also propose and evaluate an extension of the simple HMM to account for predictors, different time intervals between examinations for each subject, and different classifiers. The analyses of our motivating problem are also presented, where we look at the prevalence and incidence of CE and at the evaluation of risk factors. The paper is organized as follows. Section 4.2 introduces the ST study and the research questions. The evaluation of the properties and performance of the estimators associated with the simple HMM without validation data are presented in Section 4.3, along with the comparison with early approaches for incidence estimation in the presence of misclassified data. In Section 4.4 we propose an extension of the simple HMM and evaluate its performance under the setting of the motivating example. In Section 4.5, the analyses of the ST data are presented. A final discussion section concludes the article.

4.2 The Signal-Tandmobiel® Study and Research Questions

In this section we provide a brief description of the ST study and the associated research questions. For a more detailed description we refer to Vanobbergen et al. (2000). The ST study is a longitudinal prospective oral health screening study conducted in Flanders, Belgium, between 1996 and 2001. For this project, 4468 children were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen dental examiners. Clinical data were collected by the examiners based on visual and tactile observations (no X-rays were taken), and data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents.

Caries lesions are scored in four levels of lesion severity: d_4 (dentine caries with pulpal involvement), d_3 (dentine caries with obvious cavitation), d_2 (hidden

dentine caries) and d_1 (white or brown-spot initial lesions in enamel without cavitation). Here we consider CE as a binary variable indicating whether the tooth is decayed at d_3 level, missing or filled due to caries, which defines a progressive disease. Thus observed reversals, i.e. teeth or surfaces initially recorded as being carious subsequently recorded as caries-free, represent diagnostic errors. The diagnosis of CE might be difficult for a variety of reasons. For instance, nowadays composite materials can imitate the natural enamel so well that it is sometimes difficult to spot a restored lesion. Another reason may be that the location of the cavity e.g. 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 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 done 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 dental examiner with a lot of experience, in this case one of the authors (DD). 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. At the end of each of the three calibration exercises the sensitivity and specificity of each dental examiner vis-a-vis the benchmark examiner were determined, yielding a misclassification table for each examiner for scoring on $d3$ at tooth and surface levels. The results suggest that some examiners overscore or underscore the true CE status. It is important to stress that although children that participated in the calibration exercises were used as a validation data set in previous work of the research team (see, e.g. Mwalili et al., 2005) the validation data were not taken at random from the main data. Rather a school was selected with a presumed high prevalence for CE. Because of this, the information provided by the calibration exercises on the misclassification parameters cannot be formally used in the main analysis of the ST data. Notice also that a pure random sample would be impractical, but also a validation data set sampled in a clustered manner (first sampling schools and then children within schools) would imply a too high investment in time and personnel. Further, both sampling approaches would likely involve too few children with CE implying that the sensitivity would be poorly estimated.

The statistical findings reported below were applied to the scoring of the four permanent first molars, i.e., teeth 16, 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). The numbering of the teeth follows the FDI (Federation Dentaire Internationale) notation which indicates the position of the tooth in the mouth (an illustrative figure is available in Section C.1

of Appendix C). Position 26, for instance, means that the tooth is in quadrant 2 (upper left quadrant) and position 6 where numbering starts from the mid-sagittal plane. The primary interest of the present analysis is to evaluate the prevalence and incidence of CE and to address the influence of oral hygiene and dietary habits and of geographical information on the presence and evolution of CE over time.

4.3 The Simple Hidden Markov Model

In this section we introduce the simple HMM for the analysis of longitudinal monotone binary data and discuss its identifiability properties. The discussion is based on theoretical and empirical arguments. We evaluate the performance of the associated estimators under the absence of external information on the misclassification parameters and compare the estimates with early approaches proposed for incidence estimation.

4.3.1 The Model and Some Identification Results

Suppose that m subjects are examined at the same n time points $(t_1, \ldots, t_n)$. Let $Y_{(i,j)}$ be the true unobserved binary response for subject i at time t_j and denote the vector of n true binary responses for subject i by $\boldsymbol{Y}_i = (Y_{(i,1)}, \ldots, Y_{(i,n)})$. We assume that the vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are *iid* following a monotone inhomogeneous first-order Markov process. This process is completely characterized by the prevalence, $p = P(Y_{(i,1)} = 1)$, and by the vector of incidences $\boldsymbol{q} = (q_1, \ldots, q_{n-1})$, where $q_j = P(Y_{(i,j+1)} = 1 | Y_{(i,j)} = 0)$, $j = 1, \ldots, n-1$. Note that $P(Y_{(i,j+1)} = 1 | Y_{(i,j)} = 1) = 1$ since we assumed a monotonic binary process. Therefore, the transition matrix between time points t_j and t_{j+1}, $Q_j(q_j)$, is given by

$$Q_j(q_j) = \begin{pmatrix} 1 - q_j & q_j \\ 0 & 1 \end{pmatrix}, \qquad j = 1, \ldots, n-1.$$

Let $V \subset \{0,1\}^n$ the set of admissible monotone response patterns. For instance, for $n = 3$, $V = \{(000), (001), (011), (111)\}$. The joint probability distribution for the true latent responses is given by

$$\begin{aligned} P(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \mid p, \boldsymbol{q}) &= \prod_{i=1}^{m} P\left(Y_{(i,1)} = y_{(i,1)}, \ldots, Y_{(i,n)} = y_{(i,n)} \mid p, \boldsymbol{q}\right), \\ &= \prod_{i=1}^{m} \left\{ P\left(Y_{(i,1)} = y_{(i,1)}\right) \prod_{j=1}^{n-1} P\left(Y_{(i,j+1)} = y_{(i,j+1)} | Y_{(i,j)} = y_{(i,j)}\right) \right\}, \end{aligned}$$

$$= \prod_{i=1}^{m} \left\{ p^{y_{(i,1)}} (1-p)^{1-y_{(i,1)}} \prod_{j=1}^{n-1} \left[q_j^{y_{(i,j+1)}} (1-q_j)^{1-y_{(i,j+1)}} \right]^{1-y_{(i,j)}} \right\},$$

where $\boldsymbol{y}_i \in V$.

We assume that the response vectors $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, are prone to misclassification. Let $Y^*_{(i,j)}$ be the observed binary response at time t_j and denote the vector of corrupted binary responses for subject i by $\boldsymbol{Y}^*_i = (Y^*_{(i,1)}, ..., Y^*_{(i,n)})$. Let $\tau_{10} = P(Y^*_{(i,j)} = 1 \mid Y_{(i,j)} = 0)$ and $\tau_{01} = P(Y^*_{(i,j)} = 0 \mid Y_{(i,j)} = 1)$, $\forall\ i, j$, be the misclassification parameters. We assume that the misclassification process is characterized by the following conditional independence assumptions:

A.1) $\perp\!\!\!\perp_{1 \le i \le m} \boldsymbol{Y}^*_i \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m, \tau_{10}, \tau_{01}$, i.e. the observed response patterns for each subject are independent given the true unobserved pattern and misclassification parameters,

A.2) $\boldsymbol{Y}^*_i \perp\!\!\!\perp \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{i-1}, \boldsymbol{Y}_{i+1}, \ldots, \boldsymbol{Y}_m \mid \boldsymbol{Y}_i, \tau_{10}, \tau_{01}, \ \forall\ i$, i.e. the distribution of the observed response pattern for a subject depends only on his true unobserved pattern and the misclassification parameters,

A.3) $\perp\!\!\!\perp_{1 \le j \le n} Y^*_{(i,j)} \mid \boldsymbol{Y}_i, \tau_{10}, \tau_{01}, \forall\ i$, i.e. the observed responses for a subject are independent given his unobserved response pattern and the misclassification parameters, and

A.4) $Y^*_{(i,j)} \perp\!\!\!\perp \boldsymbol{Y}_i \mid Y_{(i,j)}, \tau_{10}, \tau_{01}, \ \forall\ i,\ j$, i.e. the observed response for a subject at time t_j depends only on the unobserved response at time t_j and the misclassification parameters.

These assumptions imply that the observed binary vectors form an *iid* process with common probability given by

$$P\left(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i \mid p, \boldsymbol{q}, \tau_{10}, \tau_{01}\right) \quad = \quad \sum_{\boldsymbol{y} \in V} P\left(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i, \boldsymbol{Y}_i = \boldsymbol{y} \mid p, \boldsymbol{q}, \tau_{10}, \tau_{01}\right), \quad (4.1)$$

where $\boldsymbol{y}^*_i \in \{0,1\}^n$, and the joint distribution for the observed and latent binary variables is given by

$$P\left(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i, \boldsymbol{Y}_i = \boldsymbol{y}_i \mid p, \boldsymbol{q}, \tau_{10}, \tau_{01}\right) = \prod_{j=1}^{n} P\left(Y^*_{(i,j)} = y^*_{(i,j)} \mid Y_{(i,j)} = y_{(i,j)}, \tau_{10}, \tau_{01}\right)$$

$$\times \left\{ P\left(Y_{(i,1)} = y_{(i,1)} \mid p\right) \prod_{j=1}^{n-1} P\left(Y_{(i,j+1)} = y_{(i,j+1)} | Y_{(i,j)} = y_{(i,j)}, \boldsymbol{q}\right) \right\}$$

$$
\begin{aligned}
&= \left[\tau_{01}^{1-y^*_{(i,1)}}\left(1-\tau_{01}\right)^{y^*_{(i,1)}} p\right]^{y_{(i,1)}} \left[\tau_{10}^{y^*_{(i,1)}}\left(1-\tau_{10}\right)^{1-y^*_{(i,1)}}\left(1-p\right)\right]^{1-y_{(i,1)}} \\
&\quad \times \prod_{j=1}^{n-1}\left[\tau_{01}^{1-y^*_{(i,j+1)}}\left(1-\tau_{01}\right)^{y^*_{(i,j+1)}}\right]^{y_{(i,j)}} \\
&\quad \times \prod_{j=1}^{n-1}\left[\tau_{01}^{1-y^*_{(i,j+1)}}\left(1-\tau_{01}\right)^{y^*_{(i,j+1)}} q_j\right]^{y_{(i,j+1)}\left(1-y_{(i,j)}\right)} \\
&\quad \times \prod_{j=1}^{n-1}\left[\tau_{10}^{y^*_{(i,j+1)}}\left(1-\tau_{10}\right)^{1-y^*_{(i,j+1)}}\left(1-q_j\right)\right]^{\left(1-y_{(i,j+1)}\right)\left(1-y_{(i,j)}\right)},
\end{aligned}
$$

where $\boldsymbol{y}_i^* \in \{0,1\}^n$ and $\boldsymbol{y} \in V$. The above model can be generalized by allowing for dropouts and intermittent missing responses under the assumption of missing at random. In the first case the number of time points differs between subjects, i.e. in the above expressions n is replaced by n_i. Allowing for intermittent missingness involves an extra summation in the likelihood contribution. Namely, suppose that the response at the k^{th} time point, Y_k^*, is missing. Then $P\left(Y_1^* = y_1^*, \ldots, Y_{k-1}^* = y_{k-1}^*, Y_{k+1}^* = y_{k+1}^*, \ldots, Y_n^* = y_n^* \mid p, \boldsymbol{q}, \tau_{10}, \tau_{01}\right)$ in the likelihood contribution is rewritten as $\sum_{y_k^*=0}^{1} P\left(Y_1^* = y_1^*, \ldots, Y_k^* = y_k^*, \ldots, Y_n^* = y_n^*\right)$ and each of the two components in the summation is then decomposed as in (4.1). When more responses are missing intermittently then the summation is done over the missing parts.

It is important to stress that there are 2^n possible observed binary patterns $\boldsymbol{y}^*$. Thus, the 2^n associated probabilities given by the corresponding evaluations of expression (4.1) are identified by the data in an equivalence to the *iid* sampling from a multinomial distribution. Because the identified parameters are functions of the parameters of interest $\boldsymbol{\theta} = (p, \boldsymbol{q}, \tau_{10}, \tau_{01})$, a possible strategy for the identification analysis of the simple HMM is to express the parameter of interest as functions of the identified parameters. Indeed, as the probability of the possible observed patterns add to one, $2^n - 1$ independent relations can be used to identify the parameters in the simple HMM.

For $n = 2$ time points, three equations are available to find the solution for four unknown parameters and the model parameters are clearly unidentified. In this case, an intuitive identifying restriction would be to assume the same misclassification errors regardless the true underlying status, i.e. $\tau = \tau_{10} = \tau_{01}$, leading to a restricted HMM. For $n > 2$ time points, the number of independent relations, $2^n - 1$, is greater than the number parameters of interest, $n + 2$, suggesting that the parameters are identified in an unrestricted HMM. However, even though the number of free identified parameters is equal or greater than

the number of parameters of interest in the restricted and unrestricted HMM, respectively, the parameters are unidentified. In fact, any point in the parameter space of the form $\boldsymbol{\theta} = (p, q_1, \ldots, q_{n-1}, \tau_{10} = 0.5, \tau_{01} = 0.5)$ induce the same probabilities $P(\boldsymbol{Y}_i^* = \boldsymbol{y}^* \mid p, q_1, \ldots, q_{n-1}, \tau_{10} = 0.5, \tau_{01} = 0.5) = 1/2^n$, $\forall\ i \in \mathcal{N}$ and $\forall\ \boldsymbol{y}^* \in \{0,1\}^n$. This result shows that identifying restrictions are needed in both situations.

The following proposition provides an identifying restriction for the first case. The proof is given in Section C.2 of Appendix C.

Proposition 4.1. *In a simple HMM with $n = 2$ time points, the assumptions i) $\tau = \tau_{10} = \tau_{01}$ and $\tau < 0.5$ or ii) $\tau = \tau_{10} = \tau_{01}$ and $\tau > 0.5$ are sufficient restrictions for the identification of the model parameters $\boldsymbol{\theta} = (p, q_1, \tau_{10}, \tau_{01})$.*

For the unrestricted HMM and $n > 2$ time points, an equivalent identification restriction would be $\tau_{10} + \tau_{01} < 1$. Although in principle a similar strategy to that in Proposition 4.1 could be considered to provide a theoretical proof, we would need to show the existence of a unique solution in a highly nonlinear equation system. Instead, we provide empirical evidence that the model parameters are identified under this restriction by means of a simulation study in Section 4.3.3. As the identification of the model parameters is a necessary condition for the existence of consistent estimators, we evaluate the behavior of the bias and the variance of maximum likelihood estimators (MLE) when the sample size increases. Therefore, a reduction in the mean square error (MSE) of the estimators would give, although no conclusive evidence, good insights on the identifiability of the model parameters. The simulation study is also used to compare the performance of the estimators with respect to early approaches proposed in the dental literature for incidence estimation. These methods are discussed in the next section.

4.3.2 Early approaches to estimate incidence in presence of misclassified data

Bias associated with misclassification of carious lesions has been documented and discussed in the dental literature for a long time. As a consequence, different methods have been proposed for correcting for misclassification of caries at tooth and surface level (Radike & Muhler, 1954; Radike, 1960; Carlos & Senning, 1968; Lu, 1968; Poole et al., 1973). Radike & Muhler (1954) and Radike (1960) discussed two approaches that deal with the exclusion or inclusion of the observed reversals in the estimation of the incidence. The reversals excluded (RE) method remove the observed reversals from the analysis and considers the apparent incidence, i.e. the observed proportion of caries-free teeth which become carious between two time points, as the estimator of the incidence. The reversals included (RI) method

considers the quantity resulting from subtracting the proportion of reversals to the apparent incidence as the estimator of the incidence.

Carlos & Senning (1968) noted that the reversals represent only partially the total number of classification errors and that both RE and RI methods ignores the fact that in the first examination the presence of caries can be misclassified. They proposed the method of moments estimator, denoted by CS, for $n = 2$ time points by assuming that both types of misclassification errors are equal, i.e. $\tau = \tau_{01} = \tau_{10}$ in our notation. Using a similar strategy, Lu (1968) parameterized the examiner's accuracy in terms of the true proportion of surfaces over which his diagnoses are certain to be correct and a random guessing misclassification. Poole et al. (1973) generalized CS and LU estimators by assuming $\tau_{10} \neq \tau_{01}$, and proposed the PCS estimator using the restricted least squares method. In order to avoid identifiability problems, the approach relies on the existence of a second sample with the same misclassification parameters but different transition probabilities.

4.3.3 The Simulation Study and Results

To explore the performance of estimators associated to the simple HMM for $n > 2$ time points and to compare them with the early approaches, we conducted a simulation study. Different settings were considered. Full results of the simulation study are available in Section C.3 of Appendix C. Here we illustrate the conclusions by discussing the results obtained by simulating two longitudinal data settings, $n = 3$ and $n = 6$ time points, and two sample sizes, $m = 2000$ and $m = 5000$. In each case, different true values for the parameters were considered, yielding 36 scenarios. Low true values for prevalence, $p = 0.02, 0.10, 0.15$, and incidences $q_1 = \ldots = q_5 = 0.04, 0.10, 0.15$ were considered in order to mimic the ST study to a certain extent. Finally, four combinations of misclassification parameters were considered: $(\tau_{10}, \tau_{01}) = (0.15, 0.15), (0.05, 0.15), (0.15, 0.05)$ and $(0.05, 0.05)$.

For each scenario we simulated 1000 data sets from the simple HMM and obtained the MLE under the constrained parameter space given by $\tau_{10} + \tau_{01} < 1$. The RE, RI, CS and LU methods where applied repeatedly for each of two consecutive time points. The PCS method was also considered by using the correlated data from an adjacent pair of time points as a second independent sample. For all methods we computed the bias and MSE of the estimators for the incidence parameters. The bias and MSE for the prevalence and the misclassification parameters were computed only for the MLE arising from the simple HMM.

For the misclassification parameters, the bias of MLE for the simple HMM, expressed as a percentage of the true value, was lower than 1% in 99% of the scenarios for $n = 3$ time points. For $n = 6$ time points the bias was lower than 0.1% of the true value of the parameters in all scenarios. In both longitudinal settings, the bias reduced with more than 50% when the sample size increased

from $m = 2000$ to $m = 5000$. Similar conclusions were obtained for the MSE of the MLE of the misclassification parameters. The largest MSE was observed in the scenarios with the lowest true prevalence and incidences and the greatest true misclassification parameters. However, the largest MSE observed value was as small as 0.012. The results for some selected and representative scenarios are displayed in Table 4.1 (see page 76).

The MSE for the MLE of the incidences obtained for $n = 6$ time points are also presented in Table 4.1 (see page 76). As expected, reductions greater than 50% were observed when the sample size increases. The bias of the MLE for the prevalence was lower than 1% of the true value in 76% and 96% of the scenarios for $n = 3$ and $n = 6$ time points, respectively. The MSE for the MLE of the prevalence showed the same behavior observed for the other model parameters.

The simulation study indicates that the maximum likelihood procedure always performs better than the other methods in terms of bias and MSE, even for scenarios where the assumptions required by the early approaches are fulfilled, e.g. when $\tau_{01} = \tau_{10}$. The results suggest that nearly unbiased estimates of the parameters can be obtained for the simple HMM. Furthermore, as important reductions in the MSE were observed when the sample size increases, the results suggest the identifiability of the parameters under the restriction $\tau_{10} + \tau_{01} < 1$. Therefore, the results suggest that the parameters can be estimated in a HMM for monotone binary data without the need of external information on the misclassification parameters.

4.4 An Extension of the Simple Hidden Markov Model

In most practical applications, the assumptions required for the simple HMM are not met. For instance, at the start of the ST study the age, but also the oral health and dietary habits differed across the children, affecting the prevalence most likely. Additionally, the timing of the examinations is not the same for all the children and the oral hygiene and dietary habits evolved differently for each child, potentially leading to different incidences across children. Finally, 16 different examiners were involved in the study who most likely have different misclassification patterns. This issue is particularly important for the ST study because the examiners switched over time. Indeed, none of the children were evaluated within the conduct of the study by the same examiner. In this section we extend the simple HMM in order to accommodate for these considerations. The performance of the proposed model is evaluated using simulated data with the same characteristics as the ST study.

Table 4.1: Mean squared error (MSE $\times 10^3$) for the incidence parameters estimated using the early approaches and the simple hidden Markov model with $n = 6$ time points for $m = 2000$ and $m = 5000$ subjects, and for different true values of the prevalence p, incidences $q_1 = ... = q_5$, and misclassification parameters τ_{01} and τ_{10}. The MSE ($\times 10^3$) for the maximum likelihood estimator of the sensitivity $(1 - \tau_{01})$ and specificity $(1 - \tau_{10})$, associated to the simple hidden Markov model are also presented.

True Values				**Parameter**	$\boldsymbol{m}$	**Method**					
p	$q_1/q_2/q_3/q_4/q_5$	τ_{01}	τ_{10}			RE	RI	CS	LU	PCS	MLE
0.03	0.04	0.15	0.15	q_1	2000	12.37	0.29	0.25	0.59	0.24	0.06
					5000	12.27	0.22	0.12	0.38	0.14	0.02
				q_2	2000	12.06	0.33	0.25	0.52	0.27	0.06
					5000	12.05	0.24	0.10	0.29	0.14	0.03
				q_3	2000	11.94	0.34	0.26	0.50	0.28	0.08
					5000	11.83	0.28	0.12	0.24	0.16	0.03
				q_4	2000	11.77	0.37	0.27	0.47	0.31	0.10
					5000	11.67	0.30	0.13	0.24	0.18	0.04
				q_5	2000	11.63	0.42	0.31	0.47	0.35	0.26
					5000	11.47	0.35	0.16	0.21	0.22	0.11
				$1 - \tau_{01}$	2000						0.37
					5000						0.14
				$1 - \tau_{10}$	2000						0.02
					5000						0.01
0.10	0.04	0.05	0.05	q_1	2000	1.50	0.11	0.08	0.07	0.09	0.03
					5000	1.49	0.08	0.04	0.03	0.05	0.01
				q_2	2000	1.41	0.14	0.10	0.08	0.11	0.03
					5000	1.39	0.10	0.06	0.04	0.06	0.01
				q_3	2000	1.33	0.16	0.11	0.09	0.12	0.03
					5000	1.30	0.13	0.08	0.05	0.08	0.01
				q_4	2000	1.24	0.19	0.14	0.11	0.15	0.04
					5000	1.23	0.15	0.09	0.05	0.10	0.01
				q_5	2000	1.17	0.21	0.16	0.12	0.17	0.07
					5000	1.14	0.18	0.12	0.07	0.13	0.03
				$1 - \tau_{01}$	2000						0.04
					5000						0.01
				$1 - \tau_{10}$	2000						0.01
					5000						0.00
0.03	0.10	0.15	0.05	q_1	2000	0.63	0.57	0.26	0.12	0.23	0.08
					5000	0.595	0.52	0.20	0.06	0.15	0.03
				q_2	2000	0.63	1.01	0.52	0.24	0.49	0.09
					5000	0.59	0.97	0.47	0.17	0.42	0.04
				q_3	2000	0.65	1.48	0.84	0.42	0.82	0.11
					5000	0.61	1.41	0.76	0.33	0.74	0.04
				q_4	2000	0.65	1.97	1.20	0.64	1.20	0.13
					5000	0.62	1.96	1.16	0.58	1.15	0.06
				q_5	2000	0.68	2.51	1.63	0.94	1.63	0.18
					5000	0.65	2.42	1.51	0.81	1.51	0.07
				$1 - \tau_{01}$	2000						0.10
					5000						0.04
				$1 - \tau_{10}$	2000						0.01
					5000						0.00

4.4.1 The Model

Suppose that subject i is examined at n time points $\left(t_{(i,1)}, \ldots, t_{(i,n)}\right)$, $i = 1, \ldots, m$. Now $Y_{(i,j)}$ is the true unobserved binary response for subject i at time $t_{(i,j)}$. As before, denote the vector of unobserved true binary responses for subject i by $\boldsymbol{Y}_i = \left(Y_{(i,1)}, \ldots, Y_{(i,n)}\right)$. Let $\boldsymbol{x}_{(i,j)} \in \mathbb{R}^k$ be a vector of covariates for subject i at examination j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. We assume that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ are independent vectors following monotone inhomogeneous first-order Markov models with parameters p_i and $\boldsymbol{q}_i = (q_{(i,1)}, \ldots, q_{(i,n-1)})$, which follow logistic regression models

$$\text{logit}(p_i) = \boldsymbol{w}_i^T \boldsymbol{\beta}_p,$$

and, for $j = 1, \ldots, n-1$,

$$\text{logit}(q_{(i,j)}) = \boldsymbol{z}_{(i,j)}^T \boldsymbol{\beta}_{q_j},$$

where $\boldsymbol{w}_i = \left(\boldsymbol{x}_{(i,1)}^T, t_{(i,1)}\right)$, $\boldsymbol{z}_{(i,j)} = \left(\boldsymbol{x}_{(i,j)}^T, t_{(i,j)}, t_{(i,j+1)} - t_{(i,j)}\right)$, and $\boldsymbol{\beta}_p \in \mathbb{R}^{k+1}$, $\boldsymbol{\beta}_{q_1} \in \mathbb{R}^{k+2}$, $\ldots$, $\boldsymbol{\beta}_{q_{n-1}} \in \mathbb{R}^{k+2}$ are regression coefficients associated with the prevalence and incidences, respectively.

Similar to the simple HMM, we assume that the response vector $\boldsymbol{Y}_i$ is prone to misclassification. Let $Y^*_{(i,j)}$ be the observed binary response at time $t_{(i,j)}$ and denote the vector of corrupted binary responses for subject i by $\boldsymbol{Y}^*_i = \left(Y^*_{(i,1)}, ..., Y^*_{(i,n)}\right)$. Here we suppose that the scoring is performed by Q examiners. Denote by $\xi_{(i,j)} \in \{1, \ldots, Q\}$ the indicator variable of examiner that scores subject i at time point $t_{(i,j)}$, and let $\boldsymbol{\xi}_i = \left(\xi_{(i,1)}, \ldots, \xi_{(i,n)}\right)$ be the vector of indicators of the examiners that score the responses of subject i over time. We also assume that the scoring behavior of the examiners is the same across the study. Let $\boldsymbol{\tau}_{01} = \left(\tau_{(1,01)}, \ldots, \tau_{(Q,01)}\right)$ and $\boldsymbol{\tau}_{10} = \left(\tau_{(1,10)}, \ldots, \tau_{(Q,10)}\right)$ be the vectors of misclassification parameters characterizing the examiners' scoring behavior. In this setting, the misclassification process is characterized by the following conditional independence assumptions:

B.1) $\perp\!\!\!\perp_{1 \le i \le m} \boldsymbol{Y}^*_i \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m, \boldsymbol{\xi}_1 \ldots, \boldsymbol{\xi}_m, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}$, i.e. the observed response patterns for each subject are independent given the true unobserved pattern, examiners indicators, and misclassification parameters,

B.2) $\boldsymbol{Y}^*_i \perp\!\!\!\perp \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{i-1}, \boldsymbol{Y}_{i+1}, \ldots, \boldsymbol{Y}_m \mid \boldsymbol{Y}_i, \boldsymbol{\xi}_i, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \forall\, i$, i.e. the distribution of the observed response pattern for a subject depends only on his true unobserved pattern, the examiners that score his responses, and the misclassification parameters,

B.3) $\perp\!\!\!\perp_{1 \le j \le n} Y^*_{(i,j)} \mid \boldsymbol{Y}_i, \boldsymbol{\xi}_i, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \forall\, i$, i.e. the observed responses for a subject are independent given his unobserved response pattern, the examiners that score his responses and the misclassification parameters, and

B.4) $Y^*_{(i,j)} \perp\!\!\!\perp \boldsymbol{Y}_i \mid Y_{(i,j)}, \xi_{(i,j)}, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \ \forall\ i,\ j$, i.e. the observed response for a subject at time $t_{i,j}$ depends only on the unobserved response and the examiner that scores that response at time $t_{(i,j)}$ and the misclassification parameters.

