Multiple Imputation for Individual Patient Data Meta-Analyses

Matteo Quartagno

A thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London

October 2016

Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Funded by Marie Curie ITN MEDIASRES

To my family
Alla mia famiglia

Declaration

Statement of Own Work

All students are required to complete the following declaration when submitting their thesis. A shortened version of the School's definition of Plagiarism and Cheating is as follows (the full definition is given in the Research Degrees Handbook):

The following definition of plagiarism will be used:

Plagiarism is the act of presenting the ideas or discoveries of another as ones own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner which may deceive the reader as to the source is plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a biography will not be deemed sufficient acknowledgement; in each instance, it must be referred specifically to its source. Verbatim quotations must be directly acknowledged, either in inverted commas or by indenting.

Declaration by candidate

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

Signed:

Date: **March 2016**

Full name: **Matteo Quartagno**

Abstract

The term meta-analysis refers to a set of statistical techniques for combining findings from different studies in order to draw more definitive conclusions about some treatment or exposure effect of interest in a particular context. Recently, meta-analyses which aim to combine the individual observations collected in each study, instead of simple summary measures, have been gaining in popularity in medical research. The main advantage of this so-called Individual Patient Data Meta-Analyses (IPD-MA) is that they have much more statistical power to investigate heterogeneity of the contributing studies and to explore treatment covariate effects.

Unfortunately, missing data are a common problem that affects nearly every dataset in clinical or epidemiological studies and therefore also the meta-analyses of such datasets. When not handled properly, missing data can lead to invalid inferences and therefore a lot of research work has focussed on deriving, implementing and disseminating appropriate methods.

The motivation for this thesis comes from two IPD-MA, called INDANA and MAGGIC. Some challenges introduced by missing data in these projects include the presence of wholly missing variables in some studies, the variety of types of partially observed variables and the presence of interactions and non-linearities in the substantive models of interest.

In this thesis we propose a Joint Modelling Multiple Imputation (JM-MI) approach to overcome these issues. Motivated by the lack of available software, in the first part of this thesis we develop and describe *jomo*, a new R package for Multilevel MI. A key feature of *jomo* compared to other packages for MI, is that it allows for the presence of random, or fixed, study-specific covariance matrices in the imputation model, therefore allowing for heteroscedasticity when imputing.

Successively we use this package to prove how our proposed method can be as good as standard methods used nowadays to treat missing data in IPD-MA with partially observed continuous variables. Furthermore we show how it performs in more challenging situations, i.e. to impute missing data in studies with few observations or even with systematically missing variables.

We then extend the method to include partially observed variables that are not continuous, developing and evaluating a strategy based on latent normal variables to impute categorical data.

Finally we use the methods introduced to impute missing data in the two motivating meta-analyses, INDANA and MAGGIC.

Acknowledgements

My sincere gratitude goes to my supervisor, professor James Carpenter, for his support and guidance throughout the three years of my PhD, for keeping me going when times were tough, asking insightful questions, and offering invaluable advice.

Ringrazio Thea, che é piombata nella mia vita poco prima che iniziassi questo percorso e che ha sempre saputo aiutarmi, supportarmi e sopportarmi nei momenti di difficoltá, facendomi dimenticare cosa sia la solitudine e mostrandomi la via della felicitá. Non avrei potuto fare niente di tutto ció che ho fatto in questi anni senza l'aiuto della mia famiglia, a cui questa tesi é dedicata. Ringrazio quindi mamma, papá e Pablo per avermi sostenuto senza pausa e per tutto ció che mi hanno trasmesso. In particolare ringrazio mamma per avermi insegnato il sacrificio e la dedizione, papá per avermi mostrato come reagire, cantando, alle difficoltá e Pablo per avermi insegnato a vedere le cose da un angolo diverso e ad abbandonarsi alla propria vena artistica.

Ringrazio anche tutto il resto della famiglia, zii e cugini tutti; in particolar modo ringrazio Stefano, che ancora in questi tre anni ha saputo essermi vicino e con cui abbiamo continuato

a crescere assieme. Ringrazio Santo e Laura, che mi hanno sempre trattato come fossi un figlio, Rina, che mi ha sopportato durante i miei esodi, e Bianca, la cui calma e tranquillitá rimane sempre per me un esempio. Infine ringrazio Mario, che mi ha fatto da chioccia nella mia nuova veste di straniero nel mondo, preparandomi ad affrontare mille problemi e trasmettendomi sempre grande fiducia.

I wish to thank all the other people that helped me during my stay in London. First of all Mel and Anower, with whom I shared an office and nice moments. Graçias Ale, you have been in London for quite a short time but your friendship has been very important to me. Obrigado George, your jokes are very much missed and were key to keep my spirits up. Thank you Tony, George and thanks to all the other Ph.D students.

Ringrazio poi tutti i miei amici che, seppur a distanza, non hanno mai mancato di farmi sentire il loro affetto. Ringrazio Michele, Stefano, Marco e Alessia, anche se senza i Michele Uo-Oh tutto é stato piú difficile. Ringrazio Amir, Monica, Rebecca, Nicoletta e Alfonso. Le nostre vacanze e le nostre cene sono state un'occasione essenziale per ricaricare le batterie e ripartire. Grazie poi a Giovanni, Linda, Marco, Lorenzo, Vincenzo e tutti gli amici con cui sono cresciuto passo dopo passo e che hanno continuato a farmi sentire la loro amicizia anche da lontano. Ringrazio tutti gli amici della montagna ed in particolar modo Lucia, sempre disponibile ad aiutarmi da 25 anni a questa parte.

In the last three years, I also had the luck to be part of a European network called MEDIASRES. I'm extremely grateful for all the opportunities that being part of this network gave me; therefore danke professor Schumacher, Nadine, Hanna and anyone else who was involved in the organization of this network! Most of all, thanks to all my fellow ph.D students. Tak Ketil, for the enjoyable discussions, for never being trivial and for helping me when I needed it. Thanks Soheila and SungWon, for sharing some nice moments in London and in Bern. *Grazie*

Federico, per avermi introdotto al mondo del disagio. Thanks Alexia, Anna(s), Corine, Hong, Leyla, Markus, Mia and Susanne.

Grazie a tutti i nuovi membri allargati della famiglia, ad Attilio, Cinzia e Clara per avermi sempre fatto sentire a casa in compagnia. A Viola, che ha portato buon umore e spensieratezza in famiglia! Grazie a Alberto e famiglia, il cui affetto sento sempre con me in ogni momento.

Thanks to anyone, both from the School and from other institutes, who helped me with comments, questions and criticism. To Chris, Harvey, Jonathan, Mike, to the reviewers of my first paper, to my upgrade and viva examiners and to anyone else who I might be forgetting.

Grazie al Politecnico di Torino ed a tutti i professori ed i compagni di corso, che in cinque bellissimi anni di studi mi hanno preparato al meglio per questa nuova esperienza.

Last but not least, I would like to thank Life, Love, Destiny and Death, already known in some literature as God. Because it is from the chain of events that occurred to me that I found the desire, and the energy, to do my best to become the person, and the statistician, that I am.

Acronyms and Abbreviations

ACE-I Angiotensin Converting Enzyme Inhibitors

AD Aggregate Data

ARB Angiotensin Receptor Blockers

BB Beta-Blocker

BMI Body Mass Index

BMI Body Mass Index

CI Confidence Interval

COPD Chronic Obstructive Pulmonary Disease

CR Complete Records

CRAN Comprehensive R Archive Network

DBP Diastolic blood pressure

DF Degrees of Freedom

DGM Data Generating Mechanism

EF Ejection Fraction

EM Expectation Maximisation

FCS Full Conditional Specification

HF Heart Failure

IVW Inverse Variance Weighting

IW Inverse Wishart

IPD Individual Patient Data

JAV Just Another Variable

JM Joint Modelling

LOCF Last Observation Carried Forward

MA Meta-Analysis

MAR Missing At Random

MARX Missing At Random depending on X

MCAR Missing Completely At Random

MCMC Markov Chain Monte Carlo

MH Metropolis-Hastings

MI Multiple Imputation

MICE Multiple Imputation by Chained Equations

ML Maximum Likelihood

MMI Multilevel Multiple Imputation

MNAR Missing Not At Random

NHS National Health Service

NYHA New York Heart Association (scale)

OR Odds Ratio

RCT Randomized Controlled Trial

RR Rubin's Rules

SBP Systolic Blood Pressure

SE Standard Error

SSLS Scottish School Leavers Survey

SYPS Scottish Young People Survey

YCS Youth Cohort Study

Table of contents

DECLARATION **1**

ABSTRACT **3**

ACKNOWLEDGEMENTS **5**

ACRONYMS AND ABBREVIATIONS **8**

I Preliminaries **24**

1 Introduction **24**

1.1 Background and Motivation 25

1.2 Outline of this thesis 27

II Literature Review **28**

2 Meta-Analysis: a Brief Overview **28**

2.1 Introduction 29

2.2 IPD vs AD 31

2.3 One-stage vs Two-stage . . . 32
2.4 Missing Data problems . . . 37
2.5 Summary . . . 38

3 Missing Data Methods 40
3.1 Introduction . . . 41
3.2 Older methods . . . 43
3.2.1 Complete Records Analysis . . . 44
3.2.2 Single Imputation . . . 45
3.2.3 EM Algorithm . . . 46
3.2.4 Direct likelihood . . . 48
3.3 Multiple Imputation . . . 49
3.3.1 Joint Modelling . . . 52
3.3.2 Full Conditional Specification . . . 61
3.4 JM vs FCS . . . 63
3.5 Summary . . . 67

4 Why Multilevel Multiple Imputation? 69
4.1 Introduction . . . 70
4.2 The general multilevel model . . . 71
4.3 Missing Data in Multilevel Datasets . . . 73
4.4 Maximum Likelihood methods . . . 75
4.5 Multilevel Multiple Imputation . . . 80
4.5.1 Joint Modelling . . . 80
4.5.2 Full Conditional Specification . . . 83
4.6 Discussion . . . 87

III Research Work **90**

5 Developing an efficient software for Joint Modelling Multiple Imputation **90**

5.1 Existing software for MI 92

5.2 Artificial Dataset 94

5.2.1 Fitting the model 95

5.3 Youth Cohort Study 98

5.3.1 Imputation model 101

5.3.2 Application to YCS model 105

5.4 Conclusions 108

6 jomo: a new R package for Multilevel Joint Modelling Multiple Imputation **109**

6.1 From Matlab to R 110

6.2 Functions 112

6.3 jomo: tutorial with single level datasets 113

6.3.1 A wrapper function: jomo1 121

6.4 jomo: tutorial with multilevel structures 122

6.4.1 Cluster-specific covariance matrices 126

6.4.2 A wrapper function: jomo1ran 130

6.5 Checking convergence of MCMC 130

6.6 Conclusions and further developments 134

7 Multiple Imputation for IPD-MA: allowing for heterogeneity and studies with missing covariates **137**

7.1 Introduction 138

7.2 Meta-analysis and multiple imputation models 141

7.2.1 Substantive analysis models . . . 142
7.2.2 Imputation models . . . 143
7.2.3 Some comments . . . 147
7.3 Simulation study . . . 151
7.3.1 Simulations with studies of equal sizes . . . 152
7.3.2 Simulations with equal size studies: results . . . 154
7.3.3 A correctly specified data-generating mechanism . . . 158
7.3.4 Simulations with studies of different sizes . . . 159
7.3.5 Simulations with systematically missing variables . . . 160
7.3.6 Summary of findings from simulation studies . . . 163
7.4 Discussion . . . 166

8 Multilevel JM MI in presence of other data types 170

8.1 Existing Software and Methods . . . 171
8.2 Methods . . . 172
8.2.1 Single Level Models . . . 173
8.2.2 Multilevel Models . . . 178
8.2.3 Substantive Meta-anaModels . . . 185
8.3 Simulation study . . . 185
8.3.1 Single level model . . . 186
8.3.2 Multilevel models . . . 189
8.3.3 Different data generating mechanisms . . . 197
8.3.4 Practical implication for use of the latent normal model . . . 208
8.4 Ordered categorical variables . . . 208
8.5 Count variables . . . 214

8.6 Discussion 216

IV Applications and Conclusions 223

9 Applications 223

9.1 Introduction 224

9.2 A first example: Indana 224

9.3 Maggic: a theoretical difficulty 228

9.3.1 Imputing covariates of survival model 230

9.3.2 Results 234

9.3.3 Interactions in the analysis model 237

9.3.4 The ideal imputation model 240

9.4 Conclusions 245

10 Summary, Discussion and Future Work 247

10.1 Missing Data in IPD meta-analysis 248

10.2 jomo: outreach and extensions 250

10.3 Imputing IPD-MA: joint multivariate normal model 252

10.4 Imputing IPD-MA: extension to different data types 254

10.5 Factorization of the Joint Model 255

10.6 Conclusion 258

BIBLIOGRAPHY **262**

V Appendices **273**

A MCMC Algorithms for Multilevel Models **273**

A.1 MCMC Algorithm for common covariance matrix 274

A.2 MCMC algorithm for random study-specific covariance matrices 278

B Further Results of Youth Cohort Study Analysis **282**

B.1 Further Results of Youth Cohort Study Analysis 283

List of tables

5.1 Posterior mean (SE) of the fixed effect estimates after running the MCMC for 1100 iterations with the different programs. 97

5.2 Times elapsed for running 1100 iterations of the Gibbs sampler with the 4 different programming approaches. The percent change in time elapsed compared to `REALCOM` is shown in the right hand column. 98

5.3 Different Missing Data Patterns for the Youth Cohort Data: '✓' observed, '×' missing. 100

5.4 Elapsed times for running 1600 updates (JM) or cycles (FCS) of the same models. 104

5.5 Results of the substantive analysis of the Youth Cohort Data. We compare Complete Records (CR), ICE, `REALCOM` and our `mex` function. We create 10 imputations with both methods (JM and FCS), with 500 updates between imputations in JM and 50 cycles in FCS. In both cases $n_{burn} = 500$. 107

7.1 Scenarios used to generate data from (7.3.1), and corresponding consistent (i) meta-analysis and (ii) imputation models, when values of X_2 are missing. . . . 153

7.2 Simulations with studies of equal sizes. Mean estimates, SE and coverages for the coefficient of variable x_1, the completely observed covariate. Scenarios 1,2 & 3: Comparison of results from imputation model (7.2.3),(7.2.5) and (7.2.6) respectively (the simplest ones compatible with the data) and imputation model (7.2.7) (the most general). Scenarios 4 & 5: Comparison of all the different models presented, starting with the simplest one (7.2.3) and ending with the most general (7.2.7). Data Generated with (7.3.2): Comparison of the 3 more general models in the previous scenario. Results in bold highlight cases where both the meta-analysis and the imputation model are compatible with the data-generating mechanism. 155

7.3 Simulations with studies of equal sizes. Mean estimates, SE and coverages for the coefficient of x_2, the partially observed covariate. Scenarios 1,2 & 3: Comparison of results from imputation model (7.2.3),(7.2.5) and (7.2.6) respectively (the simplest ones compatible with the data) and imputation model (7.2.7) (the most general). Scenarios 4 & 5: Comparison of all the different models presented, starting with the simplest one (7.2.3) and ending with the most general (7.2.7). Data Generated with (7.3.2): Comparison of the 3 more general models in the previous scenario. Results in bold highlight cases where both the meta-analysis and the imputation model are compatible with the data-generating mechanism. 156

7.4 Simulations with studies of different sizes. Mean estimates, SE and coverages for the first coefficient, the one related to x_1, the completely observed covariate. Scenario 3,5 and data generating mechanism (7.3.2): Comparison of results from imputation models (7.2.6) and (7.2.7). Once again, cases where both the imputation and the meta-analysis model are consistent with the data-generating mechanism are highlighted in bold. 161

7.5 Simulations with studies of different sizes. Mean estimates, SE and coverages for the second coefficient, the one related to x_2, the partially observed covariate. Scenario 3, 5 and data generating mechanism (7.3.2): Comparison of results from imputation models (7.2.6) and (7.2.7). Once again, cases where both the imputation and the meta-analysis model are consistent with the data-generating mechanism are highlighted in bold. 162

7.6 Simulations with studies with systematically missing variables. Scenarios 1–5 and data generated with (7.3.2): Results with imputation model (7.2.7), the only one usable in this situation. Note that with systematically missing variables, it is not possible even to run complete records analysis, unless we exclude the studies with systematically missing variables. 164

8.1 Simulations with single level data. Data are generated with model (8.3.2). Mean, SE and coverage level is reported for the four slope parameters. Missing data are introduced both with MCAR and MAR mechanisms and complete records is compared to JM imputation, using imputation model (8.2.4) with 10 imputed datasets. 189

8.2 Simulations with 2-level data. Data are generated with model (8.3.4). Mean, SE and coverage levels are reported for the four slope parameter estimates. Missing data are introduced with MCAR and MAR mechanisms. The systematically missing data case is also explored. Complete records are compared to JM imputation using imputation models (8.2.5), (8.2.6) and (8.2.7), generating 10 imputated datasets in each case. 192

8.3 Simulations with 2-level data with heterogeneous covariance matrices. Data are generated with model (8.3.5). Mean, SE and coverage levels are reported for the three parameter estimates. Missing data are introduced with MCAR and MAR mechanisms. The systematically missing data case is also explored. Complete records is compared to JM imputation using imputation models (8.2.5), (8.2.6) and (8.2.7), generating 10 imputated datasets in each case. 196

8.4 Simulation results with single level data generated with models (8.3.7), (8.3.10) and (8.3.12). Mean, SE and coverage levels are reported for the three parameter estimates. Data are MCAR and Complete records is compared to JM imputation using an imputation model with all the 3 variables as outcomes, similarly to model (8.2.4), and generating 10 imputed datasets in each case. 200

8.5 Results of the analysis of datasets with partially observed ordinal data. In both scenarios data are 20 % MCAR in all the variables and we compare complete data analysis with CR and JM-MI. We report mean, SE and coverage probabilities. 213

8.6 Results after running Poisson regression model on data generated trough (8.5.1) with 20% MCAR in all three variables. We compare mean estimates, standard errors and coverage probabilities obtained by using complete data, complete records or JM-MI, either using untransformed Y, square root or logarithmic transformation. 217

9.1 INDANA meta-analysis: extent of missing data. Number of missing items (percentage) in each study and variable. (DBP – Diastolic Blood Pressure) 225

9.2 Results of analysis of the INDANA individual patient data meta-analysis. The coefficient is the estimated reduction in mean DBP after one year due to treatment, adjusting for baseline cholesterol, age and sex. We compare complete records, JM-MI with common covariance matrix and JM with random study-specific covariance matrices. In both imputations, we use 15 imputations, 1000 burn in iterations and 100 between-imputation iterations. We analysed the data with four meta-analysis models. 227

9.3 MAGGIC datasets: Extent of missing data. Age and sex are the only fully observed variables. 235

9.4 Results of MAGGIC analysis. Outcome is time to death, or censoring. Rate ratios (91% CI) of analysis using FCS MI, JM-MI with common covariance matrix and JM-MI with study-specific covariance matrices. 236

B.1 Posterior Means for the three coefficients β in model 1 of Table 2.4 283

B.2 Posterior Means for the seven coefficients β in model 2 of Table 2.4 283

B.3 Posterior Means for the nine coefficients β in model 3 of Table 2.4 284

B.4 Posterior Means for the twenty-seven coefficients β in model 4 of Table 2.4 . . 285

List of figures

6.1 MCMC chain for one of the fixed effect parameters 132

6.2 MCMC chain for one element of the level 1 covariance matrix 133

8.1 Histogram of the distribution of p, i.e. the probability of an event $y_i = 1$, for one of the simulations. 198

8.2 Scatterplots of Y over Z, for a random selection of 1000 draws generated from (8.3.16) (on the left) and (8.3.17) (on the right). In this case $\mu_0 = 3$ and $\mu_1 = 7$, i.e. $\mu_1 - \mu_0 = 4$, four times the standard deviation of $Y|Z$. 206

8.3 Scatterplots of Y over Z, for a random selection of 1000 draws generated from (8.3.16) (on the left) and (8.3.17) (on the right). In this case $\mu_0 = 0.3$ and $\mu_1 = 0.5$, i.e. $\mu_1 - \mu_0 = 0.2$, a fifth of the standard deviation of $Y|Z$. 207

8.4 The relation between the coverage level in the estimation of a particular fixed effect estimate β_2 and the ratio $\frac{\beta_2}{\sigma_{resid}}$. The data generating model is 8.3.12 with varying values of β_2, between 0.1 and 10. 209

8.5 Imputation of Poisson data: histograms showing the distribution of Y, sqrt(Y) and log(Y+0.1) in the three scenarios considered. 218

10.1 A flow-chart explaining when multilevel MI is necessary, compared to within-study imputation and imputation with fixed cluster effect. 249

Part I

Preliminaries

1

Introduction

1.1 Background and Motivation

The motivation for this thesis comes from two meta-analyses projects, the Individual Data ANalysis of Anti-hypertensive drug intervention trials (Gueyffier *et al.*, 1995, INDANA) and the Meta-Analysis Global Group In Chronic heart failure (Pocock *et al.*, 2013, MAGGIC). These are Individual Patient Data Meta-Analyses (IPD-MA), meaning that in each of them the whole set of observations coming from the contributing studies are aggregated, instead of the simple results or summary tables produced separately (Simmonds *et al.*, 2005; Riley *et al.*, 2010).

There are a number of advantages in performing meta-analyses of IPD, among which perhaps the most important are the chance to better model heterogeneity and to undertake more powerful subgroup analyses (Stewart and Parmar, 1993). However, this presents some further challenges to the analyst. First of all, recovering all the individual patient information from all the studies may be difficult for a variety of reasons, such as the increased cost or confidentiality issues. Furthermore, even if we are actually able to collect all the available data from the authors of the constituent studies, missing data can still be a major issue:

- Missing data can occur in all the constituent studies and therefore in the MA dataset;
- Some variables may not have been observed in some studies, due to excessive cost of some measurement or simply because of a different study design;
- Some variables may be recorded on different scales in different studies.

The goal of this thesis is to devise and implement a unified approach to deal with these and other issues related to missing data that we will illustrate later, in order to be able to perform IPD-MA using data from all patients.

The method we chose to adopt is Multiple Imputation (Rubin, 1987, MI). This is a tool for dealing with missing data, roughly consisting in imputing the missing values several times from the appropriate conditional distributions given the observed data, creating a certain number of imputed datasets (with no missing values) that we can later analyse with standard techniques, before combining the results to obtain inference for the parameter(s) we are interested in.

In the context of IPD-MA, MI presents some specific and significant challenges, mainly the fact that in the whole dataset clustering cannot be ignored without introducing bias in the results (Abo-Zaid *et al.*, 2013). Therefore appropriate algorithms allowing for 2-level structures must be used. This is particularly important for handling studies which have not collected all the variables. The lack of efficient software to undertake this kind of imputation, led us to code and publish a new R package, `jomo` (Quartagno and Carpenter, 2014), that we will later introduce in this script.

Having developed this approach, and the associated software, we demonstrated its efficacy and advantages in the above settings, dividing our analysis in two situations:

1. when we have only continuous partially observed variables (Quartagno and Carpenter, 2015);

2. when we have partially observed variables of mixed data types.

Our results are encouraging, showing that joint modelling MI provides a practical and powerful tool for tackling missing data issues in the IPD-MA setting.

1.2 Outline of this thesis

This thesis is divided into ten chapters. Chapter 2 is an introduction to meta-analysis, illustrating both analysis of individual patient data (IPD) and aggregate data (AD). Chapter 3 introduces missing data issues and the main strategies to handle them in simple situations. In Chapter 4 we extend our review of these methods to the case of multilevel data, in order to justify the use of MI in the case of IPD-MA with missing observations. After having proved in Chapter 5 that it is possible to develop an efficient software for multilevel MI, we present this new program, the `R` package `jomo`, in Chapter 6. We therefore illustrate in Chapter 7 the results of simulations to test the proposed MI method in IPD-MA with partially observed continuous variables. Chapter 8 includes the results of the simulations we performed to test the method for the imputation of partially observed data of different types, mainly binary and categorical. Finally in Chapter 9 we apply the proposed methods to handle missing data in the two motivating meta-analyses (INDANA and MAGGIC), while Chapter 10 summarises the findings of this thesis and discusses extensions and future work.

Part II

Literature Review

2

Meta-Analysis: a Brief Overview

In this chapter, we introduce the concept of meta-analysis, highlighting the differences between analyses based on individual patient data (IPD) or aggregate data (AD), presenting different possible models and framing our research problem.

After a short introduction in Section 2.1, we present a comparison of IPD and AD meta-analyses in Section 2.2. Later we introduce one and two-stage models in Section 2.3, before outlining the issues arising from missing data in Section 2.4 and concluding with a short summary in Section 2.5

2.1 Introduction

In the world of clinical research, many studies are performed each year on similar topics, possibly within different populations. For this reason, it appears natural to try and synthesize the findings of these studies in order to find results that are generalizable to a larger population and to get better precision and accuracy of the estimates of interest. This process in called Meta-Analysis (MA), from the Greek word 'meta', meaning 'after', 'beyond'.

The first step in pursuing a MA is always to formulate the problem of interest; to keep things as simple as possible, imagine we want to estimate the effect of a particular exposure X over an outcome Y with a simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.1.1}$$

where i indexes the individuals and ϵ_i is the error term, that can be assumed for simplicity to be normally distributed with variance σ_e^2. In some studies the intercept β_0 and/or the residual variance σ_e^2 could be the focus of our interest, but in this case suppose we are mainly interested in the estimate $\hat{\beta}_1$ of the slope β_1.

Once we have decided what we want to estimate we need to find out which studies investigated the exposure effect of X with a careful literature search. This can be done in various ways, by simple web searching on websites like PubMed, by consultation of the experts in the field, or by means of the Cochrane Library, a collection of databases in medicine provided by some organizations, principally the Cochrane Collaboration.

After having selected all the potentially relevant studies, carefully drawn up selection criteria must be applied to choose which ones are actually worth including in the analysis.

Having short-listed the studies of interest, the first question that is encountered by the analyst is: what kind of data do we have to aggregate?

The first option is to combine study-level data obtained from publications, e.g. estimates of treatment or exposure effects together with their standard error, which in our example (2.1.1) are $\hat{\beta}_{1,s}$ with the associated $\hat{\sigma}_{1,s}$. Here, $s = (1, \ldots, S)$ indexes the studies, so that $\hat{\beta}_{1,s}$ is the estimate of β_1 from study s. This is what has been defined as Aggregate Data (AD) Meta-Analysis. However, if it is possible to collect all the individual records from the constituent studies, i.e. all the $y_{i,s}$ and $x_{i,s}$ where $i = (1, \ldots, N_s)$ and N_s is the cardinality of each study s, then we can model the Individual Patient Data (IPD) directly.

2.2 IPD vs AD

Obtaining IPD may require a lot more of effort for many reasons: some study authors may not be willing to share their data, some others may be difficult to contact, a lot of work in data cleaning may be needed, etc etc... Then, one might wonder if it is really worth doing it, or if focusing on AD meta-analyses could be enough. There are a number of papers in the literature highlighting the advantages of collecting the IPD; for example, Riley *et al.* (2010) list among other reasons:

- Inclusion and exclusion criteria can be reformulated in order to make them consistent across studies, so that in some studies some patients, when possible, can be included or excluded from the analysis;
- Results of the original publications can be verified;
- New follow-up information, unavailable at the time of publishing, can be included;
- In principle, results from unpublished studies can be incorporated, reducing the risk of publication or other selection biases;
- The same statistical analysis can be performed across studies, standardizing the procedure used; this could include running models with slightly different assumptions and including (or excluding) some covariates, and
- Subgroup analyses can be performed more easily, helping to reduce between-study heterogeneity.

This last point is key and it is probably the main reason why over the last few years IPD-MA are progressively starting to replace, or at least to complement AD-MA, which were historically the first type of meta-analyses performed (Simmonds *et al.*, 2005).

Furthermore, since our focus in this thesis will be on missing data problems, having the opportunity to directly analyse the original data can give the opportunity to handle missingness consistently across studies, using the same strategy and making it possible to try to recover not only missing information within studies, but also wholly missing variables, as we will see later in the thesis.

Of course, the only case where IPD and AD analyses will match, is when all the studies will have followed broadly the same protocol, using the same inclusion criteria, the same models and the same missing data strategies. However this will rarely be the case and therefore IPD will be usually preferable.

2.3 One-stage vs Two-stage

Once we have obtained our data, either AD or IPD, we need to choose the model to use. There are several meta-analysis models available; here our focus is on the distinction between one-stage and two-stage models.

Broadly, one-stage models correspond to the IPD-MA and two-stage models to AD-MA; however, while the former can not be used when in possession of the AD only, the latter are often used even with IPD.

In general,we talk about one-stage analysis when all the IPD from the contributing studies are modelled simultaneously in a single step; this is not done by simply running the desired model on a 'huge' dataset composed of all the IPD, a practice that can and usually does lead to invalid inferences (Abo-Zaid *et al.*, 2013). Methods taking into account correlation between observations coming from the same study must be used instead, making it possible to control for possible sources of heterogeneity between the studies. One possible solution is simply to include study ID in the model as a covariate with a fixed effect. Otherwise we can use methods allowing for random effects for clusters; these are the so-called multilevel models (Goldstein, 2011), already known as hierarchical or mixed models.

Coming back to our simple example (2.1.1), a simple hierarchical model could be:

$$y_{i,s} = (\beta_0 + u_{0,s}) + \beta_1 x_{i,s} + \epsilon_{i,s} \tag{2.3.1}$$

Here, the only substantial difference with respect to model (2.1.1) is given by the new term $u_{0,s}$, that is a study-specific random intercept that can be assumed to follow a normal distribution $N(0, \sigma_u)$. If we believe that not only the intercept, but even the slopes should be modelled allowing for a different slope in each constituent study, then we might opt for a random intercept and slope model:

$$y_{i,s} = (\beta_0 + u_{0,s}) + (\beta_1 + u_{1,s}) x_{i,s} + \epsilon_{i,s} \tag{2.3.2}$$

The random slope term $u_{1,s}$ can be modelled in different ways, for example a simple option is to suppose that $u_{0,s}$ and $u_{1,s}$ follow a bivariate normal distribution, either with an unstructured covariance matrix or with correlations set to zero.

Since in IPD-MA, variances of the residuals are usually different in the different studies, it is particularly important considering models for complex level 1 and 2 variation in this setting (Goldstein, 2014). For example, the assumption in model (2.3.2) that $\epsilon_{i,s} \sim N(0, \sigma_e^2)$ could be relaxed by allowing for study-specific residual variances, $\epsilon_{i,s} \sim N(0, \sigma_{e,s}^2)$.

Regarding the estimation method, maximum likelihood is known to lead to biased estimates of the variance components, and therefore Restricted Maximum Likelihood is often used instead, particularly with smaller sample sizes.

When we use a two-stage model, instead, in the first step we fit our analysis model of interest, i.e. (2.1.1), within each study if we have the IPD, or we collect the results, i.e. $\hat{\beta}_{1,s}$ and associated standard errors $\hat{\sigma}_{1,s}$, from the single studies s if we only have the AD, and then we synthesize the evidence in order to get some more definitive conclusions about the treatment or exposure effect of interest. A natural idea is to weight the results from the single studies according to the magnitude of the standard errors; this is what is done in the Inverse Variance Weighting (IVW) approach (DerSimonian and Laird, 1986).

There are two different approaches to IVW: a fixed-effect analysis or a random-effects analysis. In both cases, we weight the estimates for the single parameters coming from the different studies according to the inverse of their variances. The difference between the two methods is that while in the former we assume a common treatment effect across the different studies, in

the second we assume that the actual effect within each study is only a random draw from a certain distribution; therefore we make an adjustment to the study weights according to the extent of variation, or heterogeneity, among the varying intervention effects. Given estimates $\hat{\beta}_{1,s}$ and $\hat{\sigma}_{1,s}$, the fixed effect IVW model is:

$$\hat{\beta}_{1,s} = \beta_1 + \epsilon_s, \qquad \epsilon_s \sim N(0, \sigma^2_{1,s}) \tag{2.3.3}$$

The maximum likelihood estimate $\hat{\beta}_{1,FE}$ under this model is:

$$\hat{\beta}_{1,FE} = \frac{\sum_{s=1}^{S} \frac{\hat{\beta}_{1,s}}{\hat{\sigma}^2_{1,s}}}{\sum_{s=1}^{S} \frac{1}{\hat{\sigma}^2_{1,s}}} = \frac{\sum_{s=1}^{S} w_s \hat{\beta}_{1,s}}{\sum_{s=1}^{S} w_s} \tag{2.3.4}$$

Here w_s are the weights we give to the estimates and, in the case of IVW, they are just equal to the inverse of the variance for each study, i.e. $w_s = \frac{1}{\hat{\sigma}^2_{1,s}}$. Note that within-study variances $\hat{\sigma}^2_{1,s}$ are assumed to be known. An estimate of the variance of $\hat{\beta}_{1,FE}$ is:

$$\hat{Var}(\hat{\beta}_{1,FE}) = \frac{1}{\sum_{s=1}^{S} w_s}. \tag{2.3.5}$$

The corresponding random effects model is:

$$\hat{\beta}_{1,s} = \beta_1 + u_s + \epsilon_s, \qquad \epsilon_s \sim N(0, \sigma^2_{1,s}); \qquad u_s \sim N(0, \tau^2), \tag{2.3.6}$$

where u_s and ϵ_s are independent. The u_s are not intrinsically associated to each study s, but if we were able to re-run any study s we could draw a different value u_s for the same study(i.e. exchangeability). The second addition of this model, τ^2, is the between-study component of variance. The algebraic form of random effects estimates of β_1 and σ_1 are really similar to (2.3.4) and (2.3.5), with the only difference that the weights w_s are substituted by w_s^*, with:

$$w_s^* = \frac{1}{\hat{\sigma}_{1,s}^2 + \hat{\tau}^2} \tag{2.3.7}$$

There are many different methods available in the literature for finding an estimate $\hat{\tau}^2$ of the between-study component of variance τ^2, among which the most used is that proposed by Der Simonian and Laird (DerSimonian and Laird, 1986); however, in some settings this has been criticised (Hardy and Thompson, 1996; Brockwell and Gordon, 2001) and many other methods have been developed (Hunter and Schmidt, 1990; Sidik and Jonkman, 2005; DerSimonian and Kacker, 2007). The main problem with the Der-Simonian and Laird method is that often standard errors of the estimates are underestimated; therefore (Hartung and Knapp, 2001) proposed a modified estimate of the variance, and also using a t distribution for deriving confidence intervals in place of the standard normal distribution.

As we already said, with AD, only two-stage MA is possible but, as for the IPD case, we can still choose between a fixed-effect or a random-effects analysis. This is usually done both by means of appropriate strategies, like the examination of the Q and the I^2 statistics (Higgins and Thompson, 2002; Higgins *et al.*, 2003), and by an appropriate consideration of expertise knowledge in the area of the meta-analysis (Borenstein *et al.*, 2010).

On the other hand, with IPD, we also need to choose between one-stage or two-stage analysis. It has been generally thought that both methods lead to similar results and conclusions (Stewart *et al.*, 2012); however recently (Debray *et al.*, 2013) showed how the two methods can in principle lead to different conclusions in some situations and they recommend to use one-stage models, particularly in situations where few studies or few individuals per study are present.

2.4 Missing Data problems

Since missing data are ubiquitous in clinical dataset, they certainly occurr in the IPD meta-analysis of these data; when a variable is partially observed in some of the contributing studies, we say that we have **sporadically missing data**. The issues they raise have been addressed for example in (Burgess *et al.*, 2013).

Furthermore, some variables may have not been observed or collected at all in some studies, perhaps due to financial constraints or simply to different study designs. In this case we talk about **systematically missing data**; some papers where solutions to this problem have been proposed include (Resche-Rigon *et al.*, 2013; Jolani *et al.*, 2015).

As we said in Section 2.2, when considering the whole data as a unique large dataset, we need to use multilevel methods for the analysis,either with a fixed or random treatment effect. We will see in the following chapters how this will be a challenging problem when trying to find methods to handle both sporadically and systematically missing values in IPD-MA at the same time.

The problem of missing data in AD-MA has been recently addressed as well in several papers. White et al. (2008) introduced in two companion papers two methods based on the use of Informative Missingness Odds Ratios (IMOR). The idea behind these methods is to adjust the odds ratios and their associated measures of precision from contribuent studies allowing for the uncertainty due to missing data; this is done, in a Bayesian fashion, by incorporating prior belief about the missing data mechanism. More recently, Turner et al. (2015) have developed another Bayesian method allowing for priors for the parameters describing the missingness mechanism to be specified in multiple ways, not only through the IMOR but also with the probability of success given a missing subject and the response probability ratio.

2.5 Summary

In this chapter we have introduced the main features of meta-analysis, distinguishing between IPD and AD, one-stage and two-stage models and giving an overview of the missing data problems usually encountered by the analysts.

Many other key concepts and techniques related to meta-analysis have been developed, including:

- Meta-regression (Stanley and Jarrell, 1989), *"an extension to subgroup analyses that allows the effect of continuous, as well as categorical, characteristics to be investigated, and in principle allows the effects of multiple factors to be investigated simultaneously"* (Higgins and Green, 2008);

- Statistical methods to detect and adjust for the recurrent issue of publication bias (Easterbrook *et al.*, 1991), that is bias arising from the non publication of small studies that failed to prove the efficacy of a particular treatment;
- Methods for multivariate meta-analysis (Mavridis and Salanti, 2013), and
- Network meta-analysis, i.e. meta-analysis comparing several treatment effects through both direct and indirect comparisons between studies (Caldwell, 2014).

However we decided not to pursue these further here because they are not directly relevant to our work.

Some people argue that meta-analyses should be performed only with data coming from RCTs, excluding observational studies; this is because RCTs are considered a much more valid design for causal inference. However, many reviews have shown how including data from observational studies may improve the inference and that RCT-only meta-analyses often give similar results to meta-analyses of observational studies (Shrier *et al.*, 2007).

Before continuing with the presentation of the strategies for addressing the missing data issues in meta-analysis, Chapter 3 presents the main missing data methods in a general setting. We will come back to the IPD-MA setting in Chapter 4.

3

Missing Data Methods

This chapter gives an overview of the problems raised by missing data in clinical research and of the methods currently used to handle them. In Section 3.1 we outline the different missing data mechanisms, which are the key concepts underpinning the analysis of partially observed data. Then, in Section 3.2 we review the oldest and most straightforward methods for the analysis of partially observed data, discussing why we cannot always rely on them. In Section 3.3 we introduce Rubin's Multiple Imputation, describing both the Joint Modelling approach (Subsection 3.3.1) and Full Conditional Specification (Subsection 3.3.2). Finally, in Section 3.4 we compare this two imputation methods, itemizing pros and cons of both.

3.1 Introduction

Since unfortunately most of the studies carried out in medical and social sciences are unable to collect all the planned data, missing data are a common issue in statistical analyses. Not only do they cause a loss of information, and hence power, but they may also cause bias in parameter estimates and hence potentially lead to misleading inferences. The first paper to deal systematically with these issues was probably (Rubin, 1976), which is considered a landmark paper in this area. In it, Rubin suggested three different kinds of missing data mechanisms: Missing At Random, Missing Completely At Random and Missing Not At Random. In order to explain the different mechanisms, we define some notation.

Imagine that we intended to collect data on n units; we can call $\mathbf{Y}_i$ the vector of variables that we planned to observe for unit i. To reflect the missing data, we split this vector into two parts: $\mathbf{Y}_i^O$, the sub-vector of observed variables for that unit, and $\mathbf{Y}_i^M$, the sub-vector of missing variables. It is important to stress that these variables do exist, but have just not been

collected in our study. Finally, we can define a third vector, called $\mathbf{R}_i$, which is a vector of binary variables such that:

$$Y_{i,j} \text{ missing} \Rightarrow R_{i,j} = 0,$$

$$Y_{i,j} \text{ observed} \Rightarrow R_{i,j} = 1.$$

Given this notation, we can now define the three missing data mechanisms as follows:

- Missing Completely At Random (**MCAR**): Data are defined to be MCAR if the probability of a value being missing is completely unrelated to both the observed and the underlying unseen values on each unit. Algebraically:

$$\mathcal{P}(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M) = \mathcal{P}(\mathbf{R}_i).$$

- Missing At Random (**MAR**): Data are defined to be MAR if, given or conditional on the observed variables on a unit, the probability of a value being missing is completely unrelated to underlying unseen values on that unit. In formulae:

$$\mathcal{P}(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M) = \mathcal{P}(\mathbf{R}_i|\mathbf{Y}_i^O).$$

  It is quite important here to stress that this does not mean that the probability of a variable being missing is completely independent from its value. Marginally, $\mathbf{R}_i$ depends on $\mathbf{Y}_i^M$, but given $\mathbf{Y}_i^O$ this dependence is broken.

- Missing Not At Random (**MNAR**): Data are defined to be MNAR if the probability of a value being missing depends on the underlying missing values on that unit, even after conditioning on the observed variables. Algebraically:

$$\mathcal{P}(\mathbf{R}_i|\mathbf{Y}_i) = \mathcal{P}(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M).$$

We emphasize that these are just assumptions that we make about the reasons for the data being missing in the context of the analysis at hand, and not only they are un-testable, but they are not properties of the dataset itself. For this reason, even though we might often be tempted to rely on the MAR assumption, because of the simplifications that it brings to the analysis, we should always check the robustness of the results of our analysis to different assumptions with sensitivity analysis (Molenberghs *et al.*, 2014, Part V).

3.2 Older methods

Before focusing on Multiple Imputation, we give a brief overview on the main older methods to deal with missing data, trying to highlight both the advantages and the issues raised by each.

3.2.1 Complete Records Analysis

Complete Records Analysis, also commonly known as Listwise Deletion or Case Deletion, historically has been the most popular approach to deal with missing data in statistical analysis. It simply excludes from the analysis all units whose collected data are incomplete.

$$\mathbf{Y}_i^O \neq \mathbf{Y}_i \Rightarrow \text{unit } i \text{ discarded.}$$

If we are interested in simple summary statistics like the sample mean, or the standard deviation, then the only missing data mechanism that is compatible with this method is MCAR. However, if we are interested in the regression of, say, $Y_{i,1}$ on $Y_{i,2}, \ldots, Y_{i,p}$ *and* the missingness mechanism does not involve the response $Y_{i,1}$ (sometimes referred to in the literature as the MARX mechanism), then a complete records analysis will give valid inference. In both cases, of course, complete records analysis is inefficient, i.e. it does not make effective use of all the available information.

Obviously the reason for the success of this method is its extreme simplicity, plus the fact it is the default in most statistical software packages. Nonetheless it is clear that it is absolutely not a sufficient tool in a variety of situations, and hence we need some more sophisticated ones.

3.2.2 Single Imputation

Instead of dropping all data from incomplete units, we may think about 'filling in', i.e. imputing, the missing items in order to create an artificial 'complete' dataset. Data should be imputed in a way that reflects the distribution of the observed data, in order not to bias the results. There are several different possible choices, here we list some of them:

- Mean substitution: impute for each missing value the mean of the observed units on that variable. This is certainly the most straightforward single imputation approach. However, it does not preserve conditional relationships in the data and clearly leads to underestimation of the variance.
- Last observation carried forward: in a longitudinal study framework, imagine data for a patient are missing with a monotone pattern, i.e. there exists I such that, if we intended to collect N observations on subject j, $Y_{i,j}$ is missing $\forall i > I$ and observed otherwise. Then, we can copy the value of $Y_{I,j}$ for all the successive observations.However, as with mean imputation, this does not preserve conditional relationships in the data and leads to underestimation of the variance.
- Conditional mean imputation: to illustrate this method, suppose $\mathbf{Y}_1$ is partially observed. We fit a regression model of $\mathbf{Y}_1$ on the other variables for the cases for which $\mathbf{Y}_1$ is known. Then we plug the values of the other variables into the regression equation, obtaining predictions $\hat{Y}_{1,j}$ for all observations j with missing $Y_{1,j}$. We then impute the missing observations with these predictions. The main problem with this method is that it understates the variability, as we know that the missing value is not equal to its conditional mean.

- Hot Deck Single Imputation: the idea is to replace each missing value with a random draw from a suitable pool formed from the observed data. This can be viewed as a non-parametric form of conditional mean imputation

Even though when imputing from the conditional distribution both the means and the correlations may be preserved, (Rubin, 1987) and subsequently (Schafer and Graham, 2002) showed with simulations that the performance of resulting confidence intervals is poor.

In general, the problem with single imputation is that it understates the level of uncertainty, because it ignores the fact that imputations are just guesses; hence, applying standard analysis methods to the imputed data, which do not distinguish between the observed and the imputed data, will substantially overestimate the precision of the estimates. Later on in this report we will describe how Multiple Imputation overcomes this problem.

3.2.3 EM Algorithm

Drawing inferences from a likelihood function is a common procedure in statistics, but the presence of missing values complicates this. (Dempster *et al.*, 1977) formalised an algorithm for computing the ML estimates in presence of missing data. Suppose that $\log L(\boldsymbol{\beta}|\mathbf{Y})$ is the log-likelihood of the data we want to maximize with respect to $\boldsymbol{\beta}$. When $\mathbf{Y} = (\mathbf{Y}^O, \mathbf{Y}^M)$, the EM algorithm does this by iterating the following two steps:

- Expectation step (E-step): using the conditional expectation of the missing data given the observed data at the current set of parameters $\boldsymbol{\beta}^t$, calculate the expected value of the log-likelihood function:

$$\mathcal{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^t) = \mathbb{E}_{(\mathbf{Y}^M|\mathbf{Y}^O,\boldsymbol{\beta}^t)} \log L(\boldsymbol{\beta}|\mathbf{Y}^O, \mathbf{Y}^M);$$

- Maximization step (M-step): find the new set of parameters $\boldsymbol{\beta}^{t+1}$ that maximize $\mathcal{Q}$ from the previous step:

$$\boldsymbol{\beta}^{t+1} = \arg_{\boldsymbol{\beta}} \max \left\{ \mathcal{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^t) \right\}.$$

These two steps are repeated until they converge to the final solution. Aside from multi-modal distributions, where we could fall into a local maximum, the EM algorithm always reaches a final solution, because at each step it is possible to prove that the likelihood increases. However, it still has some disadvantages: it is not computationally straightforward and convergence may be very slow. Most of all, it does not provide standard errors for the parameter estimates; to obtain them we need to do model specific calculations (Louis, 1982).

For these reasons, the EM algorithm alone is not a sufficient, or especially practical, tool. Nevertheless, in the context of MI, it may be useful for finding starting values for the parameters when initializing a joint model imputation, as we will see in the next section.

3.2.4 Direct likelihood

Besides the EM algorithm, in recent years there has been increasing interest in the direct maximization of the likelihood in a single procedure, a method that has been referred to as direct likelihood or Full Information Maximum Likelihood (FIML). The likelihood contribution for each unit affected by missing data can be calculated by integrating over all the possible values of the partially observed variables; there are situations where this can be done quite easily: for example when we are interested in a linear regression and the only missing value for individual i is in a binary covariate X, the likelihood contribution for i can be obtained as:

$$L(Y_i|X_i;\theta,\phi) = \sum_{k=0,1} (L(Y_i|X_i = k;\theta)\pi(X_i = k;\phi))$$

where π indicates a marginal model for X_i with parameter ϕ.

However, with increasing numbers of variables with missing values, and hence missing data patterns, possibly with missing values in continuous variables, the calculation of the likelihood contribution can become more and more difficult. A closed form solution of the integral may not exist and in some cases we might have to rely on numerical approximations.

We will explore the possible issues of this method more in detail in Chapter 4, in the specific case of multilevel data structures.

3.3 Multiple Imputation

As we have seen in Subsection 3.2.2, imputing missing values is an attractive approach, since models can be fitted to the resulting dataset with standard software. However, this gives the same weight to the observed and imputed data, so leading to underestimation of the standard errors.

Multiple imputation is a simple, widely applicable approach to correct for this. It is based on a really simple idea: instead of just imputing a missing value once, we draw it say K times, creating K 'complete' imputed datasets. In this way, we have not a single imputed value for each missing datum, but instead K datasets that together represent the distribution of missing given the observed data. Once we have our K datasets, we then fit our substantive model to each of them with standard software and finally combine the results together in some way.

The idea of Multiple Imputation is due to Rubin (Rubin, 1987) and can be summarized in three steps:

1. Imputation step: we impute $K > 1$ different datasets using one of the different approaches we will illustrate in the next sections;

2. Analysis step: we analyse the K datasets with standard methods, as if they were complete datasets, obtaining K estimates that we will call $\hat{\beta}_k$, with $k = 1, \ldots, K$, and their associated variance estimates $\hat{\sigma}_k^2$;

3. Combination step: we combine the results for inference using Rubin's rules, defined in (Rubin, 1987).

This last step is key; it is what makes MI attractive, as the rules are simple and widely applicable. We give here Rubin's rules for a single parameter β, for vector rules see (Rubin, 1987, p.76). Starting from our K different $\hat{\beta}_k$, the point estimate is simply their average:

$$\hat{\beta}_{MI} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k.$$

The variance of $\hat{\beta}_{MI}$ is estimated by the sum of a within imputation variance and a between imputations one, corrected with a $\left(1 + \frac{1}{K}\right)$ term to account for the finite number of imputations:

$$W_{MI} = \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}_k^2$$

$$B_{MI} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{\beta}_k - \beta_{MI}\right)^2$$

$$\sigma_{MI} = W_{MI} + \left(1 + \frac{1}{K}\right) B_{MI}$$

Multiple Imputation is at heart a Bayesian procedure with some good frequentist properties; however, in order for Rubin's rules, and hence for these frequentist properties, to hold, the different imputations must be stochastically independent.

Another important feature of Multiple Imputation is its flexibility, since the imputation model and the analysis model do not have to be the same. However, in order for Rubin's variance combination rules to hold precisely, the two models need to be *congenial*, essentially meaning

that there should be one joint distribution for the data, from which both the imputation and the substantive model can be derived by appropriate conditioning. Simply speaking, there may be two possible kinds of uncongenial models:

- The imputation model is *poorer* than the analysis model: there are some variables present in the analysis model that for some reason are missing in the imputation model; this situation should be avoided because not only it causes invalidity of Rubin's rules variance formula, but it can also result in inconsistent parameter estimates for the substantive model;
- The imputation model is *richer* than the analysis model: in the imputation model there are some more variables, maybe because they are not of interest in our analysis model but they can provide further information about the missingness mechanism; in this case, we can say that there is at least one congenial imputation model nested within the imputation model we are using.

Although in this second setting Rubin's variance formula may overestimate the sampling variability of the MI estimators, this overestimation is typically small, when set against the benefits of including auxiliary variables in the imputation model and the loss of information with complete records analysis. For full details see (Meng, 1994).

In general, inference is valid with a finite number of imputations K and, in most settings, it is sufficient to rely on a relatively small number of imputations. For example, we know that even if the fraction of missing information (Rubin, 1987, p.114) is 50%, the relative efficiency for 10 imputations is 95% versus an infinite number of imputations. However, if we want to calculate

a precise estimate of the p-values, say with Monte Carlo error below 0.005, it is necessary to use a much larger number of imputations, for example (Carpenter and Kenward, 2013, p.55) suggest using at least $K = 100$.

After this short introduction, we are now ready to describe in detail two particular types of Multiple Imputation procedures. If we have missing data in multiple variables, then we might decide to define a joint multivariate distribution from which imputing these variables or we might prefer to split the problem into multiple conditional univariate models; this is the basic difference between the two methods we are going to explore in the next two subsections. At this point we should stress that MI is a Bayesian procedure, where the missing data are drawn from the posterior distribution of the missing data given the observed. Hence we use MCMC to fit imputation models.

3.3.1 Joint Modelling

Joint modelling is a theoretically natural and, at least for the Multivariate Normal distribution, computationally convenient method for multiple imputation. The idea is to set up a joint multivariate model from which to impute the missing values. The basic assumption that we make, is that the joint distribution of the partially observed vector variable $\mathbf{Y}_i$ is multivariate normal:

$$\mathbf{Y}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Omega}) \tag{3.3.1}$$

with $\boldsymbol{\Omega}$ being the unstructured $p \times p$ covariance matrix, where p is the dimension of the $p \times 1$ vector $\boldsymbol{\beta}$.

Given this, we can use a Gibbs sampler (Geman and Geman, 1984) to fit (3.3.1) and impute the missing data. The Gibbs sampler is a Markov Chain Monte Carlo algorithm (MCMC) for obtaining a series of correlated draws from a Bayesian posterior distribution, in this case the multivariate normal one. In the Bayesian setting, missing data are then considered simply as additional parameters of this joint multivariate model.

The basic idea of the Gibbs sampler is to perform 2 steps:

1. Initialize all the parameters in our model, either randomly or using some more sophisticated algorithm (i.e. EM);
2. Update: sample each variable in turn from its conditional distribution given all the other ones;

We then update the sampler a certain number of times, by repeating step 2, until it has converged to the stationary distribution, which is the desired posterior. Subsequent updates give correlated draws from the posterior distribution. So in our case the first thing to do, before considering missing data, is to derive the conditional distributions for model (3.3.1). We give a flat prior to $\boldsymbol{\beta}$ and a Wishart distribution to the inverse of the covariance matrix, following (Schafer, 1997, Chap.5). The choice of an inverse-Wishart prior is somewhat arbitrary and surely questionable, but it leads to great simplifications in the calculation of the conditional probabilities. However, even considering the least informative choice for the degrees of freedom, with small numbers of observations the inverse-Wishart prior could still affect the results

substantially and therefore other more appropriate choices might be considered. If there are no missing data, we can express the posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\Omega}^{-1}$ given $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ as a product of a normal distribution for $\boldsymbol{\beta}$ given $\boldsymbol{\Omega}^{-1}$,$\mathbf{Y}$ and a Wishart distribution for $\boldsymbol{\Omega}^{-1}$ given $\mathbf{Y}$. The former is the conditional distribution of the fixed effect parameter vector given the current draw of the covariance matrix and the observed data, while the latter is the conditional distribution of the covariance matrix given the data only. So:

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\Omega} &\sim N_p(\overline{\mathbf{Y}}, n^{-1}\boldsymbol{\Omega}) \\ \boldsymbol{\Omega}^{-1}|\mathbf{Y} &\sim W\{n + v, (\mathbf{S}_P^{-1} + \mathbf{S})^{-1}\} \end{aligned} \tag{3.3.2}$$

where n is the number of units on which we are collecting data, p is the dimension of vector $\boldsymbol{\beta}$, v and $\mathbf{S}_P$ are the parameters for the prior Wishart we are considering, $\overline{\mathbf{Y}}$ is the sample mean and $\mathbf{S}$ is $n-1$ times the sample covariance matrix.

Repeated draws from these two distributions make it possible for the MCMC to reach the stationary distribution.

Missing Data

Now suppose that elements of $\mathbf{Y}$ are missing. In the Bayesian framework, such missing values are also parameters, so we need a prior for them too. We choose again a flat improper prior.

At this point, we can start to think about how to apply the Gibbs Sampler to our example. As we said, the first step to do is to initialize the parameters. This can be done in different ways, but here we propose two different methods:

1. Calculate the starting $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ from the observed values and draw starting values for missing $Y_{i,j}$ by hot-deck imputation (Subsection 3.2.2);
2. Use the EM algorithm to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ and then use these values to draw missing $Y_{i,j}$ with the proper conditional distributions (Subsection 3.2.3).

This second approach has some advantages, since we both start from a converged state or notice immediately if there are some problems with data, if the EM algorithm has not converged. These could easily be overlooked using the first approach as MCMC would keep on running and giving incorrect results.

Once we have drawn these initial values, which we call simply $\boldsymbol{\beta}^0$, $\boldsymbol{\Omega}^0$ and $\mathbf{Y}_M^0$, we can move to the second part of the Gibbs sampler, that is the sequential update of the different variables. For each iteration t, the algorithm is the following:

1. draw $\boldsymbol{\Omega}^t$ from the conditional distribution:

$$\boldsymbol{\Omega}^t \sim W^{-1}\{n + v, (\mathbf{S}_P^{-1} + \mathbf{S}^{t-1})^{-1}\}$$

   where $\mathbf{S}_P^{-1}$ is the second parameter of the prior Wishart distribution we considered for $\boldsymbol{\Omega}$;

2. draw $\boldsymbol{\beta}^t$ from the conditional distribution:

$$\boldsymbol{\beta}^t \sim N(\overline{\mathbf{Y}}^{t-1}, n^{-1}\boldsymbol{\Omega}^t);$$

3. draw $\mathbf{Y}_M^t$ from the conditional distribution:

$$\mathbf{Y}_M^t \sim f(\mathbf{Y}_M|\boldsymbol{\beta}^t, \boldsymbol{\Omega}^t, \mathbf{Y}_O)$$

Each unit may have a different missingness pattern and they are all independent. We therefore draw missing data for each unit in turn. First of all we have to reorder all the variables so that $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,p_i}$ will be observed and $Y_{i,p_i+1}, Y_{i,p_i+2}, \ldots, Y_{i,p}$ missing, partitioning in the same way also the elements of $\boldsymbol{\beta}$ into $\boldsymbol{\beta}_O^t$ and $\boldsymbol{\beta}_M^t$ and the covariance matrix in 4 parts as:

$$\boldsymbol{\Omega}^t = \begin{pmatrix} \boldsymbol{\Omega}_{OO} & \boldsymbol{\Omega}_{OM} \\ \boldsymbol{\Omega}_{MO} & \boldsymbol{\Omega}_{MM} \end{pmatrix}.$$

Then for each unit, we draw missing data from the appropriate p_i-variate conditional normal distribution:

$$\mathbf{Y}_{M,i}^t \sim N\left(\boldsymbol{\beta}_M^t + (\mathbf{Y}_{O,i} - \boldsymbol{\beta}_O^t)^T(\Omega_{OO}^t)^{-1}\boldsymbol{\Omega}_{OM}^t, \boldsymbol{\Omega}_{MM}^t - \boldsymbol{\Omega}_{MO}^t\left(\boldsymbol{\Omega}_{OO}^t\right)^{-1}\boldsymbol{\Omega}_{OM}^t\right);$$

4. calculate the new sample mean $\overline{\mathbf{Y}}^t$ given the current draw of missing data from the previous step;

5. Update also the sum of squares and cross products $\mathbf{S}^t$ accordingly;

6. Return to step 1.

This is the general Gibbs sampler for this model. However, the set of parameters that we draw in each step is actually correlated to the one drawn at the previous step. So, in order to have stochastically independent imputations, as required by Rubin's rules, the general procedure is:

1. Initialize all the parameters;
2. Update the sampler n_{burn} times, until it has reached the stationary distribution. Running the EM algorithm at the beginning means we need a much smaller n_{burn};
3. Use the current draw of missing data to create the first imputed dataset $\mathbf{Y}^1$;
4. Run the algorithm for another $n_{between}$ times and use the final draw of missing data to form another imputed dataset;
5. Repeat the previous step until we have recorded all the desired imputations.

This completes the description of the algorithm for Joint Modelling Multiple imputation when the joint model is multivariate normal. When the partially observed data are all continuous variables, it is often the case that the multivariate normal distribution is a reasonable assumption, maybe after some proper transformation (logarithm, square root, Box-Cox...). But what if we had other types of variables, for example binary or categorical data?

Including binary or categorical variables

Following (Goldstein *et al.*, 2009), the idea is the same in both cases, to make use of *Latent Normal Variables*. Imagine we have a N-level unordered categorical variable Y_i, we can think about introducing N independent normal variables whose maximum for each record indicates

the observed category for that record:

$$V_{i,1} = \alpha_1 + \epsilon_{i,1}$$
$$V_{i,2} = \alpha_2 + \epsilon_{i,2}$$
$$\vdots$$
$$V_{i,N} = \alpha_N + \epsilon_{i,N}$$

$$\begin{pmatrix} V_{i,1} \\ V_{i,2} \\ \vdots \\ V_{i,N} \end{pmatrix} \sim N_N \left[\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix}, \Omega_e \right]$$

so, for each i, we have:

$$m = \max_j V_{i,j} \iff Y_i = m$$

Unfortunately the above model can be proven to be non-identifiable. Therefore we need to add some constraints to the model. First of all we fix the variances of the latent normals to some arbitrary value, for example 0.5:

$$\Omega_e = \begin{pmatrix} 0.5 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0.5 \end{pmatrix}$$

Then, we use this derivation:

$$Z_{i,1} = V_{i,1} - V_{i,N}$$
$$Z_{i,2} = V_{i,2} - V_{i,N}$$
$$\vdots$$
$$Z_{i,N-1} = V_{i,N-1} - V_{i,N}$$

Where we subtracted from the first $N-1$ equations the last one. Thus:

$$Z_{i,1} = \alpha_1 - \alpha_N + e_{i,1}$$
$$Z_{i,2} = \alpha_2 - \alpha_N + e_{i,2}$$
$$\vdots$$
$$Z_{i,N-1} = \alpha_N - \alpha_N + e_{i,N},$$

So that:

$$\begin{pmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,N-1} \end{pmatrix} \sim N_{N-1} \left[\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{N-1} \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.5 \\ 0.5 & \dots & 0.5 & 1 \end{pmatrix} \right] = N_{N-1}(\boldsymbol{\beta}, \boldsymbol{\Omega}).$$

Where $\beta_i = \alpha_i - \alpha_N$.

(Goldstein *et al.*, 2009) suggests, for simplicity, to set the correlation terms to 0. This is also what has been implemented in the REALCOM software (Carpenter *et al.*, 2011) that we will introduce later. In our work we will start by using this simplification in Chapter 6, but we will put back the covariance terms to 0.5 in the remaining chapters.

In our model, we have latent normals Z, but in the data we actually observed only the categories Y. So, at each iteration of our Gibbs sampling algorithm, we have to add a first step for drawing the $Z_{i,j}$, with $j = 1, \ldots, N-1$. Once we have drawn these, the Gibbs sampler proceeds as before, apart from the update of the covariance matrix (see below). We can draw the $Z_{i,j}$ using a rejection sampling approach, so for each unit i at step t of the Gibbs sampling algorithm:

1. Draw $\mathbf{Z}_i$:

$$\mathbf{Z}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Omega}).$$

2. if $Y_i = m \neq N$ then we keep drawing until $Z_{i,m} = \max_{j=1,\ldots,N-1} Z_{i,j}$ and $Z_{i,m} > 0$;

3. if $Y_i = N$ then we keep drawing until $Z_{i,j} < 0 \ \forall j = 1, \ldots, N-1$;

In presence of missing data, we draw from the appropriate conditional $N-1$-variate normal distribution given the other variables and we accept each draw, successively building the imputed dataset transforming the $N-1$ continuous variables into one categorical variable using the same rules as before.

Updating the covariance matrix

Unfortunately, because of the constraints on the covariance matrix from the latent normal model, we can no longer update the covariance matrix via draws from the appropriate inverse Wishart distribution. A good solution, though potentially more time consuming, is to perform this update element-wise using a Metropolis Hastings algorithm, as proposed by Browne (2006):

1. for each non constrained element $\Omega_{k,l}$:
   - if k=l: draw proposals for $\Omega^t_{k,l}$ from $N\left(\Omega_{l,l}, \Omega^2_{l,l}\frac{11.6}{n}\right)$, where the value for the standard deviation has been proposed in (Gelman *et al.*, 2014);
   - if $k \neq l$: draw proposals from $N\left(\Omega_{k,l}, 0.1^2\Omega_{k,k}\Omega_{l,l}\right)$;
2. if with the proposed $\Omega^t_{k,l}$, $\boldsymbol{\Omega}^t$ is still positive definite, go to step 3. Otherwise go back to step 1;
3. accept $\Omega^t_{k,l}$ with probability:
$$\min\left(1, \frac{L(\boldsymbol{\beta}, \boldsymbol{\Omega}^t)}{L(\boldsymbol{\beta}, \boldsymbol{\Omega}^{t-1})}\right)$$
where L is the N-variate normal likelihood function;
4. if the proposed $\Omega^t_{k,l}$ is rejected, then keep $\Omega^t_{k,l} = \Omega^{t-1}_{k,l}$.

Having completed this step, we have described all the Gibbs sampling steps for a JM-MI approach to impute both continuous and categorical missing data.

3.3.2 Full Conditional Specification

Full Conditional Specification, also known as Multiple Imputation using Chained Equations (MICE), is another way of imputing multivariate data with more flexibility than JM — at least in the cross-sectional setting — at the cost of an higher difficulty in establishing the algorithm's properties.

The FCS algorithm consists of imputing data variable by variable, conditional on all the other variables in the dataset. Assuming data are MAR, so we can omit $\mathbf{R}$, we define the density $f(\mathbf{Y}|\boldsymbol{\theta})$ by specifying a conditional density $f(\mathbf{Y}_j|\mathbf{Y}_{-j}, \boldsymbol{\psi}_j)$ for each variable. Here, $\mathbf{Y}_{-j}$ represents the set of responses excluding only $\mathbf{Y}_j$.

Thus the FCS algorithm is as follows:

1. Re-order the variables so that the missingness pattern is as close to monotone as possible, i.e. such that if Y_j is missing, than Y_{j+1} is missing as well;
2. Initialize the missing values in each variable by sampling with replacement from observed values of that variable;
3. for $j = 1, \ldots, p$:
   - regress the observed part of $\mathbf{Y}_j$ on all the other variables, obtaining the current set of parameters estimates $\boldsymbol{\psi}_j$;
   - draw a new value for $\boldsymbol{\psi}_j^{new}$ from the conditional distribution:

$$\boldsymbol{\psi}_j^{new} \sim P(\boldsymbol{\psi}_j|Y_O, Y_M)$$

   - Impute missing values for $\mathbf{Y}_j$ from the following conditional distribution:

$$\mathbf{Y}_j^M \sim P(\mathbf{Y}_j|\boldsymbol{\psi}_j^{new}).$$

4. Repeat step 3, which is called a *cycle*, a certain number of times n_{burn} and at the end of these n_{burn} cycles, register the current values of all the missing data to form the first imputed dataset. We choose this n_{burn} in order to guarantee that after this number of cycles the algorithm has converged to the stationary distribution;

5. Run another $n_{between}$ cycles before registering the new imputations, in order to obtain stochastically independent imputations.

If all the variables are continuous we can just use a linear regression model for each variable, given that the residuals can be assumed to be distributed normally. Otherwise for binary variables we can use logistic regression, for categorical variables multinomial logistic regression, for count data Poisson regression and so on.

3.4 JM vs FCS

We introduced two methods for multivariate MI, but do these methods really differ? And if they do, how much do they differ? Which one is preferable?

The only conceptual difference between JM and FCS is that, while with JM we specify a joint model for the partially observed data explicitly, with FCS we define the conditional univariate models and therefore the specification of the underlying joint model is only implicit. When the joint imputation model we choose is congenial with the analysis model in the sense introduced in (Meng, 1994), JM imputation is the theoretically best way of imputing the data and it is

guaranteed not to introduce any bias. When FCS is used instead, since we are defining the joint model only implicitly, finding conditions for the validity of the method is a more subtle issue.

However, recently, conditions for equivalence of FCS to JM have been explored independently from different perspectives in two studies (Hughes *et al.*, 2014; Liu *et al.*, 2013) which essentially concluded that:

- the two methods are asymptotically equivalent;
- the two methods are equivalent when used in the multivariate normal framework;
- in finite samples the two methods are equivalent when the so-called 'non-informative margins' condition holds, otherwise an order effect occurs;
- in real data the magnitude of this order effect is usually small enough to be considered negligible.

The second point actually stems from the third one. In this regard, consider the joint model (3.3.1) and imagine we want to calculate the conditional distribution for variable j, so that:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = p(Y_j|\mathbf{Y}_{-j}, \boldsymbol{\theta})p(\mathbf{Y}_{-j}|\boldsymbol{\theta})$$

If we call $\boldsymbol{\Psi}_j$ a function of the parameter $\boldsymbol{\theta}$ from which the distribution of Y_j given $\mathbf{Y}_{-j}$ depends exclusively, then if there is no information about $\boldsymbol{\Psi}_j$ in the marginal distribution of $\mathbf{Y}_{-j}$, we say that the 'non-informative margins' condition holds. Finally, regarding the last point, (Hughes *et al.*, 2014) ran some simulations in situations where the previously defined condition did not

hold, but they found that the magnitude of the order effects was really small and not inducing any practically relevant bias. Though the theoretical problem still persists, this was reassuring on the possible magnitude of this effect in real data analysis.

Therefore the two methods seems to be interchangeable in most situations. However, FCS has become more popular than JM in the last fifteen years and the main reasons for its success are:

- Its great flexibility to accommodate different kinds of variables, since for example a linear regression can be used for the univariate conditional model to impute continuous variables, while a logistic can be used for binary data, a Poisson for count data, etc etc.
- The availability of software packages like ICE (Stata) and MICE (R), easy to use and constantly updated and improved by the authors. These software packages exploit the in-built functions for regression, e.g. *glm* for generalized linear models in R, which makes them relatively simple to code and also relatively fast.

Consequently, with simple models and datasets, using the well known and established software for FCS imputation is usually good enough.

However, additional complications may arise, the main of which are:

- data are usually imputed under the MAR assumption, but what if the true mechanism is actually MNAR?
- What if the partially observed data we need to impute have a multilevel structure?

In JM imputation, there are two different possible strategies to perform sensitivity analysis to MNAR, depending on how we factorize the joint distribution of the partially observed data. If we consider for simplicity that we have two variables, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ and that $\mathbf{R}_1$ is the missingness indicator for $\mathbf{Y}_1$, then we could define for each unit i:

- A pattern-mixture model: this would mean considering a different joint model depending whether or not $Y_{i,1}$ is observed, or in a more general situation a different model for each missingness pattern,

$$f\left(Y_{i,1}, Y_{i,2}, R_{i,1}\right) = f\left(Y_{i,1}, Y_{i,2} | R_{i,1}\right) f\left(R_{i,1}\right);$$

- A selection model: in this case we would model the selection process given the joint distribution of the data, hence:

$$f\left(Y_{i,1}, Y_{i,2}, R_{i,1}\right) = f\left(R_{i,1} | Y_{i,1}, Y_{i,2}\right) f\left(Y_{i,1}, Y_{i,2}\right).$$

One model can be shown to imply the other in quite general situations, see e.g. (Molenberghs *et al.*, 1998), so the choice between modelling the different patterns or the selection process is up to the analyst.

These two types of models have been used several times in the past for sensitivity analysis to different assumption from MAR in the JM-MI framework. We are not aware of any attempt to apply the pattern-mixture approach with FCS; on the other hand, only at the time of writing, Finbarr Leacy and Ian White are developing a method for modelling the selection process within the FCS framework. Therefore, at present, FCS lacks a principled approach

for imputing data under MNAR; this limits its utility as a method for conducting sensitivity analyses to departures from MAR, which is actually one of the main attractions of the MI approach.

Extensions of MI to accommodate hierarchical structures will be a key issue in our IPD-MA setting, and therefore we will focus on these in the next Chapter.

3.5 Summary

In this chapter we introduced the main methods for handling missing data in clinical datasets. Multiple Imputation has been shown to be the gold standard approach thanks to its flexibility, and hence broad applicability, allowing both to deal with MAR data and to set up simple techniques for sensitivity analysis to different assumptions.

Two different imputation methods have been described, JM and FCS, with a discussion of the pros and cons of both.

Since in the JM-MI framework, we impute with a Bayesian method, MCMCs, one might wonder why we should not decide to go Bayesian all the way and not only at the imputation step. Methods for Bayesian inference with missing data date as back as the 1970s and a consistent literature has developed. However, the main limitation of Bayesian methods compared to JM-MI is that they assume a single model and therefore, similarly to likelihood based methods, it is not possible to include auxiliary variables only at the imputation stage or to use different analysis and imputation models; this would be preferable when we have missing data in multiple

variables, and therefore we have to use a joint Bayesian model for imputing even if we are interested in a univariate substantive model. One advantage of Bayesian methods is that, together with MI methods, they are the most easily extendible to the case of non-informative missingness, and therefore they are particualrly attractive for performing sensitivity analyses to MNAR assumptions (Mason *et al.*, 2012).

In the next chapter we are going to explore the particular situation of missing data in IPD-MA, justifying the use of MI as a research tool and our decision to pursue the JM approach.

4

Why Multilevel Multiple Imputation?

In this chapter we are going to discuss the issues raised by missing data in the IPD-MA setting. In particular we will argue for the use of Multilevel Multiple Imputation as the main tool to handle them.

After a short introduction to the problem in Section (4.1), in Section (4.2) we introduce the most general formulation of a multilevel analysis model; then, in Section (4.3) we explore the different issues raised by the presence of partially observed data in these settings, depending on where missing data occur. The successive sections are dedicated to two of the most commonly used techniques to handle missing data, i.e. Maximum Likelihood methods (Section (4.4)) and Multiple Imputation (Section (4.5)). Specifically we will explore again two different Multiple Imputation approaches, called Joint Modelling (Subsection (4.5.1)) and Full Conditional Specification (Subsection (4.5.2)). Finally, in Section (4.6) we conclude with a discussion, comparing the different methods and trying to sketch out in which circumstances one is preferable to the others.

4.1 Introduction

As already mentioned in Chapter 2, IPD-MA are considered a major improvement to AD meta-analyses and therefore, when possible, should be preferred. However, when analysing the data it is not possible to simply treat the combined dataset as data from a single huge study, because observations coming from the same study potentially have a strong source of correlation, and different variability. This needs to be appropriately taken into account when modelling.

One possibility is to use Multilevel Models. The term multilevel stands for the fact that data are clustered at different levels. For example, in the IPD-MA context, when we are aggregating cross-sectional studies we have observations (level 1) clustered in studies (level 2). However when we meta-analyse longitudinal studies, we have observations (level 1) nested in patients (level 2) nested in studies (level 3), and so on and so forth. Therefore the problem of dealing with missing data in IPD-MA can be seen as a particular case of the more general situation of missing data in datasets with a hierarchical structure.

In the next sections we will try to see how all the existing methods for handling missing data behave in this situation. However, first of all we need to introduce some general notation for multilevel models.

4.2 The general multilevel model

Consider a dataset where we have individuals i nested in clusters j. For example we may have students i nested in classes j, or, in a longitudinal setting, single observations i at different time points for the same patient j. In order to take into account the correlation between individuals in the same cluster, we consider, in matrix form, the following substantive model:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\epsilon}_j \tag{4.2.1}$$

where:

- $\mathbf{Y}_j$ is the $n_j \times 1$ response vector for observations in cluster j;

- $\mathbf{X}_j$ is the $n_j \times q$ design matrix for the fixed effects for observations in cluster j;
- $\boldsymbol{\beta}$ is the $q \times 1$ vector of fixed effects;
- $\mathbf{Z}_j$ is the $n_j \times r$ design matrix for the random effects for observations in group j, and it is typically, but not necessarily, a subset of $\mathbf{X}_j$;
- $\mathbf{u}_j$ is the $r \times 1$ vector of random effect coefficients in group j, for which we suppose that $\mathbf{u}_j \sim N_r(\mathbf{0}, \boldsymbol{\Omega}_u)$;
- $\boldsymbol{\epsilon}_j$ is the $n_j \times 1$ vector of errors for observations in group j, for which: $\boldsymbol{\epsilon}_j \sim N_{n_j}(\mathbf{0}, \boldsymbol{\Omega}_e)$, with, typically, $\boldsymbol{\Omega}_e = \sigma_e^2 \mathbf{I}_{n_j}$;

This is the general form of the univariate mixed-effects model for individuals nested in cluster j, as defined for example in (Laird and Ware, 1982). In certain settings it appears convenient to define the same model in a slightly different way, as constructed from a set of different levels equations:

$$\begin{aligned} \mathbf{Y}_j &= \mathbf{Z}_j \boldsymbol{\beta}_j + \mathbf{e}_j \qquad \text{Level 1} \\ \boldsymbol{\beta}_j &= \mathbf{W}_j \boldsymbol{\beta} + \mathbf{u}_j \qquad \text{Level 2} \end{aligned} \tag{4.2.2}$$

Here, $\boldsymbol{\beta}_j$ is a $r \times 1$ vector of coefficients that vary between classes j and we model it with the linear regression model at level 2, where $\mathbf{W}_j$ is another design matrix. It is straightforward to note that combining this two equations we obtain again (4.2.1), with $\mathbf{X}_j = \mathbf{Z}_j \mathbf{W}_j$.

4.3 Missing Data in Multilevel Datasets

Imagine we want to analyse a dataset with one of the two formulations of the multilevel model we presented in the previous section; some variables may be partially observed, so before proceeding with our substantive analysis we may need to handle the missing data, for example by multiply imputing them. But can we simply use the methods we presented for single level data or do we have to formulate specific methods in order to respect the multilevel structure?

Specialised literature in recent years seems to confirm that taking into account this structure is key in order to obtain unbiased results. In the case of Multiple Imputation, for example, (Carpenter and Kenward, 2013, p. 208), proved that, when using an algorithm correctly considering the hierarchical nature of data, you can obtain good point estimations and just slightly overestimated standard errors, relative to the complete data ones. Conversely, using a simple single level imputation algorithm leads to biased estimates and most of all to seriously underestimated cluster level variances and overestimated individual level variances. Similar conclusions were drawn using simulations by (van Buuren, 2011).

We are thus going to present and compare in the next sections various methods for imputing missing values in hierarchical, or multilevel, data.

Similarly to the single level case we can decide to use different methods depending on which data are missing in the model and on the missing data mechanism.

Considering model (4.2.1), missing data can occur in the response vector $\mathbf{Y}_j$, in the fixed effect covariates design matrix $\mathbf{X}_j$, in the random effects covariates design matrix $\mathbf{Z}_j$ or even in the class indicator j. However, the case of missing class indicator j has not been much investigated in the past and usually we just rely on complete records discarding all units with missing class indicator. Furthermore in the meta-analysis setting, this is equivalent to having some data and not knowing from which study these data come from, a situation that is unusual and hence not relevant to this thesis.

When missing data are confined to the response vector $\mathbf{Y}_j$, restricting to complete records analysis gives unbiased results, if data are MAR and all the variables included in the missing data mechanism are included as covariates in the model. However, as we saw for the single level case, complete records analysis give biased results when we have missing covariate data and the missingness mechanism depends on the outcome, Y_j. Furthermore, due to the large amount of data we waste, complete records analysis is inefficient, even if inferentially valid. In general, a (possibly) slightly biased but much more precise analysis under MAR is preferable.

With missing data in $\mathbf{X}_j$ or in $\mathbf{Z}_j$ it is therefore preferable to go beyond complete records analysis. There are two main categories of methods available for this situation: maximum likelihood methods and Multiple Imputation.

We present here details of these two approaches, considering models with continuous and/or categorical variables, and with missing outcomes and/or predictors, to see how the methods compare.

Specifically we are going to present two ways of obtaining the maximum likelihood estimates, and associated standard errors, for the parameter of our substantive multilevel model, when some data are missing. These are the EM algorithm and Full Information Maximum Likelihood (FIML). We will then present two approaches to multilevel Multiple Imputation; again, as in the single level case, these will be Joint Modelling and Full Conditional Specification.

4.4 Maximum Likelihood methods

The first step in Maximum Likelihood (ML) estimation is always to build the likelihood function. For model (4.2.1), this function is:

$$L_j = \prod_{i=1}^{n_j} f_i(Y_{i,j}|\mathbf{X}_{i,j}, \mathbf{Z}_j, \boldsymbol{\beta}, \mathbf{u}_j)$$

Now, imagine that patient i in cluster j has a missing value in covariate X^1, which is a binary predictor at level 1. In order to be able to use the information provided by this patient when calculating the likelihood function, we need to sum the likelihood term over all the possible values for the missing covariate. Normally, when calculating the maximum likelihood estimates for a generalized linear mixed model, we do not consider any distribution for the covariates, that are just considered as fixed values we condition on. However in this case, we will have to introduce a distribution for the unobserved covariate. In the above case of a binary variable:

$$L_{ij} = \sum_{k=0}^{1} f(Y_{i,j}|X^1_{i,j}, \mathbf{X}^{2,\ldots,q}_{i,j}, \mathbf{Z}_j, \boldsymbol{\beta}, \mathbf{u}_j) p(X^1_{i,j} = k, \gamma).$$

Where γ is the parameter defining the distribution of the covariate X^1 and is distinct from $(\boldsymbol{\beta}, \mathbf{u}_J)$. In this case, X^1 being a binary variable, the calculation is straightforward since we just need to sum up two terms. This approach generalises naturally to the case where X^1 is a categorical variables; in this situation we sum up over all the categories. This also readily extends to the case of more than one categorical predictor missing, so for example for observation i in cluster j, if we had a missing binary covariate X^1 and a missing 4-category covariate X^2, the likelihood term from this observation is:

$$L_{ij} = \sum_{k=0}^{1} \sum_{t=1}^{4} f(Y_{i,j}|X^1_{i,j}, X^2_{i,j}, \mathbf{X}^{3,\ldots,q}_{i,j}, \mathbf{Z}_j, \boldsymbol{\beta}, \mathbf{u}_j) p(X^1_{i,j} = k, \gamma_1) p(X^2_{i,j} = t, \gamma_2)$$

Things are analytically more complex if we have missing values in continuous covariates. In this case, we must integrate over the distribution of the missing covariate instead of summing. For example, considering again patient i in cluster j, with a missing normally distributed covariate $X^1_{i,j}$, its contribution to the likelihood is:

$$L_{ij} = \int_{-\infty}^{+\infty} f(Y_{i,j}|\mathbf{X}^{2,\ldots,q}_{i,j}, \mathbf{Z}_j, \boldsymbol{\beta}, \mathbf{u}_j) f(X^1, \mu, \sigma^2) dX^1$$

with μ and σ^2 parameters of the distribution of X^1. The problem is that a closed form solution for this integral is rarely available. Instead, we have to approximate the solution through numerical integration, for example using quadrature or adaptive quadrature, or to use Laplace approximations. The more continuous covariates with missing data we have, the more difficult it becomes to approximate this integral precisely. Furthermore, the integral needs to be recalculated as we search the parameter space for the maximum likelihood estimates.

Depending on the number and different kinds of partially observed covariates, finding a good joint model for the covariates and calculating the values of the parameters that maximize its likelihood becomes very challenging. Some situations can be solved with ad hoc methods, e.g. imagine that we have a model with covariate X^1 continuous and X^2 and X^3 binary, subject to missingness. In this case the contribution to the likelihood for subject (i, j) is:

$$L_{ij} = \int_{-\infty}^{+\infty} \sum_{k=0}^{1} \sum_{t=0}^{1} f(Y_{I,J}|X^1_{i,j}, X^2_{i,j}, X^3_{i,j}, \mathbf{X}^{4,\dots,q}_{I,J}, \mathbf{Z}_J, \boldsymbol{\beta}, \mathbf{u}_J) f(X^1_{i,j}, \mu, \sigma^2) p(X^2_{i,j} = k) p(X^3_{i,j} = t) dX^1 \tag{4.4.1}$$

This problem can be greatly simplified in some particular situations, for example if the substantive model is the general location model, introduced by (Olkin and Tate, 1961). Broadly, the idea behind this model is to define a contingency table according to the values of the binary variables. In this case:

	$X^2 = 0$	$X^2 = 1$	total
$X^3 = 0$	n_1	n_2	$n_1 + n_2$
$X^3 = 1$	n_3	n_4	$n_3 + n_4$
total	$n_1 + n_3$	$n_2 + n_4$	n

We then give to each cell c the corresponding probability p_c and, given that unit (i, j) belongs to cell c, the continuous variable follows a normal model with cell-specific mean and variance.

$$f(X^1_{IJ}|X^2_{IJ}, X^3_{IJ}) \sim N(\mu_c, \sigma^2_c)$$

Therefore, this decomposition of the problem makes the integral (4.4.1) much simpler to calculate; however, the general location model is rarely (if ever) our substantive model. In a more general situation solving (4.4.1) is not so straightforward. Finding a closed form solution for the integral is even more difficult with increasing number of partially observed variables and possibly missing data patterns.

Missing data could also occur in Zs, i.e. in covariates that are included in the multilevel model with a random effect, further complicating the situation.

Furthermore, once we have our likelihood, we still need to maximize it in order to get the desired estimates. This process of maximizing the Full Likelihood directly is called Full Information Maximum Likelihood (FIML), also known as direct maximum likelihood. In order to calculate the estimates for the standard errors, we need to calculate the second derivative of the log-likelihood at its maximum, and this is often numerically complex. There are some particular cases of multilevel models where a closed form solution can be obtained without relying on numerical approximations, and these are mainly linear mixed models and the LISREL model, essentially a latent normal variable model, which is very common in Structural Equation Modeling (Skrondal and Rabe-Hesketh, 2004, Chap. 6).

FIML is implemented for example in SAS PROC CALIS and in MPlus. The problem with PROC CALIS, is that it can only handle normal data. The only commercial software able to do FIML in presence of missing data in predictors in generalized linear models is MPlus, which can deal with missing data in continuous, categorical or counting variables, in responses and in

covariates at level 1, but still not at higher levels. In case of missing data in level 2 covariates, the software authors suggest putting these as responses in the model, turning the substantive model into a multivariate model.

Muthén and Brown (2001) and successively Little and Rubin (2002), proved that under the MAR mechanism FIML estimates and standard errors are unbiased. However, everything we said so far about maximum likelihood methods rely on the assumption that data are MAR. This is not always true, and if data are MNAR we need to model also the missing data mechanism.