The above assumptions imply that the observed binary vectors are not *iid*. In this case, only the observed responses of subjects with the same covariate vectors $\boldsymbol{w}$, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n-1}$, and with the same examiner patterns $\boldsymbol{\xi}$, form an *iid* process. Now the likelihood function is given by

$$L(\boldsymbol{\beta}_p, \boldsymbol{\beta}_{q_1}, \ldots, \boldsymbol{\beta}_{q_{n-1}}, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}) = \prod_{i=1}^{m} P(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i \mid \boldsymbol{\beta}_p, \boldsymbol{\beta}_{q_1}, \ldots, \boldsymbol{\beta}_{q_{n-1}}, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \boldsymbol{\xi}_i)$$

$$= \prod_{i=1}^{m} \left\{ \sum_{\boldsymbol{y} \in V} P(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i, \boldsymbol{Y}_i = \boldsymbol{y} \mid \boldsymbol{\beta}_p, \boldsymbol{\beta}_{q_1}, \ldots, \boldsymbol{\beta}_{q_{n-1}}, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \boldsymbol{\xi}_i) \right\},$$

where $\boldsymbol{y}^*_i \in \{0,1\}^n$, and the joint distribution for the observed and latent binary variables is given by

$$\begin{aligned}
&P(\boldsymbol{Y}^*_i = \boldsymbol{y}^*_i, \boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\beta}_p, \boldsymbol{\beta}_{q_1}, \ldots, \boldsymbol{\beta}_{q_{n-1}}, \boldsymbol{\tau}_{10}, \boldsymbol{\tau}_{01}, \boldsymbol{\xi}_i) \\
&= \left[\tau^{1-y^*_{(i,1)}}_{(\xi_{(i,1)},01)} \left(1 - \tau_{(\xi_{(i,1)},01)}\right)^{y^*_{(i,1)}} p_1\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right) \right]^{y_{(i,1)}} \\
&\times \left[\tau^{y^*_{(i,1)}}_{(\xi_{(i,1)},10)} \left(1 - \tau_{(\xi_{(i,1)},10)}\right)^{1-y^*_{(i,1)}} p_2\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right) \right]^{1-y_{(i,1)}} \\
&\times \prod_{j=1}^{n-1} \left[\tau^{1-y^*_{(i,j+1)}}_{(\xi_{(i,j+1)},01)} \left(1 - \tau_{(\xi_{(i,j+1)},01)}\right)^{y^*_{(i,j+1)}} \right]^{y_{(i,j)}} \\
&\times \prod_{j=1}^{n-1} \left[\tau^{1-y^*_{(i,j+1)}}_{(\xi_{(i,j+1)},01)} \left(1 - \tau_{(\xi_{(i,j+1)},01)}\right)^{y^*_{(i,j+1)}} \delta_1\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) \right]^{y_{(i,j+1)}(1-y_{(i,j)})} \\
&\times \prod_{j=1}^{n-1} \left[\tau^{y^*_{(i,j+1)}}_{(\xi_{(i,j+1)},10)} \left(1 - \tau_{(\xi_{(i,j+1)},10)}\right)^{1-y^*_{(i,j+1)}} \delta_2\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) \right]^{(1-y_{(i,j+1)})(1-y_{(i,j)})},
\end{aligned}$$

where $\boldsymbol{y}^*_i \in \{0,1\}^n$, $\boldsymbol{y} \in V$, $p_1\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right) = \frac{\exp\{\boldsymbol{w}_i^T \boldsymbol{\beta}_p\}}{1+\exp\{\boldsymbol{w}_i^T \boldsymbol{\beta}_p\}}$, $p_2\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right) = \frac{1}{1+\exp\{\boldsymbol{w}_i^T \boldsymbol{\beta}_p\}}$, $\delta_1\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) = \frac{\exp\{\boldsymbol{z}^T_{(i,j)} \boldsymbol{\beta}_{q_j}\}}{1+\exp\{\boldsymbol{z}^T_{(i,j)} \boldsymbol{\beta}_{q_j}\}}$, and $\delta_2\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) = \frac{1}{1+\exp\{\boldsymbol{z}^T_{(i,j)} \boldsymbol{\beta}_{q_j}\}}$.

The identification results obtained in the previous section for the simple HMM, which strongly depend on the *iid* property, are not directly applicable for this extended model. For the current model, we conjecture that sufficient identification restrictions are that the design matrices $\boldsymbol{W}$, $\boldsymbol{Z}_1 \ldots, \boldsymbol{Z}_{n-1}$ are of full rank, and that $\tau_{(i,10)} + \tau_{(i,01)} < 1$, $\forall\ i \in \{1, \ldots, Q\}$. However, a formal evaluation of these restrictions is highly complicated due to the relatively complex likelihood function and the great number of potential scenarios. We do not provide any formal proof here and restrict ourselves to the empirical evaluation of the behavior of the estimators under the same setting of the ST study. Specifically, we evaluated the behavior of a Bayesian implementation of the model using simulated data. We opted for the Bayesian implementation of the model because of simplicity. As explained latter in this section, the use of a data augmentation algorithm renders the likelihood to the one arising from the product of independent logistic regression models which simplifies the computation and solved the numerical difficulties that we have observed in the direct (not data-augmented) maximization of the likelihood function with some simulated datasets. The results of the simulation study are presented after the introduction of the Bayesian model.

4.4.2 The Bayesian Implementation

A Bayesian version of the model requires the specification of prior distributions for the model parameters. Independent normal prior distributions were assumed for the logistic regression coefficients,

$$\boldsymbol{\beta}_p \sim N_{k+1}\left(\boldsymbol{b}_p, \boldsymbol{V}_p\right),$$

and, for $j = 1, \ldots, n-1$,

$$\boldsymbol{\beta}_{q_j} \sim N_{k+2}(\boldsymbol{b}_{q_j}, \boldsymbol{V}_{q_j}).$$

As a default choice prior specification of the prior covariance matrix of the regression coefficients we consider a suitably modified version of Zellner's g-prior (Zellner, 1983), originally developed as a "reference prior" for Gaussian linear models. Specifically we assume $\boldsymbol{V}_p = g_p(\boldsymbol{W}^T\boldsymbol{W})^{-1}$ and $\boldsymbol{V}_{q_j} = g_{q_j}(\boldsymbol{Z}_j^T\boldsymbol{Z}_j)^{-1}$, where g_p, $g_{q_1}, \ldots, g_{q_{n-1}}$ are positive constants. This specification produces priors that are scale invariant in terms of the predictors and takes into account the correlation in the coefficients induce by the design matrices. In our applications of the model we set $\boldsymbol{b}_p = \boldsymbol{0}_{k+1}$, $\boldsymbol{b}_{q_j} = \boldsymbol{0}_{k+2}$, and for the g-prior constants we have taken $g_p = g_{q_1}, \ldots = g_{q_{n-1}} = 2m$.

For the misclassification parameters, we assume independent beta distributions, under the restriction $\tau_{(i,01)} + \tau_{(i,10)} < 1$, i.e.,

$$\begin{aligned}\left(\tau_{(i,01)}, \tau_{(i,10)}\right) \sim \text{Beta}\left(\alpha^1_{(i,01)}, \alpha^2_{(i,01)}\right) \times \text{Beta}\left(\alpha^1_{(i,10)}, \alpha^2_{(i,10)}\right) \times \\ I\left(\tau_{(i,01)}, \tau_{(i,10)}\right)_{\{(\tau_{(i,01)}, \tau_{(i,10)}):\, \tau_{(i,01)} + \tau_{(i,10)} < 1\}},\end{aligned} \tag{4.2}$$

where $I(\cdot)_A$ is an indicator function for the set A. We next explain the computational strategy used for posterior sampling under the model. A function implementing the Markov chain Monte Carlo (MCMC) algorithm described here was written in a compiled language and incorporated into the R-program (R Development Core Team, 2009), which is available upon request to the authors. A Metropolis within Gibbs algorithm is used to generate samples from the posterior distribution. The MCMC algorithm is based on a data augmentation step treating the unobserved true responses $\boldsymbol{Y}_i$ as unknown parameters. The full conditionals for sampling the latent data are straightforward to derive. The introduction of latent data greatly simplify the computations. Indeed, given the latent data $\boldsymbol{Y}_i$, $i = 1, \ldots, m$, the full conditionals for the regression coefficients correspond to the one arising from logistic regression models with the priors previously described and applied to the corresponding subsets of the data. For the full conditional of the regression coefficients associated to the prevalence, $\boldsymbol{\beta}_p$, the vector formed for the responses $\left(Y_{(i,1)}, \ldots, Y_{(m,1)}\right)$ is used in a logistic regression setting. For the full conditional of the regression coefficients associated to the incidences, $\boldsymbol{\beta}_{q_j}$, the coordinates of the vector $\left(Y_{(i,j+1)}, \ldots, Y_{(m,j+1)}\right)$ for which $Y_{(i,j)} = 0$ are used in a logistic regression setting. Therefore, any sampling method developed for logistic regression parameters can be used to sample from these full conditional distributions. In our function we consider Metropolis-Hastings' steps based on multivariate normal proposals. The mean and the covariance matrix of these proposals are created based on a first order Taylor series approximation to the target distribution and on one step of the Newton-Raphson algorithm (see, e.g. Gamerman, 1997). The full conditionals for the latent data and the Metropolis-Hastings' steps for the logistic regression coefficients, are available in Sections C.4 and C.5 of Appendix C, respectively.

Finally, the full conditionals for the misclassification parameters are truncated beta distributions where the prior parameters are updated by the number of different errors incurred by the corresponding examiner, i.e.

$$\begin{aligned} \tau_{(i,01)} \mid \text{rest} \quad &\sim \quad \text{Beta}\left(\alpha^1_{(i,0)1} + n^1_{(i,01)}, \alpha^2_{(i,01)} + n^2_{(i,01)}\right) \times \\ &\qquad I\left(\tau_{(i,01)}\right)_{\{\tau_{(i,01)}:\tau_{(i,01)}<1-\tau_{(i,10)}\}}, \end{aligned}$$

and

$$\begin{aligned} \tau_{(i,10)} \mid \text{rest} \quad &\sim \quad \text{Beta}\left(\alpha^1_{(i,10)} + n^1_{(i,10)}, \alpha^2_{(i,10)} + n^2_{(i,10)}\right) \times \\ &\qquad I\left(\tau_{(i,10)}\right)_{\{\tau_{(i,10)}:\tau_{(i,10)}<1-\tau_{(i,01)}\}}, \end{aligned}$$

where

$$n^1_{(i,01)} = \sum_{l=1}^{m}\sum_{j=1}^{n} I\left(Y^*_{(l,j)}, Y_{(l,j)}\right)_{\{Y^*_{(l,j)}=0, Y_{(l,j)}=1\}} I\left(\xi_{(l,j)}\right)_{\{\xi_{(l,j)}=i\}},$$

$$n^2_{(i,01)} = \sum_{l=1}^{m}\sum_{j=1}^{n} I\left(Y^*_{(l,j)}, Y_{(l,j)}\right)_{\{Y^*_{(l,j)}=1, Y_{(l,j)}=1\}} I\left(\xi_{(l,j)}\right)_{\{\xi_{(l,j)}=i\}},$$

$$n^1_{(i,10)} = \sum_{l=1}^{m}\sum_{j=1}^{n} I\left(Y^*_{(l,j)}, Y_{(l,j)}\right)_{\{Y^*_{(l,j)}=1, Y_{(l,j)}=0\}} I\left(\xi_{(l,j)}\right)_{\{\xi_{(l,j)}=i\}},$$

and

$$n^2_{(i,10)} = \sum_{l=1}^{m}\sum_{j=1}^{n} I\left(Y^*_{(l,j)}, Y_{(l,j)}\right)_{\{Y^*_{(l,j)}=0, Y_{(l,j)}=0\}} I\left(\xi_{(l,j)}\right)_{\{\xi_{(l,j)}=i\}}.$$

4.4.3 The Simulation Study and Results

We evaluated the performance of the Bayesian implementation of the model introduced in the previous section using simulated data. We consider the same covariates used in the analysis of the ST data in Section 4.5 and the same examination structure as that of the ST study. The true values of the regression coefficients and misclassification parameters are given in Tables 4.2 and 4.3 (see pages 83 and 84), respectively. The true values for the regression coefficients were motivated by estimates observed in the real analysis of the ST data.

We simulated 1000 data sets with the previously described structure and fitted the above Bayesian model in each case. For the misclassification parameters we considered independent uniform priors, under the restriction $\tau_{(i,10)} + \tau_{(i,01)} < 1$, by taking $\alpha^1_{(i,10)} = \alpha^1_{(i,01)} = \alpha^2_{(i,10)} = \alpha^2_{(i,01)} = 1$ in (4.2). As a sensitivity analysis, we also consider a concentrated beta prior by taking $\alpha^1_{(i,10)} = \alpha^1_{(i,01)} = 0.5$ and $\alpha^2_{(i,10)} = \alpha^2_{(i,01)} = 4.5$. For each of the models, one Markov chain was generated. A conservative total number of 420000 scans of the Markov chain cycle were completed. Standard convergence tests (not shown), as implemented in the BOA R library (Smith, 2007), suggested that shorter chains can be considered. Because of storage limitations the full chain was subsampled every 20 steps after a burn-in period of 20000 samples, to give a reduced chain of length 20000.

Tables 4.2 and 4.3 (see pages 83 and 84, respectively) show the means, across simulation, the bias and the MSE of the posterior means of the model parameters. The results suggest that the regression parameters as well as the misclassification parameters are estimated with only a minimal bias and with a reasonable precision.

It is further seen that the regression coefficients of the incidences are estimated with greater error towards the end of the study. As expected, a greater variability in the estimates for the regression parameters as well as for the misclassification parameters is observed under the uniform prior for the misclassification parameters. Indeed, in 30 out of 47 regression parameters (64%), the MSE is greater under the uniform prior. The same result is observed for the misclassification parameters in 23 out of 32 estimates (72%). However, concentrated information on the misclassification parameters is not needed to obtain nearly unbiased and precise estimates for the regression coefficients and the misclassification parameters. Therefore, the results show that, under the setting of the ST study, the model parameters can be estimated from the raw data without extra information on the misclassification parameters.

4.5 The Analyses of the Signal-Tandmobiel® Data

In this section we show the results of the analyses of the ST study. Here the main focus is the estimation of the prevalence and the incidence of CE in permanent first molars and the evaluation of potential risk factors for CE.

4.5.1 Global incidence estimation

We estimated the prevalence and incidences using the simple HMM studied in Section 4.3. The analysis obtained using information of 2,281 children who participated in all six examinations of the ST study, gave CE prevalence estimates (95% confidence interval, 95%CI) of 3.6% (2.9 - 4.6%), 4.1% (3.3 - 5.0%), 4.2% (3.4 - 5.2%) and 4.2% (3.4 - 5.2%), for tooth 16, 26, 36, and 46, respectively, at the age of seven.

Figure 4.1 (see page 85) shows the estimates and 95%CI for the incidences in the 5 time intervals for the four first permanent molars. For comparison purposes, the estimates obtained using the early dental approaches are also presented. The CE incidences are lower than 7 percent in all cases. Further, we observed that the CE incidence was higher in the beginning than at the end of the study period. This can be explained by the following facts: i) when the molars emerge in the mouth (starting from the age of 6) their enamel is not yet fully developed and the tooth is more vulnerable to caries than later, ii) around the emergence time the molar is still partially covered by the gingiva, which makes brushing the tooth more complicated, or iii) during eruption the tooth remains for a prolonged period (up to 12 months and more) below the occlusal plane and, as a consequence, does not (fully) participate in the chewing process increasing the risk for plaque accumulation. The results also illustrate that different estimates can be obtained

Table 4.2: Simulated Data: true values, and Monte Carlo means, bias and mean squared error (MSE) of the posterior means of the logistic regression parameters under a Beta(1,1) and Beta(0.5,4.5) prior for the misclassification parameters, respectively.

Parameter	Covariate	True Value	Beta(1,1)			Beta(0.5,4.5)		
			Mean	Bias	MSE	Mean	Bias	MSE
Prevalence	Intercept	-9.21	-8.41	0.81	21.69	-9.10	0.12	16.47
	Gender	0.33	0.34	0.01	0.17	0.41	0.08	0.16
	Age	0.93	0.87	-0.06	0.34	0.97	0.03	0.27
	x-ordinate	-0.03	-0.04	-0.00	0.15	-0.04	-0.00	0.11
	y-ordinate	-0.85	-1.01	-0.16	1.88	-1.10	-0.25	1.64
	Age Start Brushing	0.22	0.19	-0.03	0.03	0.18	-0.04	0.03
	Meals	0.08	-0.03	-0.11	0.18	0.10	0.01	0.19
Incidence 1	Intercept	-8.97	-9.42	-0.44	30.49	-8.82	0.16	24.84
	Gender	-0.07	-0.06	0.01	0.17	-0.13	-0.06	0.16
	Age	0.82	0.78	-0.04	0.37	0.82	0.00	0.26
	x-ordinate	0.34	0.35	0.02	0.16	0.38	0.04	0.14
	y-ordinate	-0.20	-0.01	0.18	1.55	-0.19	0.00	1.13
	Days Between Exam.	0.00	0.00	0.00	0.00	0.00	-0.00	0.00
	Age Start Brushing	-0.10	-0.12	-0.02	0.03	-0.14	-0.04	0.04
	Meals	0.24	0.21	-0.03	0.15	0.17	-0.07	0.17
Incidence 2	Intercept	-2.40	-2.22	0.18	31.91	-3.86	-1.46	25.58
	Gender	0.46	0.50	0.03	0.21	0.44	-0.03	0.13
	Age	-0.43	-0.49	-0.06	0.25	-0.29	0.14	0.30
	x-ordinate	-0.11	-0.09	0.02	0.15	-0.06	0.06	0.16
	y-ordinate	-0.48	-0.53	-0.04	1.39	-0.46	0.03	1.79
	Days Between Exam.	0.01	0.01	0.00	0.00	0.01	0.00	0.00
	Age Start Brushing	0.14	0.12	-0.02	0.03	0.12	-0.02	0.03
	Meals	0.14	0.04	-0.10	0.17	0.06	-0.08	0.18
Incidence 3	Intercept	-6.89	-6.76	0.13	39.02	-5.78	1.11	37.77
	Gender	0.23	0.21	-0.02	0.30	0.29	0.06	0.25
	Age	0.19	0.15	-0.04	0.43	0.18	-0.07	0.46
	x-ordinate	0.85	0.80	-0.05	0.23	0.80	-0.06	0.26
	y-ordinate	-1.15	-0.89	0.26	2.47	-1.10	0.05	1.77
	Days Between Exam.	0.01	0.01	-0.00	0.00	0.01	-0.00	0.00
	Age Start Brushing	0.04	0.00	-0.03	0.06	-0.00	-0.04	0.05
	Meals	0.42	0.27	-0.15	0.39	0.23	-0.19	0.40
Incidence 4	Intercept	-0.82	0.99	1.80	44.39	-0.24	0.58	36.70
	Gender	0.04	0.04	0.00	0.25	0.03	-0.00	0.22
	Age	-0.38	-0.45	-0.06	0.37	-0.40	-0.02	0.26
	x-ordinate	0.81	0.87	0.06	0.24	0.84	0.03	0.21
	y-ordinate	-0.65	-1.14	-0.49	2.48	-0.74	-0.09	1.76
	Days Between Exam.	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00
	Age Start Brushing	0.39	0.36	-0.03	0.03	0.33	-0.07	0.04
	Meals	-0.04	-0.10	-0.05	0.25	-0.07	-0.03	0.21
Incidence 5	Intercept	-5.07	-5.31	-0.25	101.91	-4.83	0.23	63.23
	Gender	-0.10	-0.21	-0.11	0.40	-0.08	0.02	0.39
	Age	0.24	0.23	-0.01	0.64	0.25	0.01	0.43
	x-ordinate	0.20	0.26	0.06	0.32	0.27	0.07	0.35
	y-ordinate	-1.75	-1.01	0.74	3.61	-1.43	0.32	2.26
	Days Between Exam.	0.01	0.00	-0.00	0.00	0.00	-0.00	0.00
	Age Start Brushing	-0.21	-0.17	0.04	0.06	-0.17	0.04	0.07
	Meals	1.44	0.77	-0.67	0.83	0.81	-0.62	0.72

Table 4.3: Simulated Data: true values, and Monte Carlo means, bias ($\times 10$) and mean squared error (MSE $\times 10^3$) of the posterior means of the sensitivity $(1-\tau_{01})$ and specificity $(1-\tau_{10})$ for each examiner, under a Beta(1,1) and Beta(0.5,4.5) prior for the misclassification parameters, respectively.

Examiner	**True Value**	**Beta(1,1)**			**Beta(0.5,4.5)**		
		Mean	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
			$1-\tau_{01}$				
1	0.95	0.86	0.91	12.87	0.94	0.12	1.86
2	0.95	0.89	0.62	7.06	0.94	0.13	1.67
3	0.95	0.93	0.25	1.59	0.94	0.09	0.83
4	0.95	0.92	0.31	2.42	0.95	0.04	0.99
5	0.90	0.86	0.37	4.24	0.90	0.01	2.46
6	0.90	0.87	0.26	3.07	0.90	0.05	2.67
7	0.90	0.86	0.36	4.43	0.89	0.07	2.49
8	0.90	0.86	0.40	5.63	0.91	-0.08	3.11
9	0.85	0.66	1.91	56.25	0.88	-0.34	4.75
10	0.85	0.81	0.42	5.82	0.86	-0.08	3.19
11	0.85	0.81	0.36	6.37	0.84	0.14	3.01
12	0.85	0.82	0.34	7.46	0.86	-0.05	4.07
13	0.92	0.89	0.28	2.89	0.91	0.07	1.96
14	0.92	0.86	0.60	7.08	0.92	-0.01	2.01
15	0.92	0.90	0.24	1.93	0.91	0.07	1.35
16	0.92	0.90	0.23	1.80	0.91	0.07	1.37
			$1-\tau_{10}$				
1	0.90	0.89	0.07	0.71	0.90	-0.02	0.72
2	0.90	0.90	0.04	0.50	0.90	0.00	0.30
3	0.90	0.90	0.01	0.16	0.90	0.00	0.12
4	0.90	0.90	0.00	0.22	0.90	0.01	0.24
5	0.95	0.95	0.01	0.16	0.95	0.02	0.18
6	0.95	0.95	0.04	0.11	0.95	0.02	0.13
7	0.95	0.95	0.00	0.12	0.95	0.00	0.11
8	0.95	0.95	0.02	0.19	0.95	0.01	0.24
9	0.92	0.91	0.10	0.66	0.92	0.00	0.67
10	0.92	0.92	0.00	0.23	0.92	0.00	0.25
11	0.92	0.92	0.01	0.27	0.92	0.02	0.23
12	0.92	0.92	-0.01	0.27	0.92	0.02	0.18
13	0.85	0.85	0.01	0.30	0.85	0.00	0.24
14	0.85	0.84	0.06	0.51	0.85	-0.04	0.56
15	0.85	0.85	0.01	0.16	0.85	0.01	0.20
16	0.85	0.85	0.02	0.19	0.85	0.01	0.16

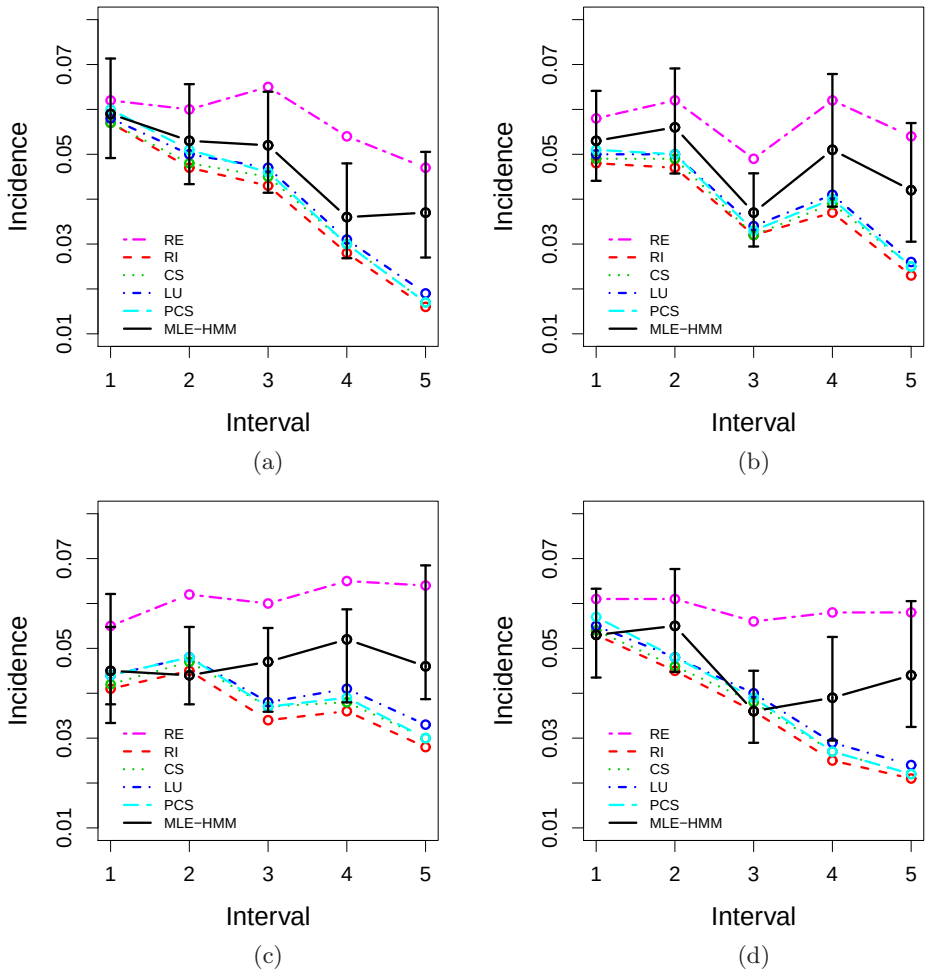


Figure 4.1: Caries experience incidence estimates for molars 16 (panel a), 26 (panel b), 36 (panel c) and 46 (panel d). The estimates and 95% confidence interval for the maximum likelihood estimator in the simple hidden Markov model (MLE-HMM) are presented in black. Point estimates associated to the early approaches (RE, RI, CS, LU and PCS) are also presented.

by using the early approaches proposed in the literature. The estimates based on the early approaches may even lie outside the asymptotic confidence intervals obtained by maximum likelihood estimation. In agreement with what we observed in the simulation study, the RE method tends to overestimate the incidence while the other early approaches underestimate the incidence.

4.5.2 Evaluation of the effect of covariates

We fitted the extended HMM proposed in Section 4.4 under the Bayesian framework using the same priors and the same MCMC specification as in Section 4.4.3. We evaluated the effect of gender (boys vs girls; `Gender`), age at start of brushing (in years; `Startbr`), the number of between-meal snacks (two or less than two a day vs more than two a day; `Meals`), the geographical location, represented by the standardized (x, y) co-ordinate of the municipality of the school to which the child belongs (`x-ordinate` and `y-ordinate`), and $t_{(i,1)}$, corresponding to the age (in years; `Age`) at first examination, on the prevalence of CE. For the incidence parameters, we evaluated the effect of gender, age at start of brushing, the x- and y-ordinate, the number of between-meal snacks at the beginning of the time interval, $t_{(i,j)}$ corresponding to the age (in years) at the beginning of the time interval, and $\left(t_{(i,j+1)} - t_{(i,j)}\right)$ corresponding to the time between examinations (in years; `Years-exam`).

Initially two models were fitted. In Model 1 we assume different intercepts and covariate effects on the incidence parameters. Model 2 is a reduced version of Model 1 with different intercepts but where the covariates are assumed to have a constant effect on the incidences. The choice between the models was based on the Bayesian information criterion (BIC) (Schwarz, 1978). The results suggest that the reduced version of the model is preferred. For instance, for tooth 26, the BIC for Model 1 and Model 2 was 3696.5 and 3526.4, respectively, suggesting no evidence of a time-dependent effect of the covariates. Therefore, we only report the results based on Model 2. For ease of exposition, we show the results for tooth 26. Similar results were observed for the other teeth under consideration.

Posterior means and 95% highest posterior density (95% HPD) credible intervals, computed as proposed by Chen & Shao (1999), for the logistic regression coefficients are displayed in Table 4.4 (see page 87). The results show that the older the child at the beginning of the study the higher the prevalence of CE. The results also show that the higher the number of between-meal snacks the higher the incidence of CE. Furthermore, a significant effect of the x-ordinate was observed on the incidences, suggesting the existence of an east-west gradient for the incidence of CE in permanent molars in Flanders. Although a non-significant effect of the x-ordinate was observed on the prevalence for permanent molars at approximately the age of seven, which can be explained by the relatively short exposition of the

Table 4.4: Posterior means and 95% highest posterior density (95% HPD) credible intervals, for the logistic regression coefficients associated to the prevalence and incidences for tooth 26.

	Covariate	Posterior Mean	95%HPD
Prevalence	Intercept	-10.13	(-17.30; -3.32)
	Gender	0.32	(-0.34; 0.97)
	Age	1.06	(0.17; 1.95)
	x-ordinate	-0.05	(-0.71; 0.61)
	y-ordinate	-0.76	(-2.87; 1.34)
	Startbr	0.20	(-0.07; 0.49)
	Meals	0.05	(-0.65; 0.75)
Incidences	Intercept 1	-3.63	(-5.26; -1.95)
	Intercept 2	-3.45	(-5.18; -1.76)
	Intercept 3	-3.73	(-5.61; -1.94)
	Intercept 4	-3.39	(-5.45; -1.51)
	Intercept 5	-3.70	(-5.91; -1.66)
	Gender	0.13	(-0.16; 0.42)
	Age	-0.05	(-0.30; 0.22)
	x-ordinate	0.33	(0.05; 0.62)
	y-ordinate	-0.78	(-1.65; 0.08)
	Years-exam	1.49	(0.30; 2.66)
	Startbr	0.11	(-0.02; 0.24)
	Meals	0.36	(0.05; 0.67)

teeth at this age, the significant regional effect on the incidences induce a significant regional effect on the prevalence at the end of the study. This geographical gradient has also been observed on the prevalence of CE in primary dentition, with a similar period of exposition, by Mwalili et al. (2005), where an analysis of aggregate data was performed using an ordinal logistic regression model for correcting for misclassification. Figure 4.2 (see page 88) shows the misclassification parameters for each examiner for molar 26. In general, the examiners were more specific than sensitive. The results suggest a greater variability in the sensitivity estimates which is explained by the low prevalence and incidences of CE. All examiners showed a sensitivity greater than 0.8, with the exception of examiner 9 who showed a rather poor scoring behavior. The latter result is explained by the lower information available for this examiner. In fact, examiner 9 was only involved in the first two years of the ST study.

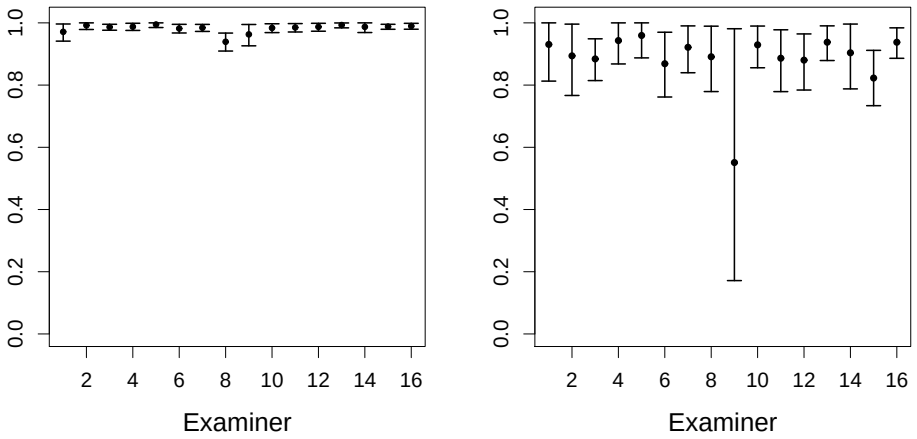


Figure 4.2: Posterior means and 95% highest posterior density credible intervals for examiners' specificity (panel a) and sensitivity (panel b), respectively.

4.6 Concluding Remarks

We have studied the properties and evaluated the performance of simple HMM for monotone binary processes. We showed that the associated MLE always outperform the early approaches proposed in the dental literature for incidence estimation under the presence of misclassified data. We also showed that restrictions in the parameter space are needed for the identification of the model parameters. Specifically, we showed that in a simple HMM for two time points and under the assumption $\tau = \tau_{10} = \tau_{01}$, and in a simple HMM for more than two time points and assuming that $\tau_{10} \neq \tau_{01}$, the parameters are unidentified by the data even though the number of free equations matches the number of parameters in the model.

We provided theoretical and empirical evidence showing that, under some restrictions on the parameter space, the parameters in a simple HMM are identified and can be estimated from the raw data, thus avoiding the need of external information on the misclassification parameters. Because the external information on the misclassification parameters is in general difficult to obtain, the simple HMM has a clear advantage over the models for cross-sectional data which require of such an input. We noted that if external information on the misclassification parameters is available, this can be easily incorporated into the HMM specification leading to more efficient estimates of the parameters. Further, the fact that the misclassification parameters can be estimated from the main data allows us to test

whether the misclassification parameters in the main and available validation data are equal, implying the existence of real internal validation data.

We proposed an extension of the simple HMM model in order to describe the relationships between covariates and the prevalence and incidence, and where different classifiers are present. In addition, we developed a Bayesian version of the extended model and showed empirically that, under the settings of our motivating example, the parameters can be estimated without any external information. The results suggest that even under the use of uniform priors on the misclassification parameters, unbiased and precise estimates of the parameters can be obtained. The formal identification analysis for this extension of the model is the subject of current research.

Several extensions of this work can be done. For instance, relaxing the assumptions A.3 or B.3 can be of interest. This was suggested by one of the reviewers and it might be justified by the existence of easy/difficult to diagnose subjects. We are currently working on a multivariate extension of the extended model, and exploring the connection between HMM and survival models. Since, the prevalence and incidences can be written as functions of the survival function for the time to event, the resulting model corresponds to a model for misclassified survival data. Finally, the extension of the results for models for multinomial data is also the subject of ongoing research.

Acknowledgements

The first author is supported by the National Scholarship for Doctoral Studies 2009, Conicyt (Chile). The first, second and last authors are supported by the Research Grant OT/05/60 and they also acknowledge the partial support from the Interuniversity Attraction Poles Program P6/03, Belgian State, Federal Office for Scientific, Technical and Cultural Affairs. The third author is partially supported by the Fondecyt grant 3095003. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (L-BioStat, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

Albert, P. S., Hunsberger, S. A. & Biro, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with

applications to studying maturation. *Journal of American Statistical Association* 92 1304–1311.

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

Carlos, J. P. & Senning, R. S. (1968). Error and bias in dental clinical trials. *Journal of Dental Research* 47 142–148.

Chen, M. H. & Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational Graphical Statistics* 8(1) 69–92.

Cook, R. J., Ng, E. T. M. & Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* 56 1109–1117.

Espeland, M. A., Murphy, W. C. & Leverett, D. H. (1988). Assessing diagnostic reliability and estimating incidence rates associated with a strictly progressive disease: dental caries. *Statistics in Medicine* 7 403–416.

Espeland, M. A., Platt, O. S. & Gallagher, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. *Journal of the American Statistical Association* 84(408) 972–979.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 7 57–68.

Geng, Z. & Asano, C. (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics* 8 2935–2954.

Jurek, A. M., Greenland, S., Maldonado, G. & Church, T. R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *International Journal of Epidemiology* 34 680–687.

Küchenhoff, H., Mwalili, S. M. & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 2 85–96.

Lu, K. H. (1968). A critical evaluation of diagnostic errors, true increment and examiner's accuracy in caries experience assessment by a probabilistic model. *Archives of Oral Biology* 13 1133–1147.

Magder, L. S. & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 146 195–203.

Mwalili, S. M., Lesaffre, E. & 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* 54(1) 77–93.

Nagelkerke, N. J. D., Chunge, R. N. & Kinot, S. N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 9 1211–1219.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86(4) 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* 58 675–683.

Pitts, N. B., Evans, D. J. & 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.

Poole, W. K., Clayton, C. A. & Shah, B. V. (1973). The estimation of examiner error and the true transition probabilities for teeth or surfaces in dental clinical trials. *Archives of Oral Biology* 18 1291–1302.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Radike, A. W. & Muhler, J. C. (1954). Effect of reversals in diagnosis on the interpretation of clinical dental caries tests. *Journal of Dental Research* 33 682 (abstract).

Radike, A. W., (1960). A study of reversals in diagnosis of carious lesions. In G. E. Green & P. R. Weinstein eds., *Caries diagnosis and experimental caries conference*. Columbus, USA: Ohio State University Research Foundation, 201–209.

Rekaya, R., Weigel, K. A. & Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics* 57 1123–1129.

Rosychuk, R. J. & Islam, M. S. (2009). Parameter estimation in a model for misclassified Markov data - a Bayesian approach. *Computational Statistics and Data Analysis* 53 3805–3816.

Rosychuk, R. J. & Thompson, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 19 394–404.

Rosychuk, R. J. & Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification *Statistics in Medicine* 22 2035–2055.

Schmid, C. H., Segal, M. R. & Rosner, B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference* 42(1–2) 1–18.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2) 461–464.

Singh, A. C. & Rao, J. N. K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey. *Journal of the American Statistical Association* 90(430) 478–488.

Smith, B. J. (2007). An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software* 21(11) 1–37.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* 65(331) 1350–1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. *Biometrics* 27 935–944.

Vanobbergen, J., Martens, L., Lesaffre, E. & 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.

Zellner, A. (1983). Applications of Bayesian analysis in Econometrics. *The Statistician* 32 23–34.

Chapter 5

Modelling of Multivariate Monotone Disease Processes in the Presence of Misclassification

A reduced version of this chapter has been published as:

García-Zattera, M. J., Jara, A., Lesaffre, E. & Marshall, G. (2010). Multivariate modelling of a monotone disease process in the presence of misclassification. In A. Bowman, ed., *Proceedings of the 25th Workshop of Statistical Modelling*. Glasgow: University of Glasgow, 221–226.

The complete version is under revision for publication as:

García-Zattera, M. J., Jara, A., Lesaffre, E. & Marshall, G. (2010). Modelling of multivariate monotone disease processes in the presence of misclassification. Invited re-submission to *Journal of the American Statistical Association*.

Abstract

Motivated by a longitudinal oral health study, the Signal-Tandmobiel®, we propose a multivariate binary inhomogeneous Markov model in which unobserved correlated response variables are subject to an unconstrained misclassification process and have a monotone behavior. The multivariate baseline distributions and Markov transition matrices of the unobserved processes are defined as a function of covariates, throughout the specification of compatible full conditional distributions. Distinct misclassification models are discussed. In all cases, the existence of different classifiers for each subject across time is taken into account. A full Bayesian implementation of the model is described and its performance is evaluated using simulated data. We show that under the settings of our motivating example, the parameters can be estimated without any external information on the misclassification parameters. Finally, the analyses of the motivating problem are presented.

Keywords: Multivariate binary data; Monotone binary processes; Misclassification; Incidence estimation; Hidden Markov model.

5.1 Introduction

Based on dental data gathered in a longitudinal oral health study, the Signal-Tandmobiel® (ST) study, we aim to asses potential risk factors for the incidence of caries experience (CE), which is defined as a binary variable indicating whether a tooth is decayed at d_3 level (see, e.g. Reis et al., 2006), missing or filled due to caries. This involves the analysis of a misclassified multivariate monotone binary process since: (i) CE, as previously defined, is a progressive or monotone disease because teeth cannot alternate between the presence or absence of CE once CE occurs over time, (ii) events on teeth of the same child are dependent, and (iii) several examiners were involved in the study and their CE scoring may not perfectly reflect the tooth's true condition and, therefore, the presence of CE can be misdiagnosed leading to misclassified outcomes.

The effect of response misclassification on estimation and hypothesis testing has been widely investigated in the literature (see, e.g. Küchenhoff, 2009; Buonaccorsi, 2010). For regression models, non-differential (covariate independent) misclassification, can cause the estimates of the regression coefficients to be attenuated strongly towards to the null and that, although the associated significance tests are still valid, its power may be drastically reduced. Under differential (covariate dependent) misclassification, the bias of the estimates can be in both directions, leading to an apparent effect or an apparent lack of effect

of the covariate when the reverse is true (see, e.g. Bross, 1954; Tenenbein, 1970, 1971; Magder & Hughes, 1997; Neuhaus, 1999; Jurek et al., 2005).

In cross-sectional studies, where the data contain no information regarding the misclassification parameters, several strategies have been proposed in the literature for correcting for misclassification (see, e.g. Geng & Asano, 1989; Magder & Hughes, 1997; Neuhaus, 1999, 2002; Rekaya et al., 2001; Mwalili et al., 2005; Küchenhoff et al., 2006). In the context of longitudinal univariate categorical data, generalized linear mixed models (see, e.g. Neuhaus, 2002), generalized estimating equation (GEE) based approaches (see, e.g. Neuhaus, 1999), and transition models (see, e.g. García-Zattera et al., 2010) have been proposed for correcting for misclassification. Due to the monotone nature of our motivating problem and because the main scientific objective here is the incidence estimation, we restrict ourselves to the latter class of models, where the parameters have a direct interpretation in terms of the conditional probabilities of developing CE in a given time-interval.

Hidden Markov models (HMM) for the analysis of misclassified alternating longitudinal responses has been considered in the literature by Cook et al. (2000), Rosychuk & Thompson (2001), Rosychuk & Thompson (2003), Nagelkerke et al. (1990), and Rosychuk & Islam (2009), whereas Espeland et al. (1988), Espeland et al. (1989), Schmid et al. (1994), Singh & Rao (1995), Albert et al. (1997), and García-Zattera et al. (2010) addressed the problem of misclassified monotone longitudinal responses. It is important to stress that in a longitudinal setting, unlike cross-sectional studies, the model parameters might be estimated without the use of external information about the misclassification parameters. For instance, García-Zattera et al. (2010) showed that under simple restrictions on the parameter space, the model parameters associated to an inhomogeneous HMM for monotone responses are identified by the available data. They also proposed a univariate model to account for predictors allowing for irregulary spaced time intervals and different classifiers. This development was motivated by the analysis of the ST data. However, the proposed model only allows for univariate longitudinal processes and, therefore, the analyses were performed for each tooth separately, with a potential loss of power in detecting significant effects.

In this paper, we propose an extension of the univariate HMM proposed by García-Zattera et al. (2010) to the multivariate case, thus providing a general framework for analysing multivariate hidden monotone processes as a function of covariates. Specifically, we define multivariate binary distributions associated to the monotone Markov processes by specifying compatible Bernoulli conditional distributions with the conditional probabilities expressed as logistic regression models (Joe & Liu, 1996; García-Zattera et al., 2007). Different misclassification models allowing for different classifiers for each subject across examinations are discussed.

The rest of the paper is organized as follows. Section 5.2 introduces the ST study

and the research questions. The proposed model is introduced in Section 5.3. The Bayesian implementation of the model along with the results of the evaluation of its performance using simulated data is given in Section 5.4. The proposed model is applied to our motivating problem in Section 5.5. A final discussion section concludes the paper.

5.2 The Signal-Tandmobiel® study and research questions

The ST study is a longitudinal prospective oral health screening study conducted in Flanders, Belgium, between 1996 and 2001. For this project, 4468 children were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen trained and calibrated dental examiners. Clinical data were collected by the examiners based on visual and tactile observations (no X-rays were taken), and data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents. For a more detailed description we refer to Vanobbergen et al. (2000).

Caries lesions were scored in four levels of lesion severity: d_4 (dentine caries with pulpal involvement), d_3 (dentine caries with obvious cavitation), d_2 (hidden dentine caries) and d_1 (white or brown-spot initial lesions in enamel without cavitation). Here we consider CE as a binary variable indicating whether the tooth is decayed at d_3 level, missing or filled due to caries, which defines a progressive disease. Thus observed reversals, i.e. teeth or surfaces initially recorded as being carious and subsequently recorded as caries-free, represent diagnostic errors. The diagnosis of CE might be difficult for a variety of reasons. For instance, nowadays composite materials can imitate the natural enamel so well that it is sometimes difficult to spot a restored lesion. Another reason may be that the location of the cavity e.g. 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.

The statistical findings reported below were applied to the scoring of the four permanent first molars, i.e., teeth 16, 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). The numbering of the teeth follows the FDI (Federation Dentaire Internationale) notation which indicates the position of the tooth in the mouth. Position 26, for instance, means that the tooth is in quadrant 2 (upper left quadrant) and position 6 where numbering starts from the mid-sagittal plane.

The purpose of the present investigation is to asses the effect of potential risk factors on the prevalence and incidence of CE, and to study the (within- and across-time) association structures. Since the main objective is the CE incidence

estimation, the model building strategy is based on transition models, where the covariates are included at the level of the Markov transition matrices instead of marginalized versions (see, e.g. Azzalini, 1994; Heagerty & Zeger, 2000; Heagerty, 2002), where the main focus is the assesment of the effect of covariates on the marginal means.

5.3 The multivariate hidden Markov model

In this section we introduce the proposed model. The elements of the model are discussed in a sequential manner. Section 5.3.1 contains the development of the multivariate Markov model, while Section 5.3.2 discusses the misclassification component of the model. Finally, these components are pooled to produce the statistical model in Section 5.3.3.

5.3.1 The multivariate Markov model

Suppose that J teeth are examined for the ith subject, $i = 1, \ldots, I$, at time points $t_{(i,k)}$, $k = 1, \ldots, K$. Let $Y_{(i,j,k)}$ be the true unobserved binary response for tooth j of subject i at time $t_{(i,k)}$ and let $\boldsymbol{Y}_{(i,k)} = (Y_{(i,1,k)}, \ldots, Y_{(i,J,k)})'$ be the J-dimensional vector of true responses for all teeth of subject i at time $t_{(i,k)}$. Let $\boldsymbol{x}_{(i,j)}$ be a p-dimensional vector of exogenous covariates associated with the first examination of jth tooth of the ith subject, and let $\boldsymbol{z}_{(i,j,k)}$, $k = 2, 3, \ldots$, be q-dimensional vectors of exogenous but possibly time-varying covariates associated to the jth tooth of the ith subject. Further, set $\boldsymbol{X}_i = \left(\boldsymbol{x}_{(i,1)}, \ldots, \boldsymbol{x}_{(i,J)}\right)'$ and $\boldsymbol{Z}_{(i,k)} = \left(\boldsymbol{z}_{(i,1,k)}, \ldots, \boldsymbol{z}_{(i,J,k)}\right)'$, $k = 1, \ldots, K$. We assume that the vectors $\boldsymbol{Y}_{(i,k)}$ follow independent multivariate monotone inhomogeneous first-order Markov processes. In order to relate the covariates to the initial distributions and elements of the transition matrices, and to evaluate the association among the responses of the same subject, multivariate distributions are defined by the specification of their full conditional distributions (see, e.g. Arnold et al., 1992).

Following Joe & Liu (1996), we assume that, for $j = 1, \ldots, J$, the conditional distribution of the corresponding binary response at the first examination, $Y_{(i,j,1)}$, given the other binary responses $Y_{(i,l,1)} = y_l$, $\forall l \neq j$, and the covariates $\boldsymbol{x}_{(i,j)}$, is a Bernoulli distribution with probability following the logistic regression model

$$\text{logit}\left[P_{\boldsymbol{X}_i}\left(Y_{(i,j,1)} = 1 \mid \boldsymbol{\beta}_j^P, \boldsymbol{\gamma}^P, Y_{(i,l,1)} = y_l, \forall l \neq j\right)\right] = \boldsymbol{x}'_{(i,j)}\boldsymbol{\beta}_j^P + \sum_{l \neq j} \gamma_{jl}^P y_l, \quad (5.1)$$

where $\boldsymbol{\beta}_j^P$ is a p-dimensional vector of logistic regression coefficients for the jth tooth, and γ_{jl}^P are conditional log-odds ratio parameters given by

$$\begin{aligned}\exp\{\gamma_{jl}^P\} &= \frac{P_{\boldsymbol{X}_i}(Y_{(i,j,1)}=1, Y_{(i,l,1)}=1 | Y_{(i,m,1)}=y_m, \forall m \neq j,l)}{P_{\boldsymbol{X}_i}(Y_{(i,j,1)}=1, Y_{(i,l,1)}=0 | Y_{(i,m,1)}=y_m, \forall m \neq j,l)} \times \\ &\quad \frac{P_{\boldsymbol{X}_i}(Y_{(i,j,1)}=0, Y_{(i,l,1)}=0 | Y_{(i,m,1)}=y_m, \forall m \neq j,l)}{P_{\boldsymbol{X}_i}(Y_{(i,j,1)}=0, Y_{(i,l,1)}=1 | Y_{(i,m,1)}=y_m, \forall m \neq j,l)}.\end{aligned}$$

Joe & Liu (1996) showed that the conditional probability distributions given by expression (5.1) are compatible if and only if $\gamma_{jl}^P = \gamma_{lj}^P$, for all $j \neq l$. Under these restrictions, the joint distribution for the initial time point is given by

$$\begin{aligned}P_{\boldsymbol{X}_i}(\boldsymbol{Y}_{(i,1)} = \boldsymbol{y} \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P) &= c_1(\boldsymbol{X}_i, \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P)^{-1} \times \\ &\quad \exp\left\{ \sum_{j=1}^{J} \left(\boldsymbol{x}'_{(i,j)} \boldsymbol{\beta}_j^P \right) y_j + \sum_{1 \leq j < l \leq J} \gamma_{jl}^P y_j y_l \right\}, \quad (5.2)\end{aligned}$$

where $\boldsymbol{y} \in \{0,1\}^J$, $\boldsymbol{\beta}^P = (\boldsymbol{\beta}_1^P, \ldots, \boldsymbol{\beta}_J^P)'$, $\boldsymbol{\gamma}^P = \{\gamma_{jl}^P : 1 \leq j < l \leq J\}$, and $c_1(\boldsymbol{X}_i, \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P)$ is a normalizing constant given by

$$c_1(\boldsymbol{X}_i, \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P) = \sum_{y_1=0}^{1} \cdots \sum_{y_J=0}^{1} \exp\left\{ \sum_{j=1}^{J} \left(\boldsymbol{x}'_{(i,j)} \boldsymbol{\beta}_j^P \right) y_j + \sum_{1 \leq j < l \leq J} \gamma_{jl}^P y_j y_l \right\}.$$

In order to model the conditional joint distributions associated with the transition matrices of the Markov processes, we extend the approach of Joe & Liu (1996). The monotone nature of the process implies that $2^J - 1$ rows of the transition matrices contain structural zeros. In fact, each row corresponds to the conditional joint distribution $P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}\right)$, with support $\mathcal{B}\{\boldsymbol{y}^{k-1}\} \subset \{0,1\}^J$ defined by the realizations $\boldsymbol{y}^{k-1}$ of the response vector in the previous examination $\boldsymbol{Y}_{(i,k-1)}$. For ease of exposition, consider the case of $J = 2$ binary response variables $\boldsymbol{Y}_{(i,k)} = \left(Y_{(i,1,k)}, Y_{(i,2,k)}\right)'$. In this case, the form of the transition matrices is shown in Figure 5.1 (see page 99).

As shown in Figure 5.1, $\mathcal{B}\{(0,0)\} = \{0,1\}^2$, $\mathcal{B}\{(0,1)\} = \{(0,1),(1,1)\}$, $\mathcal{B}\{(1,0)\} = \{(1,0),(1,1)\}$, and $\mathcal{B}\{(1,1)\} = \{(1,1)\}$ for the two-dimensional example.

Let $\boldsymbol{y}_{[-j]} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_J)$ be a $(J-1)$-dimensional vector resulting by removing the jth coordinate of $\boldsymbol{y}$. We assume that conditional on the

$$
\begin{array}{cc} & \boldsymbol{Y}_{(i,k)} \\ & \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \boldsymbol{Y}_{(i,k-1)} \begin{array}{c} (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} & \left[\begin{array}{cccc} \Pi_{11}^k & \Pi_{12}^k & \Pi_{13}^k & \Pi_{14}^k \\ 0 & \Pi_{22}^k & 0 & \Pi_{24}^k \\ 0 & 0 & \Pi_{33}^k & \Pi_{34}^k \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array}
$$

Figure 5.1: Illustration of a valid transition matrix Π^k in a bivariate monotone Markov model.

design vector of covariates $\boldsymbol{z}_{(i,j,k)}$, $\boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}$ and $\boldsymbol{Y}_{(i,k)[-j]} = \boldsymbol{y}^k_{[-j]}$, for all $\boldsymbol{y}^k \in \mathcal{B}\{\boldsymbol{y}^{k-1}\}$, $Y_{(i,j,k)}$ follows a Bernoulli distribution with probability $\pi_j\left(\boldsymbol{y}^{k-1}, \boldsymbol{y}^k_{[-j]} \mid \boldsymbol{\beta}^I_j, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right)$, with

$$
\pi_j\left(\boldsymbol{y}^{k-1}, \boldsymbol{y}^k_{[-j]} \mid \boldsymbol{\beta}^I_j, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right) = \left\{ h\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}^I_j + \sum_{l \neq j} \gamma^I_{lj} y^k_l + \sum_{l \neq j} \alpha^I_{lj} y^{k-1}_l \right) \right\}^{1-y^{k-1}_j}, \quad (5.3)
$$

where $h(\cdot) = \exp\{\cdot\}/(1+\exp\{\cdot\})$, $\boldsymbol{\beta}^I_j$ is a q-dimensional vector of logistic regression coefficients for the jth tooth, and γ^I_{lj} and α^I_{lj} are, within- and across-time conditional log-odds ratio parameters, respectively, since

$$
\begin{aligned}
\exp\{\gamma^I_{lj}\} &= \frac{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=1, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=0, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)} \times \\
&\quad \frac{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=0, Y_{(i,l,k)}=0, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=1, Y_{(i,l,k)}=0, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)},
\end{aligned}
$$

and

$$
\begin{aligned}
\exp\{\alpha^I_{lj}\} &= \frac{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=1, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=1 \mid \cdots\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=0, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=1 \mid \cdots\right)} \times \\
&\quad \frac{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=0, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,j,k)}=1, Y_{(i,l,k)}=1, Y_{(i,j,k-1)}=0, Y_{(i,l,k-1)}=0 \mid \cdots\right)}.
\end{aligned}
$$

In the context of our motivating problem, the γ^I_{lj} parameters are interpreted as the difference in the conditional log-odds of developing CE for the jth tooth between

the examinations $(k-1)$ and k, when the lth tooth has developed or not CE in the same interval. On the other hand, the α_{lj}^I parameters represent the difference in the conditional log-odds of developing CE for the jth tooth between the examinations $(k-1)$ and k, when the lth tooth had CE in the previous examination, $(k-1)$, versus when the lth tooth develops CE between the examinations $(k-1)$ and k.