One last issue, but practically very important, with maximum likelihood approaches is the difficulty in including auxiliary variables. These are variables that are not to be included in the substantive model, but which can be important to both recover information and improve the plausibility of the MAR assumption (Spratt *et al.*, 2010). If present, they need to be properly included in the model likelihood and integrated out, presenting additional difficulties.

As explained in Section 3.2.3 for the simple case of single level data, another way of maximizing the likelihood is using the EM algorithm; however the same issues arise, i.e. possibly slow convergence of the algorithm and difficulty to calculate the SEs, which make the EM algorithm hardly widely applicable.

All the discussion so far focussed on calculating maximum likelihood estimates for multilevel models in presence of sporadically missing data only. The additional problems, that are very common in IPD-MA, of handling systematically missing variables and allowing for heterogeneity between different studies, would even complicate things further.

4.5 Multilevel Multiple Imputation

Multiple Imputation (MI) is a promising approach for missing data with a multilevel structure. As we have seen in the previous chapter, nowadays there are broadly two main methods to impute data using multiple imputation: Joint Modelling and Full Conditional Specification. In this section we discuss how to handle the multilevel structure of data while imputing with a JM or an FCS approach.

4.5.1 Joint Modelling

It is quite natural to extend JM multiple imputation to account for a multilevel structure in the data. In this section we are going to replace the cross-sectional joint imputation model presented in Section 3.3.1 with a multilevel joint imputation model.

To begin with, consider a very simple situation: imagine we have model (4.2.1) and that all of the variables in the model are continuous. In this case, we can just form a multivariate normal model for the joint distribution of the data. As before, the imputation model does not need to be the same as the analysis model. So, building on what we did for single level JM MI, a good idea is to put the partially observed variables $\mathbf{Y}_{i,j}$ as responses in the imputation model and the fully observed variables $\mathbf{X}_{i,j}$ as covariates. The great advantage, with respect to likelihood methods, is that this is done only to impute the missing data; after imputation, we can fit on the imputed datasets the substantive model that we would have used in case we were able to observe all the intended data.

In this imputation model, $\mathbf{Y}_{i,j}$ may contain both level 1 and level 2 variables. We then use the following notation for the response vector of our imputation model:

$$\mathbf{Y}_{i,j} = \left(\mathbf{Y}_{i,j}^{(1)},\ \mathbf{Y}_{j}^{(2)}\right)^T,$$

where $\mathbf{Y}_{i,j}^{(1)}$ is the p_1-dimensional vector of level 1 responses and $\mathbf{Y}_{j}^{(2)}$ the p_2-dimensional vector of level 2 responses. Similarly, the vector of covariates for individual i nested in cluster j is:

$$\mathbf{X}_{i,j} = \left(\mathbf{X}_{i,j}^{(1)},\ \mathbf{X}_{j}^{(2)}\right)^T.$$

Using this notation, the joint multilevel imputation model for individual i in cluster j is:

$$\mathbf{Y}_{i,j}^{(1)} = (\mathbf{I}_p \otimes \mathbf{X}_{i,j}^{(1)})\boldsymbol{\beta}^{(1)} + (\mathbf{I}_p \otimes \mathbf{Z}_{i,j}^{(1)})\mathbf{u}^{(1)} + \mathbf{e}_{i,j}^{(1)}$$

$$\mathbf{Y}_{j}^{(2)} = (\mathbf{I}_q \otimes \mathbf{X}_{j}^{(2)})\boldsymbol{\beta}^{(2)} + \mathbf{u}_{j}^{(2)}$$

$$\mathbf{u}_j = \begin{pmatrix}\mathbf{u}_j^{(1)} \\ \mathbf{u}_j^{(2)}\end{pmatrix}, \qquad \mathbf{u}_j \sim N(\mathbf{0}, \boldsymbol{\Omega}_u)$$

$$\mathbf{e}_{i,j}^{(1)} \sim N(\mathbf{0}, \boldsymbol{\Omega}_e)$$

Where $\otimes$ is the symbol for the Kronecker product, i.e. if $p = 3$ and $\mathbf{X} = (X_1, X_2, X_3)$, then

$$I_p \otimes \mathbf{X} = \begin{pmatrix} X_1 & X_2 & X_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_1 & X_2 & X_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & X_1 & X_2 & X_3 \end{pmatrix}$$

The MCMC algorithm we described in Subsection 3.3.1 extends naturally to fit and impute from this model. We present it in detail in Appendix A. We have the following parameters:

- The two vectors of level 1 and level 2 fixed effects, $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$;
- The two vectors of level 1 and level 2 random effects, $\mathbf{u}_j^{(1)}$ and $\mathbf{u}_j^{(2)}$;
- Level 1 residuals, $\mathbf{e}_{i,j}^{(1)}$;
- Level 1 and level 2 covariance matrices: $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$;
- Possibly missing data, $\mathbf{Y}^M$, which can be both at level 1 or level 2 (note that these are outcomes of the imputation model, but they can both be outcome or covariates in the substantive analysis model);

Following our previous approach, we give inverse Wishart priors to the two covariance matrices and flat improper priors to all the other parameters. As we said in the previous chapter, the choice of an inverse-Wishart distribution as a prior for the covariance matrix is not always safe. This is particularly important for the level-2 covariance matrix, where it is not uncommon, particularly for IPD-MA, to have only few clusters and therefore where an inverse-Wishart prior may have a huge impact on the posterior. It has been shown indeed that using empirically sensible prior is usually a good idea. (Burke *et al.*, 2016)

At each step t we update each of these parameters in turn conditional on all the others, using a modification of the Gibbs sampler presented in A, which was initially sketched out by (Goldstein *et al.*, 2009).

If we introduce in our model some binary, or categorical, variables, the only thing we need to do is to implement once again the latent normal approach introduced for the single level case. So, the only differences in the algorithm are (i) the presence at each iteration of the MCMC

of an initial step for drawing the latent normals that substitute the categorical responses in the imputation model and (ii) the need to update the covariance matrix element-wise with a Metropolis-Hastings step within the Gibbs sampler.

Thus, JM multilevel multiple imputation is a natural generalization of the single level JM-MI, which retains all the advantages of MI as a generic approach for handling missing data.

4.5.2 Full Conditional Specification

The theoretical basis for multilevel imputation with full conditional specification is weaker than in the single level case. van Buuren (2011) was the first one to try this approach, adding a new function to his R package *mice*. Consider once again model (4.2.1), initially with just continuous variables and potentially with missing values in $\mathbf{Y}_j$,$\mathbf{X}_j$ and $\mathbf{Z}_j$. In order to take into account the multilevel structure in the data, the first thing to do before imputing is to run until convergence a Gibbs sampler for calculating the values of the parameters of the univariate multilevel models from which we will draw the missing data. So, our MCMC draws in turn:

$$\begin{aligned} \boldsymbol{\beta}^t &\sim N(\boldsymbol{\beta}|\mathbf{Y}_j, \mathbf{u}_j, \boldsymbol{\Omega}_e) \\ \mathbf{u}_j^t &\sim N(\mathbf{u}_j|\mathbf{Y}_j, \boldsymbol{\beta}, \boldsymbol{\Omega}_u, \boldsymbol{\Omega}_e) \\ \boldsymbol{\Omega}_u^t &\sim N(\boldsymbol{\Omega}_u|\mathbf{Y}_j, \mathbf{u}_j) \\ \boldsymbol{\Omega}_e^t &\sim N(\boldsymbol{\Omega}_e|\mathbf{Y}_j, \mathbf{u}_j, \boldsymbol{\beta}) \end{aligned} \tag{4.5.1}$$

and then, on convergence, we use the current draws of the parameters to draw the missing values. As for single level FCS imputation, the difference with JM imputation is that missing data are imputed using univariate models for each variable in turn, fully conditional on all the other variables.

As for JM MI, the imputation models are different from the analysis model, so the algorithm is the same whether data are missing in the covariates rather than in the outcomes.

The algorithm above appears to work well in practice, even though it is not possible to prove convergence in general cases (van Buuren, 2011). However it is only applicable when partially observed variables are all continuous.

Now, imagine we have a model with both a level 1 and a level 2 outcome. Ian White (personal communication) pointed out a theoretical difficulty of FCS in this situation. In a cluster j we observe a level 2 variable $Y_j^{(2)}$ and a level 1 variable $Y_{i,j}^{(1)}$ for each of the $i = 1, \ldots, n_i$ individuals in cluster j. Let $\mathbf{Y}_j^{(1)} = (Y_{1,j}^{(1)}, \ldots, Y_{n_i,j}^{(1)})$, and consider a model where we only have random intercepts; the joint model for $\mathbf{Y}_j^{(1)}$ and $Y_j^{(2)}$ is:

$$\begin{pmatrix} \mathbf{Y}_j^{(1)} \\ Y_j^{(2)} \end{pmatrix} \sim N \left[\begin{pmatrix} \beta_0^{(1)} \mathbf{j}_{n_i} \\ \beta_o^{(2)} \end{pmatrix}, \begin{pmatrix} \omega_1 \mathbf{J} + \sigma^2 \mathbf{I} & \omega_{1,2} \mathbf{j}_{n_i} \\ \omega_{1,2} \mathbf{j}_{n_i}^T & \omega_2 \end{pmatrix} \right], \tag{4.5.2}$$

where $\mathbf{J}$ is a $n_i \times n_i$ matrix of ones, $\mathbf{I}$ is the $n_i \times n_i$ identity matrix and $\mathbf{j}_{n_i}$ is a $n_i \times 1$ vector of ones. Since for FCS we need all the full conditional models, we have to consider the conditional distribution of $Y_j^{(2)} | \mathbf{Y}_j^{(1)}$. This distribution is normal with mean $\beta_0^{(2)} + \alpha(\mathbf{Y}_j^{(1)} - \beta_0^{(1)} \mathbf{j}_{n_i})$, with

$\alpha = \omega_{1,2}\mathbf{j}_{n_i}(\omega_1\mathbf{J} + \sigma^2\mathbf{I})^{-1}$. It is possible to prove that this inverse is of the form $(a\mathbf{J} + b\mathbf{I})$, with

$$a = \frac{-\omega_1}{[\sigma^2(n_i\omega_1 + \sigma^2)]}$$

and $b = \frac{1}{\sigma^2}$. So, since $e_{i,j} = Y_{i,j}^{(1)} - \beta_0^{(1)}$, we have the following expression of the conditional expectation for $Y_j^{(2)}$:

$$E[Y_j^{(2)}|\mathbf{Y}_j^{(1)}] = \left[\beta_0^{(2)} - n_i\omega_{1,2}(n_ia + b)\beta_0^{(1)}\right] + n_i\omega_{1,2}(n_ia + b)\bar{\mathbf{Y}}_j^{(1)}$$

So, the imputation model for the level 2 variable needs to include the mean of the level 1 responses as a covariate.

Next, consider a similar random intercepts model with two level 1 variables $\mathbf{Y}_{1,j}^{(1)}$ and $\mathbf{Y}_{2,j}^{(1)}$ and no level 2 variables:

$$\begin{pmatrix}\mathbf{Y}_{1,j}^{(1)}\\ \mathbf{Y}_{2,j}^{(1)}\end{pmatrix} \sim N\left[\begin{pmatrix}\beta_{0,1}^{(1)}\mathbf{j}_{n_i}\\ \beta_{0,2}^{(1)}\mathbf{j}_{n_i}\end{pmatrix}, \begin{pmatrix}\omega_{1,1}\mathbf{J} + \sigma_{1,1}^2\mathbf{I} & \omega_{1,2}\mathbf{J} + \sigma_{1,2}^2\mathbf{I}\\ \omega_{1,2}\mathbf{J} + \sigma_{1,2}^2\mathbf{I} & \omega_{2,2}\mathbf{J} + \sigma_{2,2}^2\mathbf{I}\end{pmatrix}\right] \tag{4.5.3}$$

The conditional expectation is:

$$E\left[Y_{2,i,j}^{(1)}|\mathbf{Y}_{1,j}^{(1)}\right] = \beta_{0,2}^{(1)} + \frac{\sigma_{1,2}^2}{\sigma_{1,1}^2}\left(Y_{1,i,j}^{(1)} - \beta_{0,1}^{(1)}\right) + \alpha(n_i)n_i\left(\bar{\mathbf{Y}}_{1,j}^{(1)} - \beta_{0,1}^{(1)}\right)$$

with

$$\alpha(n_i) = \frac{\omega_{1,2}\sigma_{1,1}^2 - \omega_{1,1}\sigma_{1,2}^2}{\sigma_{1,1}^2(\sigma_{1,1}^2 + n_i\omega_{1,1})}$$

So, in our imputation model for $Y^{(1)}_{2,i,j}$ we need to have as covariate $Y^{(1)}_{1,i,j}$, but also the mean of that variable in cluster j, $\bar{\mathbf{Y}}^{(1)}_{1,j}$ and also a non linear interaction between the mean and the cohort size n_i.

If instead of a random intercept model we used a random intercept and slopes model, we would have had to include in the model as covariates a function of observations and observation time instead of the mean of $\mathbf{Y}_{1,j}$.

The general point is that, unlike in the cross-sectional setting, we no longer have a generic, simple algorithm. Instead, the form of the conditional models is very different for different covariate structures.

White and Resche-Rigon (2013), evaluated the impact of including or not the mean as a covariate in the imputation model with simulations, finding that, at least in the case of equal sized cohorts, this wasn't affecting the results significantly. In their simulations, they also compared the functionality of Van Buuren's MICE with respect to two new programs doing one-step or two-step estimation of the parameters in the model. They found that each of the programs suffered some issues: Van Buuren's suffered from convergence problems in presence of systematically missing covariates, one-step approach could lead to low coverage and two-step approach was quite slow.

They suggested the two-step approach using method of moments instead of REML for the parameters estimates as a good compromise.

4.6 Discussion

In this chapter we have explored the main characteristics of methods for dealing with missing data in multilevel analyses. Discussion is ongoing in the scientific literature on which methods are the best to use in these settings.

As we have seen, maximum likelihood methods and multiple imputation are currently the best available options. Under identical MAR assumptions, both methods produce estimates that are consistent and asymptotically efficient.

Supporters of ML argue that these methods have some advantages: first of all in ML you have just one model, the substantive model, differently from MI where you have a substantive and an imputation model. This can certainly be a good thing, since it avoids possible problems of incompatibility between the two models. However, on the other hand, sometimes we may have some auxiliary variables explaining missingness which we do not wish to include in the analysis model; with MI, it is straightforward to just include them in the imputation model, without modifying your substantive model, while with ML this becomes much more complicated.

Another advantage of ML is that interactions in the model are not an issue; otherwise, in MI, and particularly in the multilevel modelling framework, imputation approaches allowing for interactions and non-linearities between partially observed variables in the substantive model are more difficult to develop. Recently, though, (Goldstein *et al.*, 2014) proposed a new method to solve this issue, that we will explore in more detail in Chapters 9 and 10 of this thesis.

Despite these advantages, ML has a really big issue: as we argued in Section 4.4, the exact calculation of the likelihood in presence of missing data in predictors in generalized linear mixed models is awkward, and it becomes almost impossible with moderate rates of missing data in a reasonable number of covariates of various types (continuous, binary, categorical).

On the other hand, Multiple Imputation is a very flexible tool and it does not become computationally substantially more difficult if we have a large number of missing values, whether they are in the outcome or in the predictors. This makes it an attractive approach in very general situations and specifically, what we are most interested in, in multilevel settings.

Furthermore, with MI it is relatively simple to carry out sensitivity analyses in order to check the robustness of our MAR assumption with respect to a possible MNAR mechanism. In ML, some software have been implemented for estimating parameters under the MNAR assumptions, for example PROC QLIM in Stata, but results may be extremely sensitive to distributional assumptions about the outcomes.

If we decide to use Multilevel Multiple Imputation, we still need to decide which imputation method to choose, JM or FCS. Both the multilevel MI approaches present advantages and disadvantages. We have seen, for example, how the JM approach extends naturally from the single level case and, provided we include all the partially observed variables in the joint imputation model as responses, the only thing we have to specify is the multilevel structure for each response. At the same time, though in principle finding a joint model for a mix of continuous and categorical variables may seem an important theoretical issue of this approach, the latent normal variables approach discussed in this chapter and in the previous one, makes JM a compelling approach.

Under the FCS approach, instead, a multilevel structure poses substantial problems because of the difficulty of setting up the appropriate conditional models, as the function of the variables we condition on depends on the assumed covariance structure of our model. This is an issue even before considering categorical data. For this reason FCS is only a good strategy in simple multilevel settings, e.g. two level models with exchangeability within clusters. It loses instead its attractiveness in more general multilevel settings, in presence of a mix of data types at both levels, for example in longitudinal studies with unbalanced times of measurements or in cross-classified structures, i.e. when we have a dataset with patients nested in different clusters, for example in hospitals and in neighbourhood. Conversely, in JM we only have to specify the appropriate covariance structure for the covariance matrix of our joint model, which the MCMC fitting algorithms can in principle handle without any further difficulties.

For all these reasons, JM MI is the most flexible tool to handle missing data with multilevel structure. Therefore it is the focus of the rest of this thesis. However REALCOM, the only software currently available for JM multilevel multiple imputation with the latent normal variables approach, presents some problems: having been developed in Matlab, it is particularly slow and therefore it presents some computational issues making it infeasible to use both for (i) relatively large problems and (ii) simulation studies, which are necessary in order to investigate finite sample properties.

For this reason, the first thing we decided to do in our research work was to develop a new software for JM-MI, in order to make it feasible to investigate the use of this strategy to handle missing data in IPD-MA with an extensive simulation work. We are going to prove that it is possible to develop an efficient JM-MI software in Chapter 5 and we will describe the resulting new R package, *jomo* (Quartagno and Carpenter, 2014), in Chapter 6.

Part III

Research Work

5

Developing an efficient software for Joint Modelling Multiple Imputation

Motivated by the reasons set out in the previous chapter, we decided to pursue multilevel JM-MI as a tool to handle missing data in IPD-MA. However, due to a lack of efficient software for performing multilevel imputation, we decided to write our own software for this purpose. In this chapter we will prove that the inefficiency of `REALCOM`, the only current program available at the moment for multilevel JM-MI with a mix of missing data types, is not due to an intrinsic limitation of the algorithm, but rather to the slowness of `Matlab`, the software used to develop it. We will see that the very substantial reduction in computational time with our new software shows that a joint modelling approach to missing data in IPD-MA is feasible.

Though our final aim is to implement JM-MI for the multilevel case, in this chapter we only investigate single level situations. This is for the sake of simplicity, since if we were able to prove the efficiency of JM-MI in the single-level case, the extension to multilevel imputation would be immediate.

In Section 5.1 we give an overview of existing software for Joint Modelling Multilevel Multiple Imputation, then in Section 5.2 we illustrate the results obtained from the first programs written to implement JM-MI, using an artificial dataset. Then, in Section 5.3 we show the results obtained from the analysis of the Youth Cohort Dataset, a huge dataset with more than $70,000$ observations on pupils' GCSE scores. We conclude with a brief discussion in Section 6.6.

5.1 Existing software for MI

We are aware of relatively few software packages for multilevel MI. Most MI software is single level, with very limited extensions to the ML case. For example, *mice* (van Buuren and Groothuis-Oudshoorn, 2011) and *ICE* (Stata, 2013) are respectively an R and a Stata software for FCS imputation, which has gained a lot of popularity over the last ten years. However, in the multilevel framework at present it is possible only to impute continuous variables and, as we stressed in the previous chapter, there are good reasons to prefer the JM approach in these settings. Furthermore, this extension of the FCS approach is based on an MCMC and therefore (i) the computational attraction of using 'in-built' code from other packages is lost and (ii) since we are using an MCMC anyway, it might be preferable to use simply the whole MCMC approach, i.e. JM.

There are a number of R packages for JM imputation:

- *norm* uses the multivariate normal model to impute single level datasets with continuous partially observed variables only;
- *cat* uses a log-linear model to impute categorical variables;
- *mix* makes use of the general location model to impute in the case of a mix of continuous and categorical variables;

However, all of these packages can deal with a quite limited range of situations and none of them extends fully to the multilevel setting. Schafer's package *pan* (Zhao and Schafer, 2013) allows for the imputation of hierarchical data. However, it only handles continuous data under

the multivariate normal model. Some macros for multilevel imputation have been written in *MLwiN* (Carpenter *et al.*, 2012) as a second option; however this framework lacks the flexibility for IPD-MA.

The third approach to multilevel JM imputation is the *REALCOM* software (Carpenter *et al.*, 2011). While this is very flexible, handling discrete data with the latent normal approach and allowing for missing values both at level 1 and 2, it is very slow, as *REALCOM* is programmed in MATLAB, a high level programming language.

The computational slowness of *REALCOM* is problematic for large IPD-MA datasets, where we may need several days to obtain even ten imputed datasets. However, comparison of *pan* and *REALCOM* suggests this is a specific problem of the software itself, rather than an intrinsic limitation of the algorithm. For this reason the first goal of this project was to write programs to speed up the *REALCOM* software. We have explored two strategies:

1. Writing C-routines for parts of the Matlab program that were running slowly (usually because they involve categorical variables);
2. Writing a standalone C program.

We began with the multivariate normal model and then moved to the latent normal model. An additional benefit of coding the algorithm has been to gain a much better understanding of the methodology.

5.2 Artificial Dataset

We decided to test our initial programs in the single level situation, to see whether the idea of running C routines within the Matlab program was actually viable. Initially we programmed Joint Modelling MI for the multivariate normal model. For brevity, we do not repeat the results here and instead we concentrate on the latent normal model for categorical variables. First of all we decided to test `REALCOM` and compare it with our programs, using an artificial dataset with no missing observations.

Data generating mechanism. We simulated 1000 observations on 2 variables, one continuous and one categorical with 4 categories. To draw this second variable, we used the latent normal model:

- for each of the 1000 observations we drew 4 values from a multivariate normal distribution with mean:

$$\mu = [0, -0.3, 0, 0.3]$$

and variance matrix:

$$\Sigma = \begin{bmatrix} 1 & 0.25 & 0.25 & 0.25 \\ 0.25 & 1 & 0 & 0 \\ 0.25 & 0 & 1 & 0 \\ 0.25 & 0 & 0 & 1 \end{bmatrix}$$

- We used the first value as the continuous variable $Y_{i,cont}$ for each observation i. The other three, denoted Z_i, with $i = 1, 2, 3$, were used as latent normals to generate the categorical variable as follows:

$$\text{if } Z_i < 0 \, \forall i \Rightarrow Y_{i,cat} = 4$$

$$\text{if } Z_m > 0 \text{ and } Z_m = \max_i Z_i \qquad i = \{1, 2, 3\} \text{ and } m < 4 \Rightarrow Y_{i,cat} = m$$

5.2.1 Fitting the model

We fit the following model:

$$\begin{aligned}
Y_{i,cont} &= \beta_1 + \epsilon_{i,cont} \\
Pr(Y_{i,cat} = 1) &= Pr(Z_{i,1} > Z_{i,2} \text{ and } Z_{i,1} > Z_{i,3} \text{ and } Z_{i,1} > 0) \\
Pr(Y_{i,cat} = 2) &= Pr(Z_{i,2} > Z_{i,1} \text{ and } Z_{i,2} > Z_{i,3} \text{ and } Z_{i,2} > 0) \\
Pr(Y_{i,cat} = 3) &= Pr(Z_{i,3} > Z_{i,1} \text{ and } Z_{i,3} > Z_{i,2} \text{ and } Z_{i,3} > 0) \\
Pr(Y_{i,cat} = 4) &= Pr(Z_{i,1} < 0 \text{ and } Z_{i,2} < 0 \text{ and } Z_{i,3} < 0), \qquad \text{where} \\
Z_{i,1} &= \beta_{cat,1} + \epsilon_{i,1} \\
Z_{i,2} &= \beta_{cat,2} + \epsilon_{i,2} \\
Z_{i,3} &= \beta_{cat,3} + \epsilon_{i,3} \\
\begin{pmatrix} \epsilon_{i,cont} \\ \epsilon_{i,1} \\ \epsilon_{i,2} \\ \epsilon_{i,3} \end{pmatrix} &= N_4 \left[\mathbf{0}, \Omega = \begin{pmatrix} \sigma^2_{cont} & \sigma_{cont,cat,1} & \sigma_{cont,cat,2} & \sigma_{cont,cat,3} \\ \sigma_{cont,cat,1} & 1 & 0 & 0 \\ \sigma_{cont,cat,2} & 0 & 1 & 0 \\ \sigma_{cont,cat,3} & 0 & 0 & 1 \end{pmatrix} \right]
\end{aligned} \tag{5.2.1}$$

We used `REALCOM` initially. We ran 1000 updates of the sampler, with a burn in of 100 updates, and the final results were really similar to the ones that we expected, since both the posterior mean of the fixed effect estimates and of the covariance matrix were within 1 /Even though this was a quite small dataset and we ran just 1000 updates, it took more than 5 minutes to run the sampler.

We then decided to try to run some routines in C, instead of a full Matlab program like `REALCOM`, in order to reduce the computational time. To do this, first of all we wrote Matlab code to repeat more or less the same procedure that `REALCOM` uses. This gave results really close to `REALCOM` ones, both in terms of precision of the estimates and of time elapsed (see Tables 5.1 and 5.2).

This gave a starting base, but we still had to choose which routines to implement in C. Analysing our Matlab program, we found that more or less 70% of the time for each iteration is spent in the rejection sampling step for drawing the latent normal variables for the categorical response, and 29% of the time is spent in the element-wise update of the covariance matrix. So these two routines seem to be the first candidates to implement with C code. Such C functions, called from within Matlab, are called `mex` functions.

We first wrote a program which implemented only the rejection sampling step with a C subroutine, obtaining again good estimates of the means and the covariance terms, but saving a lot of time, being three times faster than the full Matlab code.

Program	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
True value	0.000	-0.265	0.035	0.280
`REALCOM`	-0.005 (0.031)	-0.264 (0.054)	0.035 (0.052)	0.284 (0.049)
Matlab	-0.004 (0.036)	-0.270 (0.051)	0.035 (0.051)	0.281 (0.049)
Mex rej. samp.	-0.003 (0.034)	-0.266 (0.051)	0.040 (0.052)	0.281 (0.050)
Mex Update Cov.	-0.004 (0.031)	-0.274 (0.054)	0.031 (0.052)	0.279 (0.050)
Mex both routines	-0.002 (0.032)	-0.268 (0.053)	0.041 (0.052)	0.287 (0.049)
Full C code	-0.003 (0.032)	-0.286 (0.051)	0.037 (0.052)	0.288 (0.049)

Table 5.1: Posterior mean (SE) of the fixed effect estimates after running the MCMC for 1100 iterations with the different programs.

For the `mex` program which runs only the covariance matrix update with C code, the gain in speed was lower, as we expected, while finally for the code with both the mex-subroutines running at the same time we obtained a massive improvement. Finally we wrote a standalone C program repeating the same algorithm, to see how much time was lost for passing continuously data from Matlab to C and vice-versa. With this final program, we reached the same results in 16 seconds, gaining another 30% of speed from the previous program. The total speed up was 95%, clearly showing the potential of a more efficiently coded version of the JM approach for more complex, larger, datasets. Tables 5.1 and 5.2 summarize the results of this first example. Both the posterior means and covariance matrices were very similar across all the different programs, though still slightly different because of Monte-Carlo variation.

Program	Time Elapsed	% change compared to `REALCOM`
`REALCOM`	305 seconds	
Matlab	343 seconds	+12%
Mex rejection sampling	105 seconds	−66%
Mex Update Covariance Matrix	275 seconds	−10%
Mex both routines	23 seconds	−92%
Full C code	16 seconds	−95%

Table 5.2: Times elapsed for running 1100 iterations of the Gibbs sampler with the 4 different programming approaches. The percent change in time elapsed compared to `REALCOM` is shown in the right hand column.

5.3 Youth Cohort Study

The *Youth Cohort Time Series for England, Wales and Scotland*, 1984-2002, henceforth YCS, is a dataset available from the UK Data Archive, with Study Number SN 5765. This dataset merged data from different studies, principally one funded by the English government (YCS - Youth Cohort Study for England and Wales) and two from the Scottish Government (SYPS - Scottish Young People Survey and SSLS - Scottish School Leavers Survey). The SYPS collected data on cohorts of students in the 80s and the SSLS collected data on cohorts of students in the 90s. Unfortunately, some data are really poorly collected, mainly in the Scottish study and for the cohorts of students in the 80s. For this reason, for our analysis we decided to select only pupils from YCS who reached the end of their studies in 1990, 1993, 1995, 1997 and 1999. Restricting to these cohorts, we still have a really large dataset, containing more than $76,000$ observations and 88 variables describing both pupils' personal characteristics, results achieved in their studies, characteristics of their parents and of their families and so on. Our focus is an illustrative analysis of the relationships between pupils' GCSE scores and measures of their social stratification like sex, ethnicity and type of parental occupation. Investigating

these characteristics, we found a non-trivial proportion of missing data. For example, around 4% of observations on pupil's ethnicity are missing and the situation is considerably worse for the parental occupation variable, where nearly 15% of the values are missing (see Table 2.3).

To investigate the relationship between GCSE score and social stratification, we chose a simple substantive model, a linear regression where GCSE score is the response and with 4 covariates (plus the constant):

$$\begin{aligned} Y_i = \beta_0 + (\beta_{1,Int} X_{i,Int} + \beta_{1,Work} X_{i,Work}) + \\ + (\beta_{2,Bla} X_{i,Bla} + \beta_{2,Ind} X_{i,Ind} + \beta_{2,Pak} X_{i,Pak} + \beta_{2,Ban} X_{i,Ban} + \beta_{2,Ota} X_{i,Ota} + \beta_{2,Otr} X_{i,Otr}) + \\ + \beta_3 X_{i,Girl} + (\beta_{4,93} X_{i,93} + \beta_{4,95} X_{i,95} + \beta_{4,97} X_{i,97} + \beta_{4,99} X_{i,99}) + \epsilon_i \end{aligned} \tag{5.3.1}$$

Here,

1. GCSE score is calculated by giving 12 points to a A/A^* through to 1 point for the bottom grade. Pupils' scores are capped at 84.

2. Parental Occupation is a 3-category variable indicating to which working class belongs pupil's parents. Class 1 is for a managerial or professional work (reference), class 2 intermediate and class 3 mostly manual work. In (5.3.1) indicator variables $X_{i,Int}$ and $X_{i,Work}$ are used, so that the third category, professional, is the reference category, i.e. $X_{i,Int} = 0$ and $X_{i,Work}$.

3. Ethnicity is a 7-category variable indicating each pupil's ethnicity. The 7 categories are: White (reference), Black, Indian, Pakistani, Bangladeshi, other Asian and other response. Again, six indicator variables are used in model (5.3.1).

Pattern	GCSE	Ethnicity	Parental Occupation	N	% of Total
1	✓	✓	✓	66965	87%
2	×	✓	✓	760	1%
3	✓	×	✓	423	0.5%
4	✓	✓	×	7523	10%
5	×	×	✓	18	<< 1%
6	×	✓	×	332	0.3%
7	✓	×	×	651	1%
8	×	×	×	119	0.2%
Total:				76791	100%

Table 5.3: Different Missing Data Patterns for the Youth Cohort Data: '✓' observed, '×' missing.

4. Sex is a binary variable which is 1 for male and 2 for females.

5. Cohort is a 5-category variable indicating the year in which the pupil finished his/her 11th school year. As we previously said, this can be 1990 (reference), 1993, 1995, 1997 or 1999. Four indicator variables are used for this variable in (5.3.1).

Sex and cohort are fully observed, while for the other variables Table 5.3 shows the missing data patterns. We created 10 imputed datasets and then we fit the substantive model (5.3.1) to each imputed dataset in turn, finally combining the results via Rubin's rules. We repeated this procedure with 3 different software packages:

- ICE: Stata Package that implements the FCS approach;
- `REALCOM`: for a JM approach;
- Our new software that uses `mex` sub-routines to speed up `REALCOM`;

In this way we were able both to verify if FCS and JM led to similar results, and explore whether the new code could save time with respect to `REALCOM` — which we will see is really slow — making it computationally feasible to analyse these data with a JM approach.

5.3.1 Imputation model

As we highlighted in Section 3.3, in order to retain the relationships of interest for the substantive model in the imputed data, we should use an imputation model congenial with our substantive model, or even "richer" than it. So, for our final imputation model, we use all the variables included in our substantive model, including the ones with missing data as responses and the complete ones as covariates. However, to build this final imputation model, we decided to proceed step by step, starting from a simple imputation model with only two responses and adding the other variables one-by-one. This is not a typical procedure for choosing an imputation model, but we decided to proceed in this way here only to test the code and compare computational times with increasing number of variables.

Following this approach, the first imputation model took GCSE score and parental occupation as responses and only the constant as a covariate. Obviously, since parental occupation is a categorical variable, we adopted the latent normal strategy and so the joint model for the responses was a 3-variate normal, with an appropriately constrained covariance matrix:

$$
\begin{aligned}
Y_{i,1} &= \beta_1 + \epsilon_{i,1} \\
\Pr(Y_{i,2} = 1) &= \Pr(Z_{i,2,1} > Z_{i,2,2}\, and\, Z_{i,2,1} > 0) \\
\Pr(Y_{i,2} = 2) &= \Pr(Z_{i,2,2} > Z_{i,2,1}\, and\, Z_{i,2,2} > 0) \\
\Pr(Y_{i,2} = 3) &= \Pr(Z_{i,2,1} < 0\, and\, Z_{i,2,2} < 0) \\
Z_{i,2,1} &= \beta_{2,1} + \epsilon_{i,2,1} \\
Z_{i,2,2} &= \beta_{2,2} + \epsilon_{i,2,2} \\
\begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2,1} \\ \epsilon_{i,2,2} \end{pmatrix} &= N_3 \left[\mathbf{0}, \Omega = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2,1} & \sigma_{1,2,2} \\ \sigma_{1,2,1} & 1 & 0 \\ \sigma_{1,2,2} & 0 & 1 \end{pmatrix} \right].
\end{aligned}
\tag{5.3.2}
$$

The algorithm we followed is the same described in Section 3.3.1:

1. Choose initial value for all the parameters: we chose a vector of zeros for $\boldsymbol{\beta}$ and the identity matrix for the covariance matrix;

2. Rejection sampling and imputation step: we draw the values of the latent normals with a rejection sampling approach as explained on page 60. If one variable is missing for a certain pupil, we draw it from the proper conditional normal distribution;

3. Update covariance matrix: this is done element-wise as explained on page 60;

4. Update $\boldsymbol{\beta}$ from the conditional normal distribution given all the other variables;

5. Repeat steps 2 to 4 n times, where $n = n_{burn} + n_{between}$.

The results from Section 5.2 suggest that we will obtain a substantial computational speed up by performing steps 2 and 3 of this algorithm with a `mex` subroutine, that is a function that calls a C routine from Matlab.

Since in this first simplification of the full imputation model we are only interested in verifying if the results obtained are similar to the ones given by `REALCOM`, we did not record actual imputations, but we just run the algorithm for $n = n_{burn} + n_{between} = 100 + 1500 = 1600$ updates and we compared the posterior means of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ between `REALCOM` and our software.

The details are given in Appendix B. They show that the posterior means of the fixed effect estimates $\bar{\boldsymbol{\beta}}$ and Covariance matrix $\bar{\boldsymbol{\Sigma}}$ from both `REALCOM` and our code are within 1/

In this example the categorical variable had just 3 categories, but the more categories you have, the more difficult it becomes to draw the right values at the rejection sampling step, and so more time is demanded. For example, running the same model, but with the ethnicity variable instead of parental occupation, requires 6 latent normals. This slows the program substantially: now `REALCOM` takes almost 15 hours and our program around 1 hour.

We used the same algorithm in two further settings: the first was an imputation model with both the categorical variables as responses at the same time. The second was the desired congenial imputation model, with the same responses but with also gender and cohort as covariates. In order to do this, we had to extend the code to include covariates as described on pp.130–132 of (Carpenter and Kenward, 2013). After running the MCMC algorithm, the

Model	Responses	Covariates	REALCOM	mex	ICE
1	GCSE score Parental Occupation	Constant	472 minutes	35 minutes	19 minutes
2	GCSE score Ethnicity	Constant	863 minutes	67 minutes	36 minutes
3	GCSE score Parental Occupation Ethnicity	Constant	1247 minutes	95 minutes	65 minutes
4	GCSE score Parental Occupation Ethnicity	Constant Gender Cohort	1276 minutes	294 minutes	327 minutes
4	Writing `mex` routine for β update			127 minutes	
4	Whole C code			78 minutes	

Table 5.4: Elapsed times for running 1600 updates (JM) or cycles (FCS) of the same models.

results again agreed closely with `REALCOM`. However, again our code was much faster (Table 5.4). This time we also compared time elapsed with Stata's ICE, running 1600 cycles of the FCS algorithm.

Notice that in the last model, since the update of $\boldsymbol{\beta}$ with 3 covariates is more time consuming and involves some loops, implementing that step in C via `mex` leads to a further speed up in the programs of nearly 60% compared with Matlab.

Furthermore, writing a program that calls just one big C routine to implement the whole algorithm, in order to save the time spent in passing the arguments from Matlab to C and vice-versa resulted in further reducing the time elapsed to just 78 minutes. This is a final speed up of over 16 times, compared to the 1276 minutes necessary to run the `REALCOM` program.

It is interesting to notice that for "small" models, with only the constant, ICE seems to be faster. However, once we have "bigger" models JM overtakes FCS in time per cycle/update. We do not want to argue that our program is definitely quicker than ICE, since we are using a different algorithm and typically with ICE one would use a smaller number of cycles than we used in this example. Nonetheless we have shown that the two methods are computationally competitive for large datasets, when our software is used.

5.3.2 Application to YCS model

The final step is to compare the results of the JM imputation with those from ICE's imputation. To do this, we need to impute a certain number of datasets with both methods, say 10, and then fit the substantive model to each imputed dataset, finally applying Rubin's rules to obtain the combined results. We performed this procedure with `REALCOM` and our `mex` program running the algorithm for 500 updates between each imputation, and also with ICE, running it for 500 cycles between each imputation. We obtained the results shown in Table 5.5. We see that results obtained from `REALCOM` and our program are very similar, as expected.

If we compare these results with the ones obtained with a complete records analysis, we notice some differences, the most interesting being the fact that the coefficient for the effect of the Bangladeshi ethnicity changes direction, passing from 0.61 to -5.67. Furthermore, this effect becomes strongly significant after imputation ($p << 0.001$) while with a complete case analysis it is not significant ($p = 0.53$).

The results of MI using FCS and JM generally agree closely. The exception is for the Pakistani and Bangladeshi groups, where JM shows a stronger effect. Looking at the data more closely, more than 90% of all the pupils observed belongs to the white category, and most of the other ethnicities are not far from 1% of the total. Furthermore, if we look at the proportion of missing parental occupation values in the different ethnic groups, we see that Bangladeshi and Pakistani have really high proportions of missing data, more than 50%. Hence, perfect prediction (White *et al.*, 2010) for these two groups is more likely to occur. Indeed, *ICE* warns that it has encountered perfect prediction issues. For this reason, FCS estimates in this example are probably more accurate, since ICE uses the data augmentation solution proposed by (White *et al.*, 2010) to alleviate this problem. `REALCOM`'s results are slightly less affected by this perfect prediction problem than mex's results; this is because `REALCOM` implements its own solution to this problem by constraining all the latent normals for the categorical variables between -3 and 4:

$$-3 < Z_i < 4 \qquad \forall i,$$

a solution not implemented in our *mex* code.

Variable	CR	*ICE*	`REALCOM`	*mex*
Constant	39.35 (0.17)	38.80 (0.17)	38.76 (0.17)	38.76 (0.17)
Managerial	reference			
Intermediate	-8.49 (0.14)	-8.94 (0.14)	-8.91 (0.14)	-8.92 (0.14)
Working	-15.84 (0.16)	-16.58 (0.16)	-16.30 (0.19)	-16.34 (0.16)
White	reference			
Black	-5.77 (0.54)	-7.51 (0.50)	-7.43 (0.50)	-7.29 (0.49)
Indian	4.27 (0.39)	3.78 (0.39)	3.68 (0.37)	3.41 (0.38)
Pakistani	-2.10 (0.56)	-4.01 (0.46)	-4.56 (0.46)	-5.24 (0.45)
Bangladeshi	0.61 (1.01)	-3.87 (0.71)	-4.90 (0.81)	-5.67 (0.71)
Other Asian	6.20 (0.59)	5.34 (0.54)	5.27 (0.55)	5.02 (0.54)
Other	-1.32 (0.58)	-1.32 (0.58)	-1.14 (0.59)	-1.30 (0.58)
Girls	2.89 (0.12)	2.87 (0.12)	2.86 (0.12)	2.87 (0.12)
Cohort 1990	reference			
Cohort 1993	4.69 (0.19)	4.46 (0.18)	4.41 (0.18)	4.42 (0.18)
Cohort 1995	9.00 (0.19)	8.82 (0.19)	8.77 (0.19)	8.76 (0.19)
Cohort 1997	7.48 (0.20)	7.47 (0.19)	7.45 (0.19)	7.46 (0.19)
Cohort 1999	12.17 (0.20)	12.40 (0.19)	12.37 (0.19)	12.40 (0.20)

Table 5.5: Results of the substantive analysis of the Youth Cohort Data. We compare Complete Records (CR), ICE, `REALCOM` and our `mex` function. We create 10 imputations with both methods (JM and FCS), with 500 updates between imputations in JM and 50 cycles in FCS. In both cases $n_{burn} = 500$.

5.4 Conclusions

We have explored the potential for substantially speeding up Joint Modelling Multiple Imputation by using Matlab programs which call C subroutines. Analysing two different datasets, we found out that with respect to the older software we were able to obtain a $>$ 15-fold speed up. We also confirmed that results of our substantive analysis are similar to those obtained with an FCS approach.

Even though we would typically use more MCMC iterations than ICE cycles, we have shown that for large datasets the JM approach, with a mix of categorical and continuous data, is computationally competitive to ICE for cross sectional data. Given this, and remembering that JM is the most natural way of imputing hierarchical data, we were strongly motivated to develop a JM imputation program for Multilevel MI, in order to analyse IPD-MA data.

However before moving to the multilevel framework, we decided that it was preferable to move from Matlab to R. This was because we wanted to have the largest possible impact in the medical research world, where R is nowadays the most widely used statistical software.

The resulting package was submitted to CRAN and published under the name of *jomo* (Quartagno and Carpenter, 2014). In the next chapter, we are going to introduce the main functions available in the new package, and illustrate its use.

6

jomo: a new R package for Multilevel Joint Modelling Multiple Imputation

In Chapter 4 we argued that multilevel JM-MI is potentially the best strategy to handle missing data in IPD-MA, while in Chapter 5 we proved that the main issue related to this approach, i.e. the inefficiency of the current software available, can be overcome by developing a new software implementing some steps of (or potentially the whole) MCMC sampler with `C` subroutines.