We show that, similar to the model proposed by Joe & Liu (1996), the restrictions $\gamma_{lj}^I = \gamma_{jl}^I$, $\forall\ j \neq l$, are necessary and sufficient conditions for the full conditionals given by expression (5.3) to define a proper probability model for each row of the Markovian transition matrices. The following proposition is proved in Section D.1 of Appendix D.

Proposition 5.1. *The full conditional distributions given by expression (5.3) are compatible if and only if $\gamma_{lj}^I = \gamma_{jl}^I$, for all $j \neq l$. Under these conditions, the conditional joint distributions are given by*

$$P_{\boldsymbol{Z}_{(i,k)}}(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I) = c_2(\boldsymbol{Z}_{(i,k)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I)^{-1} \times$$

$$\exp\left\{\sum_{j\in S}\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I\right) y_j^k \ + \sum_{\{1\le l<j:j\in S, l\in S^c\}} \gamma_{lj}^I y_j^k + \sum_{\{j<l\le J:j\in S, l\in S\}} \gamma_{jl}^I y_l^k y_j^k \right.$$

$$\left. + \sum_{\{j<l\le J:j\in S, l\in S^c\}} \gamma_{jl}^I y_j^k + \sum_{\{(j,l):j\in S, l\in S^c\}} \alpha_{lj}^I\ y_j^k \right\},$$

where $\boldsymbol{y}^k \in \mathcal{B}\{\boldsymbol{y}^{k-1}\}$, $S = \{j : y_j^{k-1} = 0\}$, $S^c = \{j : y_j^{k-1} = 1\}$, $\boldsymbol{\beta}^I = (\boldsymbol{\beta}_1^I, \ldots, \boldsymbol{\beta}_J^I)'$, $\boldsymbol{\gamma}^I = \{\gamma_{jl}^I : 1 \le j < l \le J\}$, $\boldsymbol{\alpha}^I = \{\alpha_{jl}^I : j, l = 1, \ldots, J, j \neq l\}$, and $c_2(\boldsymbol{Z}_{(i,k)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I)$ is a normalizing constant given by

$$c_2(\boldsymbol{Z}_{(i,k)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I) = \sum_{\boldsymbol{y}^k\in\mathcal{B}\{\boldsymbol{y}^{k-1}\}} \exp\left\{\sum_{j\in S}\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I\right) y_j^k \ + \sum_{\{1\le l<j:j\in S, l\in S^c\}} \gamma_{lj}^I y_j^k \right.$$

$$\left. + \sum_{\{j<l\le J:j\in S, l\in S\}} \gamma_{jl}^I y_l^k y_j^k + \sum_{\{j<l\le J:j\in S, l\in S^c\}} \gamma_{jl}^I y_j^k + \sum_{\{(j,l):j\in S, l\in S^c\}} \alpha_{lj}^I\ y_j^k \right\}.$$

5.3.2 The misclassification model

We assume that the response variables $Y_{(i,j,k)}$ are prone to misclassification. Let $Y_{(i,j,k)}^*$ be the corrupted observed binary response for tooth j of subject i at time

$t_{(i,k)}$, and set $\boldsymbol{Y}^*_{(i,k)} = \left(Y^*_{(i,1,k)}, \ldots, Y^*_{(i,J,k)}\right)'$ and $\boldsymbol{Y}^*_i = \left(\boldsymbol{Y}^*_{(i,1)}, \ldots, \boldsymbol{Y}^*_{(i,K)}\right)$. Here we suppose that the scoring is performed by Q examiners. Denote by $\xi_{(i,k)} \in \{1, \ldots, Q\}$ the variable indicating the examiner that scores all teeth of subject i at time $t_{(i,k)}$, and let $\boldsymbol{\xi_i} = (\xi_{(i,1)}, \ldots, \xi_{(i,K)})$ be the vector of indicators of the examiners that score the responses of subject i over time. We further assume that the scoring behavior of the examiners is the same across the study. Let $\boldsymbol{\tau}^{00}_q = \left(\tau^{00}_{(q,1)}, \ldots, \tau^{00}_{(q,J)}\right)$ and $\boldsymbol{\tau}^{11}_q = \left(\tau^{11}_{(q,1)}, \ldots, \tau^{11}_{(q,J)}\right)$, $q = 1, \ldots, Q$, be the vectors containing the tooth-specific specificity and sensitivity parameters for the qth examiner, respectively. Finally, let $\boldsymbol{\tau}^{11} = (\boldsymbol{\tau}^{11}_1, \ldots, \boldsymbol{\tau}^{11}_Q)$ and $\boldsymbol{\tau}^{00} = (\boldsymbol{\tau}^{00}_1, \ldots, \boldsymbol{\tau}^{00}_Q)$ be the matrices containing all sensitivity and specificity parameters, respetively. In this setting, the misclassification model assumes that $\mathrm{P}\left(Y^*_{(i,j,k)} = 1 \mid Y_{(i,j,k)} = 1\right) = \tau^{11}_{(\xi_{(i,k)},j)}$ and $\mathrm{P}\left(Y^*_{(i,j,k)} = 0 \mid Y_{(i,j,k)} = 0\right) = \tau^{00}_{(\xi_{(i,k)},j)}$ and the process is characterized by the following conditional independence assumptions. Note that assumptions (A.1) - (A.6) represent natural extensions of the commonly used assumptions for the analysis of univariate and multivariate misclassified binary data (see, e.g. Geng & Asano, 1989; Magder & Hughes, 1997; Neuhaus, 1999, 2002).

(A.1) $\perp\!\!\!\perp_{1 \le i \le I} \boldsymbol{Y}^*_i \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_I, \boldsymbol{\xi}_1 \ldots, \boldsymbol{\xi}_I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}$, i.e. the observed response matrices for each subject are independent given the true unobserved responses, examiner indicators, and sensitivity and specificity parameters,

(A.2) $\boldsymbol{Y}^*_i \perp\!\!\!\perp \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{i-1}, \boldsymbol{Y}_{i+1}, \ldots, \boldsymbol{Y}_I \mid \boldsymbol{Y}_i, \boldsymbol{\xi}_i, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}, \forall\, i$, i.e. the distribution of the observed response matrix for a subject only depends on his true unobserved response matrix, the examiners that score his responses, and the sensitivity and specificity parameters,

(A.3) $\perp\!\!\!\perp_{1 \le k \le K} \boldsymbol{Y}^*_{(i,k)} \mid \boldsymbol{Y}_i, \boldsymbol{\xi}_i, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}, \forall\, (i,k)$, i.e. the observed response vectors for a subject are independent across time given his unobserved response matrix, the examiners that score his responses and the sensitivity and specificity parameters,

(A.4) $\boldsymbol{Y}^*_{(i,k)} \perp\!\!\!\perp \boldsymbol{Y}_{(i,1)}, \ldots, \boldsymbol{Y}_{(i,k-1)}, \boldsymbol{Y}_{(i,k+1)}, \ldots, \boldsymbol{Y}_{(i,K)} \mid \boldsymbol{Y}_{(i,k)}, \xi_{(i,k)}, \boldsymbol{\tau}^{00}_{\xi_{(i,k)}}, \boldsymbol{\tau}^{11}_{\xi_{(i,k)}}$, i.e., the distribution of the observed response vector in the kth examination only depends on his true unobserved response vector at the same examination, the examiner that scores his responses at examination k, and the examiner-specific sensitivity and specificity parameters,

(A.5) $\perp\!\!\!\perp_{1 \le j \le J} Y^*_{(i,j,k)} \mid \boldsymbol{Y}_{(i,k)}, \xi_{(i,k)}, \boldsymbol{\tau}^{00}_{\xi_{(i,k)}}, \boldsymbol{\tau}^{11}_{\xi_{(i,k)}}, \forall\, (i,k)$, i.e. the observed responses at the kth examination are independent given the unobserved response vector at the same examination, the examiner that scores his responses at the kth examination, and the examiner-specific sensitivity and specificity parameters,

(A.6) $Y^*_{(i,j,k)} \perp\!\!\!\perp Y_{(i,1,k)}, \ldots, Y_{(i,j-1,k)}, Y_{(i,j+1,k)}, \ldots, Y_{(i,J,k)} \mid Y_{(i,j,k)}, \xi_{(i,k)}, \tau^{00}_{(\xi_{(i,k)},j)}, \tau^{11}_{(\xi_{(i,k)},j)}$, i.e., the distribution of the jth observed response at the kth examination only depends on the true unobserved response at the same examination, the examiner that scores his responses at examination k, and the sensitivity and specificity parameters of this examiner for the jth tooth.

In order to evaluate the effect of the examiner and the tooth on the misclassification parameters, three misclassification models, varying in complexity regarding these effects, were considered. In a general model we assume unstructured examiner-tooth specific sensitivity and specificity parameters ($\boldsymbol{M}_1$). In a second version of the model, we assume the same misclassification parameters across teeth for each examiner ($\boldsymbol{M}_2$), i.e., for $q = 1, \ldots, Q$, $\tau^{00}_{(q,j)} = \tau^{00}_{(q)}$ and $\tau^{11}_{(q,j)} = \tau^{11}_{(q)}$, $\forall j$. Finally, an intermediate case assuming examiner-tooth specific misclassification parameters under an additive model was assumed ($\boldsymbol{M}_3$), by considering

$$\text{logit}\left(\tau^{00}_{(q,j)}\right) = \boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00} \quad \text{and} \quad \text{logit}\left(\tau^{11}_{(q,j)}\right) = \boldsymbol{w}'_{qj}\boldsymbol{\delta}^{11},$$

where $\boldsymbol{w}_{qj}$ is a design vector including an intercept term and dummy variables for the examiners and teeth, and $\boldsymbol{\delta}^{00}$ and $\boldsymbol{\delta}^{11}$ are regression coefficients of the additive model associated to the examiner-tooth-specific specificities and sensitivities, respectively.

Following García-Zattera et al. (2010), we consider restricted parameter spaces for the misclassification parameters in order to avoid identification problems. For the misclassification models $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, we consider the parameter space

$$\left\{\left(\tau^{00}_{(q,j)}, \tau^{11}_{(q,j)}\right) \in [0,1]^2 : \tau^{00}_{(q,j)} + \tau^{11}_{(q,j)} > 1\right\}, q = 1, \ldots, Q, \quad j = 1, \ldots, J,$$

and

$$\left\{\left(\tau^{00}_{(q)}, \tau^{11}_{(q)}\right) \in [0,1]^2 : \tau^{00}_{(q)} + \tau^{11}_{(q)} > 1\right\}, q = 1, \ldots, Q,$$

respectively. Equivalently, for the misclassification model $\boldsymbol{M}_3$ we consider the parameter space

$$\left\{\left(\boldsymbol{\delta}^{00\prime}, \boldsymbol{\delta}^{11\prime}\right)' \in \mathbb{R}^{2J+2Q-2} : h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right) + h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right) > 1\right\}, q = 1, \ldots, Q, j = 1, \ldots, J.$$

5.3.3 The implied statistical model

Regardless of the misclassification model, the assumptions (A.1) - (A.6), along with the assumptions associated to the Markov model, imply that the joint probability

model for the observed and unobserved responses for each subject is given by

$$P_{\boldsymbol{X}_i,\boldsymbol{Z}_i}\left(\boldsymbol{Y}_i^* = \boldsymbol{y}_i^*, \boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right) =$$

$$\left\{\prod_{j=1}^{J}\prod_{k=1}^{K} P\left(Y^*_{(i,j,k)} = y^*_{(i,j,k)} \mid Y_{(i,j,k)} = y_{(i,j,k)}, \xi_{i,k}, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right)\right\} \times$$

$$\left\{\prod_{k=2}^{K} P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}_{(i,k)} \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}_{(i,k-1)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right)\right\} \times$$

$$P_{\boldsymbol{X}_i}\left(\boldsymbol{Y}_{(i,1)} = \boldsymbol{y}_{(i,1)} \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P\right),$$

where

$$P\left(Y^*_{(i,j,k)} = y^*_{(i,j,k)} \mid Y_{(i,j,k)} = y_{(i,j,k)}, \xi_{(i,k)}, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right) =$$

$$\left\{\tau^{11}_{(\xi_{(i,k)},j)}{}^{y^*_{(i,j,k)}}\left(1-\tau^{11}_{(\xi_{(i,k)},j)}\right)^{1-y^*_{(i,j,k)}}\right\}^{y_{(i,j,k)}} \times$$

$$\left\{\tau^{00}_{(\xi_{i,k},j)}{}^{1-y^*_{(i,j,k)}}\left(1-\tau^{00}_{(\xi_{i,k},j)}\right)^{y^*_{(i,j,k)}}\right\}^{1-y_{(i,j,k)}},$$

and $\boldsymbol{y}_{(i,1)} \in \{0,1\}^J$, $\boldsymbol{y}_{(i,2)} \in \mathcal{B}\{\boldsymbol{y}_{(i,1)}\}, \ldots, \boldsymbol{y}_{(i,K)} \in \mathcal{B}\{\boldsymbol{y}_{(i,K-1)}\}$. Therefore, the likelihood function is given by

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I,\ \boldsymbol{\alpha}^I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right) \\
&\qquad = \prod_{i=1}^{I} P_{\boldsymbol{X}_i,\boldsymbol{Z}_i}\left(\boldsymbol{Y}_i^* = \boldsymbol{y}_i^* \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right), \\
&\qquad = \prod_{i=1}^{I} \sum_{\boldsymbol{y}_i \in \mathcal{V}} P_{\boldsymbol{X}_i,\boldsymbol{Z}_i}\left(\boldsymbol{Y}_i^* = \boldsymbol{y}_i^*, \boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11}\right),
\end{aligned} \tag{5.4}$$

where $\mathcal{V} = \{0,1\}^J \times \mathcal{B}\{\boldsymbol{y}_{(i,1)}\} \times \cdots \times \mathcal{B}\{\boldsymbol{y}_{(i,K-1)}\}$.

5.4 The Bayesian implementation

In this section the model is completed by specifying prior distributions for the parameters in Section 5.4.1. The algorithms used for posterior computation in the

proposed models are described in Section 5.4.2. Finally, the performance of the model is evaluated using simulated data in Section 5.4.3.

5.4.1 The prior specification

Independent normal prior distributions were assumed for the conditionally specified logistic regression coefficients associated with the multivariate baseline distribution and Markov model,

$$\boldsymbol{\beta}^P \sim N_{Jp}(\boldsymbol{m}_{\boldsymbol{\beta}^P}, \boldsymbol{V}_{\boldsymbol{\beta}^P}), \quad \boldsymbol{\gamma}^P \sim N_{J(J-1)/2}\left(\boldsymbol{m}_{\boldsymbol{\gamma}^P}, \boldsymbol{V}_{\boldsymbol{\gamma}^P}\right), \tag{5.5}$$

$$\boldsymbol{\beta}^I \sim N_{Jq}(\boldsymbol{m}_{\boldsymbol{\beta}^I}, \boldsymbol{V}_{\boldsymbol{\beta}^I}), \quad \boldsymbol{\gamma}^I \sim N_{J(J-1)/2}\left(\boldsymbol{m}_{\boldsymbol{\gamma}^I}, \boldsymbol{V}_{\boldsymbol{\gamma}^I}\right), \tag{5.6}$$

$$\boldsymbol{\alpha}^I \sim N_{J(J-1)}\left(\boldsymbol{m}_{\boldsymbol{\alpha}^I}, \boldsymbol{V}_{\boldsymbol{\alpha}^I}\right), \tag{5.7}$$

For the unstructured misclassification model $\boldsymbol{M}_1$, we assume that for all $q \in \{1, \ldots, Q\}$ and $j \in \{1, \ldots, J\}$,

$$\begin{aligned}\left(\tau^{00}_{(q,j)}, \tau^{11}_{(q,j)}\right) \overset{ind.}{\sim} \text{Beta}\left(\epsilon^{00}_{(1,qj)}, \epsilon^{00}_{(2,qj)}\right) \times \text{Beta}\left(\epsilon^{11}_{(1,qj)}, \epsilon^{11}_{(2,qj)}\right) \times \\ I\left(\tau^{00}_{(q,j)}, \tau^{11}_{(q,j)}\right)_{\left\{(\tau^{00}_{(q,j)}, \tau^{11}_{(q,j)}):\tau^{00}_{(q,j)}+\tau^{11}_{(q,j)}>1\right\}},\end{aligned} \tag{5.8}$$

where $I(\cdot)_A$ is an indicator function for the set A. Similarly, for the misclassification model assuming equal misclassification parameters across teeth for each examiner, $\boldsymbol{M}_2$, we assume that for all $q \in \{1, \ldots, Q\}$,

$$\begin{aligned}\left(\tau^{00}_{(q)}, \tau^{11}_{(q)}\right) \overset{ind.}{\sim} \text{Beta}\left(\epsilon^{00}_{(1,q)}, \epsilon^{00}_{(2,q)}\right) \times \text{Beta}\left(\epsilon^{11}_{(1,q)}, \epsilon^{11}_{(2,q)}\right) \times \\ I\left(\tau^{00}_{(q)}, \tau^{11}_{(q)}\right)_{\left\{(\tau^{00}_{(q)}, \tau^{11}_{(q)}):\tau^{00}_{(q)}+\tau^{11}_{(q)}>1\right\}}.\end{aligned} \tag{5.9}$$

Finally, for the additive misclassification model, $\boldsymbol{M}_3$, we assume that

$$\begin{aligned}\left(\boldsymbol{\delta}^{00\prime}, \boldsymbol{\delta}^{11\prime}\right)' \sim N_{J+Q-1}\left(\boldsymbol{m}_{\boldsymbol{\delta}^{00}}, \boldsymbol{V}_{\boldsymbol{\delta}^{00}}\right) \times N_{J+Q-1}\left(\boldsymbol{m}_{\boldsymbol{\delta}^{11}}, \boldsymbol{V}_{\boldsymbol{\delta}^{11}}\right) \times \\ I\left(\boldsymbol{\delta}^{00}, \boldsymbol{\delta}^{11}\right)_{\left\{\left(\boldsymbol{\delta}^{00\prime}, \boldsymbol{\delta}^{11\prime}\right)' \in \mathbb{R}^{2J+2Q-2}: h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)+h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)>1, \forall q,j\right\}}.\end{aligned} \tag{5.10}$$

5.4.2 The posterior computation

We next explain the computational strategy used for posterior inference, which is based on Markov chain Monte Carlo (MCMC) simulation. Functions implementing the MCMC algorithms described here for each model were written in a compiled language and incorporated into a library of the R program (R Development Core Team, 2010). This library is available upon request to the first author. Under the prior specification given in Section 5.4.1, the posterior distributions

$$p_{\boldsymbol{M}_i}\left(\boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I, \boldsymbol{\tau}^{00}, \boldsymbol{\tau}^{11} \mid \boldsymbol{Y}^*\right), i = 1, 2,$$

and

$$p_{\boldsymbol{M}_3}\left(\boldsymbol{\beta}^P, \boldsymbol{\gamma}^P, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I, \boldsymbol{\delta}^{00}, \boldsymbol{\delta}^{11} \mid \boldsymbol{Y}^*\right),$$

arising under the misclassification models $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$, are proportional to the product of expressions (5.4) - (5.7) with (5.8), (5.9), and (5.10), respectively. To explore these posterior distributions, Metropolis within Gibbs algorithms (Tierney, 1994) are used. In all cases, the MCMC algorithm is based on a data augmentation step treating the unobserved true responses $Y_{(i,j,k)}$ as unknown parameters. Although the full conditionals for the latent data $Y_{(i,j,k)}$ are straightforward to derive and sample from under each misclassification model, we propose a blocked step to update the latent vectors $\boldsymbol{Y}_{(i,k)}$ in order to improve the mixing of the chain. The full conditionals for the latent vectors $\boldsymbol{Y}_{(i,k)}$ are discrete distributions with appropriate probability parameters. Indeed, the misclassification assumptions (A.1) - (A.6), along with the assumptions of the monotone Markov model imply that for all $\boldsymbol{y}_m \in \{0,1\}^J$, the discrete probabilities are given by

$$\begin{aligned} P\left(\boldsymbol{Y}_{(i,1)} = \boldsymbol{y}_m \mid \cdots\right) \quad \propto \quad & P\left(\boldsymbol{Y}^*_{(i,1)} = \boldsymbol{y}_{(i,1)} \mid \boldsymbol{Y}_{(i,1)} = \boldsymbol{y}_m, \xi_{(i,1)}, \boldsymbol{\tau}^{00}_{\xi_{(i,1)}}, \boldsymbol{\tau}^{11}_{\xi_{(i,1)}}\right) \times \\ & P_{\boldsymbol{X}_i}\left(\boldsymbol{Y}_{(i,1)} = \boldsymbol{y}_m \mid \boldsymbol{\beta}^P, \boldsymbol{\gamma}^P\right) \times \\ & I(\boldsymbol{y}_m)_{\left\{\boldsymbol{y}_m \in \{0,1\}^J : \boldsymbol{y}_{(i,2)} \in \mathcal{B}\{\boldsymbol{y}_m\}\right\}}, \end{aligned}$$

$$\begin{aligned} P\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}_m \mid \cdots\right) \quad \propto \quad & P\left(\boldsymbol{Y}^*_{(i,k)} = \boldsymbol{y}_{(i,k)} \mid \boldsymbol{Y}_{(i,k)} = \boldsymbol{y}_m, \xi_{(i,k)}, \boldsymbol{\tau}^{00}_{\xi_{(i,k)}}, \boldsymbol{\tau}^{11}_{\xi_{(i,k)}}\right) \times \\ & P_{\boldsymbol{Z}_{(i,k+1)}}\left(\boldsymbol{Y}_{(i,k+1)} = \boldsymbol{y}_{(i,k+1)} \mid \boldsymbol{Y}_{(i,k)} = \boldsymbol{y}_m, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right) \times \\ & P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}_m \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}_{(i,k-1)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right) \times \\ & I(\boldsymbol{y}_m)_{\left\{\boldsymbol{y}_m \in \{0,1\}^J : \boldsymbol{y}_m \in \mathcal{B}\{\boldsymbol{y}_{(i,k-1)}\}, \boldsymbol{y}_{(i,k+1)} \in \mathcal{B}\{\boldsymbol{y}_m\}\right\}}, \end{aligned}$$

for $k \in \{2, \ldots, K-1\}$, and

$$P\left(\boldsymbol{Y}_{(i,K)} = \boldsymbol{y}_m \mid \cdots\right) \propto P\left(\boldsymbol{Y}^*_{(i,K)} = \boldsymbol{y}_{(i,K)} \mid \boldsymbol{Y}_{(i,K)} = \boldsymbol{y}_m, \xi_{(i,K)}, \boldsymbol{\tau}^{00}_{\xi_{(i,K)}}, \boldsymbol{\tau}^{11}_{\xi_{(i,K)}}\right) \times$$
$$P_{\boldsymbol{Z}_{(i,K)}}\left(\boldsymbol{Y}_{(i,K)} = \boldsymbol{y}_m \mid \boldsymbol{Y}_{(i,K-1)} = \boldsymbol{y}_{(i,K-1)}, \boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right) \times$$
$$I(\boldsymbol{y}_m)_{\left\{\boldsymbol{y}_m \in \{0,1\}^J : \boldsymbol{y}_m \in \mathcal{B}\{\boldsymbol{y}_{(i,K-1)}\}\right\}}.$$

In all the models, the introduction of latent data greatly simplify the computations. Given the latent data for the first examination, $\left(\boldsymbol{Y}_{(1,1)}, \ldots, \boldsymbol{Y}_{(I,1)}\right)$, the full conditionals for the parameters $\boldsymbol{\beta}^P$ and $\boldsymbol{\alpha}^P$ correspond to the one arising from the conditionally specified logistic regression model given by expression (5.2). Since these full conditionals are not standard, one-coordinate-at-a-time slice sampling (Neal, 2003) or Metropolis-Hastings (MH) algorithms (Tierney, 1994) can be used to update the elements of $\boldsymbol{\beta}^P$, $\boldsymbol{\alpha}^P$, $\boldsymbol{\beta}^I$, $\boldsymbol{\gamma}^I$ and $\boldsymbol{\alpha}^I$. Alternatively, we consider MH steps to update the joint vectors $\boldsymbol{\theta}^P = \left(\boldsymbol{\beta}^{P'}, \boldsymbol{\gamma}^{P'}\right)'$ and $\boldsymbol{\theta}^I = \left(\boldsymbol{\beta}^{I'}, \boldsymbol{\gamma}^{I'}, \boldsymbol{\alpha}^{I'}\right)'$, based on multivariate normal proposals. The mean and the covariance matrix of these proposals are created based on a first order Taylor series approximation to the logistic regressions associated to the full conditional distributions of all latent response variables at the corresponding examinations, given by expressions (5.1) and (5.3), and on one step of the Newton-Raphson algorithm (see, e.g. Gamerman, 1997). A complete description of these steps is given in Section D.2 of Appendix D.

Assumptions (A.1) - (A.6), along with the assumptions of the monotone Markov model, imply that the full conditionals for the misclassification parameters under the unstructured misclassification model $\boldsymbol{M}_1$ are truncated beta distributions given by

$$\tau^{00}_{(q,j)} \mid \cdots \sim \text{Beta}\left(\epsilon^{00}_{(1,qj)} + n^{00}_{(1,qj)}, \epsilon^{00}_{(2,qj)} + n^{00}_{(2,qj)}\right) \times$$
$$I\left(\tau^{00}_{(q,j)}\right)_{\left\{\tau^{00}_{(q,j)} : \tau^{00}_{(q,j)} > 1 - \tau^{11}_{(q,j)}\right\}},$$

and

$$\tau^{11}_{(q,j)} \mid \cdots \sim \text{Beta}\left(\epsilon^{11}_{(1,qj)} + n^{11}_{(1,qj)}, \epsilon^{11}_{(2,qj)} + n^{11}_{(2,qj)}\right) \times$$
$$I\left(\tau^{11}_{(q,j)}\right)_{\left\{\tau^{11}_{(q,j)} : \tau^{11}_{(q,j)} > 1 - \tau^{00}_{(q,j)}\right\}},$$

where

$$n^{00}_{(1,qj)} = \sum_{l=1}^{I}\sum_{k=1}^{K} I\left(Y^*_{(l,j,k)}, Y_{(l,j,k)}\right)_{\left\{Y^*_{(l,j,k)}=0, Y_{(l,j,k)}=0\right\}} I\left(\xi_{(l,k)}\right)_{\left\{\xi_{(l,k)}=i\right\}},$$

$$n^{00}_{(2,qj)} = \sum_{l=1}^{I}\sum_{k=1}^{K} I\left(Y^*_{(l,j,k)}, Y_{(l,j,k)}\right)_{\left\{Y^*_{(l,j,k)}=1, Y_{(l,j,k)}=0\right\}} I\left(\xi_{(l,k)}\right)_{\left\{\xi_{(l,k)}=i\right\}},$$

$$n^{11}_{(1,qj)} = \sum_{l=1}^{I}\sum_{k=1}^{K} I\left(Y^*_{(l,j,k)}, Y_{(l,j,k)}\right)_{\left\{Y^*_{(l,j,k)}=1, Y_{(l,j,k)}=1\right\}} I\left(\xi_{(l,k)}\right)_{\left\{\xi_{(l,k)}=i\right\}},$$

and

$$n^{11}_{(2,qj)} = \sum_{l=1}^{I}\sum_{k=1}^{K} I\left(Y^*_{(l,j,k)}, Y_{(l,j,k)}\right)_{\left\{Y^*_{(l,j,k)}=0, Y_{(l,j,k)}=1\right\}} I\left(\xi_{(l,k)}\right)_{\left\{\xi_{(l,k)}=i\right\}}.$$

Similar expressions are obtained for the model assuming the same examiner-specific misclassification parameters for each teeth. Finally, under the additive misclassification model $\boldsymbol{M}_3$, the full conditionals of the parameters correspond to the one arising from logistic regressions with normals priors, applied to the corresponding subsets of the data. Specifically, the full conditionals for $\boldsymbol{\delta}^{00}$ and $\boldsymbol{\delta}^{11}$ are given by

$$\begin{aligned} p\left(\boldsymbol{\delta}^{00} \mid \cdots\right) \quad \propto \quad & \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\exp\left\{\boldsymbol{w}_{\xi_{(i,k)},j}\boldsymbol{\delta}^{00}(1-y^*_{(i,j,k)})\right\}}{1+\exp\left\{\boldsymbol{w}_{\xi_{(i,k)},j}\boldsymbol{\delta}^{00}\right\}} \times I\left(y_{(i,j,k)}\right)_{\left\{y_{(i,j,k)}=0\right\}} \times \\ & \exp\left\{-0.5(\boldsymbol{\delta}^{00}-\boldsymbol{m}_{\boldsymbol{\delta}^{00}})'\boldsymbol{V}^{-1}_{\boldsymbol{\delta}^{00}}(\boldsymbol{\delta}^{00}-\boldsymbol{m}_{\boldsymbol{\delta}^{00}})\right\} \times \\ & I\left(\boldsymbol{\delta}^{00}\right)_{\left\{\boldsymbol{\delta}^{00}\in\mathbb{R}^{J+Q-1}: h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)+h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)>1, \forall q,j\right\}}, \end{aligned}$$

and

$$\begin{aligned} p\left(\boldsymbol{\delta}^{11} \mid \cdots\right) \quad \propto \quad & \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\exp\left\{\boldsymbol{w}_{\xi_{(i,k)},j}\boldsymbol{\delta}^{11}y^*_{(i,j,k)}\right\}}{1+\exp\left\{\boldsymbol{w}_{\xi_{(i,k)},j}\boldsymbol{\delta}^{11}\right\}} \times I\left(y_{(i,j,k)}\right)_{\left\{y_{(i,j,k)}=1\right\}} \times \\ & \exp\left\{-0.5(\boldsymbol{\delta}^{00}-\boldsymbol{m}_{\boldsymbol{\delta}^{00}})'\boldsymbol{V}^{-1}_{\boldsymbol{\delta}^{00}}(\boldsymbol{\delta}^{00}-\boldsymbol{m}_{\boldsymbol{\delta}^{00}})\right\} \times \\ & I\left(\boldsymbol{\delta}^{11}\right)_{\left\{\boldsymbol{\delta}^{11}\in\mathbb{R}^{J+Q-1}: h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)+h\left(\boldsymbol{w}'_{qj}\boldsymbol{\delta}^{00}\right)>1, \forall q,j\right\}}, \end{aligned}$$

respectively. Therefore, sampling methods for constrained logistic regression parameters can be used to sample from these full conditional distributions. In our function we consider the slice sampling algorithm proposed by Neal (2003).

5.4.3 A limited simulation study and the results

Problems in measurement error models and HMM include lack of parameter identification. Parameters in a model are nonidentified if more than one set of parameter values gives the same distribution function for the observations. To validate the proposed multivariate HMM model, we conducted the analysis of simulated datasets, which mimic to a certain extent the ST data. For this purpose the covariates and the same examination structure considered were as that of the ST study. The true values for the multivariate HMM were motivated by estimates observed in a preliminary analysis of the ST data under the more general misclassification model $\boldsymbol{M}_1$ and assuming a common effect of the predictors for each response but different intercepts terms. The true values for the Markov models parameters are given in Tables 5.1 and 5.2 (see pages 109 and 110, respectively).

We simulated 100 data sets with the previously described structure and fitted the model in each case. For the misclassification parameters we considered independent constraint uniform priors by taking $\epsilon^{11}_{(1,qj)} = \epsilon^{11}_{(2,qj)} = \cdots = \epsilon^{11}_{(1,QJ)} = \epsilon^{11}_{(2,QJ)} = \epsilon^{00}_{(1,qj)} = \epsilon^{00}_{(2,qj)} = \cdots = \epsilon^{00}_{(1,QJ)} = \epsilon^{00}_{(2,QJ)} = 1$ in (5.8). For the remaining hyper-parameters independent normal $N(0, 10^3)$ distributions were considered. For each simulated dataset, one Markov chain was generated completing a total number of 105000 scans of the Markov chain cycle described in Section 5.4 were completed. Standard tests (not shown), as implemented in the BOA R library (Smith, 2007), suggested convergence of the chains. Because of storage limitations the full chain was subsampled every 5 steps after a burn-in period of 5000 samples, to give a reduced chain of length 20000.

Tables 5.1 and 5.2 (see pages 109 and 110, respectively) show the means, across simulation, the biases and the MSEs of the posterior means of the HMM parameters. The results suggest that the regression parameters and association parameters can be estimated with only a minimal bias and with a reasonable precision. With the exception of one coefficient, which was close to zero, the sign of the regression coefficients and association parameters was correctly estimated. Similar results regarding bias and MSE were observed for the misclassification parameters (see Figures 5.2 and 5.3 in pages 111 and 112, respectively). Therefore, the results suggest that, under the setting of the ST study, concentrated information on the misclassification parameters is not needed to obtain nearly unbiased and precise estimates for the regression coefficients and

Table 5.1: Simulated data: true values, and Monte Carlo means, biases and mean squared errors (MSE) of the posterior means of the logistic regression parameters.

Covariate	True Value	Mean	Bias	MSE $\times 10$
		Prevalence		
Intercept T16	-6.094	-6.204	-0.110	24.222
Intercept T26	-5.621	-5.480	0.141	18.050
Intercept T36	-5.668	-5.474	0.195	18.070
Intercept T46	-5.453	-5.442	0.011	15.724
Startbr	0.091	0.109	0.018	0.034
Age	0.345	0.337	-0.008	0.230
Meals	0.142	0.140	-0.002	0.099
x-ordinate	0.007	-0.010	-0.017	0.137
y-ordinate	-0.614	-0.716	-0.101	1.865
		Incidence		
Intercept T16	-4.581	-4.545	0.035	1.099
Intercept T26	-4.316	-4.318	-0.003	1.096
Intercept T36	-4.709	-4.658	0.050	1.423
Intercept T46	-4.189	-4.220	-0.031	1.146
Startbr	0.086	0.092	0.006	0.004
Age	-0.041	-0.046	-0.005	0.003
Meals	0.153	0.164	0.011	0.014
x-ordinate	0.061	0.070	0.009	0.010
y-ordinate	-0.156	-0.182	-0.027	0.060
Years-exam	0.398	0.374	-0.024	0.263

misclassification parameters. Thus, the model parameters can be estimated from the raw data without extra information on the misclassification parameters.

5.5 The analysis of the Signal-Tandmobiel® data

In this section we show the results of the analyses of the ST study. Here the main focus is the assesment of potential risk factors on the prevalence and the incidence of CE in permanent first molars.

Specifically, we evaluated the effect of the gender (boys vs girls; `Gender`), age at start of brushing (in years; `Startbr`), the number of between-meal snacks (two or less than two a day vs more than two a day; `Meals`), the geographical location (in terms of the standardized (x, y) co-ordinate of the municipality of the school to which the child belongs; `x-ordinate` and `y-ordinate`), and the age (in years;

Table 5.2: Simulated data: true values, and Monte Carlo means, biases and mean squared errors (MSE) of the posterior means of the association parameters.

	Parameter	True Value	Mean	Bias	MSE
		Prevalence			
Within Time Association Parameters	$\gamma^P_{16,26}$	4.003	4.048	0.044	0.557
	$\gamma^P_{16,36}$	0.424	0.449	0.025	0.561
	$\gamma^P_{16,46}$	2.442	2.103	-0.339	1.146
	$\gamma^P_{26,36}$	1.700	1.592	-0.108	0.782
	$\gamma^P_{26,46}$	0.443	0.592	0.149	0.838
	$\gamma^P_{36,46}$	2.748	2.991	0.242	1.380
		Incidence			
Within Time Association Parameters	$\gamma^I_{16,26}$	3.824	3.919	0.095	0.135
	$\gamma^I_{16,36}$	2.729	3.081	0.352	0.308
	$\gamma^I_{16,46}$	0.950	1.330	0.380	0.357
	$\gamma^I_{26,36}$	1.003	1.213	0.210	0.696
	$\gamma^I_{26,46}$	2.288	3.354	1.066	1.518
	$\gamma^I_{36,46}$	3.912	4.072	0.160	0.435
Across Time Association Parameters	$\alpha^I_{16,26}$	-2.770	-2.854	-0.084	0.316
	$\alpha^I_{16,36}$	-1.950	-2.087	-0.138	0.342
	$\alpha^I_{16,46}$	-0.241	-0.262	-0.021	0.431
	$\alpha^I_{26,16}$	-1.918	-2.226	-0.308	0.376
	$\alpha^I_{26,36}$	1.140	1.767	0.627	0.617
	$\alpha^I_{26,46}$	-1.724	-2.580	-0.856	1.622
	$\alpha^I_{36,16}$	-2.378	-2.857	-0.479	0.537
	$\alpha^I_{36,26}$	1.515	2.426	0.911	1.396
	$\alpha^I_{36,46}$	-2.728	-3.303	-0.575	0.567
	$\alpha^I_{46,16}$	-0.426	-0.617	-0.192	0.211
	$\alpha^I_{46,26}$	-1.183	-2.123	-0.940	1.273
	$\alpha^I_{46,36}$	-1.349	-1.604	-0.255	0.497

`Age`) on the prevalence and incidence of CE for permanent first molars, teeth 16, 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). In addition, the length of time between examinations (in years; `Years-exam`) was included in a linear way in $\boldsymbol{z}_{(i,j,k)}$. The number of between-meal snacks was considered as a time dependent covariate and was included with a one year lag in the corresponding design matrices.

The inclusion of the geographical components, expressed in terms of the x- and y-coordinates, was motivated by the results of previous analyses (without correcting

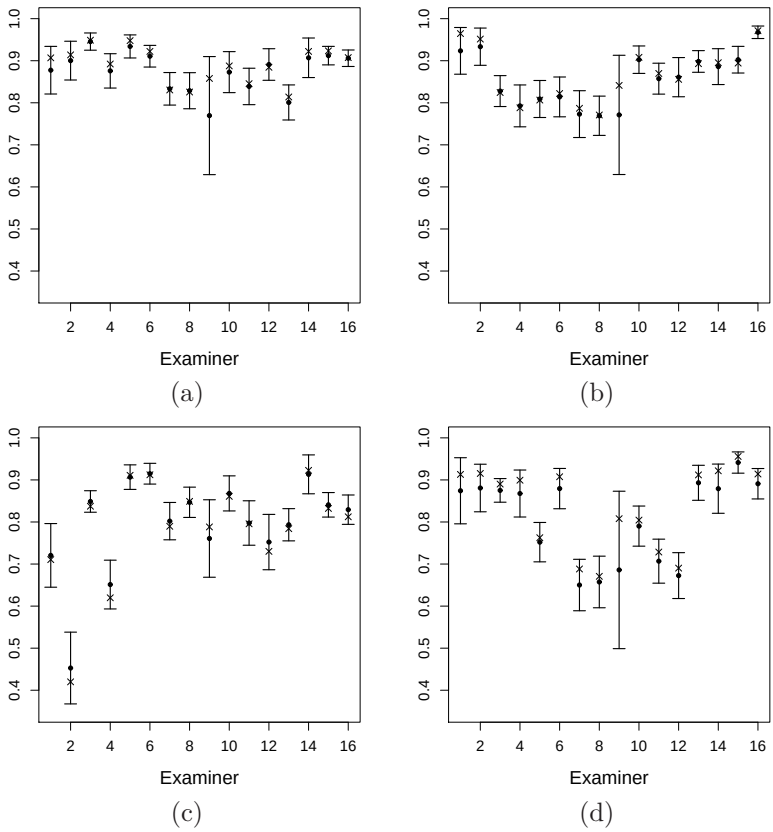


Figure 5.2: Simulated data: true value (×), mean across simulations (•) +/- $\sqrt{\text{MSE}}$ for the sensitivity of each examiner for tooth 16 (panel a), tooth 26 (panel b), tooth 36 (panel c) and , tooth 46 (panel d).

for mislcassification) that showed a significant East-West gradient in the prevalence of CE in Flanders. A possible cause for the apparent trend in CE is a different scoring behavior of the 16 dental examiners and the non-homogeneous spatial distribution of them in the study area (Mwalili et al., 2006).

We fitted the Bayesian version of the models $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$. Based on preliminary analyses, a common effect of the predictors for each tooth was assumed for the initial distributions and for the elements of the transition matrices. In all models, we assume independent $N(0, 10^3)$ prior distributions for the coordinates in $\boldsymbol{\beta}^P$, $\boldsymbol{\beta}^I$, $\boldsymbol{\gamma}^P$, $\boldsymbol{\gamma}^I$ and $\boldsymbol{\alpha}^I$, by taking $\boldsymbol{m}_{\boldsymbol{\beta}^P} = \mathbf{0}_{J+(p-1)}$, $\boldsymbol{m}_{\boldsymbol{\beta}^I} = \mathbf{0}_{J+(q-1)}$, $\boldsymbol{m}_{\boldsymbol{\gamma}^P} = \boldsymbol{m}_{\boldsymbol{\gamma}^I} = \mathbf{0}_{J(J-1)/2}$, $\boldsymbol{m}_{\boldsymbol{\alpha}^I} = \mathbf{0}_{J(J-1)}$, $\boldsymbol{V}_{\boldsymbol{\beta}^P} = \text{diag}\{10^3\}_{J+(p-1)}$, $\boldsymbol{V}_{\boldsymbol{\beta}^I} = \text{diag}\{10^3\}_{J+(q-1)}$, $\boldsymbol{V}_{\boldsymbol{\gamma}^P} = \boldsymbol{V}_{\boldsymbol{\gamma}^I} = \text{diag}\{10^3\}_{J(J-1)/2}$, and $\boldsymbol{V}_{\boldsymbol{\alpha}^I} = \text{diag}\{10^3\}_{J(J-1)}$

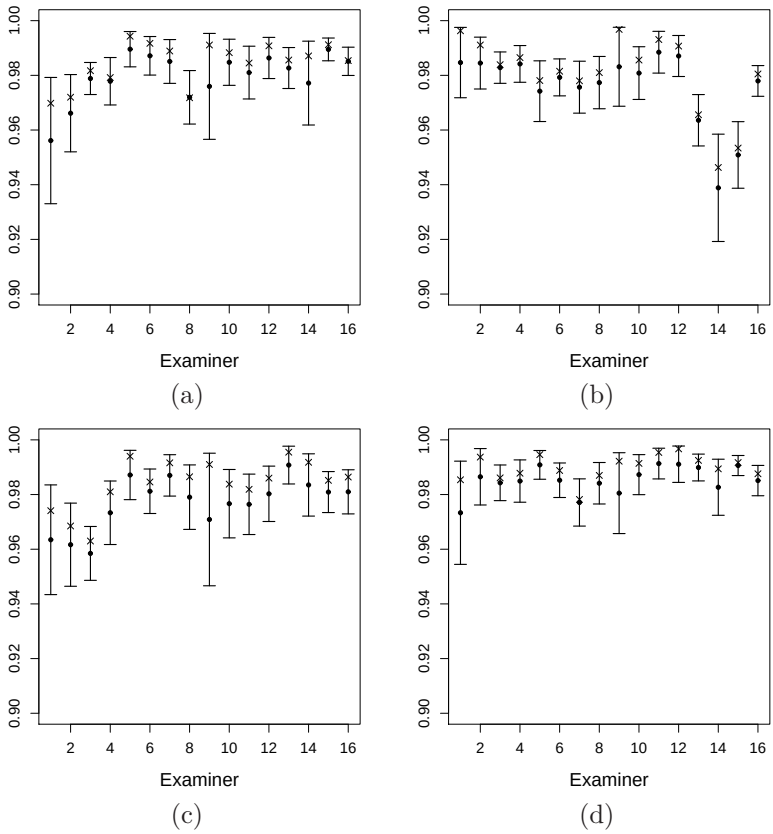


Figure 5.3: Simulated data: true value (×), mean across simulations (•) +/- $\sqrt{\text{MSE}}$ for the specificity of each examiner for tooth 16 (panel a), tooth 26 (panel b), tooth 36 (panel c) and , tooth 46 (panel d).

in the expressions (5.5), (5.6), and (5.7). For models $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, constrained independent uniform priors distributions were assumed by taking $\epsilon^{11}_{(1,qj)} = \epsilon^{11}_{(2,qj)} = \cdots = \epsilon^{11}_{(1,QJ)} = \epsilon^{11}_{(2,QJ)} = \epsilon^{00}_{(1,qj)} = \epsilon^{00}_{(2,qj)} = \cdots = \epsilon^{00}_{(1,QJ)} = \epsilon^{00}_{(2,QJ)} = 1$ in (5.8), and $\epsilon^{11}_{(1,q)} = \epsilon^{11}_{(2,q)} = \cdots = \epsilon^{11}_{(1,Q)} = \epsilon^{11}_{(2,Q)} = \epsilon^{00}_{(1,q)} = \epsilon^{00}_{(2,q)} = \cdots = \epsilon^{00}_{(1,Q)} = \epsilon^{00}_{(2,Q)} = 1$ in (5.9), respectively. Finally, for the additive misclassification model $\boldsymbol{M}_3$, constrained independent $N(0, 10^3)$ prior distributions were assumed for the misclassification parameters by taking $\boldsymbol{m}_{\boldsymbol{\delta}^{00}} = \boldsymbol{m}_{\boldsymbol{\delta}^{11}} = \mathbf{0}_{2J+2Q-2}$ and $\boldsymbol{V}_{\boldsymbol{\delta}^{00}} = \boldsymbol{V}_{\boldsymbol{\delta}^{11}} = \text{diag}\{10^3\}_{2J+2Q-2}$ in expression (5.10).

We run the Markov chain cycle described in Section 5.4 for each model. In each case, a conservative total number of 420000 samples were obtained. The full chain was subsampled every 20 steps after a burn-in period of 20000 samples,

to give a reduced chain of length 20000. Model comparison was performed using the pseudo Bayes factor (PsBF) developed by Geisser & Eddy (1979) and further considered by Gelfand & Dey (1994). The PsBF for the comparison of $\boldsymbol{M}_i$ versus $\boldsymbol{M}_j$ corresponds to the ratio between the pseudo marginal likelihood (PML) for model $\boldsymbol{M}_i$ and model $\boldsymbol{M}_j$. In our context, the PML for model $\boldsymbol{M}_i$ is defined as $\text{PML}_{M_i} = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} P_{M_i}\left(Y^*_{(i,j,k)} \mid \boldsymbol{Y}^*_{[-(i,j,k)]}\right)$, where $P_{M_i}\left(Y^*_{(i,j,k)} \mid \boldsymbol{Y}^*_{[-(i,j,k)]}\right)$ is the posterior predictive distribution for observation $Y^*_{(i,j,k)}$, based on the data $\boldsymbol{Y}^*_{[-(i,j,k)]}$, under model $\boldsymbol{M}_i$, with $\boldsymbol{Y}^*_{[-(i,j,k)]}$ being the observed data matrix that excludes the observation jth observation of subject i in examination k. Therefore, PsBF for model $\boldsymbol{M}_i$ versus model $\boldsymbol{M}_j$ is defined as $\text{PsBF}_{\boldsymbol{M}_i,\boldsymbol{M}_j} = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{P_{\boldsymbol{M}_i}\left(Y^*_{(i,j,k)}|\boldsymbol{Y}^*_{[-(i,j,k)]}\right)}{P_{\boldsymbol{M}_j}\left(Y^*_{(i,j,k)}|\boldsymbol{Y}^*_{[-(i,j,k)]}\right)}$. The individual cross-validation predictive densities known as conditional predictive ordinates (CPO) have also been used. The CPOs measure the influence of individual observations and are often used as predictive model checking tools. The method suggested by Gelfand & Dey (1994) was used to obtain estimates of CPO statistics from the MCMC output.

The log PML for model $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, and $\boldsymbol{M}_3$ was -3984, -3937 and -3956, respectively. Therefore, the $2 \times \log_{10}$PsBF for $\boldsymbol{M}_2$ versus $\boldsymbol{M}_1$ and for $\boldsymbol{M}_2$ versus $\boldsymbol{M}_3$ was 40.8 and 16.5, respectively. These results suggest no evidence for the hypothesis of tooth-specific misclassification parameters for each examiner. Because of that, only the results arising from model $\boldsymbol{M}_2$ are reported here. Table 5.3 (see page 114) shows the posterior means and 95% highest posterior density (95% HPD) credible intervals for the logistic regression coefficients. HPD intervals were computed using the method described by Chen & Shao (1999).

The results suggest that the older the child the higher the prevalence of CE in permanent molars. The lack of other significant covariates on the prevalence of CE might be due to the fact that permanent teeth have recently erupted at the age of seven, and they have not been exposed enough to infectious agents and/or to the well known loss of power associated to the presence of misclassification (see, e.g. Luan et al., 2005). Regarding the incidence of CE, the results indicate that the later the child starts brushing or the higher the number of between-meal snacks, the higher the probability of developing caries. The lack of a significant geographical trend in the prevalence and incidence of CE, supports the hypothesis that the observed geographical gradient is due to the different scoring behavior of the examiners rather than to real local geographical differences.

Posterior means and 95% HPD credible intervals for the association parameters are shown in Table 5.4 (see page 115). The posterior inferences for the within-time conditional log-odds suggest a high positive association in the presence of CE between symmetrically opponent molars and right vertically opponent molars (maxilla versus mandible) at the age of 7. At this age, a non-significant conditional

Table 5.3: Signal-Tandmobiel® data: posterior means and 95% highest posterior density (95% HPD) credible intervals, for the conditionally specified logistic regression coefficients associated to the prevalence and incidence for caries experience in permanent first molars.

	Prevalence		**Incidence**	
	Posterior Mean	**95%HPD**	**Posterior Mean**	**95%HPD**
Intercept T16	-6.16	(-8.86 ; -3.51)	-4.67	(-5.62 ; -3.71)
Intercept T26	-5.63	(-8.42 ; -3.11)	-4.39	(-5.35 ; -3.50)
Intercept T36	-5.75	(-8.38 ; -3.06)	-4.71	(-5.68 ; -3.79)
Intercept T46	-5.59	(-8.25 ; -2.98)	-4.27	(-5.20 ; -3.38)
Startbr	0.10	(-0.01 ; 0.21)	0.09	(0.04 ; 0.13)
Gender	0.23	(-0.02 ; 0.49)	0.10	(-0.01 ; 0.21)
Age	0.35	(0.02 ; 0.67)	-0.03	(-0.08 ; 0.01)
Meals	0.15	(-0.11 ; 0.41)	0.15	(0.05 ; 0.27)
x-ordinate	-0.01	(-0.25 ; 0.25)	0.07	(-0.04 ; 0.17)
y-ordinate	-0.63	(-1.50 ; 0.18)	-0.16	(-0.48 ; 0.15)
Years-exam	-	-	0.39	(-0.05 ; 0.85)

association was found between diagonally opponent teeth. High positive within-time conditional associations were found between symmetrically, right vertically opponent molars and diagonally opponent teeth as the process evolves.

The posterior inference for the across-time odds ratio parameters suggest significant and negative associations between symmetrically and diagonally opponent molars. These results suggest that the probability of developing caries on a tooth is higher when a symmetrically or diagonally opponent molar is affected at the same time but sound at the previous examination, than when it was affected in the previous examination. For instance, $\alpha^I_{16,26} = -2.81$ in Table 5.4 (see page 115) means that the log-odds of developing caries for tooth 26 is higher when tooth 16 is affected at the same time interval than when it was already affected at the previous examination. These results can be explained by the fact that once a tooth is affected by caries, it is probably treated and the infection is no longer spreading in the next examination.

In order to evaluate the posterior evidence against the hypothesis of symmetry in the across time log-odds parameters $\alpha^I_{lj} = \alpha^I_{jl}$, $\forall j \neq l$, the pseudo contour probability (PsCP) for this hypothesis was evaluated. The PsCP was computed based on simultaneous credible bands which were estimated using the method proposed by Besag et al. (1995). The PsCP for the hypothesis of symmetry, defined as 1 minus the smallest credible level for which the null hypothesis is contained

Table 5.4: Signal-Tandmobiel® data: posterior means and 95% highest posterior density (95% HPD) credible intervals of conditional log-odds ratios for caries experience in permanent first molars.

	Parameter	Posterior Mean	95% HPD
		Prevalence	
Within Time Association Parameters	$\gamma^P_{16,26}$	3.93	(2.80 ; 5.09)
	$\gamma^P_{16,36}$	0.67	(-1.43 ; 3.04)
	$\gamma^P_{16,46}$	2.23	(0.23 ; 4.01)
	$\gamma^P_{26,36}$	1.45	(-1.02 ; 3.60)
	$\gamma^P_{26,46}$	-0.16	(-2.40 ; 2.34)
	$\gamma^P_{36,46}$	2.56	(1.25 ; 3.85)
		Incidence	
Within Time Association Parameters	$\gamma^I_{16,26}$	3.84	(3.16 ; 4.53)
	$\gamma^I_{16,36}$	2.36	(1.31 ; 3.37)
	$\gamma^I_{16,46}$	1.08	(0.15 ; 2.09)
	$\gamma^I_{26,36}$	-0.63	(-1.89 ; 0.71)
	$\gamma^I_{26,46}$	2.12	(1.21 ; 3.03)
	$\gamma^I_{36,46}$	3.70	(3.07 ; 4.36)
Across Time Association Parameters	$\alpha^I_{16,26}$	-2.81	(-3.87 ; -1.76)
	$\alpha^I_{16,36}$	-1.48	(-2.75 ; -0.24)
	$\alpha^I_{16,46}$	-0.38	(-1.54 ; 0.77)
	$\alpha^I_{26,16}$	-1.91	(-2.91 ; -0.97)
	$\alpha^I_{26,36}$	0.80	(-0.67 ; 2.27)
	$\alpha^I_{26,46}$	-1.57	(-2.63 ; -0.45)
	$\alpha^I_{36,16}$	-2.09	(-3.32 ; -0.93)
	$\alpha^I_{36,26}$	1.05	(-0.39 ; 2.49)
	$\alpha^I_{36,46}$	-2.46	(-3.45 ; -1.52)
	$\alpha^I_{46,16}$	-0.46	(-1.54 ; 0.58)
	$\alpha^I_{46,26}$	-1.03	(-2.11 ; 0.01)
	$\alpha^I_{46,36}$	-1.24	(-1.98 ; -0.54)

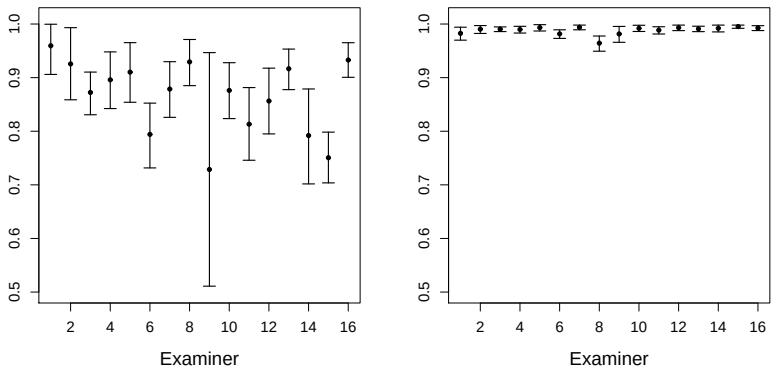


Figure 5.4: Signal-Tandmobiel® data: posterior means and 95% highest posterior density credible intervals for examiner's sensitivity (panel a) and specificity (panel b).

in the corresponding simultaneous credible bands (see, e.g. Held, 2004), was 0.059. This result suggests that there is no strong evidence against the hypothesis of symmetry in the caries process across time.