We then decided to move the 'non-C' part of the program from `Matlab` to `R`, in order to be able to reach the widest possible audience and to make our software freely available to the research community. The resulting package was submitted to `CRAN` (R Core Team, 2014) and is called *jomo* (Quartagno and Carpenter, 2014). The aim of this chapter is to introduce this package.

In Section 5.1 we briefly discuss the issues encountered when moving from `Matlab` to `R`. We then present the functions of the new `R` package in Section 6.2, followed by two short illustrations of the software for single level (Section 6.3) and 2-level (Section 6.4) imputation and for checking the convergence of the MCMC sampler (Section 6.5). We conclude with a discussion in Section 6.6.

6.1 From Matlab to R

When starting to develop the functions for multilevel JM-MI in R, we encountered some additional issues with respect to the use of `mex` functions in `Matlab`:

- The `.Call()` function in `R`, i.e. the equivalent of `mex` in `R`, is not as easy to use; this is mainly because handling `R` type variables in `C` is only possible through the `SEXP` structures (R Development Core Team, 2015). With `SEXP` variables, the `R` data types (e.g. numeric, factor, integer, etc etc) must be redefined within the `C` code for each variable, and they are not directly inherited from the `R` objects. This could have been partially simplified by using the package `Rcpp` (Eddelbuettel and Francois, 2011), that provides an interface for seamlessly accessing, extending or modifying R objects at the C++ level; however, we decided not to go down this path because it would have involved switching from `C` to `C++`, which is not an entirely different programming language but (i) it is still different in many details, and (ii) it would have required to spend a lot of time learning this new language, losing the advantage of using `Rcpp`.

- We initially encountered severe computational problems calling from `R` broadly the same `C` subroutines that we had called from `Matlab` in the functions tested in Chapter 5; we finally found out that the reason for this, was that we had to allocate the memory for arrays created internally in `C` at the beginning of the whole code, instead of doing it locally within the main body of the functions. Modifying the programs accordingly, led to a massive improvement in the performance.

- The `CRAN` policies for submitting a new `R` package are very strict and precise; therefore a lot of work was necessary for adapting the functions to the `R` guidelines and preparing the package for submission.

After having addressed all these issues, we finally submitted the package, which we decided to call `jomo` (acronym for JOint MOdelling). In the sections below we are going to introduce the main functions available in the new package, and illustrate its use. A version of the material for the remaining of this chapter is being written up for submission to the peer-reviewed `R` Journal.

6.2 Functions

jomo provides nine main functions for imputing datasets, divided into three categories:

- `jomo1con`, `jomo1cat` and `jomo1mix`: these are used in order to impute datasets with a single level structure. While `jomo1con` is very similar to the *imp.norm* function of the *NORM* package, `jomo1cat` and `jomo1mix` are used when, in the imputation model, outcomes are respectively categorical variables only or a mix of continuous and categorical data. The strategy used is the latent normal variables approach described in Chapter 3.

- `jomo1rancon`, `jomo1rancat` and `jomo1ranmix`: these are used in presence of clustered data. Again, `jomo1rancon` is very similar to *pan* while `jomo1rancat` and `jomo1ranmix` make use of latent normal variables. In these functions, a fixed common covariance matrix is assumed across all clusters in the imputation model;

- `jomo1ranconhr`, `jomo1rancathr` and `jomo1ranmixhr`: these three functions deal with clustered data as well, but considering an imputation model where each cluster has a different covariance matrix randomly distributed about an overall covariance matrix, as proposed in (Yucel, 2011).

In addition to these functions, there are two wrapper functions, called `jomo1` and `jomo1ran`. These functions have been introduced to link together all the functions for the single level and the multilevel models respectively, without the need to specify the type of variables and the number of categories. As we will highlight later in the discussion, though, it is important to be sure that the data we are using have the right format when using these two functions. In the next sections we are going to show step by step how to use these functions in order to impute datasets.

6.3 jomo: tutorial with single level datasets

In this section we introduce the functions for imputing datasets with a single level structure. The first thing we need to do when we want to use JM imputation is always to choose a joint imputation model; this needs to be compatible with the substantive model of our primary analysis.

Once we have chosen the imputation model to use, there are three functions available for running the MCMC to impute the data; we can choose to use one of these three functions depending on the type of variables that we will have as outcomes in the joint imputation model. When all of the outcomes are continuous, `jomo1con` is the right function to use, while `jomo1cat` and `jomo1mix` are intended to deal with situations where categorical or a mix of continuous and categorical data are included as outcomes in the joint model.

Consider an example based on the dataset *sldata*, included in the package. This is an artificial dataset, with four variables: age, sex, 4-category social status and an 'outcome' measure. Missing data occur in all these variables but sex. Firstly, suppose that we do not use the social status variable, so that our joint imputation model is:

$$\begin{cases} Y_{measure,i} = \beta_{0,m} + \beta_{1,m} X_{sex,i} + \epsilon_{m,i} \\ Y_{age,i} = \beta_{0,a} + \beta_{1,a} X_{sex,i} + \epsilon_{a,i} \end{cases} \quad \begin{pmatrix} \epsilon_{m,i} \\ \epsilon_{a,i} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}) \tag{6.3.1}$$

where i indexes the observations.

Both responses in this model, measure and age, are continuous variables. Sex is also present, but it is considered as a covariate, being fully observed; therefore we will need to use `jomo1con`. The code below can be used to create 5 imputed datasets with this imputation model:

```
#Upload and attach the data
data("sldata")
attach(sldata)
#Define a data.frame with the imputation model outcomes Y
Y<-data.frame(measure,age)
#Define a data.frame with the imputation model covariate X.
#A column of 1 indicates the intercept
X<-data.frame(rep(1,300),sex)
#Define the starting values for beta
beta.start<-matrix(0,2,2)
#Define the starting value for the covariance matrix (identity matrix)
```

```
l1cov.start<-diag(1,2)
#Define the scale matrix for the inverse Wishart prior for the covariance matrix
l1cov.prior<-diag(1,2);
#Assign number of burn-in, between imputation iterations and imputations
nburn<-as.integer(200);
nbetween<-as.integer(200);
nimp<-as.integer(5);
#Create imputed datasets imp with function jomo1con
imp<-jomo1con(Y=Y,X=X,beta.start=beta.start,l1cov.start=l1cov.start,
              l1cov.prior=l1cov.prior,nburn=nburn,nbetween=nbetween,nimp=nimp)
```

As we can see, we need to pass separately a data.frame, or matrix, containing the outcomes and the covariates of the imputation model considered. Then, we can give starting values to the fixed effect estimates (`beta.start`) and to the covariance matrix (`l1cov.start`). In this example we decided to start from a matrix of zeros for `beta.start` and from a simple identity matrix for the covariance matrix. These are the default choices of the function, so specifying them is actually redundant.

Regarding the choice of the priors, a flat prior in assigned to $\boldsymbol{\beta}$ in this function, while an inverse-Wishart prior is assigned to the covariance matrix. Degrees of freedom of this distribution are left to the minimum possible, in order to represent maximum uncertainty. However, we can specify the value of the scale matrix for the prior, `l1cov.prior`. This is not mandatory and the default value is the identity matrix. `nburn` is the number of iterations we want to run before registering the first imputation; this should be chosen in order to be confident of having reached the stationary distribution at the end of `nburn` iterations of the sampler. Depending on

the situation, i.e. on the complexity of the imputation model, the cardinality of the dataset, etc etc, we may have to choose higher or lower values for this number. 500 iterations are usually enough to reach the stationary distribution, but we discuss in Section 6.5 how to check the convergence of the sampler. `nbetween` is the number of iterations to be run between two successive imputations. This is done in order to obtain (approximately) stochastically independent draws. Usually we choose a value similar to `nburn` or slightly smaller. Finally, `nimp` is the number of imputations we want to create.

After running this example, on screen we can see the posterior mean of the fixed effect estimates and the posterior covariance matrix of the imputation model.

```
The posterior mean of the fixed effects estimates is:
             [,1]        [,2]
[1,]  0.888329642 67.14421571
[2,] -0.008295014 -0.03850029
The posterior covariance matrix is:
          [,1]       [,2]
[1,] 0.7843420  0.4427945
[2,] 0.4427945 52.8341167
```

In `imp`, the program places the original and the imputed datasets in long format, i.e. stacked one below the other, with a new variable, *Imputation*, indexing the imputation number. The number 0 indicates the original, non imputed data:

```
> head(imp)
```

```
     measure age rep.1..300. sex Imputation id
1         NA  59           1   0          0  1
2  0.3832350  70           1   0          0  2
3  1.2642911  63           1   0          0  3
4  0.7668186  55           1   1          0  4
5         NA  59           1   1          0  5
6 -0.1719376  58           1   0          0  6
> head(imp[imp$Imputation==1,])
       measure age rep.1..300. sex Imputation id
301  1.9278797  59           1   0          1  1
302  0.3832350  70           1   0          1  2
303  1.2642911  63           1   0          1  3
304  0.7668186  55           1   1          1  4
305  0.8018390  59           1   1          1  5
306 -0.1719376  58           1   0          1  6
```

Now, imagine that the model we were originally interested in was a simple linear regression:

$$Y_{measure,i} = \alpha_0 + \alpha_1 X_{sex,i} + \alpha_2 X_{age,i} + \epsilon_i \tag{6.3.2}$$

Once we have created our imputed datasets, it is enough to fit our substantive model 5 times and aggregate the results using Rubin's rules, which are for example implemented in function *MI.inference* of the *BaBooN* package (Meinfelder, 2011):

```
#Initialize vectors of estimates (mean,SE) for the two parameters:
```

```
estimates<-rep(0,5)
ses<-rep(0,5)
estimates2<-rep(0,5)
ses2<-rep(0,5)
#for cycle to fit substantive model for each imputation
for (i in 1:5) {
  #select imputation i
  dat<-imp[imp$Imputation==i,]
  #fit substantive model
  fit<-lm(measure~age+sex,data=dat)
  #Register values of the estimates from imputation i
  estimates[i]<-coef(summary(fit))[2,1]
  ses[i]<-coef(summary(fit))[2,2]
  estimates2[i]<-coef(summary(fit))[3,1]
  ses2[i]<-coef(summary(fit))[3,2]
}
#Upload library BaBooN
library("BaBooN")
#Apply Rubin rules to find MI estimate of both parameters
MI.inference(estimates, ses^2)
MI.inference(estimates2, ses2^2)
```

If we have binary covariates in the imputation model, these may be included in the X matrix without any problem, as we did for sex in the previous example. If there are categorical ones, these may also be included, but dummy variables have to be created first, excluding one of

the dummies for redundancy reasons. This can be easily done for example with the R package *dummies* (Brown, 2012). Typically, the most common variables level is the preferred choice for the reference level.

If we wish to include binary or categorical variables as outcomes in the imputation model, typically because they have missing values, we can use function `jomo1mix`. For example, suppose we have formulated the following imputation model:

$$\begin{cases} Y_{measure,i} = \beta_{0,m} + \beta_{1,m} X_{sex,i} + \epsilon_{m,i} \\ Y_{age,i} = \beta_{0,a} + \beta_{1,a} X_{sex,i} + \epsilon_{a,i} \\ \Pr(Y_{soc,i} = 1) = \Pr(Y_{soc,1,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\ \Pr(Y_{soc,i} = 2) = \Pr(Y_{soc,2,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\ \Pr(Y_{soc,i} = 3) = \Pr(Y_{soc,3,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\ \Pr(Y_{soc,i} = 4) = \Pr(\max_{j=(1,2,3)} Y_{soc,j,i} < 0) \\ Y_{soc,1,i} = \beta_{0,s1} + \beta_{1,s1} X_{sex,i} + \epsilon_{s1,i} \\ Y_{soc,2,i} = \beta_{0,s2} + \beta_{1,s2} X_{sex,i} + \epsilon_{s2,i} \\ Y_{soc,3,i} = \beta_{0,s3} + \beta_{1,s3} X_{sex,i} + \epsilon_{s3,i} \end{cases} \quad \begin{pmatrix} \epsilon_{m,i} \\ \epsilon_{a,i} \\ \epsilon_{s1,i} \\ \epsilon_{s2,i} \\ \epsilon_{s3,i} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}) \tag{6.3.3}$$

We need to remember that in the underlying multivariate normal imputation model, categorical variables, i.e. `social` in this case, are modelled using latent normals, i.e. $Y_{soc,1}$, $Y_{soc,2}$ and $Y_{soc,3}$. Therefore we need to specify $n-1$ starting β values for each n-category variable and a valid starting value for the covariance matrix. For example:

```
#Define data.frame with imputation model continuous outcomes
Y.con<-data.frame(measure,age)
#Define data.frame with imputation model categorical outcome
Y.cat<-data.frame(social)
#Define number of levels for categorical outcome
Y.numcat<-c(4)
#Define data.frame with imputation model covariates
X<-data.frame(rep(1,300),sex)
#Initialize fixed effect estimates and covariance matrix
beta.start<-matrix(0,2,5)
l1cov.start<-diag(1,5)
#Define scale matrix of inverse Wishart prior for covariance matrix
l1cov.prior<-diag(0,5);
#Assign number of burn-in, between imputation iterations and imputations
nburn<-as.integer(500);
nbetween<-as.integer(200);
nimp<-as.integer(5);
#Impute with function jomo1mix
imp<-jomo1mix(Y.con,Y.cat,Y.numcat,X,beta.start,l1cov.start,l1cov.prior,nburn,nbetween,nimp)
```

The output for this call is identical to the one obtained with *jomo1con* and therefore is not shown here. In the covariance matrix update step, a further complication in the method arises. As outlined in Chapter 3, for identifiability reasons, variances of the latent normals must be kept fixed to 1, while covariances of latent normals related to the same variable should be fixed to 0.5. In the *mex* functions presented in the previous chapter, we fixed these covariances to 0 as in *REALCOM*, but in *jomo* we avoid this approximation and we keep the 0.5 values.

Because of these constraints, we can no longer draw the new value for the covariance matrix at each step from an inverse Wishart distribution. The way we overcome this issue is via an element-wise update of the matrix through a Metropolis-Hastings step, as explained in Chapter 3. The same holds for function `jomo1cat`.

6.3.1 A wrapper function: jomo1

`jomo1` is the name of a function which can be used when we do not want to specify if a variable is continuous or nominal. This function links all the three previous ones, treating automatically each numeric variable as continuous and each factor as categorical, both if it is ordered or unordered. An example of the simplest possible call to this function is:

```
Y<-data.frame(measure,age)
imp<-jomo1(Y)
```

After running this code, a message is printed on screen, telling us that `jomo1con` was the function used, since both measure and age were found to be numeric variables. However, by running:

```
Y<-data.frame(measure,factor(social))
imp<-jomo1(Y)
```

we tell the program that `social` is a factor variable, and so we use function `jomo1mix`.

6.4 jomo: tutorial with multilevel structures

Suppose we want to impute a multilevel dataset,where individuals are grouped in clusters. The primary example we have in mind is patients nested in studies in IPD-MA. In this setting, we can choose between two strategies: an imputation model with a fixed common covariance matrix across clusters or with different, cluster-specific matrices, randomly distributed about an overall covariance matrix.

When choosing to adopt the first strategy, as in the single level case we can use function `jomo1rancon` if the imputation model has continuous outcomes only and `jomo1rancat` or `jomo1ranmix` otherwise.

Consider the dataset *mldata*, included in the package. It consists of simulated data from individuals clustered in different cities. Therefore, here cities play the same role of studies in an IPD-MA. Apart from this, the variable names and 'definitions' in this dataset are exactly the same as in *sldata* and missingness occurs again in all variables but `sex`.

Now, suppose we set up the following joint imputation model:

$$
\begin{cases}
Y_{measure,i,j} = \beta_{0,m} + \beta_{1,m} X_{sex,i,j} + u_{m,j} + \epsilon_{m,i,j} \\
Y_{age,i,j} = \beta_{0,a} + \beta_{1,a} X_{sex,i,j} + u_{a,j} + \epsilon_{a,i,j} \\
\Pr(Y_{soc,i} = 1) = \Pr(Y_{soc,1,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 2) = \Pr(Y_{soc,2,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 3) = \Pr(Y_{soc,3,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 4) = \Pr(\max_{j=(1,2,3)} Y_{soc,j,i} < 0) \\
Y_{soc,1,i,j} = \beta_{0,s1} + \beta_{1,s1} X_{sex,i,j} + u_{s1,j} + \epsilon_{s1,i,j} \\
Y_{soc,2,i,j} = \beta_{0,s2} + \beta_{1,s2} X_{sex,i,j} + u_{s2,j} + \epsilon_{s2,i,j} \\
Y_{soc,3,i,j} = \beta_{0,s3} + \beta_{1,s3} X_{sex,i,j} + u_{s3,j} + \epsilon_{s3,i,j}
\end{cases}
\tag{6.4.1}
$$

$$
\boldsymbol{\epsilon}_{i,j} = \begin{pmatrix} \epsilon_{m,i,j} \\ \epsilon_{a,i,j} \\ \epsilon_{s1,i,j} \\ \epsilon_{s2,i,j} \\ \epsilon_{s3,i,j} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_e) \qquad \mathbf{u}_j = \begin{pmatrix} u_{m,j} \\ u_{a,j} \\ u_{s1,j} \\ u_{s2,j} \\ u_{s3,j} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_u)
$$

In this imputation model, we have indexed with i the single observations and with j the clusters; as we can see, we have new parameters with respect to imputation model (6.3.3), specifically $\mathbf{u}_j$, the vectors of random effects, and $\boldsymbol{\Omega}_u$, the level two covariance matrix.

Here we show how to use function `jomo1ranmix` to perform MCMC imputation under the above model:

```
#Define data.frames with imputation model outcomes
Y.con<-data.frame(measure,age)
Y.cat<-data.frame(social)
#Define number of levels for categorical outcome
Y.numcat<-c(4)
#define data.frame with imputation model covariates
X<-data.frame(rep(1,1000),sex)
#Define data.frame for imputation model covariates associated with random effects
Z<-data.frame(rep(1,1000))
#Define data.frame with clustering indicator
clus<-data.frame(city)
#Initialize fixed effect estimates
beta.start<-matrix(0,2,5)
#Initialize random effect estimates
u.start<-matrix(0,10,5)
#Initialize level 1 covariance matrix
l1cov.start<-diag(1,5)
#Initialize level 2 covariance matrix
l2cov.start<-diag(1,5)
#Define scale matrix for inverse Wishart prior for level 1 covariance matrix
l1cov.prior<-diag(0,5);
#Define scale matrix for inverse Wishart prior for level 2 covariance matrix
l2cov.prior<-diag(1,5);
#Assign number of burn-in, between imputation iterations and imputations.
nburn<-as.integer(500);
nbetween<-as.integer(200);
```

```
nimp<-as.integer(5);
#Impute with function jomo1ranmix
imp<-jomo1ranmix(Y.con, Y.cat, Y.numcat, X,Z,clus,beta.start,u.start,l1cov.start, l2cov.start,
                  l2cov.prior,nburn,nbetween,nimp)
```

We need to initialize in this function our two new parameters, the matrix whose rows are the random effects vectors (`u.start`) and the level 2 covariance matrix (*l2cov.start*). Here we decided to start with default values, a vector of zeros for `u.start` and an identity matrix for l2cov.start, so specifying the default values explicitly was in fact redundant. For the level 2 covariance matrix we consider an inverse-Wishart prior, once again with the minimum possible degrees of freedom and with scale matrix `l2cov.prior`. For the level 1 covariance matrix, following (Carpenter *et al.*, 2011), we consider a flat prior instead.

In general there are some other possible new inputs to pass to function `jomo1mix`, e.g. Z, the design matrix for the random effects, and *clus*, a vector indicating cluster number for each observation.

The output is very similar to the case of single level variables, but in this case even the posterior distribution for the random effects and the level 2 covariance matrices are displayed.

6.4.1 Cluster-specific covariance matrices

In some cases, we might have some concerns about the hypothesis that all of the clusters share the same level 1 covariance matrix; for example, in the case of IPD-MA, it is often reasonable to assume that covariance matrices are different among the studies. This is an example of a multilevel dataset that we might want to impute using functions `jomo1ranconhr` and `jomo1ranmixhr`.

Consider the same dataset as in the section above, but this time imagine that it is more realistic to think that within each city (cluster), the covariance matrix is different. Algebraically:

$$
\begin{cases}
Y_{measure,i,j} = \beta_{0,m} + \beta_{1,m} X_{sex,i,j} + u_{m,j} + \epsilon_{m,i,j} \\
Y_{age,i,j} = \beta_{0,a} + \beta_{1,a} X_{sex,i,j} + u_{a,j} + \epsilon_{a,i,j} \\
\Pr(Y_{soc,i} = 1) = \Pr(Y_{soc,1,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 2) = \Pr(Y_{soc,2,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 3) = \Pr(Y_{soc,3,i} = \max_{j=(1,2,3)} Y_{soc,j,i} > 0) \\
\Pr(Y_{soc,i} = 4) = \Pr(\max_{j=(1,2,3)} Y_{soc,j,i} < 0) \\
Y_{soc,1,i,j} = \beta_{0,s1} + \beta_{1,s1} X_{sex,i,j} + u_{s1,j} + \epsilon_{s1,i,j} \\
Y_{soc,2,i,j} = \beta_{0,s2} + \beta_{1,s2} X_{sex,i,j} + u_{s2,j} + \epsilon_{s2,i,j} \\
Y_{soc,3,i,j} = \beta_{0,s3} + \beta_{1,s3} X_{sex,i,j} + u_{s3,j} + \epsilon_{s3,i,j}
\end{cases}
\tag{6.4.2}
$$

$$
\boldsymbol{\epsilon}_{i,j} = \begin{pmatrix} \epsilon_{m,i,j} \\ \epsilon_{a,i,j} \\ \epsilon_{s1,i,j} \\ \epsilon_{s2,i,j} \\ \epsilon_{s3,i,j} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{e,j}) \qquad \mathbf{u}_j = \begin{pmatrix} u_{m,j} \\ u_{a,j} \\ u_{s1,j} \\ u_{s2,j} \\ u_{s3,j} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_u)
$$

We can then decide to treat $\boldsymbol{\Omega}_{e,j}$ as fixed or random cluster-specific covariance matrices; this means that we can either just estimate in the MCMC n_{clus} different covariance matrices for each cluster separately or we can even assume that all of these covariance matrices are actually random draws from the same distribution, which in the case of *jomo* is an inverse-Wishart distribution.

For the details on the algorithm used to fit the random covariance matrices algorithm initially proposed by (Yucel, 2011), see the appendix A.

Here we present an example of how to perform JM imputation under model (6.4.2):

```
#Define data.frames with imputation model outcomes
Y.con<-data.frame(measure,age)
Y.cat<-data.frame(social)
#Define number of levels for categorical outcome
Y.numcat<-matrix(4,1,1)
#define data.frame with imputation model covariates
X<-data.frame(rep(1,1000),sex)
#Define data.frame for imputation model covariates associated with random effects
Z<-data.frame(rep(1,1000))
#Define data.frame with clustering indicator
clus<-data.frame(city)
#Initialize fixed effect estimates
beta.start<-matrix(0,2,5)
#Initialize random effect estimates
u.start<-matrix(0,10,5)
#Initialize level 1 covariance matrices. These are stacked one below the other
l1cov.start<-matrix(diag(1,5),50,5,2)
#Initialize level 2 covariance matrix
l2cov.start<-diag(1,5)
#Define scale matrix for inverse Wishart prior for level 1 covariance matrix
l1cov.prior<-diag(0,5);
```

```
#Define scale matrix for inverse Wishart prior for level 2 covariance matrix
l2cov.prior<-diag(1,5);
#Assign number of burn-in, between imputation iterations and imputations.
nburn<-as.integer(500);
nbetween<-as.integer(200);
nimp<-as.integer(5);
#Define starting value for df of the inverse Wishart distribution
#from which the covariance matrices are drawn
a<-6
# Impute with function jomo1ranmixhr
imp<-jomo1ranmixhr(Y.con, Y.cat, Y.numcat, X,Z,clus,beta.start,u.start,l1cov.start,
                   l2cov.start,l1cov.prior,l2cov.prior, nburn,nbetween,nimp, a)
```

There are two differences in the inputs for this function with respect to the previous one. First of all, as mentioned in the comments within the code, `l1cov.start` is a matrix with all the cluster-specific starting values for the covariance matrices stacked up. The same format is used in the output. Secondly, there is a new input, `a`, which is the starting value for the degrees of freedom of the inverse Wishart distribution from which all of the covariance matrices are drawn. It is usually recommendable to use the default minimum possible value p, i.e. the number of outcomes, in order to represent maximum uncertainty.

In both functions, we have an option `meth` which we can choose to set to `Fixed` or `Random` depending on whether we want to use a fixed or random cluster-specific covariance matrices approach.

6.4.2 A wrapper function: jomo1ran

In the same way that `jomo1` links the single level functions, `jomo1ran` links all the functions for multilevel imputation. Numeric variables are automatically treated as continuous and factor variables as nominal, while it is possible to choose between the functions for a common covariance matrix across clusters or study-specific covariance matrices by setting the `meth` option to either `common`, `fixed` or `random`.

6.5 Checking convergence of MCMC

When running an MCMC sampler, it is important to check the sampler has reached the stationary distribution before starting to register imputations. In order to do this, for each function in the package we have also provided a '.MCMCchain' version.

The way it works is illustrated by the following simple example:

```
Y<-data.frame(measure,age)
X<-data.frame(rep(1,300),sex)
nburn<-500

imp<-jomo1con.MCMCchain(Y,X,nburn=nburn)
```

In this case we are just running the sampler for a certain number of iterations, `nburn`, and therefore we do not need to specify the number of imputations. Secondly, the output from this function is no longer just a `data.frame` with the imputed datasets, but it is actually a list containing three elements:

- `finimp`: the final state of the dataset, which would be the first imputation if we ran the `jomo1con` function with `nburn` burn-in iterations of the MCMC sampler;
- `collectbeta`: a 3-dimensional array containing the fixed effect draws at each of the `nburn` iterations;
- `collectomega`: a 3-dimensional array containing the level 1 covariance matrix draws at each of the `nburn` iterations;

When running the functions for multilevel imputation we also have:

- *collectu*: a 3-dimensional array containing the random effects draws at each of the `nburn` iterations;
- *collectcovu*: a 3-dimensional array containing the level 2 covariance matrix draws at each of the `nburn` iterations;

We can then check the convergence of the sampler by inspection of the line graph of the evolution of each parameter value and perform more formal tests if desired. For example, Figure 6.1 shows the plot for the first element of the fixed effects vector, which we obtain by running:

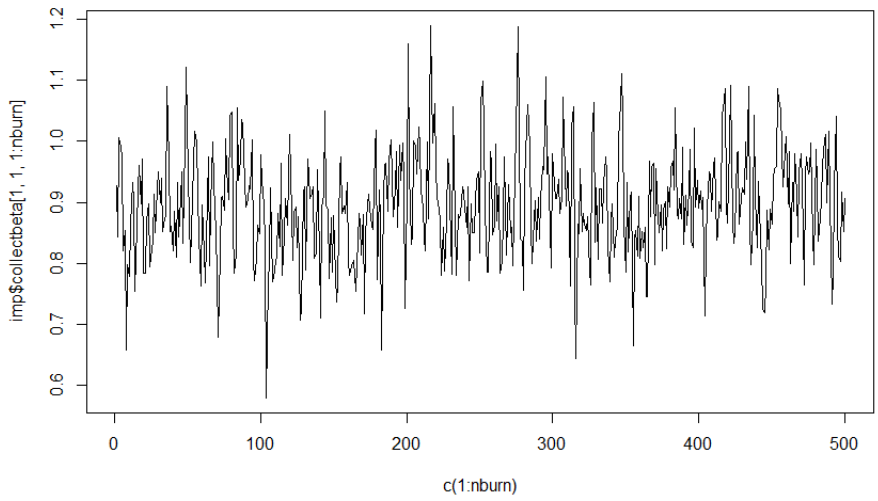

Figure 6.1: MCMC chain for one of the fixed effect parameters

```
plot(c(1:nburn),imp$collectbeta[1,1,1:nburn],type="l")
```

In this case, we can see that, based on visual check of the plot, 500 burn-in iterations appear to be quite a conservative choice, since the stationary distribution seems to have been reached much sooner.

Plots for elements of the covariance matrix updated though Metropolis-Hastings steps look quite different, because they may not be updated at each iteration. See for example in Figure 6.2 the plot obtained by running the following commands:

```
Y.con<-data.frame(measure,age)
Y.cat<-data.frame(social)
Y.numcat<-matrix(4,1,1)
```

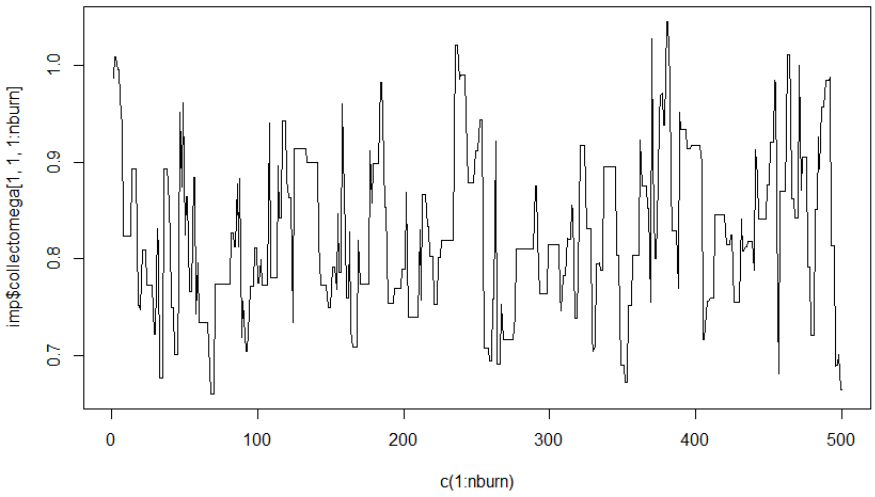

Figure 6.2: MCMC chain for one element of the level 1 covariance matrix

```
X<-data.frame(rep(1,300),sex)
nburn<-as.integer(500);

imp<-jomo1mix.MCMCchain(Y.con,Y.cat,Y.numcat,X,nburn=nburn)

plot(c(1:nburn),imp$collectomega[1,1,1:nburn],type="l")
```

Therefore, we may need to run a greater number of iterations in the burn-in phase for such models, compared to models with only continuous variables. When plotting the elements of the covariance matrix related to categorical variables, we get a straight line, since these elements are fixed to 1 (variance terms) or 0.5 (covariance terms).

6.6 Conclusions and further developments

We have written and described a new package for performing JM Multiple Imputation of clustered data, allowing for the introduction of categorical or binary variables in the imputation model as outcomes. A significant, and practically important, development is the addition of code allowing for fixed or random cluster-specific covariance matrices. This package is an important advance because (i) its computational speed enables us – for the first time – to comprehensively evaluate the performance of the JM-MI approach using simulations, and (ii) the flexibility in approaches to the cluster-specific covariance matrices will be very important in our application to IPD-MA. However, as with all statistical techniques, multiple imputation may be a great tool but it has to be used carefully. There are some important points we need to stress:

- When choosing the imputation model to use, we need to be sure that such a model is actually compatible with the substantive model of interest;
- Multiple Imputation assumes data are MAR. If data are MNAR, results of our analysis after naively imputing with *jomo* may be biased (although a small bias and substantial gain in efficiency may still be better than a complete records analysis). A proper sensitivity analysis to different assumptions to MAR is therefore recommended;
- We need to be sure that stationary distribution has been reached when we start registering imputations. This may be checked, as explained in the previous section, through a careful examination of the MCMC chains;

- Before setting the program to create the imputations, we may want to do something similar to Stata's ICE dryrun, i.e. a check to see if the model we are running is exactly the one we intend and if we are passing the correct inputs; in practice we recommend running initially the '.MCMCchain' function with $n_{burn} = 2$ and looking at the output to see whether we are running the intended model;

- As already mentioned, when using categorical variables as covariates in the imputation model, it is necessary to create dummy variables first, excluding the dummy indicator for one of the categories.

- When running certain models, functions could run quite slowly, depending mainly on these factors:

  1. Number of observations is particularly large, say $n_{obs} > 100000$;
  2. Number of variables (especially categorical) is big, say $n_{var} > 15$;
  3. Some of the categorical variables have many possible levels, say $n_{cat} > 5$;

- When using functions for random cluster-specific covariance matrices, one needs to remember that the assumption behind this method is that the covariance matrices are drawn from an inverse-Wishart distribution. Obviously this is not the only choice, but it is a natural starting point, due to the simplicity of the calculations because it is the conjugate distribution for the covariance matrix of multivariate normally distributed data.

We tested the functions written in this package extensively, and we are going to present the results of a large simulation work to investigate their use to handle missing data in IPD-MA in the next two chapters. Chapter 7 will focus on functions for imputing continuous data only, while Chapter 8 will extend to different missing data types.

One of the main advantages of our method is that, in principle, it can be used with any number of variables and with missing data in any possible pattern. However, when interactions or non-linear terms are present in the model of interest, not considering this in the imputation model may lead to bias. We will return to the issue of how to address this in Chapter 9 and 10.

In some cases, when analysing multilevel datasets, data might be missing in level 2 variables, e.g. when we have patients nested within the same surgeon and some data on the participation of surgeons to a certain training is missing. In the future we aim to introduce functions for the imputation of such variables as well.

Another possible future extension could be the inclusion of functions for multilevel imputation with random study-specific covariance matrices coming from other distributions other than the inverse-Wishart distribution.

We are continuing to maintain and update the package, which is proving popular in the research community, with at least two research papers using `jomo` that have been submitted to international scientific journals, and several other researchers that have have started to use it, though the package has been submitted to CRAN less than a year ago.

7

Multiple Imputation for IPD-MA: allowing for heterogeneity and studies with missing covariates

In this chapter we present the first part of our investigation of the use of the proposed method, multilevel JM-MI, to handle missing data in IPD-MA; we will focus here on the case of multivariate normally distributed partially observed data. Extensions to different kind of data and joint models will be discussed in Chapter 8. The research presented in this chapter forms the core of (Quartagno and Carpenter, 2015), which is in press with Statistics in Medicine.

The chapter is organized as follows. Section 7.2 presents the meta-analysis and imputation models used in the following analyses. In Section 7.3 we present the results of simulations to evaluate our proposed approach. We start with the same scenarios considered by Burgess et al., and then consider some more challenging settings. We conclude with a discussion in Section 7.4.

7.1 Introduction

Non-trivial proportions of missing covariate data are common in clinical IPD meta-analyses, and the issues they raise need to be appropriately addressed. Such missing data can take two forms: (i) missing data within studies (which we term sporadically missing data) and (ii) missing data because a particular variable was not collected in a particular study. The latter situation frequently arises, as IPD meta-analyses datasets are typically assembled post-hoc, and the individual studies have often followed different protocols and collected different variables.

With sporadically missing data, one approach to the analysis is to restrict attention to the subset of individuals with complete records. When the studies have a large number of individuals (so the resulting loss of power is acceptable), and, crucially, the chance of data being missing can be reasonably assumed not to depend on the outcome (i.e. dependent variable) in our substantive model given the covariates, so that bias is not a concern (Carpenter and Kenward, 2013, p. 28), this may be satisfactory.

However, the above assumption is typically implausible. Thus, there are many settings where the complete records analysis is insufficient. In these cases, we have argued in Chapter 4 that Multilevel Multiple Imputation is a natural approach to consider.

The specification of an appropriate imputation model (and hence appropriate predictive distribution for the missing data) raises some issues in the context of meta-analysis. First, should a separate imputation model be fitted to each study, or should some attempt be made to impute jointly across all the studies? Second, at what point in the procedure should Rubin's MI combination rules be applied? We can either:

1. apply Rubin's rules to the imputed data for each study, resulting in a single summary from each study which is then meta-analyzed. We term this Rubin's Rules then Meta-Analysis (RR then MA). *Or*

2. perform a meta-analysis of each imputed dataset, and then summarize the results using Rubin's Rules (ie MA then RR).

(Burgess *et al.*, 2013) considered the imputation of sporadically missing data, concluding that in the majority of cases the best approach is to impute separately in each study. This is because in many meta-analyses there is important between-study heterogeneity, which needs to be respected in the imputation process. If imputation is performed separately for each study contributing to the meta-analysis, then their results showed that it is best to apply Rubin's rules before meta-analysing the results. This also implicitly calls for a two-stage analysis of the IPD data, rather than a one-stage analysis. The reason for this is that a one-stage analysis has the advantage of allowing us to borrow strength across studies for the estimation of covariate and subgroup effects.

(Burgess *et al.*, 2013) adopt the FCS imputation approach, and show promising results across a range of scenarios with increasing heterogeneity.

Their results are therefore a welcome addition to the statistician's toolbox. However, there are a number of practically important scenarios where this approach falls short (Carpenter and Kenward, 2013, Chap. 9):

1. contributing studies contain a large number of variables, but relatively few individuals, so that within-study imputation is noisy;
2. contributing studies did not collect key variables, which are therefore missing for all individuals within that study;
3. data within contributing studies are hierarchical, with potentially missing values at both levels of the hierarchy, and

4. our substantive model contains interactions and non-linear effects, and we wish to impute consistently with these.

In this chapter, we address the first three points by describing a joint model approach to imputation for IPD meta-analyses. In particular, we show how introducing an inverse Wishart distribution (or potentially any other suitable distribution) for the study-specific covariance matrices allows us to respect the between study heterogeneity, yet—in common with other random effects models—borrow strength from larger studies when estimating the covariance matrix, and then imputing, for smaller studies. The ability to do this in turn naturally allows imputation of variables that are completely missing in a study—assuming missing at random—in a way that takes into account the covariance matrix of the available variables from that study. This addresses the issues raised by (Koopman *et al.*, 2008), which showed how imposing a common covariance matrix across studies when imputing missing variables within a study leads to inconsistencies when (as it is typically the case) this assumption is inappropriate. Furthermore, while in the past implementations of this approach have been relatively computationally slow, *jomo* has effectively removed this limitation.

7.2 Meta-analysis and multiple imputation models

In order to be able to compare our strategy with the one used in (Burgess *et al.*, 2013), we decided to start by running some simulations using the same data generating mechanisms they used. Therefore, we consider a continuous outcome, y, and two continuous covariates, x_1 and x_2. Let $s = 1, \ldots, S$ index studies, and i index observations within a study.

7.2.1 Substantive analysis models

Regarding the substantive meta-analysis model used to analyse the data, we opted for a two-step meta-analysis model, considering both a fixed-effect and a random-effects model. As we will see in the next subsection, all our proposed imputation methods are actually one-step methods, so a natural question is why we decided to use two-step methods as substantive models. The reason for this is that in most of our simulations, had we opted for a one-step approach, we would have had to use a hierarchical regression model allowing for heteroscedasticity, which is not easily implemented in R. However, in principle results and conclusions drawn shouldn't be much different in case we used a one-step substantive model.

Algebraically, these were the two models used to perform the meta-analyses:

1. Two-stage fixed effect model

   (a) Within each study s, simultaneously regress $y_{i,s}$ on $x_{1,i,s}$ and $x_{2,i,s}$, obtaining estimates $(\hat{\beta}_{j,s}, \hat{s}_{j,s})$ (where $\hat{s}_{j,s}$ is the estimated standard error of $\hat{\beta}_{j,s}$), for $j = 1, 2$;

   (b) Fit the fixed effects meta-analysis model,

   $$\hat{\beta}_{j,s} = \beta_{fixed,j} + \hat{s}_{j,s}\epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0, 1), \tag{7.2.1}$$

   for $j = 1, 2$.

2. Two-stage random effects model

   (a) Within each study s, simultaneously regress $y_{i,s}$ on $x_{1,i,s}$ and $x_{2,i,s}$, obtaining estimates $(\hat{\beta}_{j,s}, \hat{s}_{j,s})$ (where $\hat{s}_{j,s}$ is the estimated standard error of $\hat{\beta}_{j,s}$), for $j = 1, 2$;

(b) Fit the random effects meta-analysis model, using the DerSimonian-Laird estimate of between study heterogeneity (DerSimonian and Laird, 1986),

$$\hat{\beta}_{j,s} = \beta_{random,j} + u_{j,s} + \hat{s}_{j,s}\epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0,1); \quad u_{j,s} \sim \mathcal{N}(0, \tau_j^2) \tag{7.2.2}$$

for $j = 1, 2$.

We thought that the best way to evaluate our proposed imputation method was to apply it in combination with the most common meta-analysis method used, and this is the main reason why we used the DerSimonian and Laird estimate of between-study heterogeneity. However, though this method is undoubtedly the most used, we have already mentioned in Chapter 2 that it has its limitations; confidence intervals obtained with this method are often too narrow and using the Hartung-Knapp adjustment (Hartung and Knapp, 2001; Cornell *et al.*, 2014) could have lead to better results; in particular, the reason for some observed under-covering in the simulations could be the fact that we did not use this adjustment and this is possibly a limitation of our study.

7.2.2 Imputation models

Motivated by the reasons set out in Chapter 4, we describe a range of possible joint imputation models, which allow progressively for greater between-study heterogeneity and for missing values in one, two or all three of the variables y, x_1, x_2. In each case we highlight the substantive model that the imputation model is congenial with (Meng, 1994), (Carpenter and Kenward, 2013, p. 46).

1. Single partially observed variable, common residual variance

   As an example, suppose x_2 is the variable with missing data. In this setting, the proposed imputation model is:

$$\begin{aligned} x_{2,i,s} &= \alpha_{0,s} + \alpha_1 x_{1,i,s} + \alpha_2 y_{i,s} + \epsilon_{i,s} \\ \alpha_{0,s} &= \alpha_0 + u_s \\ u_s &\sim \mathcal{N}(0, \sigma_u^2); \quad \epsilon_{i,s} \sim \mathcal{N}(0, \sigma_e^2). \end{aligned} \tag{7.2.3}$$

   This imputation model is congenial with a homoscedastic substantive model, but the more general substantive models (7.2.1) and (7.2.2) can also be fitted. This model is equivalent to the homoscedastic stratified imputation model used by (Burgess *et al.*, 2013).

2. Single partially observed variable, study specific residual variance

   Again, suppose x_2 is the variable with missing data. In this setting, the imputation model is:

$$\begin{aligned} x_{2,i,s} &= \alpha_{0,s} + \alpha_1 x_{1,i,s} + \alpha_2 y_{i,s} + \epsilon_{i,s} \\ \alpha_{0,s} &= \alpha_0 + u_s \\ u_s &\sim \mathcal{N}(0, \sigma_u^2); \quad \epsilon_{i,s} \sim \mathcal{N}(0, \sigma_{e,s}^2). \end{aligned} \tag{7.2.4}$$

   This imputation model is congenial with (7.2.1), and the more general (7.2.2) can also be fitted.