Finally, Figure 5.4 shows the posterior means and 95% HPD credible intervals for the sensitivity and specificity for each examiner. The results suggest a greater variability in the sensitivity than in the specificity estimates, which can be explained by the low prevalence and incidence of CE. All examiners showed a sensitivity greater than 0.72, with rather narrow 95% HPD credible intervals, with one exception. The latter result is explained by the fact that this examiner was only involved in the first two years of the ST study, having less information for the estimation of his parameters. The posterior means for the specificity parameters were higher than 0.96 for all examiners.

5.6 Concluding Remarks

We have proposed a multivariate HMM for monotone binary processes. Although the methodology was motivated by an oral health application, it can be applied to any situation where correlated binary responses have an absorvent state and are subject to misclassification, such as studies about kidney failure or vision loss.

In the proposal, the multivariate initial distributions and the conditional distributions associated to the Markov transition matrices are defined by conditionally specified logistic regression models that account for the effect of

covariates on the prevalence and incidence in a conditional but population-average fashion. The association structure is taken into account by within- and across-time odds ratio parameters. Three misclassification models were proposed that consider the existence of different classifiers and different structures in the examiner-specific misclassification errors.

We provided empirical evidence showing that, under the settings of our motivating example and with simple restrictions on the parameter space, the model parameters in the proposed multivariate HMM can be estimated from the raw data only, thus avoiding the need of external information on the misclassification parameters. The results suggest that even under the use of uniform priors on the misclassification parameters, unbiased and relatively precise estimates can be obtained. We noted that if external information on the misclassification parameters is available, this can be easily incorporated into the multivariate HMM specification.

Several extensions of this work can be done. Justified by the existence of easy/difficult to diagnose subjects, the relaxation of some of the assumptions (A.1) - (A.6) could be of interest; for instance, a possible improvement of the scoring behavior of the examiners accross the study could be considered. The inclusion of time-dependent within- and across-time association parameters or their dependence on covariates can also be pursued, as well as the extension of the model for handling multinomial data.

Acknowledgements

The first author is supported by the National Scholarship for Doctoral Studies 2009, Conicyt (Chile). The second author is supported by the Fondecyt grant 3095003. The first and third authors are supported by the Research Grant OT/05/60 and they also acknowledge the partial support from the Interuniversity Attraction Poles Program P6/03, Belgian State, Federal Office for Scientific, Technical and Cultural Affairs. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel® study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (L-BioStat, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

References

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

Arnold, B., Castillo, E. & Sarabia, J. M. (1992). *Conditionally Specified Distributions*. Springer-Verlag.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* 81 767–775 (correction: 1997, 84, 989).

Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* 10 3–66.

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

Buonaccorsi, J. P. (2010). *Measurement Error*. New York, USA: Chapman & Hall/CRC.

Chen, M. H. & Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational Graphical Statistics* 8(1) 69–92.

Cook, R. J., Ng, E. T. M. & Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* 56 1109–1117.

Espeland, M. A., Murphy, W. C. & Leverett, D. H. (1988). Assessing diagnostic reliability and estimating incidence rates associated with a strictly progressive disease: dental caries. *Statistics in Medicine* 7 403–416.

Espeland, M. A., Platt, O. S. & Gallagher, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. *Journal of the American Statistical Association* 84(408) 972–979.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 7 57–68.

García-Zattera, M. J., Jara, A., Lesaffre, E. & Declerck, D. (2007). Conditional independence of a multivariate binary data with an application in caries research. *Computational Statistics and Data Analysis* 51 3223–3234.

García-Zattera, M. J., Mutsvari, T., Jara, A., Declerck, D. & Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine* (To appear).

Geisser, S. & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* 74 153–160.

Gelfand, A. E. & Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* 56 501–514.

Geng, Z. & Asano, C. (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics* 8 2935–2954.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* 58 342–352.

Heagerty, P. J. & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science* 15 1–26.

Held, L. (2004). Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics* 13 20–35.

Joe, H. & Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.

Jurek, A. M., Greenland, S., Maldonado, G. & Church, T. R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *International Journal of Epidemiology* 34 680–687.

Küchenhoff, H. (2009). Misclassification and measurement error in oral health. In E. Lesaffre, J. Feine, B. Leroux & D. Declerck, eds., *Statistical and Methodological Aspects of Oral Health Research.* Chichester, UK: Wiley, 279–294.

Küchenhoff, H., Mwalili, S. M. & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 2 85–96.

Luan, X., Pan, W., Gerberich, S. G. & Carlin, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine* 24 2221–2234.

Magder, L. S. & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 146 195–203.

Mwalili, S. M., Lesaffre, E. & 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* 54(1) 77–93.

Mwalili, S. M., Lesaffre, E. & Declerck, D. (2006). A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. *Journal of the Royal Statistical Society, Series C* 1 77–93.

Nagelkerke, N. J. D., Chunge, R. N. & Kinot, S. N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* 9 1211–1219.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics* 31 705–767.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86(4) 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* 58 675–683.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Reis, A., Medeiros Mendes, F., Angnes, V., Angnes, G., Miranda Grande, R. H. & Dourado Loguercio, A. (2006). Performance of methods of occlusal caries detection in permanent teeth under clinical and laboratory conditions. *Journal of Dentistry* 34 89–96.

Rekaya, R., Weigel, K. A. & Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics* 57 1123–1129.

Rosychuk, R. J. & Islam, M. S. (2009). Parameter estimation in a model for misclassified Markov data - a Bayesian approach. *Computational Statistics and Data Analysis* 53 3805–3816.

Rosychuk, R. J. & Thompson, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 19 394–404.

Rosychuk, R. J. & Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine* 22 2035–2055.

Schmid, C. H., Segal, M. R. & Rosner, B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference* 42(1–2) 1–18.

Singh, A. C. & Rao, J. N. K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian labour force survey. *Journal of the American Statistical Association* 90(430) 478–488.

Smith, B. J. (2007). An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software* 21(11) 1–37.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* 65(331) 1350–1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. *Biometrics* 27 935–944.

Tierney, R. M. (1994). Markov chains for exploring the posterior distribution (with discussion). *Annals of Statistics* 22 1701–1762.

Vanobbergen, J., Martens, L., Lesaffre, E. & 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.

Part III

Concluding Remarks

Chapter 6

General Conclusions and Further Research

In this thesis, we have evaluated and proposed models for the analysis of multivariate binary data. In Chapters 2 and 3, we have studied the properties of two models to analyze multivariate binary data, with respect to the interpretation and the effect of the misclassification on the inferences of their association parameters. In Chapters 4 and 5, uni- and multi-variate models for the analysis of longitudinal monotone binary data subject to misclassification were developed. In this chapter we summarize the main conclusions obtained throughout this thesis and highlight some topics for further research.

6.1 General Conclusions

The interpretation of the association parameters associated with two population-average models for the analysis of multivariate binary data were studied and illustrated in Chapter 2. On the one hand, we considered the multivariate probit model (MPM) (Ashford & Sowden, 1970), where the binary variables can be interpreted as a discretized version of correlated Gaussian underlying latent variables, which are manifested through a threshold specification. In the MPM, the association between the binary responses is induced by the correlation matrix of the multivariate normal latent random vectors. On the other hand, we considered the conditionally specified logistic regression model (CSLRM) (Joe & Liu, 1996), where the association parameters have a direct interpretation on the binary scale. In the CSLRM, the association is characterized by conditional odds ratios between pairs of binary variables, given the remaining binary responses and covariates. We showed that conditional independence assumptions on the underlying continuous latent variables, evaluated by the partial correlation matrix in the MPM, are not

transferred to the binary scale. This result implies that conditional independence structure for binary variables cannot be evaluated using the information of the tetrachoric correlation coefficients. Moreover, this result provides a possible explanation for the existence of spurious associations or associations without etiological basis. For instance, based on data obtained from the first year of the ST study, we found a high association of caries experience (CE) between diagonally opponent molars, which was believed to be the result of transitivity and to disappear by conditioning on the CE status of the other teeth. However, using models for multivariate binary data (e.g. the CSLRM), we found that this diagonal association did not disappear. When the association was explored on a latent scale, e.g. by using a MPM, conditional independence could be concluded. This contrast was confirmed when using other models. A similar phenomenon occurs when the continuous data are discretized for the sake of the analysis, a practice that is often seen in medical research.

Motivated by the lack of literature about the effect of misclassification on the estimation of the association parameters for multivariate binary data, in Chapter 3 we explored the effect of response misclassification on the small sample behavior of naive estimators of the association parameters of the MPM and CSLRM. We found that, under either non-differential or differential misclassification, the maximum likelihood estimators of the association parameters can be strongly biased towards the null of no association, if the misclassification process is ignored. Under non-differential misclassification the bias and mean squared error of naive estimators of the association parameters in the MPM and CSLRM are greater than the ones obtained under no misclassification when sample size, degree of association or misclassification errors increase. Under a differential misclassification process, the effect on the estimation of the association parameters is greater when there is a negative association between the predictors and the precision of the classification procedure.

The monotone feature of CE, as defined throughout this thesis, led us to investigate whether a monotone process subject to misclassification contains enough information to identify all the model parameters, including the misclassification parameters, without adding extra information. In Chapter 4 we gave theoretical and empirical arguments to show that, the parameters associated with simple binary hidden Markov models (HMMs), can be estimated without the need of external information. In order to take into account in a better manner the characteristics of the ST data, an extension of the simple HMM was also proposed in this chapter. This extension allowed us to assess the effect of covariates on the parameters of the Markov model and the existence of different classifiers. We showed that, under the settings of our motivating example, the parameters can be estimated without any external information in a Bayesian version of the model.

In order to gain power for the hypothesis testing and to understand the within- and across-time association structure of a multivariate binary monotone process,

in Chapter 5 we proposed an extension to the HMM described in Chapter 4. This extension was based on a generalization of the CSLRM discussed in Chapters 2 and 3. As in Chapter 4, we provided empirical evidence to show that, under the settings of the ST study and with simple restrictions on the parameter space, all the model parameters can be estimated from the main data only, avoiding the need of extra information about the misclassification parameters.

The results of Chapters 4 and 5, show that in longitudinal studies external information is not necessary to estimate all the model parameters. This shows another advantage of longitudinal over cross-sectional studies. Because, in general, cross-sectional data contain no information about the misclassification parameters, the approaches to correct for misclassification in this context rely on the existence of extra data or on expert knowledge. Since this information is usually difficult to obtain, the possibility of estimate the misclassification parameters using only the main data, is an important characteristic associated to longitudinal studies.

6.2 Further Research

Further research could focus on some of the topics considered in this thesis. As we pointed out throughout this thesis, misclassification occurs frequently in epidemiology and several approaches to correct for it have been proposed. In Chapter 3 we highlighted the impact of misclassification on the inferences about the association parameters. We found that, regardless the type of misclassification process, the maximum likelihood estimators of the association parameters are strongly biased towards the null. Based on this result, the conclusions obtained in Chapter 2 regarding conditional independence on the latent and observed scales, should be verified. The lack of significance found in the partial tetrachoric correlation coefficients could be due to the underestimation induced by the misclassification of CE. In other words, it might be that the same association structure is concluded using either the latent or the observed scale, once correcting for misclassification. Thus, in general, correction for misclassification would be an intuitively solution to obtain better inferences. However, the benefits of correcting for misclassification should be explored before applying correction methods since they can reduce the bias of the estimators, but increase their variability (see, e.g. Luan et al., 2005).

In Chapters 4 and 5 we proposed HMMs for monotone binary data. These models can be extended to handle covariate measurement error and multinomial or multi-state data to model, for instance, different degrees of CE. In addition, the existence of easy/difficult to diagnose subjects and possible improvements of the scoring behavior of the examiners across the study can be considered in the specification of the misclassification process. Given the hierarchical structure of dental data, i.e surfaces in teeth, teeth in jaws and jaws in mouths, the extension of the proposed

HMM of Chapter 4 considering a multilevel modelling which takes into account the spatial structure, could also be considered.

The HMM of Chapter 5 allows us to study the within- and across-time association structure of monotone multivariate binary data. In this context, the inclusion of time-dependent association parameters can be explored as well as their dependence on covariates.

An important aspect associated with models for data subject to measurement error is the identifiability of the parameters. The results of the simulation studies in Chapters 4 and 5 suggest that the regression coefficients and the misclassification parameters can be estimated from the data, when they have a similar structure than the ST study. The formal study of this property in general contexts seems to be needed. The identification study of hidden alternating binary and multinomial Markov models can also be considered.

The connection between HMMs and survival models is also subject of further research. Time to caries, which is defined as the time length of onset of caries since emergence, can be modeled as a function of covariates. Since, the prevalence and incidences can be written as functions of the survival function for the time to event, the resulting model corresponds to a model for misclassified survival data. This involves the analyzes of misclassified uni- and multivariate doubly-interval censored data, since time to emergence and time to caries are both censored. Very little has been written on models for multivariate doubly-interval-censored data with the exception of the frailty models of Komárek & Lesaffre (2008) and Jara et al. (2010). However, none of these models have taken into account the misclassification problem.

References

Ashford, J. R. & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics* 26 535–546.

Komárek, A. & Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association* 103(482) 523–533.

Jara, A., Lesaffre, E., De Iorio, M. & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Annals of Applied Statistics* (To appear).

Joe, H. & Liu, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.

Luan, X., Pan, W., Gerberich, S. G. & Carlin, B. P. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine* 24 2221–2234.

Part IV

Supplementary Material

Appendix A

Supplementary Material for Chapter 2

A.1 Proof that conditional independence on the latent scale does not imply conditional independence on the binary scale

Let $\boldsymbol{V} \sim N_3(\boldsymbol{\mu}, \boldsymbol{R})$, where,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{V_1} \\ \mu_{V_2} \\ \mu_{V_3} \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

When $V_1 \perp\!\!\!\perp V_2 | V_3$ holds, then e.g.,

$$\begin{aligned} P(V_1 > 0, V_2 > 0 | V_3 > 0) &= \frac{\int_0^\infty \int_0^\infty \int_0^\infty f(v_1|v_3)\ f(v_2|v_3)\ f(v_3)\ dv_1\ dv_2\ dv_3}{\int_0^\infty f(v_3)\ dv_3} \\ &= \frac{\int_0^\infty \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right) f(v_3)\ dv_3}{\int_0^\infty f(v_3)\ dv_3} \\ &= \int_{\mathcal{A}} \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right)\ h(x)\ dx, \end{aligned} \tag{A.1}$$

where, $\mathcal{A} = [0, \infty)$, $\mu_1^* = \mu_{V_1} + \rho_{13}(x - \mu_X)$, $\mu_2^* = \mu_{V_2} + \rho_{23}(x - \mu_X)$, $\sigma_1^{2*} = \sqrt{1 - \rho_{13}^2}$, $\sigma_2^{2*} = \sqrt{1 - \rho_{23}^2}$, and $h(x) \equiv TN_{(0,\infty)}(\mu_X, \sigma_X^2)$, which means Truncated Normal between zero and infinity with location μ_X and scale σ_X^2.

On the other hand, when $Y_1 \perp\!\!\!\perp Y_2 | Y_3$ holds,

$$\begin{aligned} P\left(Y_1 = 1, Y_2 = 1 | Y_3 = 1\right) &= \mathrm{P}(Y_1 = 1 | Y_3 = 1)\mathrm{P}(Y_2 = 1 | Y_3 = 1) \\ &= \mathrm{P}(V_1 > 0 | V_3 > 0)\mathrm{P}(V_2 > 0 | V_3 > 0) \\ &= \int_{\mathcal{A}} \Phi\left(\frac{\mu_1^*}{\sigma_1^*}\right) h(x)dx \times \int_{\mathcal{A}} \Phi\left(\frac{\mu_2^*}{\sigma_2^*}\right) h(x)dx. \quad \text{(A.2)} \end{aligned}$$

Expression (A.1) is greater than expression (A.2), because

$$E\left(g_1\left(X\right) g_2\left(X\right)\right) \geq E\left(g_1\left(X\right)\right) E\left(g_2\left(X\right)\right), \tag{A.3}$$

holds for all real-valued functions g_1 and g_2 which are nondecreasing (in each component) and are such that the expectations in (A.3) exist. The equality holds iff $g_1(X) = c$ or $g_2(X) = c$ (a.s.) (see, e.g., Esary et al. 1967). This shows that conditional independence on the latent scale does not imply conditional independence on the observed scale and vice versa.

■

References

Esary, J. D., Proschan, F. & Walkup, D. W. (1967). Association of random variables, with applications. *Annals of Mathematical Statistics* 38 1466–1474.

Appendix B

Supplementary Material for Chapter 3

B.1 Results of the Multivariate Probit Model under Non-Differential Misclassification

Table B.1: Bias of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = 0.85$ and $\tau^{00} = 0.95$. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	5.5	4.7	11.9	6.5	4.3	40.0
	400	9.3	7.9	5.3	17.3	10.3	4.7
	1000	12.3	21.3	16.4	13.0	32.0	20.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	10.5	14.6	10.8	11.1	53.0	9.6
	400	12.5	19.7	12.8	14.3	51.5	11.7
	1000	43.0	28.3	28.5	100.0	33.0	24.8
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	12.5	18.7	9.2	15.1	59.5	4.7
	400	41.8	35.6	32.3	57.0	28.0	25.0
	1000	60.0	48.4	60.3	54.5	28.3	110.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	24.5	20.3	21.3	17.0	6.1	11.7
	400	34.1	51.0	76.3	6.2	57.5	37.3
	1000	74.0	296.0	98.7	12.1	36.7	110.0

Table B.2: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = 0.85$ and $\tau^{00} = 0.95$. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	1.4	1.5	1.4	1.1	1.2	1.1
	400	1.5	1.7	1.5	1.2	1.3	1.1
	1000	2.4	2.4	2.2	1.8	1.8	1.6
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	3.0	2.5	2.9	1.4	1.2	1.2
	400	4.1	4.6	3.8	1.6	1.6	1.5
	1000	8.5	8.3	8.5	3.0	2.3	2.5
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	5.4	6.1	6.1	1.1	1.1	1.2
	400	11.8	11.8	10.9	1.5	1.5	1.8
	1000	30.5	20.7	20.7	2.8	2.8	2.8
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	16.6	17.3	16.7	0.9	0.9	0.9
	400	34.3	33.7	33.3	1.1	1.3	1.2
	1000	90.0	90.0	90.0	2.1	2.1	2.4

Table B.3: Bias of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = 0.95$ and $\tau^{00} = 0.85$. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	6.1	5.0	13.1	7.4	4.6	46.0
	400	10.8	9.0	6.5	20.5	11.7	6.0
	1000	14.9	25.5	20.6	15.8	38.5	26.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	12.3	18.2	12.7	13.0	73.0	11.3
	400	16.1	25.3	16.3	19.4	70.0	15.3
	1000	54.5	36.3	36.7	131.0	43.7	33.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	16.3	24.3	11.9	21.1	83.0	6.4
	400	55.5	47.1	42.4	81.5	39.8	34.4
	1000	81.5	64.8	81.5	80.0	39.5	160.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	33.6	26.8	28.9	26.1	8.0	17.5
	400	46.2	70.0	105.0	8.6	85.0	57.0
	1000	102.0	406.0	135.0	18.3	53.7	159.0

Table B.4: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = 0.95$ and $\tau^{00} = 0.85$. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	1.2	1.2	1.2	0.9	0.9	0.9
	400	1.5	1.6	1.6	1.1	1.2	1.1
	1000	2.8	2.8	2.8	2.0	2.0	2.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	3.3	3.1	3.3	1.3	1.3	1.2
	400	5.9	6.7	5.5	1.9	2.1	1.8
	1000	12.8	12.5	12.8	4.0	3.3	3.5
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	8.3	9.4	8.9	1.4	1.4	1.3
	400	19.7	19.3	17.4	2.3	2.1	2.4
	1000	54.5	36.0	36.3	4.8	4.7	4.8
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	29.3	28.1	28.9	1.1	0.9	1.2
	400	59.7	61.0	60.7	1.5	1.9	1.8
	1000	169.0	167.0	166.0	3.9	3.6	4.1

Table B.5: Bias of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.95$. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	4.1	3.3	7.8	4.8	3.1	25.5
	400	6.2	5.3	3.9	11.3	6.6	3.6
	1000	8.9	14.3	11.6	9.2	20.5	13.7
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	7.5	10.8	7.3	7.5	41.0	5.8
	400	9.4	14.6	8.7	10.7	38.0	7.1
	1000	30.0	19.7	19.5	68.0	22.0	16.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	9.2	13.6	6.5	11.3	42.0	3.0
	400	29.5	25.6	21.8	39.0	20.5	14.6
	1000	43.3	34.0	42.3	39.0	18.5	72.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	17.4	13.8	14.9	12.0	3.6	8.2
	400	23.4	34.7	52.5	4.1	35.5	25.3
	1000	49.5	201.0	67.0	7.3	23.7	72.0

Table B.6: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.95$. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	**Sample Size**	**Correlation**			**Partial Correlation**		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	1.2	1.1	1.0	1.0	0.9	0.9
	400	1.2	1.3	1.3	1.0	1.1	1.1
	1000	1.6	1.4	1.6	1.4	1.2	1.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	2.0	1.8	1.8	1.1	1.0	0.9
	400	2.7	3.0	2.2	1.2	1.2	1.0
	1000	4.5	4.5	4.3	1.8	1.5	1.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	3.5	3.8	3.4	1.0	1.0	0.8
	400	6.5	6.7	5.4	1.1	1.2	1.0
	1000	16.5	10.7	10.7	1.8	1.7	1.7
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	9.1	8.7	9.0	0.8	0.7	0.9
	400	17.0	16.3	16.7	0.8	0.9	0.9
	1000	41.0	42.0	43.0	1.1	1.3	1.4

B.2 Results of the Conditionally Specified Logistic Regression Model under Non-Differential Misclassification

Table B.7: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under non-differential misclassification and $\tau^{11} = 0.85$ and $\tau^{00} = 0.95$. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification, B^* and MSE^* respectively, and the corresponding values of the MLE when there is no misclassification, B and MSE respectively.

True Values	Sample Size	$\lvert B^*/B \rvert$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	14.3	10.1	6.7	1.1	1.3	1.2
	400	23.2	21.9	153.9	1.4	1.6	1.6
	1000	16.0	403.7	31.7	2.2	2.4	2.2
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.1$	200	59.1	245.7	103.3	3.3	3.2	3.0
	400	336.3	185.2	474.2	6.2	4.8	5.4
	1000	8391.9	496.8	869.8	12.7	13.5	11.9
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	27.4	31.5	40.8	6.0	6.4	5.7
	400	49.8	42.5	325.7	12.1	12.2	14.3
	1000	67692.4	518.1	118.4	33.9	36.1	34.6
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	19.5	23.4	17.4	7.7	7.3	7.8
	400	134.6	331.7	37.1	17.6	16.8	17.6
	1000	118.2	253.7	131.5	46.6	48.1	45.3

Table B.8: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under non-differential misclassification and $\tau^{11} = 0.95$ and $\tau^{00} = 0.85$. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification, B^* and MSE^* respectively, and the corresponding values of the MLE when there is no misclassification, B and MSE respectively.

True Values	**Sample Size**	$\|B^*/B\|$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	19.0	11.5	8.1	1.0	1.2	1.1
	400	27.3	25.4	175.5	1.5	1.7	1.5
	1000	18.7	457.3	36.4	2.4	2.5	2.3
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.1$	200	61.6	243.8	103.7	3.2	3.1	2.9
	400	323.5	188.8	475.6	5.6	4.9	5.3
	1000	8155.2	474.6	847.1	12.0	12.3	11.2
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	22.1	25.5	32.7	4.1	4.4	3.9
	400	39.7	34.0	259.5	8.0	8.0	9.4
	1000	52704.3	403.7	93.3	20.8	22.3	21.9
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	13.4	16.3	11.8	3.9	3.7	3.8
	400	93.5	230.1	25.4	8.7	8.4	8.5
	1000	82.2	177.8	91.6	22.8	23.9	22.2

Table B.9: Bias and mean squared error (MSE) of the estimators of the association parameters of the conditionally specified logistic regression model under non-differential misclassification and $\tau^{11} = \tau^{00} = 0.95$. The results correspond to the ratio between the bias and MSE of the naive maximum likelihood estimator (MLE) under misclassification, B^* and MSE^* respectively, and the corresponding values of the MLE when there is no misclassification, B and MSE respectively.

True Values	**Sample Size**	$\|B^*/B\|$			MSE^*/MSE		
		γ_{12}	γ_{13}	γ_{23}	γ_{12}	γ_{13}	γ_{23}
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 0.4$	200	10.1	6.1	4.1	1.0	1.1	1.0
	400	14.8	13.9	96.0	1.1	1.1	1.2
	1000	10.1	253.5	20.0	1.4	1.5	1.4
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.1$	200	34.9	137.8	59.6	1.7	1.5	1.5
	400	186.9	104.7	281.0	2.6	2.1	2.5
	1000	4581.8	266.6	485.2	4.4	4.5	4.2
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 1.8$	200	14.1	16.7	21.1	2.1	2.3	2.0
	400	25.7	22.4	171.7	3.7	3.9	4.6
	1000	35725.4	274.1	63.1	10.0	10.8	10.5
$\gamma_{12} = \gamma_{13} = \gamma_{23} = 2.5$	200	10.7	12.9	9.6	2.7	2.5	2.7
	400	75.8	187.3	20.8	6.0	5.7	5.9
	1000	68.1	146.5	75.7	15.8	16.4	15.3

B.3 Results of the Multivariate Probit Model under Differential Misclassification

Table B.10: Bias of the estimators of the association parameters of the multivariate probit model under differential misclassification with negative association between the precision of the classification and the continuous predictor. The results correspond to the absolute ratio between the bias of the naive maximum likelihood estimator, B^*, and the bias of the maximum likelihood estimator when there is no misclassification, B, i.e. $|B^*/B\ |$.

True Values	Sample Size	Correlation			Partial Correlation		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	7.8	6.9	17.3	9.7	6.6	62.0
	400	14.7	12.1	8.4	29.5	16.4	7.9
	1000	20.4	35.3	28.0	22.6	55.5	36.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	16.4	23.3	16.6	18.7	97.5	16.0
	400	20.6	31.8	20.6	26.6	91.5	20.6
	1000	70.8	47.0	47.2	181.0	60.0	45.3
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	20.2	30.5	14.7	27.9	113.5	8.4
	400	69.5	59.6	52.8	110.0	54.8	45.0
	1000	104.5	82.4	104.3	111.0	53.8	221.0
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	42.5	34.1	36.0	35.9	11.5	23.3
	400	59.7	88.8	133.8	12.5	116.0	78.3
	1000	133.5	530.0	176.7	26.2	77.0	231.0

Table B.11: Mean squared error (MSE) of the estimators of the association parameters of the multivariate probit model under differential misclassification with negative association between the precision of the classification and the continuous predictor. The results correspond to the ratio between the MSE of the naive maximum likelihood estimator, MSE^* under misclassification and the MSE of the maximum likelihood estimator when there is no misclassification, MSE, i.e. MSE^*/MSE .