3. Two partially observed variables, study specific covariance matrix

   Now suppose y and x_2 have missing data. Then – as described in (Carpenter and Kenward, 2013, p. 81–85) and extended to the multilevel setting in Chapter 9 of the same

book – we can use a bivariate response model to impute missing values of y and x_2 :

$$\begin{aligned}
x_{2,i,s} &= \alpha^1_{0,s} + \alpha^1_1 x_{1,i,s} + \epsilon^1_{i,s} \\
y_{i,s} &= \alpha^2_{0,s} + \alpha^2_1 x_{1,i,s} + \epsilon^2_{i,s} \\
\alpha^1_{0,s} &= \alpha^1_0 + u^1_s \\
\alpha^2_{0,s} &= \alpha^2_0 + u^2_s \\
\begin{pmatrix} u^1_s \\ u^2_s \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_u \right) \\
\begin{pmatrix} \epsilon^1_{i,s} \\ \epsilon^2_{i,s} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_{e,s} \right)
\end{aligned} \tag{7.2.5}$$

if we are only interested in meta-analysing β_1, this imputation model is congenial with (7.2.1), and the more general (7.2.2) can also be fitted. When β_2 is also of interest, only (7.2.2) is congenial with this imputation model.

In our simulation studies below, only x_2 will be affected by missingness. Nevertheless we will see in Subsection 7.3.2 that, in some cases, modelling y as a response is useful.

4. Three partially observed variables, study specific covariance matrix

   In general, the imputation model for this setting is not congenial with (7.2.1); however it is congenial with (7.2.2):

$$
\begin{aligned}
x_{2,i,s} &= \alpha^1_{0,s} + \epsilon^1_{i,s} \\
y_{i,s} &= \alpha^2_{0,s} + \epsilon^2_{i,s} \\
x_{1,i,s} &= \alpha^3_{0,s} + \epsilon^3_{i,s} \\
\alpha^1_{0,s} &= \alpha^1_0 + u^1_s \\
\alpha^2_{0,s} &= \alpha^2_0 + u^2_s \\
\alpha^3_{0,s} &= \alpha^3_0 + u^3_s \\
\begin{pmatrix} u^1_s \\ u^2_s \\ u^3_s \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_u \right) \\
\begin{pmatrix} \epsilon^1_{i,s} \\ \epsilon^2_{i,s} \\ \epsilon^3_{i,s} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_{e,s} \right)
\end{aligned}
\tag{7.2.6}
$$

   Again, in our simulation study, only x_2 has missing values but nevertheless modelling all three variables as outcomes has advantages in general.

5. Three partially observed variables, random study specific covariance matrix

   This model is the same as (7.2.6), apart from the important extension that the 'fixed' study-specific covariance matrix is replaced by a 'random' covariance matrix, that is:

$$
\Omega_{e,s} \sim \mathcal{IW}(a, A) \tag{7.2.7}
$$

where a and A are the parameters of an Inverse-Wishart distribution. We take the identity matrix with minimum scale parameter as the inverse Wishart prior for distribution of the study specific covariance matrices throughout. In principle, other distributions could be used; we return to this point in the discussion. With this imputation model, also systematically missing predictors can be imputed.

7.2.3 Some comments

Below—in order to assist with interpreting the simulation results—we distinguish between *correctly specified*, *compatible* and *incompatible* imputation and analysis models.

In the *correctly specified* case, the data generating model is the same as the imputation model, and the imputation model is congenial with the analysis model (essentially, fixed effects for coefficients with no across-study heterogeneity, and random effects otherwise).

In the *compatible* case, the imputation model attempts to accommodate the heterogeneity present, and the imputation and the substantive model are congenial. However the data generating model does not match the imputation model.

In the *incompatible* case, either the imputation model does not allow for between study heterogeneity present in the data generation model, *or* the analysis model and imputation model are uncongenial.

With this in mind, we note that imputation model (7.2.6) allows us to share some information across studies, through the random effects distribution for $(\alpha^1_{0,s}, \alpha^2_{0,s}, \alpha^3_{0,s})$. It is the congenial imputation model for the multivariate random effects meta-analysis of $(\beta_0, \beta_1, \beta_2)$. Models (7.2.5)–(7.2.3) make increasing restrictions on the between study heterogeneity, as well as the number of variables with missing data.

If we modify (7.2.6) to have a separate fixed effects vector $(\alpha^1_{0,s}, \alpha^2_{0,s}, \alpha^3_{0,s})$ for each study, it is equivalent to imputing separately in each study using the multivariate normal distribution. For multivariate normal data this is known to be equivalent to FCS (Carpenter and Kenward, 2013, p87–88) and practically equivalent in other settings (Hughes *et al.*, 2014). Therefore, within-study joint-model imputation should give very similar results to those reported by (Burgess *et al.*, 2013).

Importantly, in the light of our original motivation, none of (7.2.3)–(7.2.6) share information on the covariance matrix across studies. Thus, none can impute a variable that is wholly missing in a study, because for such studies the study-specific covariance matrix is not estimable. To overcome this, we need to introduce a random covariance matrix effect, i.e. (7.2.7). A similar addition could be made to (7.2.4)–(7.2.6) if desired.

Even if we have a number of missing variables spread across the studies in the meta-analysis, model (7.2.7) is still estimable. In general, provided we have information from two or more studies about each term in the covariance matrix, (7.2.7) should be estimable with minimal prior information. In fact, as this is a Bayesian imputation model, we can in principle fit the model and impute however weak the information in the observed data.

Using MI in the context of IPD meta-analysis begs the question: at which point in the procedure should Rubin's MI combination rules be applied? We can either:

1. apply Rubin's rules to the imputed data for each study, resulting in a single summary from each study which is then meta-analysed (we term this Rubin's Rules then Meta-Analysis (RR then MA) *or*
2. perform a meta-analysis of each imputed dataset, and then summarize the results using Rubin's Rules (ie MA then RR).

As we already mentioned earlier, in a recent study Burgess et al. considered the imputation of sporadically missing data, concluding that in the majority of cases the best approach was to impute separately in each study. This is because in many meta-analyses there is important between-study heterogeneity, which needs to be respected in the imputation process. If imputation is performed separately for each study contributing to the meta-analysis, then their results showed that it is best to apply Rubin's rules before meta-analysing the results. This also implicitly calls for a two-stage analysis of the IPD data, rather than a one-stage analysis. However, a one-stage analysis has the potential advantage of allowing us to borrow strength across studies, which may be important for estimation of covariate and subgroup effects.

However, suppose we view the data as a whole, and that if there were no missing values we intended to fit a one-stage (hierarchical) analysis model. Now suppose we have missing data. If the data are imputed together using our joint model, the usual MI justification tells us to fit the analysis model to each imputed data set, and then apply Rubin's rules. Although in practice we may often replace the one-stage analysis by the more common two-stage analysis,

the same principle applies. In other words, the more we share information across studies in the imputation process, the more we should prefer to apply Rubin's rules *after* meta-analysis. This was nicely explained in (Burgess *et al.*, 2013):

If we consider a stratified imputation model [...] then a congenial analysis requires the meta-analysis to be performed before the application of Rubin's rules. This is because missing data in each study is imputed conditional on data in other studies, inducing a dependence between the imputed data values in different studies, which is not accounted for when Rubin's rules are applied at the study level. The inverse-variance weighted analysis models assume that estimates of the parameter of interest from each study are independent. In this case, the fixed-effect analysis still has slightly low coverage, whereas the random-effects analysis has correct coverage levels.

For all the simulations in this chapter, we used 5 imputed datasets. This number was initially chosen according to general recommendations from the literature, since it has been shown several times that 5 imputations are enough in most situations (Carpenter and Kenward, 2013, Chap.2). Furthermore we tested this empirically by looking at the magnitude of the Monte Carlo SEs with some simulations, and we found that increasing the number of imputations above 5 did not improve inference significantly.

7.3 Simulation study

Here we present the results of three simulation studies. In Subsection 7.3.1 we consider the same data generation mechanisms and scenarios as considered by (Burgess *et al.*, 2013). We use these to explore whether the joint modelling imputation strategy performs comparably to the within-study imputation approach. Subsection 7.3.4 goes beyond the scenarios discussed by (Burgess *et al.*, 2013), exploring settings where in some studies there are few complete records. The aim is to compare the random study-specific covariance matrix to the fixed study-specific covariance matrix approach (itself one step removed from imputing separately in each study) in this setting. Finally, in Subsection 7.3.5 we consider the setting of systematically missing variables.

For each simulated data set, the MI analysis used 5 imputations, with the MCMC sampler burned in for 500 iterations, and with 100 between-imputation iterations. In more complex settings, more iterations and imputations may be needed. However, in our setting, examination of the MCMC chains showed these numbers were sufficient for convergence of the sampler and stochastic independence of the imputed data sets. Depending on the imputation model used, imputing the data took from 3 to 10 seconds per simulation. This means that we were able to create 5 imputed datasets for 1000 simulations in each scenario in less than 3 hours.

Preliminary results showed that applying meta-analysis before Rubin's rules was the best approach, as expected given the discussion at the end of Subsection 7.2.3. We therefore use this order for the results below.

7.3.1 Simulations with studies of equal sizes

Here we use the data generating mechanism considered by (Burgess *et al.*, 2013), that is

$$\begin{aligned} y_{i,s} &= \beta_{0s} + \beta_{1,s}x_{1,i,s} + \beta_{2,s}x_{2,i,s} + \epsilon_{i,s} \\ \begin{pmatrix} x_{1,i,s} \\ x_{2,i,s} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_s \\ \rho_s & 1 \end{pmatrix} \right) \\ \epsilon_{i,s} &\sim \mathcal{N}(0, \sigma_s^2). \end{aligned} \tag{7.3.1}$$

Exactly as in the above paper, we generate data for 30 studies, each with 200 patients. We use the same five scenarios shown in Table 7.1 with increasing levels of heterogeneity between studies.

For each scenario we generated 1000 simulated datasets, and then made 50% of the values of x_2 MAR dependent on x_1. This covariate-dependent missingness mechanism means that the complete records analysis will be unbiased, so we can readily assess the extent of the information loss due to missing data, and the extent to which this is recovered using multiple imputation.

Since in scenarios 2–5 a consistent meta-analysis model requires study-specific residual variances to be considered, one-stage meta-analysis is only appropriate for scenario 1, unless we use methods which deal with complex level 1 variation in hierarchical models. As such models are still relatively rarely used by meta-analysts, we present the results of two-stage meta-analysis only, returning to this point in the discussion. For each scenario, we report both the estimates from a fixed-effects and from a random-effects meta-analysis. The penultimate row of Table

	Scenario 1 Homosced.	Scenario 2 Heterosc.	Scenario 3 Heterogeneity	Scenario 4 Random slope	Scenario 5 2 random slopes
β_{0s}	$\mathcal{N}(3, 1^2)$	$\mathcal{N}(3, 1^2)$	$\mathcal{N}(3, 1^2)$	$\mathcal{N}(3, 1^2)$	$\mathcal{N}(3, 1^2)$
σ_s^2	1	$\mathcal{N}(1, 0.3^2)$	$\mathcal{N}(1, 0.3^2)$	$\mathcal{N}(1, 0.3^2)$	$\mathcal{N}(1, 0.3^2)$
ρ_s	0.2	0.2	$\mathcal{N}(0.2, 0.2^2)$	$\mathcal{N}(0.2, 0.2^2)$	$\mathcal{N}(0.2, 0.2^2)$
β_{2s}	-0.6	-0.6	-0.6	$\mathcal{N}(-0., 0.2^2)$	$\mathcal{N}(-0., 0.2^2)$
β_{1s}	0.3	0.3	0.3	0.3	$\mathcal{N}(0.3, 0.2^2)$
consistent meta- analysis model	β_1 : fixed β_2 : fixed	β_1 : fixed β_2 :fixed	β_1 : fixed β_2 :fixed	β_1 : fixed β_2 : random	β_1 : random β_2 : random
simplest compatible imputation model	(7.2.3)	(7.2.5)	(7.2.6)	(7.2.6)	(7.2.6)

Table 7.1: Scenarios used to generate data from (7.3.1), and corresponding consistent (i) meta-analysis and (ii) imputation models, when values of X_2 are missing.

7.1 shows for both β_1 and β_2 which one of the two is consistent with the data. We expect some overestimation of the standard errors with random-effects meta-analysis when the fixed-effects model is the consistent one. Conversely, when a random effects meta-analysis is consistent, a fixed effects meta-analysis may underestimate the standard errors and hence result in poor confidence interval coverage.

The bottom row of Table 7.1 shows which imputation model is the simplest compatible with the data generation mechanism. Imputation models that are incompatible (too restrictive) are expected to lead to bias. Imputation models that are more flexible than the data generation mechanism should not lead to bias, but may lead to some loss of information.

7.3.2 Simulations with equal size studies: results

Tables 7.2 and 7.3 show the results. Those from the correct meta-analysis model and simplest imputation model compatible with the data generating mechanism are highlighted in bold. We now comment briefly on each scenario.

Sccenario 1: This is the most homogeneous scenario, since the only source of between-study heterogeneity comes from the different study-specific intercepts, $\beta_{0,s}$, in the data-generating algorithm. In this scenario, the simplest imputation model (7.2.3) is sufficient. Analysis of datasets imputed with this model gives good results both in terms of bias, precision and confidence interval coverage. We see some gain in precision compared with the complete records analysis. Reassuringly, using the most general imputation model (7.2.7) gives similar results under fixed-effects meta-analysis (the consistent analysis for this scenario). However, when imputing with (7.2.7) and using the random effects analysis, additional variability allowed for by the (7.2.7) is picked up by the random effects analysis, resulting in larger standard errors and mild over-coverage.

Scenario 2: Here, imputation model (7.2.3) is not compatible with the data generation mechanism, as it assumes a common variance σ_e across studies. Further, imputation model (7.2.4) is still not compatible. This is because the data generating model allows the variance of y to vary across studies, whereas imputation model (7.2.4) only allows the variance of x_2 to vary across studies. Therefore, to have sufficient flexibility for this scenario, we need to include y as a response in the imputation model, by using (7.2.5). Using this imputation model, or the most general imputation model (7.2.7), the findings are similar to those from scenario 1.

	Coefficient 1					
	Fixed Effect meta-analysis			Random Effects meta-analysis		
	Mean	SE	Cov	Mean	SE	Cov
True value:	0.300		95.0	0.300		95.0
Scenario 1 (homoscedasticity):						
Complete Data	0.300	0.013	94.9	0.300	0.014	95.7
Complete Records	0.300	0.020	93.6	0.300	0.022	95.0
Imp. Model (7.2.3)	**0.300**	**0.016**	**93.3**	0.300	0.016	94.1
Imp. Model (7.2.7)	0.299	0.016	93.8	0.299	0.020	98.1
Scenario 2 (heteroscedasticity):						
Complete Data	0.300	0.012	96.0	0.300	0.012	97.1
Complete Records	0.300	0.018	93.6	0.300	0.019	94.7
Imp. Model (7.2.5)	**0.299**	**0.015**	**95.2**	0.299	0.016	96.2
Imp. Model (7.2.7)	0.296	0.015	93.2	0.297	0.019	98.1
Scenario 3 (heterogeneity):						
Complete Data	0.300	0.012	94.4	0.300	0.013	95.5
Complete Records	0.299	0.018	95.6	0.299	0.020	94.8
Imp. Model (7.2.6)	**0.296**	**0.015**	**95.1**	0.296	0.019	98.5
Imp. Model (7.2.7)	**0.296**	**0.015**	**94.0**	0.296	0.019	97.6
Scenario 4 (One random slope):						
Complete Data	0.300	0.012	94.5	0.300	0.013	95.9
Complete Records	0.300	0.018	93.8	0.300	0.020	94.9
Imp. Model (7.2.3)	0.293	0.015	88.7	0.296	0.020	97.0
Imp. Model (7.2.4)	0.301	0.015	93.5	0.302	0.019	97.2
Imp. Model (7.2.5)	0.303	0.015	92.8	0.302	0.018	97.4
Imp. Model (7.2.6)	**0.296**	**0.015**	**92.2**	0.297	0.019	97.5
Imp. Model (7.2.7)	**0.299**	**0.015**	**94.2**	0.300	0.019	97.7
Scenario 5 (Two random slopes):						
Complete Data	0.299	0.012	65.5	0.299	0.023	95.1
Complete Records	0.299	0.018	77.7	0.299	0.027	94.1
Imp. Model (7.2.3)	0.293	0.015	72.5	0.295	0.026	94.6
Imp. Model (7.2.4)	0.301	0.015	74.4	0.303	0.025	94.2
Imp. Model (7.2.5)	0.303	0.015	74.8	0.300	0.023	94.2
Imp. Model (7.2.6)	0.296	0.015	73.5	**0.295**	**0.027**	**96.3**
Imp. Model (7.2.7)	0.298	0.015	72.4	**0.298**	**0.027**	**97.0**
Scenario (7.3.2) (Inverse Wishart):						
Complete Data	0.300	0.013	50.1	0.301	0.036	93.2
Complete Records	0.301	0.029	77.3	0.301	0.046	95.0
Imp. Model (7.2.5)	0.284	0.022	66.3	0.282	0.039	91.0
Imp. Model (7.2.6)	0.297	0.021	66.6	**0.301**	**0.043**	**96.1**
Imp. Model (7.2.7)	0.301	0.021	65.9	**0.302**	**0.040**	**94.0**

Table 7.2: Simulations with studies of equal sizes. Mean estimates, SE and coverages for the coefficient of variable x_1, the completely observed covariate.
Scenarios 1,2 & 3: Comparison of results from imputation model (7.2.3),(7.2.5) and (7.2.6) respectively (the simplest ones compatible with the data) and imputation model (7.2.7) (the most general).
Scenarios 4 & 5: Comparison of all the different models presented, starting with the simplest one (7.2.3) and ending with the most general (7.2.7).
Data Generated with (7.3.2): Comparison of the 3 more general models in the previous scenario.
Results in bold highlight cases where both the meta-analysis and the imputation model are compatible with the data-generating mechanism.

	Coefficient 2					
	Fixed Effect meta-analysis			Random Effects meta-analysis		
	Mean	SE	Cov	Mean	SE	Cov
True value:	-0.600		95.0	-0.600		95.0
Scenario 1 (Homoscedastic):						
Complete Data	-0.600	0.013	95.0	-0.600	0.014	95.8
Complete Records	-0.600	0.019	93.3	-0.600	0.020	95.1
Imp. Model (7.2.3)	**-0.595**	**0.018**	**92.6**	-0.595	0.018	93.9
Imp. Model (7.2.7)	-0.603	0.017	91.8	-0.601	0.023	97.5
Scenario 2 (Heteroscedastic):						
Complete Data	-0.600	0.012	94.3	-0.600	0.012	95.5
Complete Records	-0.600	0.016	93.8	-0.600	0.018	94.7
Imp. Model (7.2.5)	**-0.594**	**0.015**	**92.1**	-0.593	0.021	97.9
Imp. Model (7.2.7)	-0.596	0.015	93.3	-0.596	0.021	97.9
Scenario 3 (Heterogeneity):						
Complete Data	-0.600	0.012	95.0	-0.600	0.013	95.8
Complete Records	-0.599	0.017	94.4	-0.599	0.018	95.2
Imp. Model (7.2.6)	**-0.596**	**0.016**	**93.6**	-0.595	0.022	98.4
Imp. Model (7.2.7)	**-0.596**	**0.015**	**93.0**	-0.595	0.022	98.2
Scenario 4 (One random slope):						
Complete Data	-0.602	0.012	38.0	-0.602	0.039	94.8
Complete Records	-0.602	0.017	51.5	-0.602	0.041	94.4
Imp. Model (7.2.3)	-0.538	0.018	26.3	-0.577	0.035	83.9
Imp. Model (7.2.4)	-0.559	0.018	38.4	-0.599	0.040	93.7
Imp. Model (7.2.5)	-0.590	0.016	48.8	-0.588	0.043	95.1
Imp. Model (7.2.6)	-0.598	0.018	51.4	**-0.596**	**0.043**	**95.8**
Imp. Model (7.2.7)	-0.604	0.016	49.5	**-0.602**	**0.043**	**95.4**
Scenario 5 (Two random slopes):						
Complete Data	-0.599	0.012	39.4	-0.598	0.039	93.4
Complete Records	-0.600	0.017	50.0	-0.599	0.041	94.5
Imp. Model (7.2.3)	-0.535	0.018	25.1	-0.572	0.035	82.7
Imp. Model (7.2.4)	-0.555	0.018	38.4	-0.594	0.040	93.4
Imp. Model (7.2.5)	-0.585	0.016	47.6	-0.583	0.042	93.7
Imp. Model (7.2.6)	-0.595	0.016	49.6	**-0.592**	**0.043**	**95.4**
Imp. Model (7.2.7)	-0.601	0.015	50.1	**-0.599**	**0.043**	**96.2**
Scenario (7.3.2) (Inverse Wishart):						
Complete Data	-0.600	0.013	46.6	-0.601	0.036	91.6
Complete Records	-0.600	0.018	60.2	-0.600	0.039	91.9
Imp. Model (7.2.5)	-0.474	0.018	0.3	-0.479	0.037	8.4
Imp. Model (7.2.6)	-0.603	0.017	56.9	**-0.600**	**0.041**	**94.3**
Imp. Model (7.2.7)	-0.600	0.017	58.0	**-0.600**	**0.038**	**93.0**

Table 7.3: Simulations with studies of equal sizes. Mean estimates, SE and coverages for the coefficient of x_2, the partially observed covariate.
Scenarios 1,2 & 3: Comparison of results from imputation model (7.2.3),(7.2.5) and (7.2.6) respectively (the simplest ones compatible with the data) and imputation model (7.2.7) (the most general).
Scenarios 4 & 5: Comparison of all the different models presented, starting with the simplest one (7.2.3) and ending with the most general (7.2.7).
Data Generated with (7.3.2): Comparison of the 3 more general models in the previous scenario.
Results in bold highlight cases where both the meta-analysis and the imputation model are compatible with the data-generating mechanism.

Scenario 3: Because this scenario has study specific correlations, ρ_s, the simplest compatible imputation model is (7.2.6). Using this imputation model, we once again obtain unbiased estimates and good confidence interval coverage with the fixed-effects analysis, while the random-effects analysis leads to mild over-coverage for the confidence intervals. The more general imputation model (7.2.7) gives very similar results.

Scenarios 4–5: Scenario 4 generalizes scenario 3 by adding a random effect for β_2; scenario 5 further adds a random effect for β_1. The simplest compatible imputation model is therefore (7.2.6). First, we see that—even with no missing data—using a fixed effects model for a random effect leads underestimation of the standard error and under-coverage of the confidence interval. Secondly, the results for β_2 in scenarios 4 and 5 show that using an imputation model that is incompatible (i.e. insufficiently flexible) can lead to bias (and some reduction in coverage) particularly for the parameter estimate for the partially observed variable. Once again, when used with the appropriate meta-analysis model, results from imputation model (7.2.6), and the more general (7.2.7) show virtually no bias and good confidence interval coverage. Unlike with scenarios 1–3, though, little or no information is gained compared to the complete records analysis. The likely reason for this is that the data generating mechanism implies (i) a different across-study distribution of the parameters from the tri-variate normality of (7.2.6), and (ii) a very different joint distribution of the covariance matrices to the inverse-Wishart used in imputation model (7.2.7). We investigate this point in the next subsection, by exploring the performance of (7.2.6) and (7.2.7) with a correctly specified data generating mechanism.

7.3.3 A correctly specified data-generating mechanism

We simulated data from a correctly specified data generating mechanism for imputation models (7.2.6) and (7.2.7). We used a trivariate normal distribution with an inverse Wishart distribution for the study specific covariance, parametrized to be consistent with (7.3.1). This gives the following:

$$
\begin{aligned}
&\begin{pmatrix} y_{i,s} \\ X_{1,i,s} \\ X_{2,i,s} \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} \beta_{0,s} \\ 0 \\ 0 \end{pmatrix}, \Omega_s \right) \\
&\Omega_s \sim \mathcal{W}^{-1} \left(a, A_s = \begin{pmatrix} \sigma_{1,s} & \sigma_{2,s} & \sigma_{3,s} \\ \sigma_{2,s} & 1 & 0.2 \\ \sigma_{3,s} & 0.2 & 1 \end{pmatrix} \right) \\
&\sigma_{3,s} = \frac{1.042\beta_{2,s} + 0.208\beta_{1,s}}{(1.042^2 - 0.208^2)} \\
&\sigma_{2,s} = \frac{\beta_{1,s} + 0.208\sigma_3}{1.042} \\
&\sigma_{1,s} = 1 + (\sigma_{2,s}, \sigma_{3,s}) \begin{pmatrix} 1.042 & -0.208 \\ -0.208 & 1.042 \end{pmatrix} \begin{pmatrix} \sigma_{2,s} \\ \sigma_{3,s} \end{pmatrix}
\end{aligned}
\tag{7.3.2}
$$

We explored different values for a, the degrees of freedom of the inverse Wishart distribution:

1. $a = 3$, i.e. the minimum possible;
2. $a = 200$, i.e. the number of patients in each study, and
3. $a = 30$, i.e. the number of studies.

The conclusions were unchanged by the choice of a, so we report results for $a = 30$ in the bottom five rows of Tables 7.2 and 7.3. Once again—even with no missing data—fitting a fixed effects meta-analysis in this random effects setting gives poor results. With missing data, and a random effects meta-analysis, imputation model (7.2.5) now gives poor results for β_2, because it does not allow for a study specific dependency of y and x_2 on x_1. However, imputation using (7.2.6) gives unbiased results with good coverage and gain in information for β_1, while imputation using (7.2.7), which allows for sharing of information on the covariance matrix across studies, additionally recovers some information on β_2. We discuss the results further in Subsection 7.3.6 below.

7.3.4 Simulations with studies of different sizes

In the simulation settings discussed so far, we have 200 individuals in each of the 30 studies, and a missingness rate for x_2 of around 50%. Thus there is sufficient information for a reasonable estimate of the covariance matrix within each study.

Here we allow the study sizes to vary. We still have 30 studies overall, and the same MAR missingness mechanism, so that around 50% of the observations are missing for x_2 only, with an MAR missingness mechanism dependent on the outcome only. However, now

- 10 studies have between 10 and 50 individuals (average sample size ~ 25);
- 10 studies have between 50 and 100 individuals (average sample size ~ 75), and
- 10 studies have between 100 and 200 individuals (average sample size ~ 150).

As a consequence, studies with only 10 individuals may have only 4 or 5 complete records, which are not enough for reasonable estimation of the covariance matrix. In such situations imputation model (7.2.7) should have a clear advantage over imputation model (7.2.6).

Tables 7.4 and 7.5 show the results when data are simulated under scenarios 3, 5 and using the correctly specified data generating mechanism (7.3.2). Here, the complete records analysis struggles because of the small number of observations from some studies. For the same reason, the results for imputation model (7.2.6) show considerable bias and poor confidence interval coverage.

However, imputation model (7.2.7) gives results that are essentially unbiased, and with good coverage both when the distribution of study specific covariance matrices is not inverse Wishart (scenarios 3, 5) and when it is (data generated using (7.3.2)). Information is recovered for the coefficient of the fully observed covariate in all cases, and for the coefficient of the partially observed variable when the data generation model is the same as the imputation model.

7.3.5 Simulations with systematically missing variables

Here we again simulate data from 30 studies, each with 200 patients. We use scenarios 1–5 shown in Table 7.1, and then the scenario where the data generating mechanism is the same as the imputation model, (7.3.2). We make values of x_2 missing using the same MAR mechanism as before, so about 50% are missing. In addition, for each simulated dataset, out of 30 studies, we completely remove x_2 from two randomly selected studies.

	Coefficient 1					
	Fixed Effect meta-analysis			Random Effects meta-analysis		
	Mean	SE	Cov	Mean	SE	Cov
True value:	0.300		95.0	0.300		95.0
Scenario 3 (heterogeneity):						
Complete Data	0.300	0.021	93.6	0.300	0.023	95.3
Complete Records	0.296	0.030	80.9	0.297	0.041	88.4
Imp. Model (7.2.6)	**0.287**	**0.026**	**91.3**	0.284	0.036	97.4
Imp. Model (7.2.7)	**0.296**	**0.028**	**92.5**	0.295	0.041	98.5
Scenario 5 (two random slopes):						
Complete Data	0.299	0.021	49.9	0.300	0.036	93.1
Complete Records	0.297	0.031	75.6	0.297	0.046	92.8
Imp. Model (7.2.6)	0.286	0.027	78.2	**0.282**	**0.043**	**94.7**
Imp. Model (7.2.7)	0.296	0.026	78.8	**0.297**	**0.041**	**96.0**
Scenario (7.3.2) (Inverse Wishart):						
Complete Data	0.299	0.021	57.0	0.300	0.044	93.7
Complete Records	0.301	0.048	75.5	0.301	0.073	94.4
Imp. Model (7.2.6)	0.208	0.029	24.9	0.202	0.051	49.9
Imp. Model (7.2.7)	0.296	0.033	71.4	**0.297**	**0.053**	**94.9**

Table 7.4: Simulations with studies of different sizes. Mean estimates, SE and coverages for the first coefficient, the one related to x_1, the completely observed covariate.
Scenario 3,5 and data generating mechanism (7.3.2): Comparison of results from imputation models (7.2.6) and (7.2.7). Once again, cases where both the imputation and the meta-analysis model are consistent with the data-generating mechanism are highlighted in bold.

	Coefficient 2					
	Fixed Effect meta-analysis			Random Effects meta-analysis		
	Mean	SE	Cov	Mean	SE	Cov
True value:	-0.600		95.0	-0.600		95.0
Scenario 3:						
Complete Data	-0.600	0.021	93.3	-0.600	0.023	94.5
Complete Records	-0.604	0.028	80.7	-0.600	0.038	89.3
Imp. Model (7.2.6)	**-0.575**	**0.034**	**89.4**	-0.564	0.049	96.9
Imp. Model (7.2.7)	**-0.590**	**0.039**	**92.2**	-0.591	0.052	98.8
Scenario 5:						
Complete Data	-0.599	0.021	53.7	-0.599	0.046	94.3
Complete Records	-0.599	0.028	56.5	-0.598	0.055	92.7
Imp. Model (7.2.6)	-0.571	0.037	66.1	**-0.557**	**0.064**	**92.2**
Imp. Model (7.2.7)	-0.603	0.029	62.3	**-0.601**	**0.059**	**95.1**
Scenario (7.3.2)						
Complete Data	-0.599	0.021	55.7	-0.599	0.043	92.5
Complete Records	-0.582	0.029	64.8	-0.595	0.053	93.4
Imp. Model (7.2.6)	-0.313	0.042	0.0	-0.293	0.053	0.0
Imp. Model (7.2.7)	-0.595	0.029	70.3	**-0.594**	**0.049**	**93.0**

Table 7.5: Simulations with studies of different sizes. Mean estimates, SE and coverages for the second coefficient, the one related to x_2, the partially observed covariate.
Scenario 3, 5 and data generating mechanism (7.3.2): Comparison of results from imputation models (7.2.6) and (7.2.7). Once again, cases where both the imputation and the meta-analysis model are consistent with the data-generating mechanism are highlighted in bold.

In this setting we do not present the complete records analysis, as this excludes studies with no data on x_2. For imputation in this setting, we need to be able to draw study specific parameter values to impute x_2 when it is systematically missing. For scenario 1, we can do this using (7.2.3); for the other scenarios we need (7.2.7). Table 7.6 shows the results.

Recall that scenarios 1–5 use a different data generation model from the imputation model. For scenarios 1–3, where the fixed effects model is correct, we see minimal bias and good coverage if the fixed effects model is used, and somewhat conservative inference if the random effects model is used. For scenarios 4–5, using the fixed effects model when the random effects model is correct results in poor coverage, even with no missing data. Using the correct analysis model appears to give a little bias, and confidence intervals coverage is always too high.

With data simulated from the correctly specified data generation model, (7.3.2), we see no bias, correct coverage, and minimal loss of information compared to the complete records analysis.

7.3.6 Summary of findings from simulation studies

Drawing together the results from the simulation studies, we conclude the following:

Choice of meta-analysis model

Even without missing data, using a fixed effects model when random effects are present leads to underestimation of the standard error and under-coverage of the confidence intervals.

	Coefficient 1						Coefficient 2					
	Fixed Effect meta-analysis			Random Effects meta-analysis			Fixed Effect meta-analysis			Random Effects meta-analysis		
	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov
True value:	0.300		95.0	0.300		95.0	-0.600		95.0	-0.600		95.0
Scenario 1:												
Complete Data	0.300	0.013	94.9	0.300	0.014	95.7	-0.600	0.013	95.0	-0.600	0.014	95.8
Imp. Mod. (7.2.3)	0.300	0.016	94.2	0.300	0.017	95.5	-0.600	0.018	93.8	-0.600	0.019	94.6
Scenario 2:												
Complete Data	0.300	0.012	96.0	0.300	0.012	97.1	-0.600	0.012	94.3	-0.600	0.012	95.5
Imp. Mod. (7.2.7)	0.296	0.021	96.0	0.297	0.026	99.6	-0.562	0.055	95.7	-0.589	0.041	99.8
Scenario 3:												
Complete Data	0.300	0.012	94.4	0.300	0.013	95.5	-0.600	0.012	95.0	-0.600	0.013	95.8
Imp. Mod. (7.2.7)	0.296	0.022	96.5	0.297	0.028	99.5	-0.565	0.054	96.6	-0.588	0.042	99.6
Scenario 4:												
Complete Data	0.300	0.012	94.5	0.300	0.013	95.9	-0.602	0.012	38.0	-0.602	0.039	94.8
Imp. Model (7.2.7)	0.277	0.33	99.0	0.297	0.038	99.6	-0.569	0.059	81.3	-0.592	0.064	97.6
Scenario 5:												
Complete Data	0.299	0.012	65.5	0.299	0.023	95.1	-0.599	0.012	39.4	-0.598	0.039	93.4
Imp. Mod. (7.2.7)	0.297	0.021	84.1	0.297	0.033	99.0	-0.566	0.056	81.5	-0.589	0.055	97.2
Scenario (7.3.2)												
Complete Data	0.302	0.013	48.7	0.301	0.036	95.3	-0.602	0.013	49.7	-0.602	0.036	93.8
Imp. Mod. (7.2.7)	0.301	0.015	58.0	0.301	0.037	95.6	-0.603	0.014	55.6	-0.604	0.037	94.4

Table 7.6: Simulations with studies with systematically missing variables.
Scenarios 1–5 and data generated with (7.3.2): Results with imputation model (7.2.7), the only one usable in this situation. Note that with systematically missing variables, it is not possible even to run complete records analysis, unless we exclude the studies with systematically missing variables.

Studies of equal sizes

When a fixed effects model is valid—and used—imputation using the most general model (7.2.7) gives comparable results to imputing using the simplest compatible model.

When a fixed effects model is valid, but a random effects model is used, the most general imputation model (7.2.7) leads to some loss of efficiency relative to the simplest compatible model.

In applications, we are most likely to be faced with study specific correlation matrices. In this case, unless we use imputation models that allow this, i.e. (7.2.6) or (7.2.7), we may get bias and poor coverage, especially for the coefficients of variables with a lot of missing data. When the across-study distribution of covariance matrices is close to inverse-Wishart, then we gain information on both parameters with (7.2.7). When it is not, bias remains negligible and coverage good, but there is little gain in information compared to a complete records analysis.

Studies with unequal sample sizes

Here, sharing information on the covariance structure across studies using (7.2.7) improved markedly on the complete records analysis, with more information recovered the closer the data generation mechanism is to the imputation model.

Studies with systematically missing variables

Imputation using (7.2.7) is the only possible general approach in this setting. It does not introduce practically important bias in any of the scenarios considered, though it was a little conservative in the random effects setting when the data generation mechanism was very differ-

ent to the imputation model. The closer the data generation mechanism is to the imputation model, the greater the gain in information through multiple imputation, with excellent results when they are the same.

Overall

In applications, we therefore advocate (7.2.7), as the simulations suggest the worst that can happen is that inferences are mildly conservative.

7.4 Discussion

In this chapter we have considered a joint modelling (JM) approach to multiple imputation for IPD meta-analysis where data are continuous. Our results show that, for sporadically missing data within studies—ie the situations considered by (Burgess *et al.*, 2013)—the JM approach performs comparably. However, our JM approach allows each study's covariance matrix to itself be drawn from an overall Inverse Wishart distribution (i.e. model (7.2.7) and its extensions). This flexibility gives improved results when studies are of different sizes, and/or some variables have a lot of missing values in particular studies. This is because in such settings the study specific covariance matrices are poorly estimated, and our approach allows appropriate pooling of the covariance matrices across studies, in exactly the same way as multilevel/hierarchical models.

The flexibility of our approach also allows us to naturally handle the case where some studies have completely missing data on particular variables. This will frequently happen in IPD meta-analyses, where the IPD data set is often assembled post-hoc, and different studies have used different protocols.

The FCS imputation approach is typically preferred to joint modelling imputation by practitioners because of the availability of software, and the fact that the REALCOM software for multilevel MI (Carpenter *et al.*, 2011) can be relatively slow. However, using the `jomo` package I wrote, eliminated this problem since imputed datasets were created in seconds.

As our approach derives from a proper joint model, fitted by MCMC, it does not rely on any problem-specific adaptations. Further, it naturally extends to settings with more levels in the hierarchy, and variables defined at the second and higher levels of the hierarchy. The approach described in this chapter dovetails with that proposed by (Goldstein *et al.*, 2014) for handling non-linear relationships and interactions in the substantive model.

We now consider the question of which imputation model to use, and how to construct it. Our results show that, even when the study specific covariance matrices are drawn from a highly unusual (implicit) distribution—ie in the scenarios 1–5 which are those considered by (Burgess *et al.*, 2013)—our general imputation model (7.2.7) does not introduce bias. However, when the study specific covariance matrices come from an Inverse Wishart (the conjugate distribution for the multivariate normal) then the simulation results show imputation using (7.2.7) gains efficiency, in line with statistical theory.

Since in applications appropriate handling of heterogeneity is typically a key concern, we therefore suggest using the random covariance matrix approach exemplified by (7.2.7) as the default, unless we can be confident that a simpler approach is justifiable in the context at hand.

A further decision is whether to include fully observed variables as covariates in the imputation model, as suggested in (Carpenter and Kenward, 2013, pp. 132–133), or include them all as responses. In the meta-analysis context, we believe it is typically best to include all the variables as responses (both partially and fully observed) so that they can have their own study specific covariance matrix—ie study specific associations with the other variables. If it is the case that associations across studies are similar, the model will pool information across studies appropriately.

When imputing systematically missing variables, we have to remember that we are assuming a multivariate normal distribution for the random-effects. This is a particularly simple and convenient assumption, and it is in line with the theory of generalized linear mixed effects models, where random effects are generally assumed to be multivariate normally distributed as well. However, this is quite a restrictive assumption, and recently it has been proposed the use of some other more flexible distributions (Lee and Thompson, 2008). In theory, it is possible to consider different distributions for the random effects in the imputation model as well and therefore we may consider adding this extension in our software in the future.

In all the simulations in this chapter, we had 30 studies. This is quite a large number, while meta-analyses often aggregate the data of 10-15 studies, if not even less. When a lot of variables have to be imputed, our imputation model may have quite a large number of level-2 parameters

to estimate and, therefore, with smaller number of studies (level-2 units) estimating all these parameters may be difficult. Also, it has been suggested (Gelman *et al.*, 2014, Chap.11) that with small number of clusters, MCMCs could lead to overestimation of the level-2 variances. Therefore, in the next chapter we focus on a situation where we only have 15 studies.

As briefly discussed in Subsection 7.2.1, random-effects meta-analyses throughout this chapter were performed with the DerSimonian and Laird model; confidence intervals reported are therefore obtained simply by adding and subtracting to the point estimates 1.96 times the standard error. However, Hartung and Knapp method (Hartung and Knapp, 2001), based on a different estimate of the standard error and on the use of a t distribution to calculate the confidence interval, is nowadays tought to be a superior approach (Cornell *et al.*, 2014). Therefore, undercovering observed in some scenarios could be explained in this way and this is surely one of the limitations of our study.

Finally we need to remember that all of the techniques described so far assume that data are MAR and therefore a proper sensitivity analysis to different assumptions should be considered (Carpenter *et al.*, 2007).

In conclusion, the main reason for MI's growing popularity for handling missing data is its flexibility in a wide range of situations. In this chapter we argued that the additional flexibility of joint modelling multiple imputation makes it a natural choice for missing data in IPD meta-analyses, since our results show good performance across a range of scenarios that are common in practice. We believe a key barrier to the more widespread use of this approach has been the lack of fast, general software. We have sought to remedy this by providing the `jomo` package for R, which is available from `CRAN` and was used for all the analyses in this chapter.

8

Multilevel JM MI in presence of other data types

We showed in the previous chapter the strengths and the usefulness of Multilevel JM Imputation as a tool to handle partially observed continuous data in individual patient data meta-analysis. In those settings, we simply used a multivariate normal model for the joint distribution of the partially observed data. However, finding a proper joint model in cases where other kind of variables, for example binary or categorical, are present appears more challenging. In this chapter we are going to explore solutions to this issue.

In Section 8.1 we give a very brief overview of the existing software and methods to include nominal data in the JM MI framework. In Section 8.2 we present the methods we chose to adopt and the imputation models considered. In Section 8.3 we explore the results of the simulations to demonstrate the validity and utility of our proposed approach. Sections 8.4 and 8.5 are dedicated to the inclusion of ordered categorical variables and count variables in the imputation models. Finally we conclude with a discussion in Section 8.6.

8.1 Existing Software and Methods

Some work has already been undertaken to try to impute partially observed categorical variables through JM imputation. most notably, two R packages have been uploaded to CRAN:

- with `cat` (Ted Harding and Schafer, 2012) it is both possible to impute categorical data through a saturated multinomial model or through log-linear models; the first model is very general and allows for any kind of associations among the categorical variables, while log-linear models make it possible to simplify some of these associations. The main

limitations of this package are (i) that it does not handle missingness in a mix of different data types; and (ii) that the saturated multinomial model often has too many parameters and it is not obvious what the appropriate parametrization is.

- `mix` (Schafer, 2015) uses the so-called general location model to impute in the case of a mix of continuous and categorical data. However, it can only handle a very limited number of variables and in many cases users need to choose an appropriate restricted model in order to get sensible estimates.

Overall, both packages deal with a somewhat limited range of situations and they do not allow for multilevel imputation, which is the situation where JM has clear advantages over FCS. In our package *jomo*, we decided to use a different strategy, using latent normal variables to model categorical and binary data, as proposed in (Goldstein *et al.*, 2009) and as described in Chapter 3. After having outlined the model, and the substantial work entailed in the corresponding extension of the `jomo` package, the main aim of this chapter is to prove through a proper simulation work the practical utility and validity of this approach.

8.2 Methods

In this Section, we will briefly review JM Imputation, presenting the latent normal imputation models used to handle binary and categorical data. We will start with single level models in Subsection 8.2.1 and extend to the multilevel case in Subsection 8.2.2.