True Values	**Sample Size**	**Correlation**			**Partial Correlation**		
		ρ_{12}	ρ_{13}	ρ_{23}	$\rho_{12.3}$	$\rho_{13.2}$	$\rho_{23.1}$
$\rho_{12} = \rho_{13} = \rho_{23} = 0.2$	200	1.7	1.9	1.7	1.2	1.4	1.2
	400	2.5	2.6	2.3	1.8	1.8	1.6
	1000	4.8	4.6	4.8	3.4	3.2	3.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.4$	200	5.6	4.9	5.2	2.1	1.9	1.8
	400	9.4	10.2	8.5	3.2	3.1	2.9
	1000	21.0	20.8	21.0	7.4	6.0	6.2
$\rho_{12} = \rho_{13} = \rho_{23} = 0.6$	200	12.4	14.4	13.4	2.1	2.1	2.0
	400	30.5	30.5	26.9	3.7	3.6	3.8
	1000	89.0	57.7	59.3	8.8	8.3	8.8
$\rho_{12} = \rho_{13} = \rho_{23} = 0.8$	200	46.6	45.1	44.4	1.8	1.6	1.8
	400	99.3	97.3	98.0	2.8	3.2	3.0
	1000	288.0	285.0	285.0	7.4	7.1	8.1

Appendix C

Supplementary Material for Chapter 4

C.1 Federation Dentaire Internationale Notation for Permanent Teeth

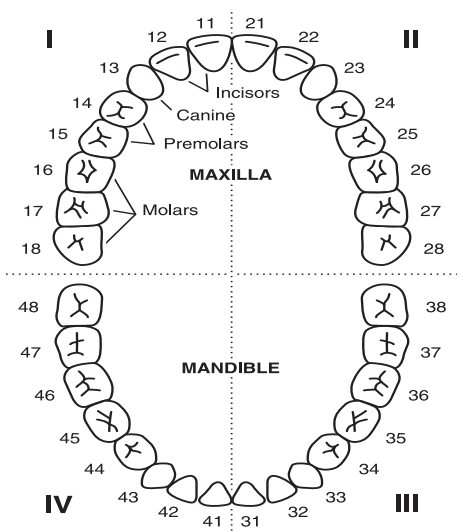


Figure C.1: Federation Dentaire Internationale notation for the position of permanent teeth. Maxilla = upper jaw, mandible = lower jaw. The first and the fourth quadrants are at the right-hand side of the subject, the second and the third quadrants are at the left-hand side of the subject.

C.2 Proof of Proposition 4.1

The proof is based on the expression of the parameters of interest as functions of identified parameters and follows similar arguments to the ones used by Carlos and Sening (1968) in the context of the derivation of method of moments estimators. It is easy to show that in the HMM for two time points the sampling probabilities, the probabilities of the observed response patterns, are identified parameters. Let $Q_1 = P(Y_1^* = 0, Y_2^* = 0)$, $Q_2 = P(Y_1^* = 0, Y_2^* = 1)$, $Q_3 = P(Y_1^* = 1, Y_2^* = 0)$ and $Q_4 = P(Y_1^* = 1, Y_2^* = 1)$ be the corresponding sampling probabilities. Under the restriction $\tau = \tau_{01} = \tau_{10}$, these sampling probabilities are the following functions of the parameters of interest

$$\begin{aligned}
Q_1 &= (1-p)(1-q_1)(1-\tau)^2 + (1-p)q_1(1-\tau)\tau + p\tau^2, \\
Q_2 &= (1-p)q_1(1-\tau)^2 + (1-p)(1-q_1)(1-\tau)\tau + p(1-\tau)\tau, \\
Q_3 &= (1-p)(1-q_1)(1-\tau)\tau + (1-p)q_1\tau^2 + p(1-\tau)\tau, \\
Q_4 &= (1-p)q_1(1-\tau)\tau + (1-p)(1-q_1)\tau^2 + p(1-\tau)^2.
\end{aligned}$$

Notice that the probability of an observed reversal, Q_3, can be rewritten as follows

$$\begin{aligned}
Q_3 &= (1-p)(1-q_1)(1-\tau)\tau + (1-p)q_1\tau^2 + p(1-\tau)\tau, \\
&= \tau[1-\tau-\underbrace{(1-p)q_1(1-2\tau)}], \\
&= \tau[1-\tau- \quad (Q_2-Q_3)], \\
&= -\tau^2 + (1-Q_2+Q_3)\tau. \qquad \text{(C.1)}
\end{aligned}$$

Expression (C.1) corresponds to a two degree polynomial in τ which can be used to express the misclassification parameter as a function of the identified sampling probabilities. By assuming $\tau < 0.5$ this polynomial has a unique solution which is given by

$$\tau = 0.5\left[(1-Q_2+Q_3) - \sqrt{(1-Q_2+Q_3)^2 - 4Q_3}\right]. \qquad \text{(C.2)}$$

The restriction $\tau < 0.5$ implies the existence of a unique solution since it is equivalent to consider the part of the quadratic function smaller than its inflection point, $\frac{1-(Q_2-Q_3)}{2}$. In fact, as $\frac{1-(Q_2-Q_3)}{2} = 1-(1-p)q_1(1-2\tau)$, $\tau < \frac{1}{2} \Longrightarrow \tau < \frac{1-(Q_2-Q_3)}{2}$. Therefore, since τ is a function of the identified parameters Q_2 and Q_3 under the restriction $\tau < 0.5$, the misclassification parameter is identified.

Now notice that $Q_2 - Q3 = (1-p)q_1(1-2\tau)$. It follows that $\Lambda = (1-p)q_1 = \frac{Q_2-Q3}{(1-2\tau)}$ is a function of the identified parameters τ, Q_2, and Q_3, and, therefore, is also an identified parameter. By noticing that

$$\begin{aligned} Q_1 - Q_4 &= [1-(1-p)q_1 - 2p](1-2\tau), \\ &= (1-\Lambda-2p)(1-2\tau), \end{aligned}$$

it is possible to show that p is a function of the identified parameters Q_1, Q_4, τ, and Λ, given by

$$\begin{aligned} p &= \frac{1}{2}\left[1-\Lambda-\frac{Q_1-Q_4}{1-2\tau}\right], \\ &= \frac{1}{2}\left[1-\frac{Q_2-Q_3}{(1-2\tau)}-\frac{Q_1-Q_4}{1-2\tau}\right]. \end{aligned}$$

Therefore, p is also an identified parameter. Finally, as the incidence is a function of the identified parameters Λ and p, given by $q_1 = \frac{\Lambda}{(1-p)}$, it is an identified parameter. Thus, we conclude that the constraints $\tau = \tau_{10} = \tau_{0,1}$ and $\tau < 0.5$, are sufficient identifying restrictions for the parameters in the simple HMM. The proof showing that the constraints $\tau = \tau_{10} = \tau_{0,1}$ and $\tau > 0.5$ are sufficient for the identification of the HMM parameters follows the same arguments. However, in this case τ is the following function of the identified parameters Q_2 and Q_3,

$$\tau = 0.5\left[(1-Q_2+Q_3) + \sqrt{(1-Q_2+Q_3)^2 - 4Q_3}\right]. \tag{C.3}$$

■

C.3 Full results for Section 4.3.3

Table C.1: Mean squared error (MSE $\times 10^3$) for the maximum likelihood estimator of the sensitivity $(1-\tau_{01})$ and specificity $(1-\tau_{10})$, associated to the simple hidden Markov model with $n = 3$, for $n = 6$ time points for $m = 2000$ and $m = 5000$ subjects, and for different true values of the prevalence p, incidences $q = q_1 = \ldots = q_{n-1}$, and misclassification parameters τ_{10} and τ_{01}.

True Values				$n = 3$				$n = 6$			
				$1-\tau_{01}$		$1-\tau_{10}$		$1-\tau_{01}$		$1-\tau_{10}$	
p	q	τ_{01}	τ_{10}	$m=2000$	5000	2000	5000	$m=2000$	5000	2000	5000
0.03	0.04	0.15	0.15	12.33	5.92	0.10	0.04	0.37	0.14	0.02	0.01
0.03	0.10	0.15	0.15	7.80	3.68	0.12	0.05	0.16	0.06	0.03	0.01
0.03	0.15	0.15	0.15	4.53	1.86	0.13	0.05	0.12	0.04	0.03	0.01
0.03	0.04	0.05	0.05	1.26	0.58	0.02	0.01	0.08	0.03	0.01	0.00
0.03	0.10	0.05	0.05	1.14	0.52	0.03	0.01	0.04	0.01	0.01	0.03
0.03	0.15	0.05	0.05	0.86	0.38	0.03	0.01	0.02	0.01	0.01	0.00
0.10	0.04	0.15	0.15	1.85	0.73	0.08	0.04	0.17	0.07	0.02	0.01
0.10	0.10	0.15	0.15	1.98	0.84	0.12	0.05	0.12	0.04	0.03	0.01
0.10	0.15	0.15	0.15	1.78	0.66	0.15	0.06	0.08	0.03	0.04	0.02
0.10	0.04	0.05	0.05	0.26	0.10	0.02	0.01	0.04	0.01	0.01	0.00
0.10	0.10	0.05	0.05	0.25	0.11	0.03	0.01	0.02	0.01	0.01	0.00
0.10	0.15	0.05	0.05	0.28	0.10	0.03	0.01	0.02	0.01	0.01	0.00
0.15	0.04	0.15	0.15	1.01	0.36	0.09	0.03	0.12	0.05	0.03	0.01
0.15	0.10	0.15	0.15	0.99	0.39	0.12	0.05	0.10	0.03	0.04	0.01
0.15	0.15	0.15	0.15	0.94	0.40	0.14	0.06	0.07	0.03	0.04	0.02
0.15	0.04	0.05	0.05	0.14	0.06	0.02	0.01	0.03	0.01	0.01	0.00
0.15	0.10	0.05	0.05	0.15	0.06	0.02	0.01	0.02	0.01	0.01	0.00
0.15	0.15	0.05	0.05	0.17	0.06	0.03	0.01	0.02	0.01	0.01	0.00
0.03	0.04	0.15	0.05	3.22	1.28	0.02	0.01	0.22	0.09	0.01	0.00
0.03	0.10	0.15	0.05	2.58	1.02	0.04	0.01	0.10	0.04	0.01	0.00
0.03	0.15	0.15	0.05	1.59	0.65	0.04	0.02	0.07	0.03	0.01	0.01
0.03	0.04	0.05	0.15	0.63	0.27	0.02	0.01	0.11	0.04	0.01	0.00
0.03	0.10	0.05	0.15	0.70	0.26	0.03	0.01	0.07	0.03	0.01	0.00
0.03	0.15	0.05	0.15	0.68	0.24	0.04	0.02	0.06	0.02	0.01	0.01
0.10	0.04	0.15	0.05	0.43	0.17	0.03	0.01	0.08	0.03	0.01	0.00
0.10	0.10	0.15	0.05	0.42	0.18	0.03	0.01	0.06	0.02	0.01	0.00
0.10	0.15	0.15	0.05	0.41	0.17	0.04	0.02	0.05	0.02	0.02	0.01
0.10	0.04	0.05	0.15	5.36	2.54	0.07	0.03	0.12	0.05	0.02	0.01
0.10	0.10	0.05	0.15	3.58	1.86	0.09	0.04	0.06	0.02	0.02	0.01
0.10	0.15	0.05	0.15	2.17	1.23	0.09	0.05	0.04	0.01	0.03	0.01
0.15	0.04	0.15	0.05	0.92	0.40	0.07	0.03	0.05	0.02	0.02	0.01
0.15	0.10	0.15	0.05	0.94	0.40	0.09	0.04	0.04	0.01	0.02	0.01
0.15	0.15	0.15	0.05	0.90	0.43	0.11	0.05	0.03	0.01	0.03	0.01
0.15	0.04	0.05	0.15	0.47	0.20	0.07	0.03	0.04	0.02	0.02	0.01
0.15	0.10	0.05	0.15	0.50	0.19	0.09	0.04	0.03	0.01	0.03	0.01
0.15	0.15	0.05	0.15	0.46	0.19	0.11	0.05	0.02	0.01	0.03	0.01

C.4 Full Conditional for the Latent Data

The full conditional for the true status $Y_{(i,j)}$ of subject i in examination j is a Bernoulli distribution with probability $\pi_{(i,j)}$ depending on the position in the sequence. The different cases are described in page 148.

For $j = 1$, $\pi_{(i,1)} = P\left(Y_{(i,1)} = 1 \mid \text{rest}\right)$ is given by,

$$\pi_{(i,1)} = \begin{cases} 0 & \text{if } Y_{(i,2)} = 0, \\ \frac{\left(1-\tau_{\xi_{i,1},10}\right)p_1\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)}{\left(1-\tau_{\xi_{i,1},10}\right)p_1\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)+\tau_{\xi_{i,1},01}p_2\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)\delta_1\left(\boldsymbol{z}_{(i,1)},\boldsymbol{\beta}_{q_1}\right)} & \text{if } Y^*_{(i,1)} = 1, \text{and } Y_{(i,2)} = 1, \\ \frac{\tau_{\xi_{i,1},10}p_1\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)}{\tau_{\xi_{i,1},10}p_1\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)+\left(1-\tau_{\xi_{i,1},01}\right)p_2\left(\boldsymbol{w}_i,\boldsymbol{\beta}_p\right)\delta_1\left(\boldsymbol{z}_{(i,1)},\boldsymbol{\beta}_{q_1}\right)} & \text{if } Y^*_{(i,1)} = 0, \text{and } Y_{(i,2)} = 1. \end{cases}$$

For $j \in \{2, 3, 4, n-1\}$, $\pi_{(i,j)} = P\left(Y_{(i,j)} = 1 \mid \text{rest}\right)$ is given by,

$$\pi_{(i,j)} = \begin{cases} 0 & \text{if } Y_{(i,j-1)} = 0, \text{and } Y_{(i,j+1)} = 0, \\ \frac{\left(1-\tau_{\xi_{i,j},10}\right)\delta_1\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)}{\left(1-\tau_{\xi_{i,j},10}\right)\delta_1\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)+\tau_{\xi_{i,j},01}\delta_2\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)\delta_1\left(\boldsymbol{z}_{(i,j)},\boldsymbol{\beta}_{q_j}\right)} & \text{if } Y^*_{(i,j)} = 1, Y_{(i,j-1)} = 0, Y_{(i,j+1)} = 1, \\ \frac{\tau_{\xi_{i,j},10}\delta_1\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)}{\tau_{\xi_{i,j},10}\delta_1\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)+\left(1-\tau_{\xi_{i,j},01}\right)\delta_2\left(\boldsymbol{z}_{(i,j-1)},\boldsymbol{\beta}_{q_{j-1}}\right)\delta_1\left(\boldsymbol{z}_{(i,j)},\boldsymbol{\beta}_{q_j}\right)} & \text{if } Y^*_{(i,j)} = 0, Y_{(i,j-1)} = 0, Y_{(i,j+1)} = 1, \\ 1 & \text{if } Y_{(i,j-1)} = 1. \end{cases}$$

For $j = n$, $\pi_{(i,n)} = P\left(Y_{(i,n)} = 1 \mid \text{rest}\right)$ is given by,

$$\pi_{(i,n)} = \begin{cases} \frac{\left(1-\tau_{\xi_{i,n},10}\right)\delta_1\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)}{\left(1-\tau_{\xi_{i,n},10}\right)\delta_1\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)+\tau_{\xi_{i,n},01}\delta_2\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)} & \text{if } Y^*_{(i,n)} = 1, \text{and } Y_{(i,n-1)} = 0, \\ \frac{\tau_{\xi_{i,n},10}\delta_1\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)}{\tau_{\xi_{i,n},10}\delta_1\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)+\left(1-\tau_{\xi_{i,n},01}\right)\delta_2\left(\boldsymbol{z}_{(i,n-1)},\boldsymbol{\beta}_{q_{n-1}}\right)} & \text{if } Y^*_{(i,n)} = 0, \text{and } Y_{(i,n-1)} = 0, \\ 1 & \text{if } Y_{(i,n-1)} = 1. \end{cases}$$

C.5 Weighted Least Squares MH Step

In the weighted least square normal proposal, the candidates for the regression coefficients associated to the prevalence are generated from the multivariate normal distribution $\boldsymbol{\beta}_p^* \sim N_{k+1}\left(\boldsymbol{m}^p(\boldsymbol{\beta}_p), \boldsymbol{C}^p(\boldsymbol{\beta}_p)\right)$, where

$$\boldsymbol{m}^p(\boldsymbol{\beta}_p) = \boldsymbol{C}^p(\boldsymbol{\beta}_p)\left\{\boldsymbol{V}_p^{-1}\boldsymbol{b}_p + \boldsymbol{W}^T\boldsymbol{\Gamma}^p(\boldsymbol{\beta}_p)\,\widetilde{\boldsymbol{\eta}}^p(\boldsymbol{\beta}_p)\right\},$$

and

$$\boldsymbol{C}^p(\boldsymbol{\beta}_p) = \left(\boldsymbol{V}_p^{-1} + \boldsymbol{W}^T\boldsymbol{\Gamma}^p(\boldsymbol{\beta}_p)\boldsymbol{W}\right)^{-1},$$

where $\widetilde{\boldsymbol{\eta}}^p(\boldsymbol{\beta}_p)$ is a vector of "pseudo-data" with coordinates

$$\widetilde{\eta}_i^p(\boldsymbol{\beta}_p) = \boldsymbol{w}_i^T\boldsymbol{\beta}_p + \frac{Y_{(i,1)} - p_1\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right)}{p_1\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right)p_2\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right)},$$

and $\boldsymbol{\Gamma}^p(\boldsymbol{\beta}_p) = \mathrm{diag}(\gamma_{1,1}^p(\boldsymbol{\beta}_p), \ldots, \gamma_{m,m}^p(\boldsymbol{\beta}_p))$, with $\gamma_{i,i}^p(\boldsymbol{\beta}_p) = p_1\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right)p_2\left(\boldsymbol{w}_i, \boldsymbol{\beta}_p\right)$, $i = 1, \ldots, m$. The candidate is accepted with probability

$$1 \wedge \frac{\phi_{k+1}\left(\boldsymbol{\beta}_p^*|\boldsymbol{b}_p, \boldsymbol{V}_p\right)}{\phi_{k+1}\left(\boldsymbol{\beta}_p|\boldsymbol{b}_p, \boldsymbol{V}_p\right)}\frac{\phi_{k+1}\left(\boldsymbol{\beta}_p|\boldsymbol{m}^p(\boldsymbol{\beta}_p^*), \boldsymbol{C}^p(\boldsymbol{\beta}_p^*)\right)}{\phi_{k+1}\left(\boldsymbol{\beta}_p^*|\boldsymbol{m}^p(\boldsymbol{\beta}_p), \boldsymbol{C}^p(\boldsymbol{\beta}_p)\right)} \left[\prod_{i=1}^{m}\frac{\exp\{\boldsymbol{w}_i^T(\boldsymbol{\beta}_p^* - \boldsymbol{\beta}_p)\}^{Y_{(i,1)}}}{[1+\exp(\boldsymbol{w}_i^T\boldsymbol{\beta}_p^*)]/[1+\exp(\boldsymbol{w}_i^T\boldsymbol{\beta}_p)]}\right], \tag{C.4}$$

where $\phi_{k+1}\left(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the density of a $k+1$-variate normal distribution with mean and covariance matrix $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.

Equivalently, the candidates for the regression coefficients associated to the incidences are generated from the multivariate normal distribution $\boldsymbol{\beta}_{q_j}^* \sim N_{k+2}\left(\boldsymbol{m}^{q_j}(\boldsymbol{\beta}_{q_j}), \boldsymbol{C}^{q_j}(\boldsymbol{\beta}_{q_j})\right)$, where

$$\boldsymbol{m}^{q_j}(\boldsymbol{\beta}_{q_j}) = \boldsymbol{C}^{q_j}(\boldsymbol{\beta}_{q_j})\left\{\boldsymbol{V}_{q_j}^{-1}\boldsymbol{b}_{q_j} + \boldsymbol{Z}_j^T\boldsymbol{\Gamma}^{q_j}(\boldsymbol{\beta}_{q_j})\,\widetilde{\boldsymbol{\eta}}^{q_j}(\boldsymbol{\beta}_{q_j})\right\},$$

and

$$\boldsymbol{C}^q(\boldsymbol{\beta}_{q_j}) = \left(\boldsymbol{V}_{q_j}^{-1} + \boldsymbol{Z}_j^T\boldsymbol{\Gamma}^{q_j}(\boldsymbol{\beta}_{q_j})\boldsymbol{Z}_j\right)^{-1}.$$

Now $\widetilde{\boldsymbol{\eta}}^{q_j}(\boldsymbol{\beta}_{q_j})$ is the vector of "pseudo-data" with coordinates

$$\widetilde{\eta}_i^{q_j}(\boldsymbol{\beta}_{q_j}) = \boldsymbol{z}_{i,j}^T\boldsymbol{\beta}_{q_j} + \frac{Y_{(i,j+1)} - q_1\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right)}{q_2\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right)q_2\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right)},$$

and $\mathbf{\Gamma}^{q_j}(\boldsymbol{\beta}_{q_j}) = \text{diag}(\gamma_{1,1}^{q_j}(\boldsymbol{\beta}_{q_j}), \ldots, \gamma_{m,m}^{q_j}(\boldsymbol{\beta}_{q_j}))$. However, the weights now are defined by

$$\gamma_{i,i}^{q_j}(\boldsymbol{\beta}_{q_j}) = \begin{cases} 0 & \text{if } Y_{(i,j)} = 1, \\ q_1\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) q_2\left(\boldsymbol{z}_{(i,j)}, \boldsymbol{\beta}_{q_j}\right) & \text{if } Y_{(i,j)} = 0. \end{cases}$$

Finally, the candidate is accepted with probability

$$1 \wedge \frac{\phi_{k+2}\left(\boldsymbol{\beta}_{q_j}^* | \boldsymbol{b}_{q_j}, \boldsymbol{V}_{q_j}\right)}{\phi_{k+2}\left(\boldsymbol{\beta}_{q_j} | \boldsymbol{b}_{q_j}, \boldsymbol{V}_{q_j}\right)} \frac{\phi_{k+2}\left(\boldsymbol{\beta}_{q_j} | \boldsymbol{m}^{q_j}(\boldsymbol{\beta}_{q_j}^*), \boldsymbol{C}^{q_j}(\boldsymbol{\beta}_{q_j}^*)\right)}{\phi_{k+2}\left(\boldsymbol{\beta}_{q_j}^* | \boldsymbol{m}^{q_j}(\boldsymbol{\beta}_{q_j}), \boldsymbol{C}^{q_j}(\boldsymbol{\beta}_{q_j})\right)} \times \left[\prod_{i=1}^{m} \left\{ \frac{\exp\{\boldsymbol{z}_{(i,j)}^T(\boldsymbol{\beta}_{q_j}^* - \boldsymbol{\beta}_{q_j})\}^{Y_{(i,j+1)}}}{[1+\exp(\boldsymbol{z}_{(i,j)}^T\boldsymbol{\beta}_{q_j}^*)]/[1+\exp(\boldsymbol{z}_{(i,j)}^T\boldsymbol{\beta}_{q_j})]} \right\}^{1-Y_{(i,j)}}\right]. \tag{C.5}$$

Appendix D

Supplementary Material for Chapter 5

D.1 Proof of Proposition 5.1

The proof of Proposition 1 is similar to the one provided by Joe & Liu (1996) for the compatibility of the full conditionals of their conditionally specified logistic regression model. The proof of the necessary condition is based on the compatibility conditions provided by Gelman & Speed (1993).