8.2.1 Single Level Models

Suppose we intended to collect a certain number N of observations on K variables y_i, but we finally end up with some missing values in each (or at least in some) of these variables. If we want to use JM Imputation to impute these missing data, the first thing we have to do is to set up our joint model for the missing data. In the simple case of normal data only, we can simply define the following multivariate normal model:

$$\begin{aligned} & y_{i,1} = \beta_{0,1} + \epsilon_{i,1} \\ & \dots \\ & y_{i,m} = \beta_{0,m} + \epsilon_{i,m} \\ & \dots \\ & y_{i,K} = \beta_{0,K} + \epsilon_{i,K} \\ & \begin{pmatrix} \epsilon_{i,1} \\ \dots \\ \epsilon_{i,m} \\ \dots \\ \epsilon_{i,K} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ \dots \\ 0 \\ \dots \\ 0 \end{pmatrix}, \Omega_e \right) \end{aligned} \tag{8.2.1}$$

Here we termed Y all the partially observed variables, because they are outcomes in the imputation model; however, this is irrespective of what it is their role in the substantive analysis model, where they could be instead included as X (covariate) of a regression model.

When some of the K variables are binary, or categorical, a possible extension to this model has been proposed by (Goldstein *et al.*, 2009). The basic idea is that there is a latent normal variable for each level of the categorical variables. For example, each binary Y_m will be substituted by two normal variables $Z^*_{1,m}$ and $Z^*_{2,m}$, the maximum of which will indicate the level of Y_m.

$$
\begin{aligned}
& y_{i,1} = \beta_{0,1} + \epsilon_{i,1} \\
& \dots \\
& z^*_{i,m,1} = \beta^*_{0,m,1} + \epsilon^*_{i,m,1} \\
& z^*_{i,m,2} = \beta^*_{0,m,2} + \epsilon^*_{i,m,2} \\
& \dots \\
& y_{i,K} = \beta_{0,K} + \epsilon_{i,K} \\
& \begin{pmatrix} \epsilon_{i,1} \\ \dots \\ \epsilon_{i,m,1} \\ \epsilon_{i,m,2} \\ \dots \\ \epsilon_{i,K+1} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \begin{pmatrix} 0 \\ \dots \\ 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \Omega_e \end{pmatrix} \right.
\end{aligned}
\tag{8.2.2}
$$

Unfortunately, the model defined in such a way can be shown to be non-identifiable. Informally, we can see this by noting that a binary variable is defined by one single parameter, the probability of the event π, but our two latent normals have two parameters each (mean

and variance), making four in total. We address this by doing two things: (i) we fix the value of the variance for the normals to an arbitrary value, say for example to 0.5, so that $\epsilon^*_{i,m,1}, \epsilon^*_{i,m,2} \sim N(0, 0.5)$, and (ii) we subtract the equation for $Z^*_{2,m}$ from the one for $Z^*_{1,m}$:

$$(z^*_{i,m,1} - z^*_{i,m,2}) = (\beta^*_{0,m,1} - \beta^*_{0,m,2}) + (\epsilon^*_{i,m,1} - \epsilon^*_{i,m,2}).$$

Setting $z_{i,m} = z^*_{i,m,1} - z^*_{i,m,2}$, $\beta_{0,m} = \beta^*_{0,m,1} - \beta^*_{0,m,2}$ and $\epsilon_{i,m} = \epsilon^*_{i,m,1} - \epsilon^*_{i,m,2}$, we get:

$$\begin{aligned}
&y_{i,1} = \beta_{0,1} + \epsilon_{i,1} \\
&\dots \\
&z_{i,m} = \beta_{0,m} + \epsilon_{i,m} \\
&\dots \\
&y_{i,K} = \beta_{0,K} + \epsilon_{i,K} \\
&\begin{pmatrix} \epsilon_{i,1} \\ \dots \\ \epsilon_{i,m} \\ \dots \\ \epsilon_{i,K} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \dots \\ 0 \\ \dots \\ 0 \end{pmatrix}, \Omega_e \right) \qquad \Omega_{e,m,m} = 1.
\end{aligned} \tag{8.2.3}$$

Here, since $\epsilon_{i,m} = \epsilon^*_{i,m,1} - \epsilon^*_{i,m,2}$ and $\epsilon^*_{i,m,1}, \epsilon^*_{i,m,2} \sim N(0, 0.5)$, we have that $\epsilon_{i,m} \sim N(0, 1)$ and therefore $\Omega_{e,m,m} = 1$. In this situation, the single latent normal remaining is greater than 0 when the binary variable is 1 and lower than zero otherwise. The same reasoning can be applied to any T–level categorical variable. For $T = 3$ we have initially three latent normals:

$$z^*_{i,m,1} = \beta^*_{0,m,1} + \epsilon^*_{i,m,1}$$

$$z^*_{i,m,2} = \beta^*_{0,m,2} + \epsilon^*_{i,m,2}$$

$$z^*_{i,m,3} = \beta^*_{0,m,3} + \epsilon^*_{i,m,3}$$

We then fix the variances of the residuals to a particular value, e.g. 0.5, so that $\epsilon^*_{i,m,1}, \epsilon^*_{i,m,2}, \epsilon^*_{i,m,3} \sim N(0, 0.5)$; subsequently, we subtract one of the equations, for example the one for $z^*_{i,m,3}$, from the other two:

$$(z^*_{i,m,1} - z^*_{i,m,3}) = (\beta^*_{0,m,1} - \beta^*_{0,m,3}) + (\epsilon^*_{i,m,1} - \epsilon^*_{i,m,3})$$

$$(z^*_{i,m,2} - z^*_{i,m,3}) = (\beta^*_{0,m,2} - \beta^*_{0,m,3}) + (\epsilon^*_{i,m,2} - \epsilon^*_{i,m,3})$$

And we set $z_{i,m,1} = z^*_{i,m,1} - z^*_{i,m,3}$, $z_{i,m,2} = z^*_{i,m,2} - z^*_{i,m,3}$, $\beta_{0,m,1} = \beta^*_{0,m,1} - \beta^*_{0,m,3}$, $\beta_{0,m,2} = \beta^*_{0,m,2} - \beta^*_{0,m,3}$, $\epsilon_{i,m,1} = \epsilon^*_{i,m,1} - \epsilon^*_{i,m,3}$ and $\epsilon_{i,m,2} = \epsilon^*_{i,m,2} - \epsilon^*_{i,m,3}$. Again, being $\epsilon_{i,m,1}$ and $\epsilon_{i,m,2}$ obtained from the sum of two normals with variance 0.5, they have variance 1. It is also possible to see that the covariance term between them is equal to 0.5, basically because of the term $\epsilon^*_{i,m,3}$ being in common between the two expressions.

When, among the $T-1$ latent normals, $z_{i,m,M} = \max_{j=(1,\ldots,T-1)} z_{i,m,j}$ and $z_{i,m,M} > 0$, then the observed category is M. Otherwise, if $z_{i,m,M} < 0$, the observed category is T.

We can use the same strategy to deal with, potentially, any number of categorical variables with any number of levels each. Therefore, for example in a situation where we have a continuous, a 4-level categorical and a binary variable, a possible joint imputation model is:

$$
\begin{aligned}
y_{i,1} &= \beta_{0,1} + \epsilon_{i,1} \\
z_{i,2,1} &= \beta_{0,2,1} + \epsilon_{i,2,1} \\
z_{i,2,2} &= \beta_{0,2,2} + \epsilon_{i,2,2} \\
z_{i,2,3} &= \beta_{0,2,3} + \epsilon_{i,2,3} \\
z_{i,3} &= \beta_{0,3} + \epsilon_{i,3} \\
\begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2,1} \\ \epsilon_{i,2,2} \\ \epsilon_{i,2,3} \\ \epsilon_{i,3} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_e = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} & \sigma_{1,5} \\ \sigma_{2,1} & 1 & 0.5 & 0.5 & \sigma_{2,5} \\ \sigma_{3,1} & 0.5 & 1 & 0.5 & \sigma_{3,5} \\ \sigma_{4,1} & 0.5 & 0.5 & 1 & \sigma_{4,5} \\ \sigma_{5,1} & \sigma_{5,2} & \sigma_{5,3} & \sigma_{5,4} & 1 \end{pmatrix} \right).
\end{aligned}
\tag{8.2.4}
$$

To apply JM Imputation, we run a Gibbs sampler to update the values of the parameters of the model, $\boldsymbol{\beta} = (\beta_{0,1}, \beta_{0,2,1}, \beta_{0,2,2}, \beta_{0,2,3}, \beta_{0,3})$, $\boldsymbol{\Omega}_e$ and then the missing data, which are also parameters from the Bayesian prospective. First of all we need to choose the priors; it is a good idea to start by considering flat, non-informative priors, to give the greatest weight to the data, compared to the prior. When we have strong belief in some different priors, the extensions of the MCMC to accommodate these are straightforward, so we will focus on the flat prior case.

Successively we need to find the proper conditional distribution from which to draw all the parameters. Unfortunately, because of the constraints in the covariance matrix, we cannot draw a new value for the covariance matrix from a known distribution. Therefore, we rely on a Metropolis-Hastings step to update it element-wise, as suggested in (Browne, 2006). Furthermore, we need to update at each step of the sampler the values of the latent normal variables, $z_{i,2,\cdot}$ and $z_{i,3}$, for patients whose categorical variables were actually observed; this is done with a rejection sampling step, where a new value for each of the observed T–level categorical variables is drawn from the proper conditional distribution until the maximum of the latent normals indicates the observed category, or all the latent normals are lower than zero if the observed category is T.

8.2.2 Multilevel Models

When the data are clustered, or otherwise multilevel, some more care is needed. Consider for example the situation of individual patient data meta-analyses; in this case level 1 of this multilevel structure is represented by patients while level 2 are the studies. It is highly likely that data from patients within the same study are correlated, which is not accounted for by models (8.2.1)–(8.2.4). We therefore have to set up a different joint imputation model.

Building from model (8.2.4), considering the case of J level 2 units with n_J observations each, so that $\Sigma_{j=1}^{J} n_J = N$, we continue with the example of a continuous, a 4-level categorical and a binary variable, and we extend to the multilevel setting as follows:

$$
\begin{aligned}
y_{i,j,1} &= \beta_{0,1} + u_{0,1,j} + \epsilon_{i,j,1} \\
z_{i,j,2,1} &= \beta_{0,2,1} + u_{0,2,1,j} + \epsilon_{i,j,2,1} \\
z_{i,j,2,2} &= \beta_{0,2,2} + u_{0,2,2,j} + \epsilon_{i,j,2,2} \\
z_{i,j,2,3} &= \beta_{0,2,3} + u_{0,2,3,j} + \epsilon_{i,j,2,3} \\
z_{i,j,3} &= \beta_{0,3} + u_{0,3,j} + \epsilon_{i,j,3}
\end{aligned}
$$

$$
\begin{pmatrix} \epsilon_{i,j,1} \\ \epsilon_{i,j,2,1} \\ \epsilon_{i,j,2,2} \\ \epsilon_{i,j,2,3} \\ \epsilon_{i,j,3} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_e = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} & \sigma_{1,5} \\ \sigma_{2,1} & 1 & 0.5 & 0.5 & \sigma_{2,5} \\ \sigma_{3,1} & 0.5 & 1 & 0.5 & \sigma_{3,5} \\ \sigma_{4,1} & 0.5 & 0.5 & 1 & \sigma_{4,5} \\ \sigma_{5,1} & \sigma_{5,2} & \sigma_{5,3} & \sigma_{5,4} & 1 \end{pmatrix} \right) \tag{8.2.5}
$$

$$
\begin{pmatrix} u_{0,1,j} \\ u_{0,2,1,j} \\ u_{0,2,2,j} \\ u_{0,2,3,j} \\ u_{0,3,j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Omega}_u \right).
$$

Here j indexes the clusters, i.e. the studies in our example, while i indexes the individuals.

With respect to the previous model, we have two more sets of parameters to update: J vectors of random effects $\mathbf{u}_j$ and the corresponding level 2 covariance matrix $\mathbf{\Omega}_u$. We choose a flat prior for the former and an inverse-Wishart prior for the latter before running the Gibbs sampler. Again, choosing other more informative priors, the modifications to the algorithm used would be straightforward.

The beauty of the latent normal formulation is that no additional complications are introduced in the algorithm by this addition.

In the above model, the same level 1 covariance matrix is used across all the clusters. In many situations this may not be reasonable. For example, in our motivating setting of IPD meta-analysis, on which our simulation study is based, it is often reasonable to assume that the residual error variances, and possibly the covariance structure itself, differ by study are different in different studies. This motivates us to consider two further models, still in the case of one continuous, one 4-level categorical and one binary variable. The first is as follows:

$$
\begin{aligned}
y_{i,j,1} &= \beta_{0,1} + u_{0,1,j} + \epsilon_{i,j,1} \\
z_{i,j,2,1} &= \beta_{0,2,1} + u_{0,2,1,j} + \epsilon_{i,j,2,1} \\
z_{i,j,2,2} &= \beta_{0,2,2} + u_{0,2,2,j} + \epsilon_{i,j,2,2} \\
z_{i,j,2,3} &= \beta_{0,2,3} + u_{0,2,3,j} + \epsilon_{i,j,2,3} \\
z_{i,j,3} &= \beta_{0,3} + u_{0,3,j} + \epsilon_{i,j,3}
\end{aligned}
$$

$$
\begin{pmatrix} \epsilon_{i,j,1} \\ \epsilon_{i,j,2,1} \\ \epsilon_{i,j,2,2} \\ \epsilon_{i,j,2,3} \\ \epsilon_{i,j,3} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}_{e,j} = \begin{pmatrix} \sigma^j_{1,1} & \sigma^j_{1,2} & \sigma^j_{1,3} & \sigma^j_{1,4} & \sigma^j_{1,5} \\ \sigma^j_{2,1} & 1 & 0.5 & 0.5 & \sigma^j_{2,5} \\ \sigma^j_{3,1} & 0.5 & 1 & 0.5 & \sigma^j_{3,5} \\ \sigma^j_{4,1} & 0.5 & 0.5 & 1 & \sigma^j_{4,5} \\ \sigma^j_{5,1} & \sigma^j_{5,2} & \sigma^j_{5,3} & \sigma^j_{5,4} & 1 \end{pmatrix} \right), \tag{8.2.6}
$$

$$
\begin{pmatrix} u_{0,1,j} \\ u_{0,2,1,j} \\ u_{0,2,2,j} \\ u_{0,2,3,j} \\ u_{0,3,j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}_u \right).
$$

In this model, the single Ω_e of model (8.2.4) is substituted by J cluster-specific $\Omega_{e,j}$. The same flat prior is assumed for all the new matrices and therefore all the covariance matrices are updated separately in the same way, while all the remaining steps of the MCMC are unchanged.

However, following on from what we did in Chapter 7, we introduce a further extension, i.e. that the study-specific covariance-matrices are actually random draws from a certain probability distribution. This can be really helpful in cases where a variable is wholly missing from some studies, so that a fixed study-specific covariance matrix would not be estimable, provided that the distribution from which we assume covariance matrices are drawn is well defined and plausible.

Finding out a proper distribution for the covariance matrices in these settings is not so easy, mainly because of the constraints in some values of the matrices imposed by the presence of latent normals for categorical variables. To start with, extending both (Yucel, 2011) and the work in Chapter 7, we consider the case of a (constrained) inverse-Wishart distribution:

$$\Omega_{e,j} \sim \mathcal{IW}(a, A). \tag{8.2.7}$$

We are aware that this might not be the best choice, and it might just represents a starting point. Nonetheless, the inverse-Wishart distribution remains the conjugate distribution for the covariance matrix of multivariate normally distributed data and therefore it is attractive because of the simplifications that it brings to the calculations. A proper investigation to find out which other distributions would be better choices could be the topic for a future research paper.

Having opted for a constrained Inverse Wishart distribution for the random study-specific covariance matrices, we still have to choose how to update this within the Gibbs sampler, since, as in the case of a single common covariance matrix, the conditional distribution of these matrices given all the other parameters in the model is not of a known form.

Finding a proper way to perform this step took several weeks of work, as we investigated the use of many different strategies. To mention some of these:

1. Updating all the covariance matrices element-wise with an extension of the Metropolis-Hastings step introduced in Chapter 3 and used in the case of a common covariance matrix across clusters.

   - **Pros**: the extension from the common covariance matrix case is straightforward, and the only thing we need to do is to modify the log-likelihood of the data used to calculate the acceptance ratio of the Metropolis-Hastings step. This was, in the case of single level data:

   $$logL(\boldsymbol{\beta}, \boldsymbol{\Omega}_e) \propto -\frac{1}{2} log|\boldsymbol{\Omega}| - \frac{1}{2} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}) \boldsymbol{\Omega}^{-1} (Y_i - \boldsymbol{\beta})^T .$$

   with n the number of observations, and it becomes in the new settings:

   $$logL(\boldsymbol{\beta}, \boldsymbol{\Omega}_{e,j}, \mathbf{u}_j, \boldsymbol{\Omega}_u) \propto -\frac{1}{2} tr(\boldsymbol{\epsilon}_j \boldsymbol{\epsilon}_j^T + A^{-1}) \boldsymbol{\Omega}^{-1} - (a + p + 1) log|\boldsymbol{\Omega}|$$

   where p is the dimension of the p-variate normal distribution considered.

- **Cons**: updating the covariance matrices element-wise may be computationally quite time consuming, especially when a lot of clusters, and therefore covariance matrices, are present.

2. Using the conditional probabilities of the IW distribution to update the matrices in blocks. For example, considering a matrix A we could divide it in blocks as follows:

$$A = \left(\begin{array}{c|c} A_1 & A_2 \\ \hline A_2 & A_3 \end{array}\right)$$

the idea is to initially update the block related to continuous variables, i.e. A_1, as in Chapter 7 and then, given that A_3 is constrained, use the conditional probability given A_1 and A_3 to update A_2.

- **Pros**: By updating the covariance matrices in blocks we could potentially gain in computational time, with respect to the MH step where we update each element of the matrix separately.
- **Cons**: Unfortunately, there is not a known form for the distribution of A_2 given A_1 and A_3, and after trying hard to find a solution to this problem we decided to give up on this idea.

3. Updating the covariance matrices in blocks, similarly to the previous point, but using rejection sampling (Ripley, 2008) to update A_2.

- **Pros**: Using rejection sampling, we can update A_2 even if we can't draw directly from the actual conditional distribution.

- **Cons**: In some cases, the sampler may be extremely slow in finding an acceptable draw for some of the matrices, losing the advantage of updating the covariance matrices in blocks instead of element-wise.

By looking at the computational time of the MCMC sampler using the MH step both for a common covariance matrix and for study-specific covariance matrices, we realized that the possible issue related with the computational time necessary to update all the covariance matrices element-wise was not actually so important. We then decided that the first method, involving the extension of the element-wise MH step to the random study-specific covariance matrices case, was the best one among the ones investigated and we used this method in all the analyses in this chapter.

8.2.3 Substantive Meta-anaModels

8.3 Simulation study

In this section, we present the results of a simulation study to demonstrate the efficacy of multilevel multiple imputation to handle missing data when a mix of continuous, binary and categorical data are to be imputed. For simplicity, we start in Subsection 8.3.1 with a simple single level model. Later on we extend to the multilevel case (Subsection 8.3.2), considering both a setting where there is and where there is not heterogeneity in the covariance matrices across studies. We conclude in Subsection 8.3.3 with a sensitivity analysis to data generating mechanisms different from the ones hypothesized in the algorithm. Investigation of the Monte

Carlo SEs showed that 5 imputations were not enough for some of the simulation settings in this Chapter. Therefore, in order to be consistent throughout the whole chapter, we decided to use 10 imputations in all the examples.

8.3.1 Single level model

Imagine we intended to collect 1000 observations on three variables, continuous Y, 4-level categorical X_1 and binary X_2. Suppose the substantive scientific model is as follows:

$$Y_i = \beta_0 + \beta_1(X_{1,i} == 2) + \beta_2(X_{1,i} == 3) + \beta_3(X_{1,i} == 4) + \beta_4 X_{2,i} + \epsilon_i. \tag{8.3.1}$$

Where:

$$(X == x) = 1 \Leftrightarrow X = x$$

$$(X == x) = 0 \Leftrightarrow X \neq x$$

If we have missing values on some or all of the variables, we might consider using imputation model (8.2.4) and running our Gibbs sampler with this model to impute the missing values.

To explore the statistical properties of this approach, we generated data through a joint latent normal data generating mechanism:

$$\begin{pmatrix} y_{i,1} \\ x_{i,1,1} \\ x_{i,1,2} \\ x_{i,1,3} \\ x_{i,2} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 2 \\ -0.2 \\ 0.6 \\ 0 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \right). \tag{8.3.2}$$

After having generated these data, 4–level categorical X_1 and binary X_2 are generated with the same mechanism explained while presenting the imputation models, i.e. for the binary variable:

$$X_{i,2,bin} = 1 \Leftrightarrow X_{i,2} > 0$$

$$X_{i,2,bin} = 0 \Leftrightarrow X_{i,2} < 0,$$

While for categorical X_1:

$$X_{1,i,cat} = M = \max_{k=1,2,3} X_{i,1,k} \Leftrightarrow M > 0$$

$$X_{1,i,cat} = 4 \Leftrightarrow M < 0.$$

This is basically the model we are later using as imputation model when running our Gibbs sampler; it is a multivariate normal model, where categorical variables are handled through a joint multivariate multinomial probit model.

We generated 1000 simulations and we introduced missingness in all the variables. We introduced 20% MCAR in Y and X_2, while regarding X_1 we explored two different situations:

- 40 % MCAR;
- Around 35 % MAR given Y; in particular, the probability of $x_{1,1}$ being missing was given by:

$$(1 + (\exp(3 - y_i))^{-1}$$

We ran JM imputation using our R package jomo, now extended to fit and impute from model (8.2.4). We generated 10 imputed datasets, using a burn-in of 500 updates of the Gibbs sampler and 500 between imputation iterations. In both missingness scenarios we fitted model (8.3.1)to the fully observed data first, and later to the complete records and to the 10 imputed datasets, combining results with Rubin's rules.

Results Table 8.1 shows the results. When data are MCAR, complete records analysis is valid and therefore gives consistent estimates of the parameters. However, when applying MI we can gain some information, as it is noticeable by looking at the estimates of the standard errors.

Another good reason for preferring MI to complete records is given by looking at the results for MAR data: here the complete records estimates for β_1 and β_4 are biased, as it is usually the case when data are MAR depending on the outcome of the analysis model. On the other hand, MI is valid under MAR as well, and therefore estimates of the parameters are again unbiased, and coverage is good. Comparing standard errors in this case is not too much informative, since complete records analysis is not a well specified model.

	β_1			β_2			β_3			β_4		
	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov
True Value	0.624		95.0	-0.063		95.0	-0.028		95.0	0.808		95.0
Complete Data	0.624	0.089	95.0	-0.063	0.137	95.2	-0.027	0.160	95.3	0.807	0.163	96.3
MCAR Data:												
Complete Records	0.624	0.143	95.0	-0.066	0.222	94.4	-0.033	0.259	95.5	0.814	0.266	95.6
MI Model (8.2.4)	0.626	0.116	94.0	-0.064	0.206	94.7	-0.031	0.239	95.4	0.807	0.240	93.9
MAR Data:												
Complete Records	0.499	0.122	71.3	-0.049	0.191	95.3	-0.022	0.223	93.8	-0.673	0.219	88.6
MI Model (8.2.4)	0.628	0.117	95.0	-0.049	0.224	94.9	-0.023	0.266	95.0	-0.793	0.249	94.1

Table 8.1: Simulations with single level data. Data are generated with model (8.3.2). Mean, SE and coverage level is reported for the four slope parameters. Missing data are introduced both with MCAR and MAR mechanisms and complete records is compared to JM imputation, using imputation model (8.2.4) with 10 imputed datasets.

8.3.2 Multilevel models

Suppose now we observe data with a multilevel structure. We consider the example of an IPD meta-analysis, where the source of clustering is the study to which a patient belongs. There is no reason why this should not apply to other example multilevel structures.

First example: common covariance matrix. Suppose we collected data on the same 3 variables of the previous subsection, continuous Y, 4-level categorical X_1 and binary X_2, in 15 studies. Suppose each of these studies is affected by missing data. Within each study, the substantive model of interest is still (8.3.1), but this time we want to meta-analyse the results. Suppose we chose to run a two-stage random-effects meta-analysis, using the Der Simonian and Laird estimate of the between-study heterogeneity (DerSimonian and Laird, 1986); for each parameter β_j of (8.3.1) we will have to fit the following model:

$$\hat{\beta}_{j,s} = \beta_{random,j} + u_{j,s} + \hat{s}_{j,s}\epsilon_s, \quad \epsilon_s \sim \mathcal{N}(0,1); \quad u_{j,s} \sim \mathcal{N}(0, \tau_j^2) \tag{8.3.3}$$

for $j = 0, \ldots, 4$.

To impute the missing data, we could consider Model (8.2.5) if we believe a common covariance matrix is to be used among clusters, or Model (8.2.6) or (8.2.7) otherwise. Once again, we generated 1000 simulations for each scenario, generating 400 observations for each of the 15 studies under investigation.

The data generating mechanism we used was the following one:

$$\begin{aligned}
\begin{pmatrix} y_{i,1} \\ x_{i,1,1} \\ x_{i,1,2} \\ x_{i,1,3} \\ x_{i,3} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 2 + u_{1,s} \\ 0.05 + u_{2,s} \\ -0.02 + u_{3,s} \\ 0 + u_{4,s} \\ 0.01 + u_{5,s} \end{pmatrix}, \begin{pmatrix} 2 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \right) \\
\mathbf{u}_s = \begin{pmatrix} u_{1,s} \\ u_{2,s} \\ u_{3,s} \\ u_{4,s} \\ u_{5,s} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Omega}_u = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{pmatrix} \right).
\end{aligned} \tag{8.3.4}$$

where $s = 1, \ldots, 15$ indexes the 15 studies. The binary and the categorical variable are then obtained from the latent normals exactly as in the single level case. Again, we introduced 20% MCAR in Y and X_2, while for X_1 we explored three different situations:

- 40 % MCAR;

- Around 35 % MAR given Y; in particular, the probability of $x_{1,1}$ being missing was given by:

$$(1 + (\exp(3 - y_{i,1}))^{-1};$$

- X_1 systematically missing in one of the 15 studies.

We ran JM imputation using an extension of R package jomo (Quartagno and Carpenter, 2014), to fit models like (8.2.6) and (8.2.7). We used imputation models (8.2.5), (8.2.6) and (8.2.7) for the first two missingness scenarios and only models (8.2.5) and (8.2.7) for the third case, where, since X_1 was completely missing from some studies, it was impossible to estimate the corresponding study-specific covariance matrix without hypothesizing an across-study distribution for the matrices. We imputed 10 datasets, with 500 burn-in and between-imputation iterations. We fitted model (8.3.3) to the complete records and to the 10 imputed datasets from each imputation model, combining results with Rubin's rules.

Results. Table 8.2 shows the results. Starting with the MCAR data case, we can see how using imputation model (8.2.5), which is the correct one to use in these settings since a common covariance matrix was used to generate the data, all the estimates are unbiased and standard errors are smaller than in the complete records analysis. Confidence interval coverage is also close to the nominal level.

	β_1			β_2			β_3			β_4		
	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov
True Value	0.593		95.0	0.005		95.0	0.005		95.0	-0.749		95.0
Complete Data	0.593	0.040	93.9	0.006	0.056	97.5	0.004	0.055	96.9	-0.749	0.056	96.5
MCAR Data:												
CR	0.596	0.065	94.8	0.005	0.091	94.6	0.000	0.090	96.2	-0.747	0.091	94.1
MI Model (8.2.5)	0.596	0.051	94.1	0.006	0.081	95.5	0.003	0.080	95.3	-0.746	0.079	95.5
MI Model (8.2.6)	0.595	0.066	97.4	0.008	0.110	99.2	0.004	0.110	99.3	-0.736	0.108	99.1
MI Model (8.2.7)	0.602	0.063	97.9	0.008	0.103	98.1	0.001	0.103	98.3	-0.752	0.100	98.1
MAR Data:												
CR	0.481	0.049	33.6	0.002	0.068	95.5	0.002	0.068	96.0	-0.634	0.066	57.7
MI Model (8.2.5)	0.597	0.046	95.8	0.004	0.073	94.2	0.005	0.072	96.7	-0.754	0.069	94.3
MI Model (8.2.6)	0.596	0.053	97.9	0.002	0.097	98.5	0.005	0.097	99.1	-0.749	0.088	98.0
MI Model (8.2.7)	0.598	0.051	97.2	0.003	0.093	98.4	0.006	0.093	98.8	-0.757	0.084	97.4
Syst. missing:												
MI Model (8.2.5)	0.594	0.041	93.9	0.007	0.058	95.5	0.006	0.058	95.7	-0.747	0.059	96.1
MI Model (8.2.7)	0.595	0.042	97.7	0.008	0.075	99.1	0.006	0.075	98.9	-0.745	0.073	98.8

Table 8.2: Simulations with 2-level data. Data are generated with model (8.3.4). Mean, SE and coverage levels are reported for the four slope parameter estimates. Missing data are introduced with MCAR and MAR mechanisms. The systematically missing data case is also explored. Complete records are compared to JM imputation using imputation models (8.2.5), (8.2.6) and (8.2.7), generating 10 imputated datasets in each case.

On the other hand, when using imputation models (8.2.6) and (8.2.7), the estimates of the fixed effects are still unbiased, but standard errors are overestimated and therefore confidence interval coverage exceeds the nominal level. This is in line with what we were expecting, since the imputation model is too general in this situation, allowing for study-specific covariance matrices.

Results of applying MI are similar when data are MAR, while complete records analysis is severely biased in two of the four parameters estimated. Once again, the fact that SEs are smaller in complete records analysis rather than in MI is not important, since complete records in this case is just an invalid analysis and therefore the comparison with MI is inappropriate.

Finally, it is really interesting to see how even in the analysis with systematically missing X_1 in one study, conclusions about the efficacy of MI remains unaltered, with model (8.2.5) giving unbiased parameter estimates and standard errors really close to the complete data analysis and model (8.2.7) yielding an overestimation of the standard errors but not introducing any bias.

Second example: study-specific covariance matrices. In many situations with real data it is more plausible to assume that different clusters have different level 1 covariance matrices. This is particularly likely in the setting of IPD meta-analysis. For this reason we tried to generate data with a slightly different model, to see how imputation using models (8.2.6) and (8.2.7) performs in these settings.

For simplicity we considered a slightly simpler situation, where we only have two variables, continuous Y and 3-level categorical X, again in 15 studies. The substantive model is a linear regression of Y over X in each study, followed by a random-effects meta-analysis as defined by (8.3.3).

We used the following data generating mechanism:

$$\begin{aligned}
\begin{pmatrix} y_{i,1} \\ x_{i,1,1} \\ x_{i,1,2} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 2+u_{1,s} \\ -0.05+u_{2,s} \\ 0+u_{3,s} \end{pmatrix}, \Omega_{e,s}\right) \\
\Omega_{e,s} &\sim \mathcal{IW}\left(df = 15, A = \begin{pmatrix} 2 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}\right) \\
\mathbf{u}_s = \begin{pmatrix} u_{1,s} \\ u_{2,s} \\ u_{3,s} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{\Omega}_u = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}\right).
\end{aligned} \tag{8.3.5}$$

where the categorical variable X was again generated with the usual mechanism from the latent normals X_1 and X_2:

$$x_{i,1} = \begin{cases} 1 & \text{if } x_{i,1,1} > 0 \text{ and } x_{i,1,1} > x_{i,1,2} \\ 2 & \text{if } x_{i,1,2} > 0 \text{ and } x_{i,1,2} > x_{i,1,1} \\ 3 & \text{if } x_{i,1,2} < 0 \text{ and } x_{i,1,1} < 0. \end{cases} \tag{8.3.6}$$

Again we considered MCAR data for Y and three possibilities for X:

- 40 % MCAR;
- Around 35 % MAR given Y; in particular, the probability of x_i being missing was given by:

$$(1 + (\exp(3 - y_i))^{-1}, \text{ and}$$

- X systematically missing in one of the 15 studies.

Results Looking at the results displayed in Table 8.3, we see how in this case MI with imputation model (8.2.5) fails to reach a satisfactory coverage level. This is because this model uses a common covariance matrix, which is incorrect in this situation. Therefore the final estimate for the standard error is even smaller than in the complete data analysis. By contrast, both models (8.2.6) and (8.2.7) are now well specified and so they present unbiased estimates and good coverage levels, pretty close to the levels attained with the complete data.

The advantage of the imputation models over simple complete records analysis is given by the fact that, as usual, these models work for MAR data as well, while the advantage of model (8.2.7) over model (8.2.6) is that it can be used with good results even in the more challenging situation of data systematically missing, provided we can assume that covariance matrices are distributed according to a certain distribution, that in this case is the inverse-Wishart distribution.

	β_1			β_2			β_3		
	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov
True Value	2.284		95.0	0.000		95.0	-0.850		95.0
Complete Data	2.284	0.262	94.0	-0.002	0.105	92.2	-0.843	0.106	91.4
MCAR Data:									
CR	2.283	0.262	93.8	-0.002	0.107	91.4	-0.841	0.109	91.6
MI Model (8.2.5)	2.286	0.259	76.1	-0.002	0.077	94.1	-0.843	0.090	85.2
MI Model (8.2.6)	2.280	0.262	92.6	-0.002	0.111	94.1	-0.835	0.111	92.1
MI Model (8.2.7)	2.283	0.262	92.2	-0.002	0.109	94.1	-0.843	0.109	91.6
MAR Data:									
CR	2.101	0.228	85.9	-0.002	0.099	92.2	-0.762	0.097	81.8
MI Model (8.2.5)	2.292	0.258	93.4	-0.002	0.077	78.7	-0.847	0.094	87.0
MI Model (8.2.6)	2.282	0.262	94.1	-0.002	0.109	93.2	-0.838	0.110	91.8
MI Model (8.2.7)	2.284	0.262	93.9	-0.002	0.108	93.0	-0.843	0.109	91.4
Syst. missing:									
MI Model (8.2.5)	2.319	0.262	90.6	-0.002	0.102	87.2	-0.846	0.104	91.2
MI Model (8.2.7)	2.290	0.262	92.0	-0.002	0.109	92.3	-0.838	0.111	92.0

Table 8.3: Simulations with 2-level data with heterogeneous covariance matrices. Data are generated with model (8.3.5). Mean, SE and coverage levels are reported for the three parameter estimates. Missing data are introduced with MCAR and MAR mechanisms. The systematically missing data case is also explored. Complete records is compared to JM imputation using imputation models (8.2.5), (8.2.6) and (8.2.7), generating 10 imputated datasets in each case.

8.3.3 Different data generating mechanisms

In the previous subsections we always generated our data from a multivariate normal distribution, assuming it was possible to model categorical data via latent normal models. One might wonder how much these results are sensitive to this assumption and how would they vary in cases where real data follow different distributions.

In this subsection we explore this, deciding for simplicity to focus on the single level case. Similar results and conclusions could be drawn for the multilevel case.

Conditional logit model First, we investigate the case of data generated with the logit model, in place of the probit. We decide to generate the data with a logit model conditional on values of the covariates, instead of using a joint model as we did in the previous sections.

In particular this is the data generating mechanism used:

$$\begin{aligned} \beta_1 = \beta_2 = \beta_3 &= 3 \\ X_1, X_2 &\sim N(0,1) \\ Z &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ p &= \frac{1}{1+\exp(-Z)} \\ Y &\sim Bernoulli(p). \end{aligned} \tag{8.3.7}$$

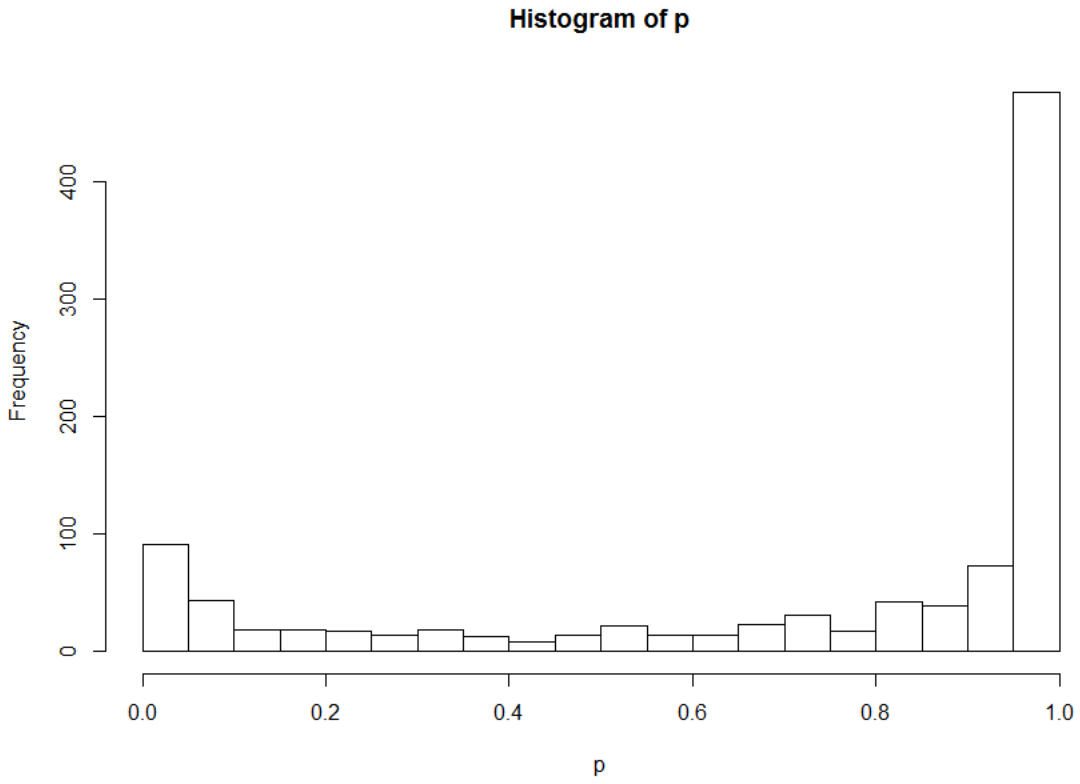

Figure 8.1: Histogram of the distribution of p, i.e. the probability of an event $y_i = 1$, for one of the simulations.

We simulate 1000 datasets with 1000 observations in each. The implied marginal distribution for Z is therefore normal with mean 3 and variance 18; Figure 8.1 shows the histogram of the distribution of p for one of the simulations.

We then make around 20% of the variables MCAR. Therefore, having 3 variables with probability of missingness 0.2 each, the probability of a complete record is $(1 - 0.2)^3 = 0.51$.

Then, we generate 10 imputations using an imputation model similar to (8.2.4), i.e. including all the variables in the model as outcomes:

$$
\begin{aligned}
z_i &= \beta_{0,1} + \epsilon_{i,1} \\
x_{i,1} &= \beta_{0,2} + \epsilon_{i,2} \\
x_{i,2} &= \beta_{0,3} + \epsilon_{i,3} \\
y_i &= 1 \,\text{if}\, z_i > 0 \\
y_i &= 0 \,\text{if}\, z_i < 0 \\
\begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \epsilon_{i,3} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_e = \begin{pmatrix} 1 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} \end{pmatrix} \right).
\end{aligned}
\tag{8.3.8}
$$

Table 8.4 shows the results. We can see that the estimates of the fixed effects after imputation are almost perfect, with minimal bias towards the null ($< 1\%$) and standard errors which are always around 10% smaller than in complete records analysis.

Categorical covariates with small effect Another situation in which we might want to impute a mix of categorical and normal data types, is when we have a simple linear regression model with categorical covariates. Imagine for example we observed continuous variables Y, X_1, and binary X_2, and that our substantive analysis model is:

$$
Y_i = \beta_{0,i} + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i, \tag{8.3.9}
$$

	β_1			β_2			β_3		
	Mean	SE	Cov	Mean	SE	Cov	Mean	SE	Cov
Conditional logit model (8.3.7)									
True Value	3.00		95.0	3.00		95.0	3.00		95.0
Complete Data	3.03	0.226	94.5	3.02	0.239	95.0	3.03	0.239	95.8
Complete Records	3.08	0.376	95.9	3.08	0.397	95.7	3.10	0.399	95.4
MI	2.99	0.343	93.5	2.98	0.369	92.3	2.97	0.368	93.0
Categorical variable with small effect Model (8.3.10)									
True Value	3.000		95.0	0.500		95.0	0.500		95.0
Complete Data	2.998	0.045	94.0	0.499	0.032	94.8	0.502	0.063	95.4
Complete Records	3.000	0.063	95.2	0.499	0.044	94.7	0.500	0.089	95.0
MI	3.000	0.055	95.4	0.500	0.040	95.1	0.499	0.082	95.0
Categorical variable with large effect Model (8.3.12)									
True Value	3.000		95.0	3.000		95.0	3.000		95.0
Complete Data	3.000	0.045	93.5	3.000	0.032	94.6	3.000	0.063	94.6
Complete Records	3.000	0.063	95.3	3.000	0.044	93.3	3.000	0.088	95.7
MI	2.991	0.068	93.8	3.012	0.049	89.7	2.960	0.094	85.4

Table 8.4: Simulation results with single level data generated with models (8.3.7), (8.3.10) and (8.3.12). Mean, SE and coverage levels are reported for the three parameter estimates. Data are MCAR and Complete records is compared to JM imputation using an imputation model with all the 3 variables as outcomes, similarly to model (8.2.4), and generating 10 imputed datasets in each case.

with residual error following a normal distribution.

Suppose we have the following data generating mechanism:

$$\begin{aligned} \beta_1 = \beta_2 = \beta_3 &= 0.5 \\ X_1 &\sim Bernoulli(0.5) \\ X_2 &\sim N(0,1) \\ X_1 &\perp\!\!\!\perp X_2 \\ Y &\sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma = 1). \end{aligned} \tag{8.3.10}$$

As in the previous case of the logit model, we simulate 1000 datasets with 1000 observations each, introducing again 20% missing data in all the variables, so that the probability of a complete record is again around 50%, and we create 10 imputations with a model similar to (8.2.4), with all the variables as outcomes and with a latent normal with variance fixed to 1 for the latent normal corresponding to the binary variable:

$$
\begin{aligned}
y_{i,1} &= \beta_{0,1} + \epsilon_{i,1} \\
x_{i,1} &= \beta_{0,2} + \epsilon_{i,2} \\
z_{i,2} &= \beta_{0,3} + \epsilon_{i,3} \\
x_{i,2} &= 1 \,\text{if}\, z_{i,2} > 0 \\
x_{i,2} &= 0 \,\text{if}\, z_{i,2} < 0 \\
\begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \epsilon_{i,3} \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Omega_e = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & 1 \end{pmatrix} \right).
\end{aligned}
\tag{8.3.11}
$$

Looking at the results for this analysis (see Table 8.4, 'Categorical variables with small effects'), we see that again MI seems to behave well, leading to good mean estimates for the fixed effects and to standard errors smaller than with complete records, reflecting the amount of information recovered by imputing the missing data.

Categorical covariates with large effects In the previous example all the fixed effects parameters were quite 'small', since $\beta_i = 0.5 \; \forall i = 0, 1, 2$, so that they were a half of the residual standard deviation. Here, we consider much stronger effects, $\beta_i = 3 \; \forall i = 0, 1, 2$, i.e. three times the residual standard deviation of $Y|X_1, X_2$. We repeat the same analysis as before, i.e. the same number of observations, the same proportion of missing data and the same imputation model. The results, shown in Table 8.4, are surprising at first sight.

The means of the estimates of the fixed effects are not too bad, but the standard errors are larger than in the complete records analysis. Despite this, coverage falls short of the nominal 95% level for both β_2 and β_3.

We will now consider an even simpler situation to understand the reasons for this.

General Location Model Imagine we have two variables, continuous Y and binary X, distributed according to a general location model:

$$\begin{aligned} X &\sim Bernoulli(0.5) \\ Y|X=0 &\sim N(\mu_0, 1) \\ Y|X=1 &\sim N(\mu_1, 1) \\ Y|X &\sim N(\mu_0 + (\mu_1 - \mu_0)X, \sigma = 1). \end{aligned} \tag{8.3.12}$$

If both Y and X are partially observed, we might want to use JM imputation and therefore we have to set up a joint model:

$$\begin{pmatrix} y_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \Omega_e\right) \tag{8.3.13}$$

where z_i is a latent normal variable for x_i and, as usual, $\Omega_{e,2,2}$ is constrained to 1. Now imagine for a certain observation i, only y_i is missing. To impute it with JM, we calculate the conditional distribution of y_i given z_i. From the general location model, the conditional distribution of $y_i|z_i$ is:

$$\begin{aligned} y_i|z_i < 0 &\sim N(\mu_0, 1) \\ y_i|z_i > 0 &\sim N(\mu_1, 1) \\ y_i|z_i &\sim N(\mu_0 + (\mu_1 - \mu_0)\mathcal{I}_{z_i>0}, 1). \end{aligned} \tag{8.3.14}$$

However, in JM we are actually imputing from the conditional distribution for one element of a bivariate normal model given the second, that is:

$$y_i|z_i \sim N\left(\beta_1 + \sigma_{12}\sigma_{22}^{-1}(z_i - \beta_2), \sigma_{11} - \sigma_{12}^2\sigma_{22}^{-1}\right) = N\left(\mu_3 + \mu_4 z_i, \sigma_2^2\right) \tag{8.3.15}$$

Therefore, the performance of the latent normal imputation model in this setting depends on whether is it possible to find some values of μ_3 and μ_4 for which model (8.3.15) can be a good approximation of (8.3.14).

We therefore simulated a huge dataset from the following model:

$$\begin{aligned} z_i &\sim N(0, 1) \\ y_i|z_i < 0 &\sim N(\mu_0 = 3, 1) \\ y_i|z_i > 0 &\sim N(\mu_1 = 7, 1) \\ y_i|z_i &\sim N(\mu_0 + (\mu_1 - \mu_0)\mathcal{I}_{z_i>0}, 1) \end{aligned} \tag{8.3.16}$$

Successively, we investigated the linear relationship between Y and Z, finding the values of μ_3, μ_4 and σ_2 giving the best fit for model (8.3.15) given the observed y_i and z_i:

$$y_i|z_i \sim N\left(\mu_3 + \mu_4 z_i, \sigma_2^2\right) = N\left(6.5 + 2.79 z_i, 2.33^2\right) \tag{8.3.17}$$

If (8.3.17) was a good approximation of (8.3.16), then simulating from both models should give similar results in terms of the substantive model of interest, that is the regression of Y over binary X, whose values are dependent on the values of the latent normal Z. However, while fitting this model to the data generated through (8.3.16) gives – as expected – the right values for the fixed effect estimates of $\mu_0 = 3$, $\mu_1 = 7$ and $\sigma = 1$, fitting the same model to the data generated from (8.3.17) gives biased estimates $\mu_0 = 4.27$, $\mu_1 = 4.45$ and $\sigma = 2.87$.

In Figure 8.2 we can see the comparison between the scatter-plot of a random selection of 1000 observations of Y and Z from models (8.3.16) and (8.3.17). This illustrates that the latter is not a good approximation of the former when the effects we want to estimate are particularly large, relative to the standard deviation of $Y|X$. On the other hand, for smaller effects, for example $\mu_0 = 0.3$ and $\mu_1 = 0.5$, the two scatter-plots are much more similar (Figure 8.3).

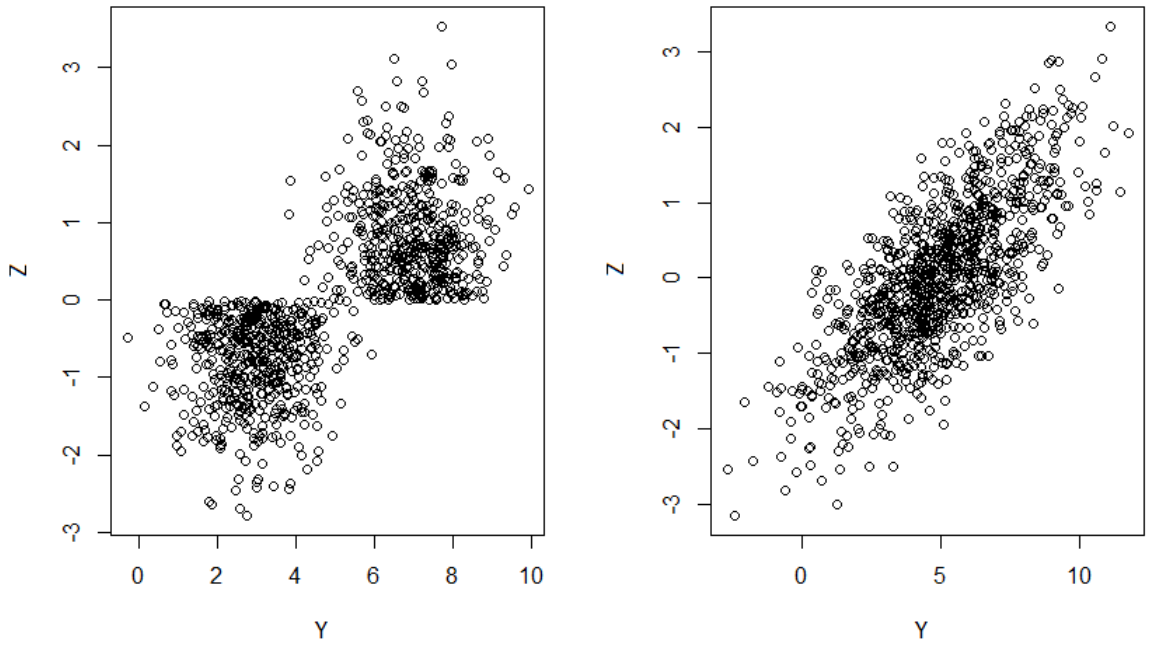

Figure 8.2: Scatterplots of Y over Z, for a random selection of 1000 draws generated from (8.3.16) (on the left) and (8.3.17) (on the right). In this case $\mu_0 = 3$ and $\mu_1 = 7$, i.e. $\mu_1 - \mu_0 = 4$, four times the standard deviation of $Y|Z$.

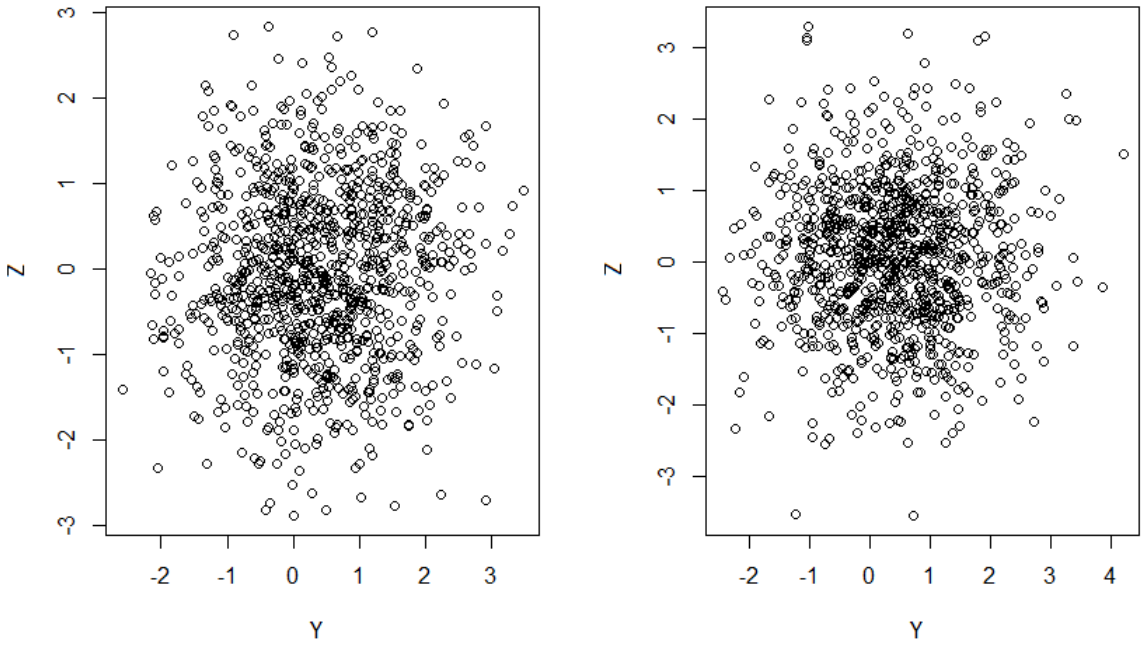

Figure 8.3: Scatterplots of Y over Z, for a random selection of 1000 draws generated from (8.3.16) (on the left) and (8.3.17) (on the right). In this case $\mu_0 = 0.3$ and $\mu_1 = 0.5$, i.e. $\mu_1 - \mu_0 = 0.2$, a fifth of the standard deviation of $Y|Z$.

8.3.4 Practical implication for use of the latent normal model

The above results suggest that in some situations the latent normal variables approach will give biased parameter estimates. This is basically when the categorical variable we want to impute is present in the substantive model of interest as a covariate, with (at least some of) the levels having a fixed effect parameter particularly large with respect to the residual standard error of the model. As a rule of thumb, we found that results of the imputation of variables whose effect is larger than twice the residual standard error of the substantive model leads to invalid imputation. For example, we repeated the simulation study where data are simulated from 8.3.10, i.e. from a linear regression model with a binary covariate, using several different values of β_2, the fixed effect estimate for the covariate effect. Figure 8.4 shows the coverage level attained in the estimation of β_2 after using the latent normal model, for the different values of the ratio $\frac{\beta_2}{\sigma_{resid}}$.

8.4 Ordered categorical variables

There are two possible kinds of categorical data that we might observe for a clinical study. An example of the first type is a pain scale, whose levels might be 'mild', 'moderate' and 'severe'. In this case, we can clearly see an ordering in the levels of the categorical variables. Such

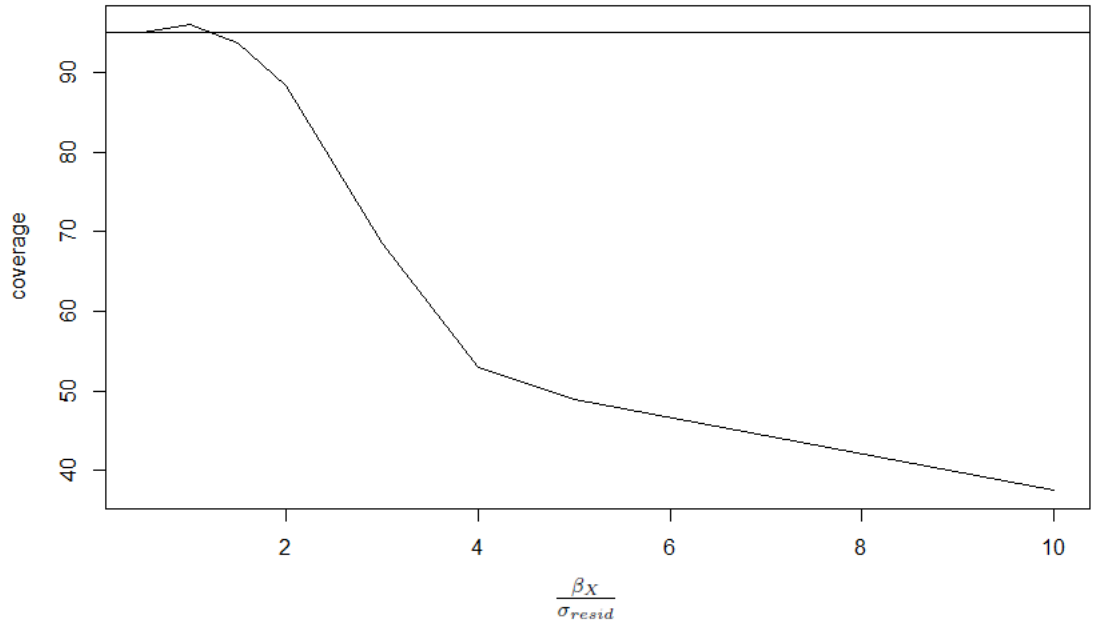

Figure 8.4: The relation between the coverage level in the estimation of a particular fixed effect estimate β_2 and the ratio $\frac{\beta_2}{\sigma_{resid}}$. The data generating model is 8.3.12 with varying values of β_2, between 0.1 and 10.

data are therefore known as *ordinal data*, or *ordered categorical variables*. A counterexample is ethnicity, that is a categorical variable where we cannot find any ordering in the levels 'white', 'black' and 'asian'; this is therefore an *unordered categorical variable*.

The methods we presented so far in this chapter were developed with unordered categorical variables in mind; It is therefore important to explore whether they work well even when there is an order among the levels of the categorical variables or if it would be better to design specific methods.

In REALCOM, Goldstein et al. handled ordinal data with a slightly different algorithm but still making use of latent normal variables. In particular, instead of defining $N-1$ latent normals for each N-*level* variable, they consider a single latent normal, using the proportional probit model (McCullagh, 1980). For a single ordinal N-level variable Y, this is:

$$probit(P(Y \leq n)) = \Phi^{-1}(P(Y \leq n)) = \alpha_n \qquad n = 1, \ldots, N$$

We arbitrarily set α_0 to $-\infty$ and α_N to $+\infty$ and define a latent normal $Z \sim N(0,1)$ so that:

$$\begin{aligned} Z \leq \alpha_1 &\Leftrightarrow Y = 1 \\ \alpha_{n-1} < Z \leq \alpha_n &\Leftrightarrow Y = n \\ \alpha_{N-1} < Z &\Leftrightarrow Y = N \end{aligned} \tag{8.4.1}$$

The α cut-off values are then updated within the Gibbs sampler, for example as proposed in (Albert and Chib, 1993) or (Chib and Greenberg, 1998). A Metropolis Hastings step for updating these values might also be possible, speeding up the update process.

Before deciding on whether to include this algorithm in the *jomo* package, we decided to explore the adequacy of imputing ordered categorical variables as if they were unordered. To test this, we generate data from two distributions and we check the relative performances of complete records analysis and JM-MI.

1. Multivariate proportional probit data generating mechanism: this is the model used in REALCOM to handle ordinal data. We generate 1000 simulations from the following multivariate normal distribution:

$$\begin{pmatrix} y_{i,1} \\ z_{i,1} \\ z_{i,2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 2 \\ 0 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}\right). \tag{8.4.2}$$

   Then, we transform the latent normal Z_1 into a 4-category variable X_1 according to the following rule:

$$x_{i,1} = \begin{cases} 1 & \text{if } z_{i,1} < -0.65 \\ 2 & \text{if } -0.65 < z_{i,1} < 0 \\ 3 & \text{if } 0 < z_{i,1} < 0.65 \\ 4 & \text{if } z_{i,1} > 0.65 \end{cases}, \tag{8.4.3}$$

and similarly we transform Z_2 into a binary variable X_2:

$$x_{i,2} = \begin{cases} 1 & \text{if } z_{i,2} > 0 \\ 0 & \text{if } z_{i,2} < 0 \end{cases} . \tag{8.4.4}$$

The substantive model of interest in this case is a simple linear regression of Y over X_1 and X_2:

$$y_i = \beta_0 + \beta_1(X_{i,1} == 1) + \beta_2(X_{i,2} == 2) + \beta_3(X_{i,3} == 3) + \beta_4 X_{i,2} + \epsilon_i,$$

where $(X_{i,1} == k) = 1$ if $X_{i,1} = k$ and $(X_{i,1} == k) = 0$ otherwise.

2. Univariate ordered probit model: we simply generate two independent normal variables X_1 and X_2 and a residual error term ϵ drawing from a standard normal and then we calculate the latent normal Z_i as follows:

$$z_i = 4 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i \tag{8.4.5}$$

Successively, 4-category y_i is created according to this rule:

$$y_{i,1} = \begin{cases} 1 & \text{if } z_{i,1} < 3.5 \\ 2 & \text{if } 3.5 < z_{i,1} < 4 \\ 3 & \text{if } 4 < z_{i,1} < 4.5 \\ 4 & \text{if } z_{i,1} > 4.5 \end{cases} . \tag{8.4.6}$$

	β_0		β_1		β_2		β_3	
Multivariate prop. probit DGM (8.4.2)	Mean (SE)	Cov	Mean (SE)	Cov	Mean (SE)	Cov	Mean (SE)	Cov
True Value	0.517	95.0	0.377	95.0	0.620	95.0	0.985	95.0
Complete Data	0.517 (0.090)	95.8	0.377 (0.119)	94.4	0.620 (0.121)	95.2	0.984 (0.125)	94.3
Complete Records	0.518 (0.146)	95.0	0.375 (0.192)	93.6	0.615 (0.196)	93.7	0.988 (0.202)	93.3
Joint Model. MI	0.531 (0.122)	95.1	0.369 (0.180)	93	0.608 (0.182)	93.9	0.963 (0.183)	93.5
Univariate ordered probit DGM (8.4.5)								
True Value	0.500	95.0	0.500	95.0				
Complete Data	0.500 (0.040)	95.1	0.502 (0.040)	95.1				
Complete Records	0.503 (0.064)	95.0	0.502 (0.064)	95.0				
Joint Model. MI	0.481 (0.058)	94.7	0.481 (0.058)	94.7				

Table 8.5: Results of the analysis of datasets with partially observed ordinal data. In both scenarios data are 20 % MCAR in all the variables and we compare complete data analysis with CR and JM-MI. We report mean, SE and coverage probabilities.

The substantive model in this setting is an ordered probit regression of Y over X_1 and X_2.

In both these scenarios, we simulate 1000 datasets with 1000 observations, making 20% of all variables MCAR. Ten imputed datasets are generated using JM-MI.

In Table 8.5 we can see the results obtained by running the substantive model when handling missing data with CR or with JM-MI with the latent normal approach designed for unordered categorical variables. We can see that mean estimates and coverage levels are good for all parameters and in both scenarios; furthermore we are always gaining some information with respect to CR analysis.

Looking at these results, it seems that using the model for general unordered categorical data will often perform very well in applications. Of course, there are cases where introducing an order between the categories can improve imputations because such an approach makes more

efficient use of the data, using fewer latent normals. However, in the majority of applications, these results suggest the unordered categorical data algorithm is sufficient. Nevertheless, we will consider adding the specific algorithm for ordinal variables to the *jomo* package in future.

8.5 Count variables

Another type of variables that one may wish to include in a statistical analysis are count variables; this includes for example cases where we want to model a count process through a Poisson regression. The simplest approach to impute such variables, is just to include them in the imputation model as continuous variables, possibly after an appropriate variance-stabilizing transformation in order to make the normal distribution assumption more plausible. The variance-stabilizing transformation for the Poisson distribution is the square root, so this is typically used.

We test this strategy by generating data from a Poisson model, using a log-link function, with two covariates, X_1 and X_2, drawn from the standard normal distribution:

$$\log \mathcal{E}[y_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} \tag{8.5.1}$$

First, we consider $\beta_0 = 3$. As for the categorical data case, we explore two different scenarios: one where the fixed effects β_1 and β_2 are particularly strong and one where they are not. Thus, we repeat the same analysis for $\beta_1 = \beta_2 = 0.1$ and for $\beta_1 = \beta_2 = 0.3$, the second of which, since we are working on a logarithmic scale with $\beta_0 = 3$, is a quite strong effect.

In Table 8.6 we can see the results obtained by running a Poisson model on the complete data, on the CR and after handling missing data with JM-MI. For this strategy, we include Y in the joint imputation model either untransformed, square rooted or taking the logarithm. For the untransformed case, we truncate to 0 negative imputed values and for the logarithmic transformation, we add 0.1 to each observation, to allow for the transformation of zero counts. Finally in all three cases, we round imputed values to the nearest integer.

We see that, for all the three MI methods, estimates are (almost) unbiased and some information is recovered when the magnitude of the fixed effects is small; when the fixed effects are stronger, using untransformed Y in the imputation model leads to the worst results, both in terms of bias and coverage level, but even using a variance stabilizing transformation like the square root improves the results only marginally. This is because, even though the square root transformation stabilized the variance and made the distribution of the counts more similar to a normal distribution, by modelling jointly the transformed counts and the covariates with a multivariate normal distribution, we are imputing missing data in Y from a conditional normal distribution, which does not resemble the real conditional distribution from which Y was created in the data generating mechanism, because of the presence of the log-link function. This is roughly the same problem we encountered for categorical data with strong effects.

Using the log-transformation, results are much better, probably thanks to the fact that the logarithm is both the function used for the transformation and the link function in the data generating model.

However, when repeating the same analysis in a setting where the mean of the Poisson distribution is smaller, e.g. with $\beta_0 = \beta_1 = \beta_2 = 1$, the transformation seems not to be able to make the distribution of Y to resemble the normal distribution, and therefore results are not optimal with any of the strategies adopted. The best results in terms of bias are those for the log-transformation again, but standard errors are hugely overestimated, even with respect to the complete records analysis.

In Figure 8.5, we can see the histograms showing the distribution of Y, either untransformed, square rooted or on the logarithmic scale, for the three scenarios presented. We can see that while in the first two examples, the transformations are able to improve normality of the data, this is not quite possible in the last scenario, mainly because of the high number of zero counts in this situation.

8.6 Discussion

In this chapter we investigated the validity of the latent normal approach to include binary and categorical variables in the Joint Modelling imputation framework. We explored in particular the multilevel structure case, that is the situation where JM appears to be preferable to FCS, showing that, provided we choose an appropriate imputation model, congenial with the analysis model of interest, it is possible to impute the data coherently in a multivariate multinomial probit situation.

	β_0		β_1		β_2	
Poisson model	Mean (SE)	Cov	Mean (SE)	Cov	Mean (SE)	Cov
$\beta_0 = 3, \beta_1 = \beta_2 = 0.3$						
True Value	3.000	95.0	0.100	95.0	0.100	95.0
Complete Data	3.000 (0.007)	93.5	0.100 (0.007)	94.7	0.100 (0.007)	95.8
Complete Records	3.000 (0.010)	94.5	0.100 (0.010)	94.6	0. 099 (0.010)	94.1
JM-MI, untransformed Y	3.000 (0.008)	94.0	0.100 (0.009)	94.6	0.099 (0.009)	94.7
JM-MI, square root of Y	3.000 (0.008)	94.2	0.099 (0.009)	93.6	0.099 (0.008)	94.7
JM-MI, log(Y+0.1)	3.000 (0.009)	94.2	0.099 (0.009)	94.3	0.099 (0.009)	94.9
Poisson model						
$\beta_0 = 3, \beta_1 = \beta_2 = 0.3$						
True Value	3.000	95.0	0.300	95.0	0.300	95.0
Complete Data	3.000 (0.007)	95	0.300 (0.007)	94.9	0.300 (0.007)	94.8
Complete Records	3.000 (0.012)	94.8	0.300 (0.011)	95.1	0.300 (0.011)	94.6
JM-MI, untransformed Y	3.008 (0.013)	92.0	0.286 (0.011)	79.2	0.286 (0.012)	79.9
JM-MI, square root of Y	3.005 (0.011)	92.9	0.291 (0.010)	85.3	0.291 (0.010)	85.8
JM-MI, log(Y+0.1)	3.001 (0.010)	95.4	0.301 (0.010)	95.3	0.300 (0.010)	95.9
Poisson model						
$\beta_0 = 1 = \beta_1 = \beta_2 = 1$						
True Value	1.000	95.0	1.000	95.0	1.000	95.0
Complete Data	0.999 (0.021)	95.1	1.000 (0.012)	95.1	1.001 (0.012)	94.6
Complete Records	1.000 (0.029)	94.5	0.999 (0.018)	94.8	1.000 (0.018)	95.1
JM-MI, untransformed Y	1.647 (0.050)	0.0	0.609 (0.047)	0.0	0.607 (0.048)	0.0
JM-MI, square root of Y	1.391 (0.046)	0.0	0.714 (0.045)	0.7	0.712 (0.046)	0.8
JM-MI, log(Y+0.1)	1.009 (0.083)	78.7	1.001 (0.072)	98.6	1.001 (0.075)	98.3

Table 8.6: Results after running Poisson regression model on data generated trough (8.5.1) with 20% MCAR in all three variables. We compare mean estimates, standard errors and coverage probabilities obtained by using complete data, complete records or JM-MI, either using untransformed Y, square root or logarithmic transformation.

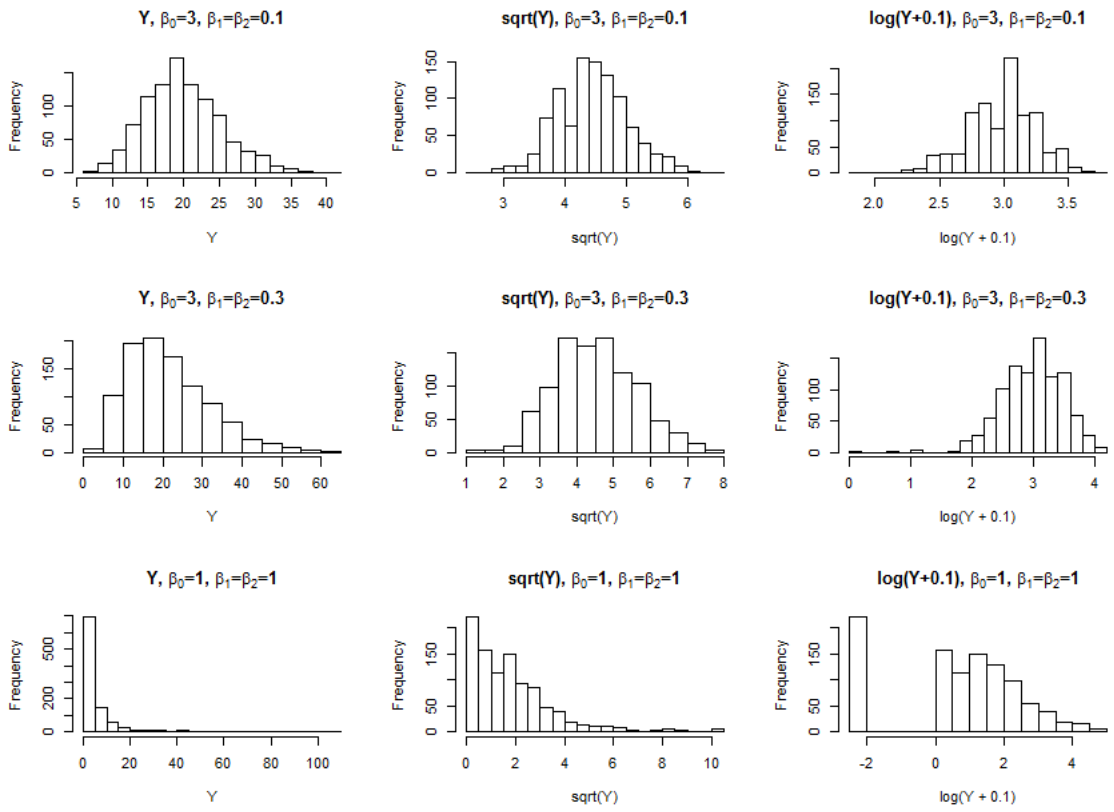

Figure 8.5: Imputation of Poisson data: histograms showing the distribution of Y, sqrt(Y) and log(Y+0.1) in the three scenarios considered.

We showed how a model like (8.2.5), with a common covariance matrix considered across clusters, can lead to biased results when the assumption about the covariance matrix is actually wrong, while for models like (8.2.6) and (8.2.7) we need to distinguish between three situations:

- if the imputation model matches exactly the data generating model they work well, i.e. there is no bias in the parameter estimates and we recover information on all parameters;
- if the imputation model is moderately misspecified, still we get unbiased estimates but we may not recover information. This happens for example when in the data generating mechanism there is a common covariance matrix across studies, while in the imputation model we assume the covariance matrices are study-specific;
- if the imputation model is badly misspecified, some bias may be introduced. This happens for example when the categorical variables we are imputing is a covariate in the substantive model, whose effect is at least twice the residual standard error of the regression model.

Therefore, in most IPD applications, where it is usually the case that covariance matrices are actually quite different due to heterogeneity among the aggregated studies, models assuming random covariance matrices, like (8.2.7), are preferable.

However, in the relatively uncommon setting of strong effects for some levels of a categorical variable included as a covariate in the imputation model, the method could fail to work. A possible solution to this issue is to split the joint model in two components; suppose we have as substantive model a linear regression with outcome Y, continuous covariates X_1 and X_2 and binary X_3, whose effect is three times the residual error of the model. Instead of defining

the joint model for $f(Y, X_1, X_2, X_3)$, we might define a marginal model for X_3, for example a multinomial model, and then a conditional model for the remaining variables:

$$f(Y, X_1, X_2, X_3) = f(X_3)f(Y, X_1, X_2|X_3)$$

This is also the strategy used to overcome another problem: when some of the partially observed variables are included in the substantive model of interest as covariates, imputation without considering this is invalid (Carpenter and Kenward, 2013, Chap. 7–8). Therefore a possible solution is again to split the model in two parts, defining a model for the covariates with non-linearities and interactions and a conditional model for the other variables (Goldstein *et al.*, 2014). We will talk about this more extensively in the next chapter.

Throughout this whole chapter, when presenting results of meta-analyses, the substantive model used was always a random-effects meta-analysis with the DerSimonian and laird estimate of between-study heterogeneity. This has the same limitations we discussed in Chapter 7 and was chosen for the same reasons set out in that chapter, i.e. to check the performance of our imputation methods in combination with the most commonly used meta-analysis model.

Another possible issue we did not mention in this chapter is that of perfect prediction (White *et al.*, 2010). As we said in Chapter 5, this may occur when we have more than one categorical variable and the strata formed by the covariates create cells in which all the responses are either 1 or 0 (or potentially any other value for categorical variables). For example, with binary X and 5-category Z, it occurs if, for all observations for which $Z = 5$, $X = 0$ and therefore, the maximum likelihood estimate of the probability is either 1 or 0 and the corresponding parameter estimates tend to $+\infty$; the associated standard errors also become very large. The

easiest way to deal with perfect prediction with JM-MI is to bound the values of the latent normals Z_i to lie in a pre-specified range; This can prevent the associated estimates to tend to infinity.

Another possible solution, is to insist, at the update step for the fixed effects $\boldsymbol{\beta}$, that the implied marginal probabilities lie in another pre-specified range. For example, for a binary variable, this entails:

- Proposing a new β^*;
- Calculating the marginal probability of the latent normal being greater than zero, given this new value of β:
$$Pr(Z > 0|\beta^*) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\beta^*)^2} dz$$
- Rejecting β^* if this is too large or too small, for example if it lies outside the range $(0.05, 0.95)$, and possibly drawing a new proposed value for β until we are able to accept it.

This process is easily generalizable to the case of multi-category variables; we therefore plan to implement it in a nearby future, comparing it to the other method, consisting in simply bounding the latent normals, to see which one is worth including in *jomo* to avoid perfect prediction bias.

As we mentioned in Section 8.2.2, when considering the random study-specific covariance matrices imputation models, the inverse Wishart distribution might not be the best distribution to use in a number of cases. A sensitivity analysis comparing different distributions for the covariance matrices when generating the data with different distributions could be the topic of further research.

We have also explored the case of ordered categorical and count variables, showing how to include such variables in this framework and demonstrating how in some applications these work well. Nevertheless, it would be nice to add the special case of ordinal variables into *jomo* in future.

To conclude, we believe we have shown how JM multiple imputation can be successfully applied in many contexts, even when some of the observed variables are not continuous. This represents a great advantage, particularly when the data we want to impute have a multilevel structure, as with IPD meta-analyses. In the next chapter we therefore present the results of using JM-MI on real data applications.

Part IV

Applications and Conclusions

9

Applications

9.1 Introduction

In Chapter 1 we motivated our research project with two IPD-MA projects, INDANA and MAGGIC, whose analyses are complicated by a large proportion of missing data. Therefore, to conclude, after having argued that JM-MI is the most appropriate method to handle missing data in this setting, we apply this method to the above examples. In this chapter we show the results of these analyses.

In Section 9.2 we present the results of the analysis of INDANA handling missing data with our *jomo* package, and broadly using the methods described in Chapter 7. In Section 9.3 we impute and analyse MAGGIC, introducing a theoretical issue and outlining possible strategies to overcome it. Finally we conclude with a short summary in Section 9.4.

9.2 A first example: Indana

INDANA is the name of an individual patient data meta-analysis project which aimed to investigate the overall effect of an anti-hypertensive drug treatment on patients at high cardiovascular risk. INDANA's dataset is made up of IPD data from 10 studies, comprising altogether 53271 patients. We will start in the next subsection with a small illustrative example showing how to use JM-MI when all the variables affected by missing data are continuous and can be assumed to follow a multivariate normal distribution.

Study	Missing items (%)			Total
	DBP after 1 year	Baseline DBP	Baseline Cholesterol	
ANBP	492 (12.5%)	0 (0%)	38 (1%)	3931
COOP	89 (10 %)	0 (0 %)	120 (13.6 %)	884
EWPH	136 (16.2 %)	0 (0 %)	18 (2.1 %)	840
HDSP	886 (8.1 %)	0 (0%)	186 (1.7 %)	10940
MRC1	993 (5.7 %)	0 (0 %)	388 (3.2 %)	17354
MRC2	285 (6.5 %)	9 (0.2 %)	21 (0.5 %)	4396
MRFT	8012 (100%)	1 (< 0.1%)	1 (< 0.1%)	8012
SHEP	512 (10.8 %)	15 (0.3 %)	1910 (40.3%)	4736
SHPP	551 (100 %)	551 (100 %)	551 (100 %)	551
STOP	89 (5.5 %)	15 (0.9%)	650 (40 %)	1627
TOTAL	4033 (7.6 %)	590 (1.1 %)	3882 (7.3 %)	53271

Table 9.1: INDANA meta-analysis: extent of missing data. Number of missing items (percentage) in each study and variable. (DBP – Diastolic Blood Pressure)

As an illustration, suppose we are interested in a simple linear regression to analyse the effect of treatment on diastolic blood pressure after one year, adjusting for sex, age, baseline cholesterol value and baseline diastolic blood pressure. Sporadically missing data occurs in all of these variables, except sex and age. In one of the studies, diastolic blood pressure and cholesterol data are not available. In practice we would typically omit this study, which (for this analysis) has age and sex only. However, we will keep them in our study in order to illustrate how our program behaves in presence of systematically missing data. Table 9.1 shows the number and the percentages of missing data for each variable in each study.

Table 9.2 shows the results of our analysis. We decided to try both one-step and two-step meta-analysis; for the one-step regression, we dealt with heteroscedasticity using SAS *proc mixed*, which allows for a different residual variance for each study, and we tried to fit both a model

with a fixed treatment effect and a random one. The estimation method used was REML, since maximum likelihood estimates are known to be biased for the variance components, especially with small sample sizes.

Concerning missing data, we dealt with them in three different ways:

1. Complete records analysis; in this case, since we didn't have any data on diastolic blood pressure in two studies, we had to exclude these studies from our analysis.
2. JM Multiple imputation with a common covariance matrix among the studies, using an imputation model with partially observed variables as outcomes and sex, age and treatment as covariates;
3. JM Multiple Imputation with an imputation model similar to the previous one but with random study-specific covariance matrices.

We created 15 imputations, with 1000 burn in iterations and 500 between imputation iterations. Results show as both imputation models seem to gain some information with respect to complete records analysis, although this is limited for the fixed effect models. In this example, probably because of the substantial heterogeneity, the random covariance matrices algorithm is able to recover even slightly more information than the common covariance matrix imputation model, at least for the random-effects meta-analyses, leading to smaller standard errors of the estimates.

	One-step heterosc. regression, fixed treatment effect Estimate (S.E.)	One-step heterosc. regression, random treatment effect Estimate (S.E.)	IVW - Fixed-effect meta-analysis Estimate (S.E.)	IVW - Random-effects meta-analysis Estimate (S.E.)
CR	-5.69 (0.10)	-6.65 (0.60)	-5.70 (0.10)	-6.72 (0.62)
MI common: covariance matrix	-5.69 (0.09)	-6.38 (0.49)	-5.70 (0.09)	-6.43 (0.50)
MI random: covariance matrices	-5.65 (0.10)	-6.27 (0.46)	-5.66 (0.10)	-6.38 (0.48)

Table 9.2: Results of analysis of the INDANA individual patient data meta-analysis. The coefficient is the estimated reduction in mean DBP after one year due to treatment, adjusting for baseline cholesterol, age and sex. We compare complete records, JM-MI with common covariance matrix and JM with random study-specific covariance matrices. In both imputations, we use 15 imputations, 1000 burn in iterations and 100 between-imputation iterations. We analysed the data with four meta-analysis models.

Regarding the clinical interpretation of these results, what we can conclude is that, after having handled missing data with our JM-MI approach, there is a very strong evidence that one year of treatment is able to decrease blood pressure substantially. The conclusion is the same we would have got by using simply complete records, but the estimates of the treatment effect are slightly smaller and more precise.

9.3 Maggic: a theoretical difficulty

The MAGGIC meta-analysis aggregated IPD from 30 cohort studies, only six of which were randomised clinical trials; overall, it includes data on 39372 patients with heart failure, of both reduced and preserved left-ventricular ejection function. 40% of patients died during a median follow-up time of 2.5 years, with deaths equally distributed between RCTs and registries data.

(Pocock *et al.*, 2013) analyzed this meta-analysis finding 13 predictors of mortality in HF. The statistical methods used included:

- Poisson survival regression models to relate baseline variables to time to death, clustering on studies (with a random intercept) to account for the correlation between observations coming from the same study;
- Since mortality risk is higher initially, the follow-up was divided into 3 time bands: less than 3 months, between three and six months and more than six months;
- The final model was built using forward stepwise regression, using as inclusion criterion a p-value lower than 0.01.

- Missing data were handled with FCS MI, creating 25 imputations before running the substantive model on the imputed datasets and combining the results with Rubin's rules.

We can immediately see that there are some theoretical problems in the strategy used by the authors to handle missing data:

1. In the MI, clustering was accounted for simply by including a fixed effect for the study indicator; therefore, this method lacked of a principled approach to impute systematically missing variables.

2. Residual variance was considered to be the same across all the studies, both for an RCT recruiting nearly 5000 patients (DIAMOND) and for registry data consisting of as few as 5 observations (NPC-I). This seems to be an unrealistic assumption;

3. In the final substantive model, two interactions were present. However, this method did not consider interactions in the imputation model. This may lead to biased results (Carpenter and Kenward, 2013, Chap. 7).

Using our hierarchical JM-MI approach, we could in principle overcome the first two issues. We therefore repeated the same analysis, but using our stratified JM-MI strategy to create the 25 imputations. However, differently from the examples we ran with the INDANA data and from all the situations considered so far in this thesis, partially observed variables in MAGGIC are actually covariates of a survival regression analysis. This raises the issue of which joint imputation model is the best one to consider in this situation.

9.3.1 Imputing covariates of survival model

When the partially observed variables $\mathbf{X}$ that we need to impute are covariates of a survival model with survival times T and event indicator D (or censoring indicators $C = 1 - D$), finding a joint model for $\mathbf{X}, T$ and D does not have a clear-cut solution.

The most natural solution is to factor the joint model, setting up a marginal model for the covariates only and the conditional model for the survival times:

$$f(T, D, X_1, \ldots, X_N) = f(T, D|X_1, \ldots, X_N) f(X_1, \ldots, X_N). \tag{9.3.1}$$

In order to describe the algorithm that can be used to impute the missing covariates using this factorization, imagine for simplicity we have a survival model with only 2 covariates X_1 and X_2. We recall here, that for a survival time distribution $f(t)$ with cumulative distribution function $F(t)$ and survival function $S(t) = 1 - F(t)$, the hazard rate $h(t)$ is defined as the death rate at time t conditional on survival until t and it holds that:

$$h(t) = \frac{f(t)}{S(t)}.$$

Furthermore the cumulative hazard rate is the integral from time 0 to t of the hazard rate and it holds that:

$$H(t) = \int_0^t h(s)ds = -\log(S(t).$$

With this in mind, the log-likelihood contribution for an individual with event time T_i with censoring indicator C_i, given covariate values $X_{i,1}$ and $X_{i,2}$ is:

$$\begin{aligned} l_{PH} &= C_i \log f(T_i|X_{i,1}, X_{i,2}) + (1 - C_i) \log(S(T_i|X_{i,1}, X_{i,2})) = \\ &= C_i \log h(T_i|X_{i,1}, X_{i,2}) - H(T_i|X_{i,1}, X_{i,2}) = \\ &= C_i(\log(h_0(T_i)) + \beta_1 X_{i,1} + \beta_2 X_{i,2}) - H_0(T_i) \exp(\beta_1 X_{i,1} + \beta_2 Xi, 2) \end{aligned} \tag{9.3.2}$$

where in the last line we have assumed that the proportional hazards assumption holds, without specifying explicitly a form for the baseline hazard so that both the general Cox model or some parametric model can be assumed.

Assuming a joint multivariate normal model for the two covariates marginally, with parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Omega}})$ and density f_X, after having initialized all the parameters in the model, missing data included, we can then divide the JM imputation algorithm into three steps:

1. Update β_1, β_2 and the baseline hazard $h_0(t)$; this should be done in principle with a fully Bayesian approach (Clayton, 1991);

2. Draw the proposed new values for the missing data $X^*_{i,1}$ and/or $X^*_{i,2}$ from the bivariate normal distribution and accept or reject them with probability p:

$$p = \min\left(\frac{L(X^*_{i,1}, X^*_{1,2})}{L(X^*_{i,1}, X^*_{1,2})}\right)$$

   where $L = \exp(l_{PH}) f_X(X_1, X_2)$;

3. Update $\tilde{\boldsymbol{\theta}}$ as usual.

This method is congenial with the proportional hazards models, because it is directly derived from the joint distribution. Furthermore, it can be extended to impute categorical covariates through latent normals.