D.1.1 Sufficient condition

Set $\boldsymbol{Y}_{(i,k)[-m]} = \left(Y_{(i,1,k)}, \ldots, Y_{(i,m-1,k)}, Y_{(i,m+1,k)}, \ldots, Y_{(i,J,k)}\right)$, for all $m \in \{1, \ldots, J\}$ and $k \in \{2, \ldots, K\}$, and $\boldsymbol{\theta}^I = \left(\boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right)$ and let $P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)[-m]} \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right)$ be the marginal distribution of $\boldsymbol{Y}_{(i,k)}^{[-m]}$. If $y_m^{k-1} = 0$, the joint distribution given by expression (5.4), with $\gamma_{jl}^I = \gamma_{lj}^I$, $j \neq l$, implies that

$$
\begin{aligned}
&P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}_{[-m]}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \\
&\quad = \sum_{y_m^k=0}^{1} P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right), \\
&\quad = \sum_{y_m^k=0}^{1} c_{(2,i)}^{-1} \exp\left\{ \boldsymbol{z}'_{(i,m,k)} \boldsymbol{\beta}_m^I y_m^k + \sum_{\{j \in S : j \neq m\}} \boldsymbol{z}'_{(i,j,k)} \boldsymbol{\beta}_j^I y_j^k + \sum_{l \in S^c} \gamma_{lm}^I y_m^k + \right.
\end{aligned}
$$

$$\sum_{\{(j,l):j\in S,l\in S^c,j\neq m\}} \gamma^I_{lj} y^k_l \; + \sum_{\{l\in S:m<l\leq J:\}} \gamma^I_{ml} y^k_l y^k_m \; +$$

$$\sum_{\{j\in S:j<m\}} \gamma^I_{jm} y^k_m y^k_j \; + \sum_{\{j<l\leq J:j\in S,l\in S,l\neq m,j\neq m\}} \gamma^I_{jl} y^k_l y^k_j \; +$$

$$\sum_{l\in S^c} \alpha^I_{lm} y^k_m \; + \sum_{\{(j,l):j\in S,l\in S^c,j\neq m\}} \alpha^I_{lj} y^k_j \Bigg\},$$

$$= c^{-1}_{(2,i)} \Bigg[\exp\Bigg\{ \sum_{\{j\in S:j\neq m\}} \boldsymbol{z}'_{(i,j,k)} \boldsymbol{\beta}^I_j y^k_j + \sum_{\{(j,l):j\in S,l\in S^c,j\neq m\}} \gamma^I_{lj} y^k_l \; +$$

$$\sum_{\{j<l\leq J:j\in S,l\in S,l\neq m,j\neq m\}} \gamma^I_{jl} y^k_l y^k_j + \sum_{\{(j,l):j\in S,l\in S^c,j\neq m\}} \alpha^I_{lj} y^k_j \Bigg\} \times$$

$$\left(1 + \exp\left\{ \boldsymbol{z}'_{(i,m,k)} \boldsymbol{\beta}^I_m + \sum_{l\neq m} \gamma^I_{lm} y^k_l + \sum_{l\neq m} \alpha^I_{lm} y^{k-1}_l \right\} \right) \Bigg],$$

where $c_{(2,i)} \equiv c_2(\boldsymbol{Z}_{(i,k)}, \boldsymbol{\theta}^I)$. Thus, if $y^{k-1}_m = 0$ then for all $m \in \{1,\ldots,J\}$, for all $k \in \{2,\ldots,K\}$, and for all $\boldsymbol{y}^{k-1}_{[-m]} \in \{0,1\}^{J-1}$, it follows that,

$$\begin{aligned} P_{\boldsymbol{Z}_{(i,k)}} &\left(Y_{(i,m,k)} = y^k_m \mid \boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]}, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I \right) \\ &= \frac{P_{\boldsymbol{Z}_{(i,k)}} \left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I \right)}{P_{\boldsymbol{Z}_{(i,k)}} \left(\boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]} \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I \right)}, \\ &= \frac{\exp\left\{ \left(\boldsymbol{z}'_{(i,m,k)} \boldsymbol{\beta}^I_m + \sum_{l\neq m} \gamma^I_{lm} y^k_l + \sum_{l\neq m} \alpha^I_{lm} y^{k-1}_l \right) y^k_m \right\}}{1 + \exp\left\{ \boldsymbol{z}'_{(i,m,k)} \boldsymbol{\beta}^I_m + \sum_{l\neq m} \gamma^I_{lm} y^k_l + \sum_{l\neq m} \alpha^I_{lm} y^{k-1}_l \right\}}. \end{aligned} \tag{D.1}$$

Now, if $y^{k-1}_m = 1$, then $y^k_m = 1$ and

$$
\begin{aligned}
&P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]} \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \\
&\quad = \sum_{y^k_m=1}^{1} P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right), \\
&\quad = \sum_{y^k_m=1}^{1} c^{-1}_{(2,i)} \exp\left\{ \boldsymbol{z}'_{(i,m,k)}\boldsymbol{\beta}^I_m y^k_m + \sum_{\{j\in S: j\neq m\}} \boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}^I_j y^k_j + \sum_{l\in S^c} \gamma^I_{lm} y^k_m + \right. \\
&\qquad\qquad \sum_{\{(j,l): j\in S, l\in S^c, j\neq m\}} \gamma^I_{lj} y^k_l + \sum_{\{l\in S: m<l\leq J:\}} \gamma^I_{ml} y^k_l y^k_m + \\
&\qquad\qquad \sum_{\{j\in S: j<m\}} \gamma^I_{jm} y^k_m y^k_j + \sum_{\{j<l\leq J: j\in S, l\in S, l\neq m, j\neq m\}} \gamma^I_{jl} y^k_l y^k_j + \\
&\qquad\qquad \left. \sum_{l\in S^c} \alpha^I_{lm} y^k_m + \sum_{\{(j,l): j\in S, l\in S^c, j\neq m\}} \alpha^I_{lj} y^k_j \right\}, \\
&\quad = P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right).
\end{aligned}
$$

Therefore, if $y^{k-1}_m = 1$ then for all $m \in \{1, \ldots, J\}$, for all $k \in \{2, \ldots, K\}$, and for all $\boldsymbol{y}^{k-1}_{[-m]} \in \{0,1\}^{J-1}$, it follows that,

$$
\begin{aligned}
&P_{\boldsymbol{Z}_{(i,k)}}\left(Y_{(i,m,k)} = y^k_m \mid \boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]}, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \\
&\qquad = \frac{P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]} \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right)}, \qquad \text{(D.2)} \\
&\qquad = 1.
\end{aligned}
$$

From expressions (D.1) and (D.2), it follows that $Y_{(i,m,k)}$ conditional on the design vector $\boldsymbol{z}_{(i,j,k)}$, $\boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}$ and $\boldsymbol{Y}_{(i,k)[-m]} = \boldsymbol{y}^k_{[-m]}$, for all $\boldsymbol{y}^k \in \mathcal{B}\{\boldsymbol{y}^{k-1}\}$, follows a Bernoulli distribution with probability

$$
\left\{ h\left(\boldsymbol{z}'_{(i,m,k)}\boldsymbol{\beta}^I_m + \sum_{l\neq m} \gamma^I_{lm} y^k_l + \sum_{l\neq m} \alpha^I_{lm} y^{k-1}_l \right) \right\}^{1-y^{k-1}_m}, \qquad \text{(D.3)}
$$

where $h(\cdot) = \exp\{\cdot\}/(1+\exp\{\cdot\})$. ■

D.1.2 Necessary condition

Let $\boldsymbol{\theta}^I = \left(\boldsymbol{\beta}^I, \boldsymbol{\gamma}^I, \boldsymbol{\alpha}^I\right)$ be the vector of parameters. From the result of Gelman & Speed (1993), it follows that

$$P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \propto \frac{A}{B}$$

where

$$A = \prod_{j=1}^{J} P\left(Y_{(i,j,k)} = y_j^k \mid Y_{(i,1,k)} = \hat{y}_1^k, \ldots, Y_{(i,j-1,k)} = \hat{y}_{j-1}^k, \right.$$
$$\left. Y_{(i,j+1,k)} = y_{j+1}^k, \ldots, Y_{(i,J,k)} = y_J^k, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right)$$

and

$$B = \prod_{j=1}^{J} P\left(Y_{(i,j,k)} = \hat{y}_j^k \mid Y_{(i,1,k)} = \hat{y}_1^k, \ldots, Y_{(i,j-1,k)} = \hat{y}_{j-1}^k, \right.$$
$$\left. Y_{(i,j+1,k)} = y_{j+1}^k, \ldots, Y_{(i,J,k)} = y_J^k, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right).$$

Thus,

$$\begin{aligned} & P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \\ & = \frac{\prod_{j\in S} \exp\left\{\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l<j}\gamma_{lj}^I \hat{y}_l^k + \sum_{l>j}\gamma_{lj}^I y_l^k + \sum_{l\neq j}\alpha_{lj}^I y_l^{k-1}\right) y_j^k\right\}}{\prod_{j\in S} \exp\left\{\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l<j}\gamma_{lj}^I \hat{y}_l^k + \sum_{l>j}\gamma_{lj}^I y_l^k + \sum_{l\neq j}\alpha_{lj}^I y_l^{k-1}\right) \hat{y}_j^k\right\}} \\ & \propto \exp\left\{\sum_{j\in S}\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l>j}\gamma_{lj}^I y_l^k + \sum_{l\neq j}\alpha_{lj}^I y_l^{k-1}\right) y_j^k + \right. \\ & \qquad \left. \sum_{j\in S}\left(\sum_{l<j}\gamma_{lj}^I \hat{y}_l^k\right) y_j^k - \sum_{j\in S}\left(\sum_{l>j}\gamma_{lj}^I y_l^k\right)\hat{y}_j^k\right\}, \end{aligned} \tag{D.4}$$

for any arbitrary vector $\hat{\boldsymbol{y}}^k = \left(\hat{y}_1^k, \ldots, \hat{y}_J^k\right) \in \mathcal{B}(\boldsymbol{y}^{k-1})$.

Note that by taking $\hat{y}_j^k = y_j^{k-1}$, $j = 1, \ldots, J$, it follows that

$$
\begin{aligned}
& P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} = \boldsymbol{y}^k \mid \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \\
& \quad \propto \exp\left\{\sum_{j\in S}\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l>j}\gamma_{lj}^I y_l^k + \sum_{l\neq j}\alpha_{lj}^I y_l^{k-1}\right) y_j^k + \right. \\
& \qquad \left. \sum_{j\in S}\left(\sum_{l<j}\gamma_{lj}^I \hat{y}_l^k\right) y_j^k - \sum_{j\in S}\left(\sum_{l>j}\gamma_{lj}^I y_l^k\right)\hat{y}_j^k\right\}, \\
& \quad = \exp\left\{\sum_{j\in S}\left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{\{j<l\leq J : l\in S\}}\gamma_{lj}^I y_l^k + \right.\right. \\
& \qquad \left.\left. \sum_{\{j<l\leq J : l\in S^c\}}\gamma_{lj}^I + \sum_{\{l<j : l\in S^c\}}\gamma_{lj}^I + \sum_{l\in S^c}\alpha_{lj}^I\right) y_j^k\right\},
\end{aligned}
\tag{D.5}
$$

which corresponds, up to a normalizing constant, to the joint marginal distribution given by expression (5.4) Chapter 5.

Now, let

$$
\begin{aligned}
C = \prod_{j=1}^{J} P\big(Y_{(i,j,k)} = \tilde{y}_j^k \mid Y_{(i,1,k)} = \tilde{y}_1^k, \ldots, Y_{(i,j-1,k)} = \tilde{y}_{j-1}^k, \\
Y_{(i,j+1,k)} = y_{j+1}^k, \ldots, Y_{(i,J,k)} = y_J^k, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\big)
\end{aligned}
$$

and

$$
\begin{aligned}
D = \prod_{j=1}^{J} P\big(Y_{(i,j,k)} = y_j^k \mid Y_{(i,1,k)} = \tilde{y}_1^k, \ldots, Y_{(i,j-1,k)} = \tilde{y}_{j-1}^k, \\
Y_{(i,j+1,k)} = y_{j+1}^k, \ldots, Y_{(i,J,k)} = y_J^k, \boldsymbol{Y}_{(i,k-1)} = \boldsymbol{y}^{k-1}\boldsymbol{\theta}^I\big)
\end{aligned}
$$

For arbitrary vectors $\hat{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ and $\tilde{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$, set

$$
\mathcal{R}\left(\boldsymbol{y}^k \mid \hat{\boldsymbol{y}}^k, \tilde{\boldsymbol{y}}^k, \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \quad = \quad \frac{A}{B} \times \frac{C}{D}
$$

$$
\begin{aligned}
&= \prod_{j\in S} \exp\left\{ \left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l<j} \gamma_{lj}^I \hat{y}_l^k + \sum_{l>j} \gamma_{lj}^I y_l^k + \sum_{l\neq j} \alpha_{lj}^I y_l^{k-1} \right) \left(y_j^k - \hat{y}_j^k \right) \right\} \times \\
&\quad \prod_{j\in S} \exp\left\{ \left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{l<j} \gamma_{lj}^I \tilde{y}_l^k + \sum_{l>j} \gamma_{lj}^I y_l^k + \sum_{l\neq j} \alpha_{lj}^I y_l^{k-1} \right) \left(\tilde{y}_j^k - y_j^k \right) \right\}, \\
&= \prod_{j\in S} \exp\left\{ \left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{\{l<j : l\in S\}} \gamma_{lj}^I \hat{y}_l^k + \sum_{\{l<j : l\in S^c\}} \gamma_{lj}^I + \sum_{\{l>j : l\in S\}} \gamma_{lj}^I y_l^k + \right. \right. \\
&\qquad \left. \left. \sum_{\{l>j : l\in S^c\}} \gamma_{lj}^I + \sum_{l\in S^c} \alpha_{lj}^I \right) \left(y_j^k - \hat{y}_j^k \right) \right\} \times \\
&\quad \prod_{j\in S} \exp\left\{ \left(\boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}_j^I + \sum_{\{l<j : l\in S\}} \gamma_{lj}^I \tilde{y}_l^k + \sum_{\{l<j : l\in S^c\}} \gamma_{lj}^I + \sum_{\{l>j : l\in S\}} \gamma_{lj}^I y_l^k + \right. \right. \\
&\qquad \left. \left. \sum_{\{l>j : l\in S^c\}} \gamma_{lj}^I + \sum_{l\in S^c} \alpha_{lj}^I \right) \left(\tilde{y}_j^k - y_j^k \right) \right\}, \\
&\propto \prod_{j\in S} \exp\left\{ \sum_{\{l<j : l\in S\}} \gamma_{lj}^I \hat{y}_l^k y_j^k - \sum_{\{l<j : l\in S\}} \gamma_{lj}^I \tilde{y}_l^k y_j^k + \sum_{\{l>j : l\in S\}} \gamma_{lj}^I y_l^k \left(\tilde{y}_j^k - \hat{y}_j^k \right) \right\}, \\
&= \prod_{j\in S} \exp\left\{ \sum_{\{l<j : l\in S\}} \gamma_{lj}^I y_j^k \left(\hat{y}_l^k - \tilde{y}_l^k \right) - \sum_{\{l<j : l\in S\}} \gamma_{jl}^I y_j^k \left(\hat{y}_l^k - \tilde{y}_l^k \right) \right\}, \\
&= \exp\left\{ \sum_{\{l<j : (j,l)\in S\times S\}} \left[\left(\gamma_{lj}^I - \gamma_{jl}^I \right) y_j^k \left(\hat{y}_l^k - \tilde{y}_l^k \right) \right] \right\}. \qquad \text{(D.6)}
\end{aligned}
$$

The full conditional compatibility condition of Gelman & Speed (1993) implies that, for arbitrary vectors $\hat{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ and $\tilde{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ such that $\hat{\boldsymbol{y}}^k \neq \tilde{\boldsymbol{y}}^k$, the ratio $\mathcal{R}\left(\boldsymbol{y}^k \mid \hat{\boldsymbol{y}}^k, \tilde{\boldsymbol{y}}^k, \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I \right)$ must not depend on $\boldsymbol{y}^k$. Convenient choices of $\hat{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ and $\tilde{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ show that, for all $\boldsymbol{y}^{k-1} \in \{0,1\}^J$, expression (D.6) does not depend on $\boldsymbol{y}^k$ only if $\gamma_{jl}^I = \gamma_{lj}^I$ for all $j \neq l$. For instance, for $\boldsymbol{y}^{k-1} = \mathbf{0}_J$,

the choices $\tilde{\boldsymbol{y}}^k = \mathbf{0}_J$ and $\hat{\boldsymbol{y}}^k = (0, \ldots, 0, 1, 0)$ leads to

$$\mathcal{R}\left(\boldsymbol{y}^k \mid \hat{\boldsymbol{y}}^k, \tilde{\boldsymbol{y}}^k, \boldsymbol{y}^{k-1}, \boldsymbol{\theta}^I\right) \quad \propto \quad \exp\left\{\left(\gamma^I_{J-1,J} - \gamma^I_{J,J-1}\right) y^k_J\right\}, \tag{D.7}$$

which is independent of $\boldsymbol{y}^k$ only if $\gamma^I_{J-1,J} = \gamma^I_{J,J-1}$. Other convenient choices of $\hat{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ and $\tilde{\boldsymbol{y}}^k \in \mathcal{B}(\boldsymbol{y}^{k-1})$ show the need of the symmetry constraint in the remaining conditional log-odds parameters.

■

D.2 Metropolis-Hastings steps for the Markov model parameters

D.2.1 Updating $\boldsymbol{\theta}^P$

Let $\boldsymbol{\theta}^P = \left(\boldsymbol{\beta}^{P'}_1, \ldots, \boldsymbol{\beta}^{P'}_J, \boldsymbol{\gamma}^{P'}\right)'$ be the R-dimensional vector of parameters associated to the initial distribution, with $R = Jp + J(J-1)/2$, and $\widetilde{\boldsymbol{\eta}}_P$ be a $I \times J$-dimensional vector of "pseudo-data" with coordinates

$$\widetilde{\boldsymbol{\eta}}_P\left(\boldsymbol{\theta}^P\right) = \left(\widetilde{\eta}^P_{(1,1)}\left(\boldsymbol{\theta}^P\right), \ldots, \widetilde{\eta}^P_{(1,J)}\left(\boldsymbol{\theta}^P\right), \ldots, \widetilde{\eta}^P_{(I,1)}\left(\boldsymbol{\theta}^P\right), \ldots, \widetilde{\eta}^P_{(I,J)}\left(\boldsymbol{\theta}^P\right)\right)',$$

where

$$\widetilde{\eta}^P_{(i,j)}\left(\boldsymbol{\theta}^P\right) = \boldsymbol{w}^{P'}_{(i,j)}\boldsymbol{\theta}^P + \frac{Y_{(i,j,1)} - h\left(\boldsymbol{w}^{P'}_{(i,j)}\boldsymbol{\theta}^P\right)}{\left\{h\left(\boldsymbol{w}^{P'}_{(i,j)}\boldsymbol{\theta}^P\right)\left[1 - h\left(\boldsymbol{w}^{P'}_{(i,j)}\boldsymbol{\theta}^P\right)\right]\right\}},$$

with $\boldsymbol{w}^P_{(i,j)}$ being an appropriate design vector created such that $\boldsymbol{w}^{P'}_{(i,j)}\boldsymbol{\theta}^P = \boldsymbol{x}'_{(i,j)}\boldsymbol{\beta}^P_j + \sum_{l \neq j}\gamma^P_{jl}Y_{(i,l,1)}$ and $h(\cdot) = \exp\{\cdot\}/(1 + \exp\{\cdot\})$. The candidates for the parameters associated to the initial distribution are generated from the multivariate normal distribution

$$\boldsymbol{\theta}^{P*} \sim N_R\left(\boldsymbol{m}_P\left(\boldsymbol{\theta}^P\right), \boldsymbol{C}_P\left(\boldsymbol{\theta}^P\right)\right),$$

with mean vector given by

$$\boldsymbol{m}_P\left(\boldsymbol{\theta}^P\right) = \boldsymbol{C}_P\left(\boldsymbol{\theta}^P\right)\left\{\boldsymbol{V}_P^{-1}\boldsymbol{b}_P + \boldsymbol{W}_P\boldsymbol{\Gamma}_P\left(\boldsymbol{\theta}^P\right)\widetilde{\boldsymbol{\eta}}_P\left(\boldsymbol{\theta}^P\right)\right\},$$

and covariance matrix given by

$$\boldsymbol{C}^P\left(\boldsymbol{\theta}^P\right) = \left(\boldsymbol{V}_P^{-1} + \boldsymbol{W}_P\boldsymbol{\Gamma}_P\left(\boldsymbol{\theta}^P\right)\boldsymbol{W}_P'\right)^{-1},$$

where $\boldsymbol{b}_P = \left(\boldsymbol{m}_{\boldsymbol{\beta}^P}', \boldsymbol{m}_{\boldsymbol{\gamma}^P}'\right)'$, $\boldsymbol{V}_P = \boldsymbol{V}_{\boldsymbol{\beta}^P} \oplus \boldsymbol{V}_{\boldsymbol{\gamma}^P}$, with $\oplus$ denoting the direct product,

$$\boldsymbol{W}_P = \left(\boldsymbol{w}_{(1,1)}^P, \ldots, \boldsymbol{w}_{(1,J)}^P, \ldots, \boldsymbol{w}_{(I,1)}^P, \ldots, \boldsymbol{w}_{(I,J)}^P\right)',$$

and

$$\boldsymbol{\Gamma}_P\left(\boldsymbol{\theta}^P\right) = \operatorname{diag}\left\{\zeta_{(1,1)}^P\left(\boldsymbol{\theta}^P\right), \ldots, \zeta_{(1,J)}^P\left(\boldsymbol{\theta}^P\right), \ldots, \zeta_{(I,1)}^P\left(\boldsymbol{\theta}^P\right), \ldots, \zeta_{(I,J)}^P\left(\boldsymbol{\theta}^P\right)\right\},$$

with

$$\zeta_{(i,j)}^P\left(\boldsymbol{\theta}^P\right) = h\left(\boldsymbol{w}_{(i,j)}^{P'}\boldsymbol{\theta}^P\right)\left[1 - h\left(\boldsymbol{w}_{(i,j)}^{P'}\boldsymbol{\theta}^P\right)\right], i = 1, \ldots, I, j = 1, \ldots, J.$$

The candidate $\boldsymbol{\theta}^{P*}$ is accepted with probability

$$1 \wedge \frac{\phi_R\left(\boldsymbol{\theta}^{P*}|\boldsymbol{b}^P, \boldsymbol{V}^P\right)}{\phi_R\left(\boldsymbol{\theta}^P|\boldsymbol{b}^P, \boldsymbol{V}^P\right)} \frac{\phi_R\left(\boldsymbol{\theta}^P|\boldsymbol{m}_P\left(\boldsymbol{\theta}^{P*}\right), \boldsymbol{C}_P\left(\boldsymbol{\theta}^{P*}\right)\right)}{\phi_R\left(\boldsymbol{\theta}^{P*}|\boldsymbol{m}_P\left(\boldsymbol{\theta}^P\right), \boldsymbol{C}_P\left(\boldsymbol{\theta}^P\right)\right)} \left[\prod_{i=1}^{I} \frac{P_{\boldsymbol{X}_i}\left(\boldsymbol{Y}_{(i,1)} \mid \boldsymbol{\theta}^{P*}\right)}{P_{\boldsymbol{X}_i}\left(\boldsymbol{Y}_{(i,1)} \mid \boldsymbol{\theta}^P\right)}\right],$$

where $\phi_R\left(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the density of a R-variate normal distribution with mean and covariance matrix $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, and $P_{\boldsymbol{X}_i}\left(\boldsymbol{Y}_{(i,1)} \mid \boldsymbol{\theta}^P\right)$ is defined as in expression (2) in the paper.

D.2.2 Updating $\boldsymbol{\theta}^I$

Let $\boldsymbol{\theta}^I = \left(\boldsymbol{\beta}_1^{I'}, \ldots, \boldsymbol{\beta}_J^{I'}, \boldsymbol{\gamma}^{I'}, \boldsymbol{\alpha}^{I'}\right)'$ be the D-dimensional vector of parameters associated to the initial distribution, with $D = Jq + 3J(J-1)/2$, and $\widetilde{\boldsymbol{\eta}}_I$ be a $I \times J \times (K-1)$-dimensional vector of "pseudo-data" with coordinates

$$\begin{aligned}\widetilde{\boldsymbol{\eta}}_I\left(\boldsymbol{\theta}^I\right) = {} & \left(\widetilde{\eta}_{(1,1,2)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(1,1,K)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(1,J,2)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(1,J,K)}^I\left(\boldsymbol{\theta}^I\right), \ldots,\right. \\ & \left.\widetilde{\eta}_{(I,1,2)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(I,1,K)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(I,J,2)}^I\left(\boldsymbol{\theta}^I\right), \ldots, \widetilde{\eta}_{(I,J,K)}^I\left(\boldsymbol{\theta}^I\right)\right)',\end{aligned}$$

where

$$\widetilde{\eta}_{(i,j,k)}^I\left(\boldsymbol{\theta}^I\right) = \boldsymbol{w}_{(i,j,k)}^{I'}\boldsymbol{\theta}^I + \frac{Y_{(i,j,k)} - h\left(\boldsymbol{w}_{(i,j,k)}^{I'}\boldsymbol{\theta}^I\right)}{\left\{h\left(\boldsymbol{w}_{(i,j,k)}^{I'}\boldsymbol{\theta}^I\right)\left[1 - h\left(\boldsymbol{w}_{(i,j,k)}^{I'}\boldsymbol{\theta}^I\right)\right]\right\}},$$

with $\boldsymbol{w}^I_{(i,j,k)}$ being an appropriate design vector created such that

$$\boldsymbol{w}^{I'}_{(i,j,k)}\boldsymbol{\theta}^I = \boldsymbol{z}'_{(i,j,k)}\boldsymbol{\beta}^I_j + \sum_{l\neq j}\gamma^I_{lj}Y_{(i,l,k)} + \sum_{l\neq j}\alpha^I_{lj}Y_{(i,l,k-1)}.$$

The candidates for the parameters associated to the initial distribution are generated from the multivariate normal distribution

$$\boldsymbol{\theta}^{I*} \sim N_D\left(\boldsymbol{m}_I\left(\boldsymbol{\theta}^I\right), \boldsymbol{C}_I\left(\boldsymbol{\theta}^I\right)\right),$$

with mean vector given by

$$\boldsymbol{m}_I\left(\boldsymbol{\theta}^I\right) = \boldsymbol{C}_I\left(\boldsymbol{\theta}^I\right)\left\{\boldsymbol{V}_I^{-1}\boldsymbol{b}_I + \boldsymbol{W}_I\boldsymbol{\Gamma}_I\left(\boldsymbol{\theta}^I\right)\widetilde{\boldsymbol{\eta}}_I\left(\boldsymbol{\theta}^I\right)\right\},$$

and covariance matrix given by

$$\boldsymbol{C}^I\left(\boldsymbol{\theta}^I\right) = \left(\boldsymbol{V}_I^{-1} + \boldsymbol{W}_I\boldsymbol{\Gamma}_I\left(\boldsymbol{\theta}^I\right)\boldsymbol{W}'_I\right)^{-1},$$

where $\boldsymbol{b}_I = \left(\boldsymbol{m}'_{\boldsymbol{\beta}^I}, \boldsymbol{m}'_{\boldsymbol{\gamma}^I}, \boldsymbol{m}'_{\boldsymbol{\alpha}^I}\right)'$, $\boldsymbol{V}_I = \boldsymbol{V}_{\boldsymbol{\beta}^I} \oplus \boldsymbol{V}_{\boldsymbol{\gamma}^I} \oplus \boldsymbol{V}_{\boldsymbol{\alpha}^I}$,

$$\begin{aligned}\boldsymbol{W}_I \;=\; & \left(\boldsymbol{w}^I_{(1,1,2)}, \ldots, \boldsymbol{w}^I_{(1,1,K)}, \ldots, \boldsymbol{w}^I_{(1,J,2)}, \ldots, \boldsymbol{w}^I_{(1,J,K)}, \ldots,\right.\\ & \left.\boldsymbol{w}^I_{(I,1,2)}, \ldots, \boldsymbol{w}^I_{(I,1,K)}, \ldots, \boldsymbol{w}^I_{(I,J,2)}, \ldots, \boldsymbol{w}^I_{(I,J,K)}\right)',\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Gamma}_I\left(\boldsymbol{\theta}^I\right) \;=\; & \operatorname{diag}\left\{\zeta^I_{(1,1,2)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(1,1,K)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(1,J,2)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(1,J,K)}\left(\boldsymbol{\theta}^I\right), \ldots,\right.\\ & \left.\zeta^I_{(I,1,2)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(I,1,K)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(I,J,2)}\left(\boldsymbol{\theta}^I\right), \ldots, \zeta^I_{(I,J,K)}\left(\boldsymbol{\theta}^I\right)\right\},\end{aligned}$$

with

$$\zeta^I_{(i,j,k)}\left(\boldsymbol{\theta}^I\right) = \left\{h\left(\boldsymbol{w}^{I'}_{(i,j,k)}\boldsymbol{\theta}^I\right)\left[1 - h\left(\boldsymbol{w}^{I'}_{(i,j,k)}\boldsymbol{\theta}^I\right)\right]\right\}^{1-Y_{(i,j,k-1)}},$$

$i = 1, \ldots, I, j = 1, \ldots, J, k = 2, \ldots, K.$

The candidate $\boldsymbol{\theta}^{I*}$ is accepted with probability

$$1 \wedge \frac{\phi_D\left(\boldsymbol{\theta}^{I*}|\boldsymbol{b}^I, \boldsymbol{V}^I\right)}{\phi_D\left(\boldsymbol{\theta}^I|\boldsymbol{b}^I, \boldsymbol{V}^I\right)} \frac{\phi_D\left(\boldsymbol{\theta}^I|\boldsymbol{m}_I\left(\boldsymbol{\theta}^{I*}\right), \boldsymbol{C}_I\left(\boldsymbol{\theta}^{I*}\right)\right)}{\phi_D\left(\boldsymbol{\theta}^{I*}|\boldsymbol{m}_I\left(\boldsymbol{\theta}^I\right), \boldsymbol{C}_P\left(\boldsymbol{\theta}^I\right)\right)}$$

$$\left[\prod_{i=1}^{I}\prod_{k=2}^{K}\frac{P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} \mid \boldsymbol{Y}_{(i,k-1)}, \boldsymbol{\theta}^{I*}\right)}{P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} \mid \boldsymbol{Y}_{(i,k-1)}, \boldsymbol{\theta}^{I}\right)}\right],$$

where $P_{\boldsymbol{Z}_{(i,k)}}\left(\boldsymbol{Y}_{(i,k)} \mid \boldsymbol{Y}_{(i,k-1)}, \boldsymbol{\theta}^I\right)$ is defined as in expression (5.4) in Chapter 5.

References

GELMAN, A & SPEED, T. P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society, Series B* 55 185–188.

JOE, H. & LIU, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. *Statistics and Probability Letters* 31 113–120.