However, in a popular paper, (White and Royston, 2009) analysed the form of the conditional distribution of the covariates given the survival times and found that:

- For a single binary partially observed covariate Y_1, the missing values can be imputed from a logistic regression of Y_1 provided the censoring indicator C and the baseline cumulative hazard $H_0(t)$ are included in the imputation model;
- If there are other covariates in the model, we need to include them in the logistic regression together with their interactions with $H_0(t)$, but then the results of the imputation will be true only if all the covariates are binary, otherwise they will hold only approximately;
- If the partially observed covariates are normal, it is possible to fit a linear regression instead of a logistic regression, but results will be approximate as well.

In the paper, they tested the use of FCS-MI including different estimates of the baseline cumulative hazard and the censoring indicator, finding that best results were usually obtained using the Nelson-Aalen estimator of $H_0(t)$ and that inclusion of the interactions of $H_0(t)$ with the other observed covariates did not seem to be crucial.

This is valid for a generic Cox proportional hazards model; in the case of MAGGIC, where an exponential distribution is used for the hazard of the survival model, we might think that the cumulative hazard can be simply approximated with survival times T. However, in this example the hazard rate was divided in three time bands, so that the hazard and the cumulative hazard are respectively:

$$h_0(t) = \begin{cases} k_1 & t < t_1 \\ k_2 & t_1 \leq t < t_2 \\ k_3 & t \geq t_2 \end{cases}$$

$$H_0(t) = \begin{cases} k_1 t & t < t_1 \\ k_1 t_1 + k_2(t - t_1) & t_1 \leq t < t_2 \\ k_1 t_1 + k_2(t_2 - t_1) + k_3(t - t_2) & t \geq t_2 \end{cases}$$

Therefore, survival time T is not a good estimate of the cumulative hazard and we decided to keep the Nelson-Aalen estimator. Including an estimate of the cumulative hazard and the censoring indicator in the imputation model, can then lead to imputations that are approximately correct, even though these approximations may not be precise with increasing number of covariates in the imputation model, with large values of variance for the covariates and for large values of the cumulative hazard.

In principle the first method we proposed will always be superior, since it does not entail any approximation, but this second method has the advantage that it can be used with standard software simply introducing the cumulative hazard estimate and the censoring indicator in the imputation model.

9.3.2 Results

We decided to use for our analysis of the MAGGIC data the White and Royston method, including censoring indicator and the Nelson-Aalen estimator as an estimate of the cumulative hazard in the imputation model. This was because the survival model used was a piecewise Poisson regression, which is known to be equivalent to a piecewise exponential survival model (Holford, 1980; Laird and Olivier, 1981). Similarly to what had been done in the original paper, we created 25 imputations, using our *jomo* software with two different imputation models:

1. A model with all partially observed variables as outcomes, completely observed as covariates, including the Nelson-Aalen estimates and the event indicator D, and a random effect for study-effect;

2. A model similar to the previous one, with the important distinction that study-specific covariance matrices were considered.

Variable	Studies with no data		Studies with some data		Total partially observed
	Studies	Missing	Studies	Missing	observations
Age	0	0	0	0	0
Gender	0	0	0	0	0
BMI	16	6898	13	2686	9584
Current smoker	5	1549	24	448	1997
SBP	9	12016	20	273	12289
Diabetes	1	348	28	340	688
NYHA class	5	2503	24	1121	3624
Ejection fraction	6	3279	24	6186	9465
COPD	9	9171	20	253	9424
HF Duration	20	11679	9	1034	12713
Creatinine	5	2800	24	13529	16329
Beta-blocker	2	273	27	709	982
ACE-I/ARB	1	97	28	159	256

Table 9.3: MAGGIC datasets: Extent of missing data. Age and sex are the only fully observed variables.

The MAGGIC dataset consists of 39372 observations divided between 6 RCTs and 24 registries, but we had to exclude from our analysis data coming from one of the RCTs, because the data owners did not provide permission for us to use them. We therefore have 31755 observations. Table 9.3 shows the number of missing observations, both sporadic and systematic, in the whole aggregated IPD dataset.

In the original paper, the final model was built using forward stepwise regression, with inclusion criteria $p < 0.01$. We did not repeat this process, taking the same variables selected in their work, in order to explore possible differences in the results with alternative imputation models.

In Table 9.4 we can see the results of our analysis. Rate ratios calculated in the paper with respective confidence intervals are compared to rate ratios calculated on data imputed with our two imputation models.

Variable	Rate ratio (CI) from (Pocock *et al.*, 2013)	MI using *jomo* with common cov.	MI using *jomo* with random covs.
Age (10 years)	1.154 (1.092,1.220)	1.220 (1.142,1.304)	1.229 (1.144,1.320)
Gender	1.115 (1.073,1.159)	1.133 (1.086,1.181)	1.133 (1.086,1.181)
BMI	0.965 (0.959,0.972)	0.965 (0.958,0.972)	0.968 (0.961,0.975)
Current smoker	1.159 (1.109,1.210)	1.163 (1.112,1.217)	1.164 (1.111,1.220)
SBP (10mmHg)	0.882 (0.855,0.910)	0.927 (0.893,0.962)	0.935 (0.899,0.972)
Diabetes	1.422 (1.365,1.481)	1.420 (1.359,1.483)	1.410 (1.347,1.476)
NYHA 1	0.788 (0.732,0.848)	0.778 (0.706,0.857)	0.768 (0.698,0.845)
NYHA 3	1.410 (1.354,1.467)	1.418 (1.283,1.568)	1.431 (1.297,1.579)
NYHA 4	1.684 (1.580,1.796)	1.701 (1.513,1.912)	1.723 (1.545,1.920)
Ejection fraction (5%)	0.581 (0.539,0.627)	0.649 (0.587,0.718)	0.659 (0.597,0.727)
COPD	1.228 (1.152,1.310)	1.219 (1.151,1.290)	1.230 (1.160,1.303)
HF duration < 18 m	1.188 (1.139,1.240)	1.247 (1.180,1.318)	1.247 (1.176,1.321)
Creatinine (10 μmol/L)	1.039 (1.035,1.042)	1.036 (1.028,1.043)	1.032 (1.027,1.038)
Beta-Blocker	0.760 (0.726,0.796)	0.769 (0.735,0.805)	0.771 (0.736,0.807)
ACE-I/ARB	0.908 (0.856,0.963)	0.933 (0.896,0.971)	0.937 (0.898,0.977)
Interaction EF*Age	1.040 (1.031,1.049)	1.038 (1.027,1.048)	1.037 (1.026,1.049)
Interaction EF*SBP	1.012 (1.008,1.017)	1.005 (1.000,1.011)	1.005 (1.001,1.011)

Table 9.4: Results of MAGGIC analysis. Outcome is time to death, or censoring. Rate ratios (91% CI) of analysis using FCS MI, JM-MI with common covariance matrix and JM-MI with study-specific covariance matrices.

First of all, all the covariates were found to be highly significant with all the three imputation methods, the only exception being the interaction between ejection fraction and SBP, that was only borderline significant when using JM-MI. Estimates of the rate ratio were similar for most of the variables, with some important exceptions, like for example ejection fraction (0.581 vs. 0.649/0.659). Standard errors did not seem to be systematically larger or narrower between FCS and JM methods and between the two JM methods. However, we need to remember here that we were using three different models for imputing the data. Therefore we cannot compare the magnitude of the standard errors for the three methods as we do with complete records analysis when data are assumed to be MCAR, both because data might be MAR and because (some of) the models could be wrong, so that smaller estimates of the standard errors could be due to model misspecification rather than because of more effective recovery of information.

Finally, it is important to note that the imputation process with *jomo* took altogether around thirty minutes. With REALCOM, imputing such a large dataset would have taken much longer time, even more than 24 hours; this was once again a proof of the importance of having created a software package that is able for the first time to impute such big datasets in a feasible amount of time.

9.3.3 Interactions in the analysis model

As we already noted at the beginning of this section, there is a further theoretical problem in the imputation of MAGGIC: in the substantive model of interest we have two interactions. However, these are not included in the imputation model. Therefore, results of imputation not

considering this will be a little biased. This is likely to take the form of an attenuation of the estimate of the interaction effect, and correspondingly stronger estimates of the main effect parameters.

In the original paper, interactions were simply addressed with passive imputation, a method that consists in just calculating the values for the interactions passively from the imputed values of the two variables without setting up a model for the interactions. This method is known to lead to bias (Seaman *et al.*, 2012).

The Just Another Variable (JAV) approach (Von Hippel, 2009), simply consists in including the interactions in the imputation model as if they were another variable, without bothering respecting the relationship between the imputed values for the variables themselves and for the interactions. This is also not an optimal method, relying on a moment based justification which only holds if data are MCAR and the substantive model is a linear regression, neither of which apply here. Nonetheless, it has been shown to be usually less biased than the passive imputation approach, apart from the case of logistic regressions (Seaman *et al.*, 2012).

Since both these methods lack a theoretical basis and have been shown to perform poorly in simulation studies, other methods should be used for dealing with interactions in a theoretically well-formulated way. Using a similar approach to that proposed for categorical data in the discussion section of Chapter 8 and to the first method we proposed for imputing covariates of a survival analysis, we propose factoring the joint distribution into two components.

For example, if we have four partially observed variables to impute, two of which, Y_3 and Y_4, are included in the substantive model as covariates together with their interactions, then we can factor the joint distribution for the four variables as follows:

$$f(Y_1, Y_2, Y_3, Y_4) = f(X_3, X_4) f(X_1, X_2 | X_3, X_4) \tag{9.3.3}$$

Then, we have to set up a joint model for Y_3 and Y_4 marginally and another joint model for Y_1 and Y_2 conditional on the draws from the first model. For the first model, we can simply consider a multivariate normal distribution:

$$\begin{pmatrix} Y_3 \\ Y_4 \end{pmatrix} \sim N_2 \left(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Omega}} \right) \tag{9.3.4}$$

When this assumption is not proper, we can simply transform variables to approximate normality. The joint model for Y_1 and Y_2 will then be:

$$\begin{aligned} Y_{i,1} &= \beta_{0,1} + \beta_{3,1} Y_{i,3} + \beta_{4,1} Y_{i,4} + \beta_{34,1} Y_{i,3} Y_{i,4} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_{0,2} + \beta_{3,2} Y_{i,3} + \beta_{4,2} Y_{i,4} + \beta_{34,2} Y_{i,3} Y_{i,4} + \epsilon_{i,2} \\ \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix} &\sim N(\mathbf{0}, \boldsymbol{\Omega}) \end{aligned} \tag{9.3.5}$$

To fit this model by MCMC, we have to initialize and find priors for all the parameters of the model: the parameters of (9.3.4), $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Omega}}$, and the parameters of (9.3.5), $\boldsymbol{\beta} = (\beta_{0,1}, \beta_{3,1}, \beta_{4,1}, \beta_{34,1}, \beta_{0,2}, \beta_{3,2}, \beta_{4,2}, \beta_{34,2})$ and $\boldsymbol{\Omega}$.

The next step is to run the sampler by iterating the following steps:

1. Update $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$ and possibly missing $Y_{i,1}$ and $Y_{i,2}$ from the usual conditional distributions of a bivariate normal model where we treat Y_3, Y_4 and their interaction as completely observed covariates using their current imputed values;

2. Given $\boldsymbol{\beta}$, $\boldsymbol{\Omega}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Omega}}$ update the missing values in $Y_{i,3}$ and $Y_{i,4}$; this can be done either by the use of a Metropolis-Hastings step or by rejection sampling, as outlined for example in (Carpenter and Kenward, 2013, Chap. 6-7);

3. Update the values of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Omega}}$.

Among the methods presented here, this is the only strategy to handle missing data in covariates included with interactions in the substantive analysis model using a strictly congenial imputation model. Passive imputation and JAV are methods that gained in popularity because of their simplicity but they should be avoided in principle.

9.3.4 The ideal imputation model

Following from the previous subsection, here we describe the ideal joint imputation model for MAGGIC, consistent with the survival model, the multilevel structure and the interactions.

First, being the substantive model a survival model, we factor the joint distribution of all the variables in two components:

$$f(T, D, \mathbf{X}, \mathbf{X}_{int}) = f(T, D|\mathbf{X}, \mathbf{X}_{int}) f(\mathbf{X}, \mathbf{X}_{int}) \tag{9.3.6}$$

where T is the survival time, D the event indicator (death from all causes) and:

$$\mathbf{X} = (\text{Sex, BMI, Smoke, Diabetes, NYHA, COPD, HF Duration,}$$
$$\text{Creatinine, Beta-Blocker, ACE-I/ARB})$$

$$\mathbf{X}_{int} = (\text{Age,Ejection Fraction, SBP})$$

Here, $\mathbf{X}_{int}$ includes all the covariates that are included in the substantive model with interactions, while $\mathbf{X}$ includes all the remaining covariates. It is even possible to further partition the joint distribution, introducing a third component as follows:

$$f(T, D, \mathbf{X}, \mathbf{X}_{int}) = f(T, D|\mathbf{X}, \mathbf{X}_{int}) f(\mathbf{X}|\mathbf{X}_{int}) f(\mathbf{X}_{int}) \tag{9.3.7}$$

However, this is not strictly necessary, as we will see that model (9.3.6) is enough to set up a substantive model compatible JM-MI approach. Therefore, we have to define two joint distributions, one for the covariates and one for survival time and event indicator, conditional on the values of the covariates.

Survival distribution Regarding the survival distribution, a good idea is simply to use the same model as the substantive model. Therefore, following (Pocock *et al.*, 2013), we choose a piece-wise exponential proportional hazards model. This is known to be equivalent to fitting a Poisson regression model to the event indicator, using the logarithm of the survival time as an offset, i.e. as a covariate with parameter estimate constrained to 1 (Laird and Olivier, 1981).

More precisely, we first divide survival time in three intervals, with thresholds after 90 and 180 days. k_1 is the constant baseline hazard in the first interval, k_2 in the second and k_3 in the third. Then, we create a bunch of pseudo-observations, one for each combination of individual and interval. For example, for an individual i, whose survival time t_i was 187 days, we create three pseudo-observations as follows:

$$\begin{cases} t_{i,1} = 90, d_{i,1} = 0 \\ t_{i,2} = 180 - 90 = 90, d_{i,2} = 0 \\ t_{i,1} = 187 - 180 = 7, d_{i,3} = 1 \end{cases}$$

Then, we fit the piece-wise exponential model to data by treating the event indicators $d_{i,j}$, for $j = 1, 2, 3$, as if they were independent Poisson observations with means:

$$\mu_{i,j} = t_{i,j} k_{i,j}$$

where $k_{i,j}$ is the hazard for individual i at time j, defined as:

$$k_{i,j} = k_j \exp(\boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\beta}_{int}\mathbf{X}_{i,int} + u_s)$$

with s indexing the studies and u_s indicating a random effect. Taking logs, and recalling that the hazard rates satisfy the proportional hazards assumption, we finally obtain:

$$\log(\mu_{i,j}) = \log(t_{i,j}) + log(k_j) + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\beta}_{int}\mathbf{X}_{i,interactions} + \mathbf{u}_s, \tag{9.3.8}$$

with $u_s \sim N(0, \sigma_u)$ and:

$$\mathbf{X}_{interactions} = (\mathbf{X}_{int}, \text{Age*Ejection Fraction, SBP*Ejection Fraction}).$$

Thus, the piece-wise exponential proportional hazards model is equivalent to a Poisson log-linear model for the pseudo observations, one for each combination of individual and interval, where the death (or event) indicator is the response and the log of survival time enters as an offset. This is the ideal model we would like to use for the first part of (9.3.6), i.e. the conditional distribution of survival time and death indicator, given the covariates.

Joint distribution of the covariates Regarding the joint imputation model for the covariates of the survival substantive model, the idea is not to define explicitly the joint distribution f, but to use a proposal distribution g to draw the missing values, accepting or rejecting the proposals either with a Metropolis-Hastings step or with rejection sampling, depending on the

value of the log-likelihood of model (9.3.8). The easiest option is again to use a multivariate normal model, where fully observed variables enter as covariates, while partially variables enter as outcomes. Therefore the proposal distribution is as follows:

$$
\begin{cases}
X_{BMI,i} = \alpha_{0,BMI} + u_{BMI,s} + \alpha_{1,BMI} X_{sex,i} + \alpha_{2,BMI} X_{age,i} + \epsilon_{BMI,i} \\
X_{Smoke,i} = \alpha_{0,Smoke} + u_{Smoke,s} + \alpha_{1,Smoke} X_{sex,i} + \alpha_{2,Smoke} X_{age,i} + \epsilon_{Smoke,i} \\
X_{SBP,i} = \alpha_{0,SBP} + u_{SBP,s} + \alpha_{1,SBP} X_{sex,i} + \alpha_{2,SBP} X_{age,i} + \epsilon_{SBP,i} \\
X_{Diabetes,i} = \alpha_{0,Diabetes} + u_{Diabetes,s} + \alpha_{1,Diabetes} X_{sex,i} + \alpha_{2,Diabetes} X_{age,i} + \epsilon_{Diabetes,i} \\
X_{NYHA1,i} = \alpha_{0,NYHA1} + u_{NYHA1,s} + \alpha_{1,NYHA1} X_{sex,i} + \alpha_{2,NYHA1} X_{age,i} + \epsilon_{NYHA1,i} \\
X_{NYHA3,i} = \alpha_{0,NYHA3} + u_{NYHA3,s} + \alpha_{1,NYHA3} X_{sex,i} + \alpha_{2,NYHA3} X_{age,i} + \epsilon_{NYHA3,i} \\
X_{NYHA4,i} = \alpha_{0,NYHA4} + u_{NYHA4,s} + \alpha_{1,NYHA4} X_{sex,i} + \alpha_{2,NYHA4} X_{age,i} + \epsilon_{NYHA4,i} \\
X_{EF,i} = \alpha_{0,EF} + u_{EF,s} + \alpha_{1,EF} X_{sex,i} + \alpha_{2,EF} X_{age,i} + \epsilon_{EF,i} \\
X_{COPD,i} = \alpha_{0,COPD} + u_{COPD,s} + \alpha_{1,COPD} X_{sex,i} + \alpha_{2,COPD} X_{age,i} + \epsilon_{COPD,i} \\
X_{HFDur,i} = \alpha_{0,HFDur} + u_{HFDur,s} + \alpha_{1,HFDur} X_{sex,i} + \alpha_{2,HFDur} X_{age,i} + \epsilon_{HFDur,i} \\
X_{Creat,i} = \alpha_{0,Creat} + u_{Creat,s} + \alpha_{1,Creat} X_{sex,i} + \alpha_{2,Creat} X_{age,i} + \epsilon_{Creat,i} \\
X_{BB,i} = \alpha_{0,BB} + u_{BB,s} + \alpha_{1,BB} X_{sex,i} + \alpha_{2,BB} X_{age,i} + \epsilon_{BB,i} \\
X_{ACE-I,i} = \alpha_{0,ACE-I} + u_{ACE-I,s} + \alpha_{1,ACE-I} X_{sex,i} + \alpha_{2,ACE-I} X_{age,i} + \epsilon_{ACE-I,i},
\end{cases}
\tag{9.3.9}
$$

with the $\epsilon_{x,i}$ jointly following a normal distribution with mean zero and covariance matrix $\Omega_{e,s}$, and the random intercepts $u_{x,s}$ jointly following a multivariate normal distribution with mean zero and covariance matrix Ω_u.

Missing data in these variables, are drawn from the proper conditional distribution obtained from this joint model, and later the draws are accepted or rejected depending on the value of the log-likelihood of model (9.3.8) with the new draw of the missing values, compared to the previous draw in the MCMC, as explained in Subsection 9.3.1 and 9.3.3.

In conclusion, the ideal joint imputation model for MAGGIC, would be (9.3.6), with the first member of the right-hand side defined by (9.3.8) and the second as (9.3.9). It is still not possible to consider such a model in *jomo*, but work is on-going to extend the package accordingly.

9.4 Conclusions

In this chapter, we applied our multilevel JM-MI strategy to handle missing data in two large IPD-MA aggregating data from trials and, in the case of MAGGIC, data from both trials and registries. In the INDANA meta-analysis, we performed an illustrative analysis, obtaining promising results and confirming the potentiality of the imputation method based on random study-specific covariance matrices to respect heterogeneity in the covariance structures of different studies, and recover information including patients with partially observed data.

However, the MAGGIC meta-analysis, posed new challenges. Specifically, the fact that the substantive model of interest was a survival model and the presence of interactions between partially observed variables in the substantive model. While for the former problem, a simple approximate solution exists, namely including an estimate of the cumulative hazard and the censoring indicator in the imputation model, for the second problem simple methods like passive imputation or JAV should be possibly avoided; we therefore introduced a possible solution based on the factorization of the joint distribution, a solution that, as we will see in the next chapter, could be potentially used even to deal with non-linearities in the substantive analysis model.

This led us to propose our ideal imputation model for the MAGGIC study, whose analysis provided the initial motivation for this thesis. Unfortunately, at the time of writing, we have still not implemented the algorithm for fitting this kind of model in *jomo*.

Nevertheless, even in the MAGGIC data, our JM proposal worked successfully, providing a computationally feasible approach which properly respects the between-study heterogeneity and allows us – for the first time – to include in a MA studies where relatively few values are recorded on key covariates.

10

Summary, Discussion and Future Work

Multiple Imputation is a strategy to handle missing data that has been gaining a lot of popularity in recent years in the world of clinical research for a variety of reasons: among the others we may list its broad applicability, its flexibility, the wide availability of well implemented statistical software packages and the fact that it is possible to use the standard software to analyse the data after imputation.

The aim of this thesis was to implement and evaluate a strategy for MI in the particular case of missing data in individual patient data meta-analysis, focusing on an approach based on the specification of a joint multilevel imputation model for the whole IPD dataset. In this final chapter we review the research chapters of this thesis, highlighting the main messages that we believe should be taken from our work and setting out a possible plan of future research to be undertaken.

10.1 Missing Data in IPD meta-analysis

As we set out in Chapter 2, missing data introduce many issues in IPD-MA; variables may be both sporadically or systematically missing, multilevel imputation models may be required to impute consistently with the analysis model and heteroscedasticity may be important as well. In Figure 10.1 we can see a flow-chart explaining how to approach missing data in IPD-MA with MI depending on the presence of the aforementioned problems.

The key factors that should lead to prefer multilevel MI in most situations are the presence of systematically missing variables, random slopes and heteroscedasticity. All of these issues have been addressed in our thesis, particularly in Chapter 7.

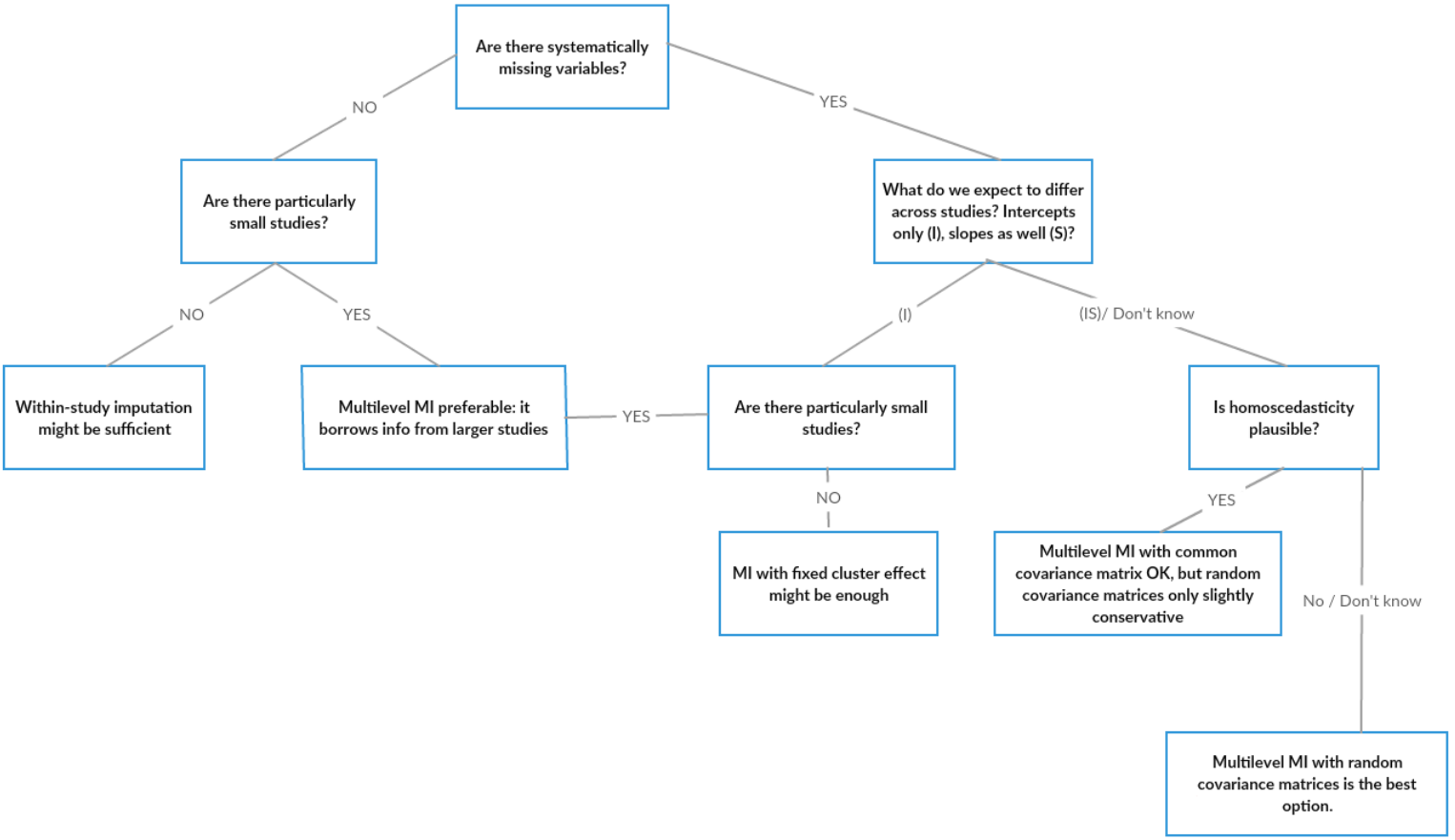

Figure 10.1: A flow-chart explaining when multilevel MI is necessary, compared to within-study imputation and imputation with fixed cluster effect.

10.2 jomo: outreach and extensions

As we explained in detail in Chapter 4, the difficulty of extending standard methods for MI (based on FCS) to allow for multilevel structure in the IPD-MA setting – i.e. patients nested within studies – resulted in adopting a Joint Modelling strategy. Unfortunately, the only two software packages available to apply this method were R package *pan* and standalone program REALCOM, but the former lacks the flexibility to impute non-normal data, and the latter has problems of computational efficiency which made it infeasible to use for both simulations studies and the applications at hand. Further, neither of the existing packages implemented an approach to address systematically missing covariates.

We therefore created a new R package, called *jomo*, which is now available on CRAN and that is programmed in C; its algorithm is similar to that used in REALCOM, with some practically important extensions. The first is to allow for an imputation model with study-specific covariance matrices, in order to respect heterogeneity between different aggregated studies. The second main extension is the possibility to impute systematically missing variables.

Inevitably, there are still a number of developments that could be added to make *jomo* more flexible:

- At the moment, it is possible to use multilevel imputation models, but only level 1 variables can be imputed. We plan to include functions for the imputation of level 2 variables soon after the completion of the Ph.D;

- *jomo* treats all categorical variables as nominal using the latent normal model. A possible extension in future will be to include an algorithm for ordered categorical variables as well, similar to that used in REALCOM and introduced in Section 8.4;

- *jomo* still lacks a principled approach to face perfect prediction issues. We are planning to include one of the approaches presented in Chapter 8 to address this;

- In the case of an imputation model with random study-specific covariance matrices, the assumption of the inverse-Wishart distribution for the covariance matrices is a good starting point, since it is the conjugate distribution for the covariance matrix of multivariate normally distributed data, and is likely to be a sensible choice in many situations, as we illustrated throughout this thesis. However it may well not be the best option in all scenarios. Different assumptions, e.g. based on the multivariate-t distribution, may be investigated and made available in *jomo*;

- Functions for substantive model compatible imputations (Section 9.3.3) would be a great addition. Currently the *jomo* package assumes that partially observed variables are not included in interactions or non-linearities in the substantive model. We return to this in Section 10.5.

The first two points do not present theoretical difficulties, since they have been already implemented in REALCOM. However, there may be computational challenges implementing them efficiently. The last two entail novel strategies and accordingly will need to be evaluated using a simulation study.

Another important on-going aspect of this project consists in continuing outreach activities to make researchers more aware of the availability of our software. In order to do this, we have drafted a paper for the R Journal and we presented results from this thesis at the ISCB and IBC conferences.

10.3 Imputing IPD-MA: joint multivariate normal model

Having written and submitted the new package to the Comprehensive R Archive Network (CRAN), the next part of my Ph.D involved an intensive simulation study to test the performance of our proposed imputation method to address missing data issues in IPD-MA. We focused initially on the situation where a joint model was easiest to find, i.e. when partially observed data are approximately multivariate normally distributed.

In this setting, first of all we compared our method to the standard method of imputing with FCS within the contributing studies; to do this, we used the same simulation settings used in (Burgess *et al.*, 2013), showing that our approach works at least as good as theirs in the settings considered in their paper. Then, we explored some more challenging situations, in particular the presence in the meta-analysis of studies with a very small number of observations, and even the imputation of systematically missing variables.

We demonstrated how, when an IPD-MA contains studies with few observations, our imputation strategy is able to mitigate the noise by sharing information across studies, at the same time dealing properly with the heterogeneity between the different studies. Situations like the one considered in Subsection 7.3.4, where in some studies just over 5 observations were

considered, might seem extreme but indeed in one of our motivating IPD-MA, MAGGIC, one of the contributing studies recorded information only for five patients for some of the variables included in the analysis model.

The method gave good results even when imputing systematically missing variables, both in the case of fixed and random covariance matrices, particularly when the assumption regarding the distribution of the covariance matrices was met by the data.

In real data, we would not know whether the common or study-specific covariance matrix assumption holds; however, the simulation study demonstrated how, wrongly assuming that there is a common covariance matrix usually leads to biased results, while wrongly assuming there are study-specific covariance matrices usually leads only to a slight overestimation of the standard errors, i.e. mildly conservative inference. Therefore in applications, we recommend using imputation models with study-specific covariance matrices, since in IPD-MA it is usually the most plausible assumption. Partially observed variables then enter in the multivariate imputation model as responses, while completely observed ones enter the model as covariates.

When the joint distribution is not plausibly multivariate normal, we may consider a transformation. For example we could use a Box-Cox transformation, the parameters of which are updated via a further MCMC step, as suggested in (Goldstein *et al.*, 2009). However, a simpler approach, that is likely to be just as effective, would be to use the observed values to identify marginal transformations to a normal scale.

10.4 Imputing IPD-MA: extension to different data types

Having demonstrated the utility of our approach for continuous data, the next step in our thesis was to try to extend the method to cases where the multivariate normal joint distribution was not directly usable for the joint model, starting with the case of partially observed binary or categorical data. We showed how the latent normal variables approach proposed in (Goldstein *et al.*, 2009), implemented in REALCOM and which we implemented much more efficiently in *jomo*, can be used to impute consistently partially observed nominal variables, making use of a multivariate multinomial probit model. The very substantial increase in computational speed of *jomo* with respect to REALCOM allowed us to undertake – for the first time – a comprehensive simulation study to explore its performance in a range of practically relevant situations. We found that the approach gave good results when data were generated with many different distribution; the only exception was when data were generated from the general location model and effect sizes were strong. For this situation, we found a condition for the approximate validity of the latent normal imputation model; we further proposed – and are currently implementing – a solution for when this condition does not hold.

In general, the conclusions from the previous chapters about the choice of the imputation model hold in this case too; in other words, a multivariate model with all partially observed variables as outcomes and with study-specific covariance matrices is preferable in the IPD-MA setting.

Finally we undertook smaller simulation studies to explore the behaviour of our joint imputation model with ordered categorical variables and count data. For the former case, we showed how using the latent normal variables approach and treating the ordered categorical variables

as unordered, still gives good results. The main advantage of the specific method for ordinal variables introduced in REALCOM appears therefore to be that it requires far fewer latent normals (one for each ordinal variable, instead of $n-1$, with n being the number of levels of the variable). This is likely to be practically important for sparse data, so we intend to implement it in *jomo* in due course. For Poisson data, we showed the results of transforming the data with either a square root or a logarithmic transformation, obtaining better results with the second, but still observing bad results in some situations, e.g. mainly when the Poisson rate is quite small, resulting in a large number of zero counts.

10.5 Factorization of the Joint Model

When facing the problem of interactions in the substantive model, and the issue of the difference between the general location model and the latent normal model for categorical data, we ended up proposing a similar solution based on an appropriate factorization of the joint distribution in two (or potentially more) components:

$$f(\mathbf{Y}_1, \mathbf{Y}_2) = f(\mathbf{Y}_1|\mathbf{Y}_2)f(\mathbf{Y}_2).$$

Let $\boldsymbol{\theta}$ denote the set of parameters underlying the conditional distribution $f(\mathbf{Y}_1|\mathbf{Y}_2)$ and $\tilde{\boldsymbol{\theta}}$ the set of parameters underlying the marginal distribution $f(\mathbf{Y}_2)$. We proposed this factorization strategy when we had:

1. Categorical covariates X for which $\frac{\beta_X}{\sigma_{resid}} > 2$: in this case a marginal multinomial distribution could be used for X, and the usual multivariate normal model for the other variables in the model;

2. Missing covariates in survival models: covariates may be assumed to follow the usual multivariate normal model, drawing the new imputed values consistently with the survival model of substantial interest;

3. Interactions between partially observed variables: we could set up two different multivariate normal models for variables with or without interactions in the analysis model;

4. Partially observed variables involved with non-linearities in the analysis model: again, similarly to the case of interactions, we could just set up two different multivariate normal models. It is important to stress here that in this case, for non-linearities we intend the presence of a non-linear component of the partially observed data in the linear model, and not a non-linearity in the parameters like we usually assume in statistical models.

The algorithm to run our Gibbs sampler in all of these cases is similar to the ones proposed in Subsections 9.3.1 and 9.3.3, consisting of three essential steps:

1. Update $\boldsymbol{\theta}$ and the missing values in $\mathbf{Y}_1$. In most cases this is done with a regular Gibbs sampling step where we assume a (partially latent) multivariate normal model for $f(\mathbf{Y}_1|\mathbf{Y}_2)$ considering Y_2 as covariate(s), so there are no theoretical differences from the situations we considered in Chapters 7 and 8. The only exception is the survival model case, where we already said (9.3.1) that a fully Bayesian approach should be performed to update $\boldsymbol{\theta}$;

2. Draw the new values for the missing data in $\mathbf{Y}_2$ from an appropriate distribution compatible with our substantive model. Since this will not be possible directly, two options are available for performing this step: a Metropolis-Hastings procedure and rejection sampling;

3. Update $\tilde{\boldsymbol{\theta}}$ as usual, with the Gibbs sampling steps for the marginal multivariate normal distribution of $f(\mathbf{Y}_2)$.

The only step presenting substantial differences from what we have presented so far, is the second one; as we said, there are (at least) two options to impute missing Y_2 values coherently with the substantive model: Metropolis-Hastings and rejection sampling. We illustrated in Subsection 9.3.1 the Metropolis-Hastings step in the case of survival analysis and similar reasoning is required to build the same step for the other three cases. However, the problem with the MH algorithm is that we are not necessarily imputing new values at each iteration of the sampler, as the proposals may be rejected. Therefore, it may be necessary to 'tune' the parameters of the proposal distribution in an initial adaptive phase.

For this reason, another strategy, making use of rejection sampling, has been proposed recently (Bartlett *et al.*, 2014). If we wanted to draw the missing values for $\mathbf{Y}_2$ from the actual distribution f, but this does not have a known distribution form, we may use g as a proposal distribution. Given the appropriate bound M over values of y for the ratio $\frac{f(y)}{g(y)}$, the algorithm:

- draws a proposed value y_2 from g;
- Accepts y_2 with probability $\frac{f(y_2)}{Mg(y_2)}$;
- If y_2 was rejected at the previous step, draws a new proposal until it is accepted.

Using this algorithm, it is in principle always possible to draw a new value for all the missing data at each iteration of the sampler; however, in some cases, a large rejection rate can result in many proposals being required.

(Bartlett *et al.*, 2014) proposed and tested this method, opting for the FCS approach to imputation. A Stata and an R package, both named *smcfcs*, are available to use it. Their results seem to be promising and therefore we aim to create the functions to implement this approach in the joint modelling framework, in order to be able to use it in the multilevel IPD-MA situation as well.

In conclusion, writing some functions for substantive model compatible imputation through factorization of the joint distribution, we could solve lots of different issues; we therefore believe that this is the next important extension to be added to *jomo* which we hope to implement as part of a post-doc research project.

10.6 Conclusion

Handling missing data correctly is one of the main challenges faced by an analyst wishing to perform an Individual Patient Data Meta-Analysis aggregating thousands of observations from diverse studies. Such an analysis may have as its main goal estimates of treatment effects, or estimation of a prognostic model.

Motivated by such data, in this thesis we have argued for, developed, implemented and finally evaluated an approach based on Joint Modelling Imputation. The approach provides a flexible strategy for imputing missing values, sharing information across studies (clusters), while at the same time respecting the heterogeneity between contributing studies. The results of the extensive simulation studies we have performed suggest the approach is reliable and robust. Moreover, using the software we have written it is –for the first time – computationally feasible in datasets of the size that typically occur. Further, we have described a unified, computationally feasible strategy for addressing the remaining issues.

Other methods were developed recently with the aim of handling missing data in IPD-MA. Two important examples of strategies to handle systematically missing covariates were the studies by (Fibrinogen Studies Collaboration, 2009) and (Jolani *et al.*, 2015). Both methods are based on FCS imputation; the method proposed in (Fibrinogen Studies Collaboration, 2009) can deal only with continuous data and (Jolani *et al.*, 2015) extended the method to the case of a mix of categorical and continuous data; both methods do not deal with heteroscedasticity in the analysis model but have been proven to handle really well systematically missing predictors in homoscedastic generalized linear mixed models. However, these approaches do not cover the common situation of a mix of sporadically and systematically missing data.

An issue that was not covered in this thesis, is that of wholly missing studies (Ahmed *et al.*, 2012; Riley *et al.*, 2008). These are studies that were possibly performed, but whose results were not published, most likely because significant results were not found. This may lead to the so called publication bias. Extending our approach to address this issue might be an interesting plan of future research.

Another thing we did not consider in our thesis is the model fit evaluation; if we are unsure about the model we want to fit on the data and we want to check and compare model fits for different models, in principle we should use a different, congenial, imputation model for each analysis model we want to compare. However, even though including auxiliary variables in the imputation model make Rubin's variance rules invalid, the amount of information usually recovered by including these additional variables usually overshadows the overestimation of the variances with Rubin's rules and therefore it is usually recommended to use as many variables as possible in the imputation model; therefore, it is possible to use the most general imputation model and then to check model fit using different models on the same imputed datasets.

Though our Multilevel MI method has been proposed in this thesis for the particular situation of missing data in IPD-MA, it could be applied whenever we have any clustered data, from longitudinal studies to cluster-randomized trials and E-health data. Also, it [our approach] could be easily extended to impute data with a multilevel structure with more than two levels, simply by defining the joint multi-level imputation model accordingly; an example of this could be a dataset where children may be clustered both in classes and in schools. However, when we do not have missing variables at the top level or systematically missing variables at level-2, e.g. in the previous example when we do not have missing school-level variables or systematically missing class-level variables, it might be enough to impute separately the missing data with a different imputation model for each level-2 cluster, e.g. to impute missing data for pupils with a different imputation model within each school, using a 2-level model clustering by class only.

In conclusion, we believe this method can improve the way many IPD-MA with missing data are analysed and for this reason, in order to maximize the impact of our research, we made available and are continuing to support and develop an R package that we hope will contribute in improving the quality of meta-analyses.

Bibliography

Abo-Zaid, G., Guo, B., Deeks, J. J., Debray, T. P., Steyerberg, E. W., Moons, K. G. and Riley, R. D. (2013) Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol*, **66**(8), 865–873.

Ahmed, I., Sutton, A. J. and Riley, R. D. (2012) Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ*, **344**.

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

Bartlett, J. W., Seaman, S. R., White, I. R. and Carpenter, J. R. (2014) Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*.

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. (2010) A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*, **1**(2), 97–111.

Brockwell, S. E. and Gordon, I. R. (2001) A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, **20**(6), 825–840.

Brown, C. (2012) *dummies: Create dummy/indicator variables flexibly and efficiently.* R package version 1.5.6.

Browne, W. J. (2006) MCMC algorithms for constrained variance matrices. *Computational Statistics and Data Analysis*, **50**(7), 1655 – 1677.

Burgess, S., White, I. R., Resche-Rigon, M. and Wood, A. M. (2013) Combining multiple imputation and meta-analysis with individual participant data. *Stat Med*, **32**(26), 4499–4514.

Burke, D. L., Bujkiewicz, S. and Riley, R. D. (2016) Bayesian bivariate meta-analysis of correlated effects: Impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences. *Statistical Methods in Medical Research.*

Caldwell, D. M. (2014) An overview of conducting systematic reviews with network meta-analysis. *Syst Rev*, **3**, 109.

Carpenter, J. and Kenward, M. (2013) *Multiple Imputation and its Application.* Wiley. ISBN: 978-0-470-74052-1.

Carpenter, J., Kenward, M. and White, I. (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research.*, **16**(3), 259–275. DOI: 10.1177/0962280206075303.

Carpenter, J., Goldstein, H. and Kenward, M. (2011) Realcom-impute software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software.*, **45**(5), 1–14.

Carpenter, J., Goldstein, H. and Kenward, M. (2012) Statistical modelling of partially observed data using multiple imputation: Principles and practice. In *Modern Methods for Epidemiology* (Eds Y.-K. Tu and D. C. Greenwood), pp. 15–31. Springer Netherlands.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**(2), 347–361.

Clayton, D. G. (1991) A monte carlo method for bayesian inference in frailty models. *Biometrics*, **47**(2), pp. 467–485.

Cornell, J. E., Mulrow, C. D., Localio, R., Stack, C. B., Meibohm, A. R., Guallar, E. and Goodman, S. N. (2014) Random-effects meta-analysis of inconsistent effects: A time for change. *Annals of Internal Medicine*, **160**(4), 267–270.

Debray, T. P., Moons, K. G., Abo-Zaid, G. M., Koffijberg, H. and Riley, R. D. (2013) Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS ONE*, **8**(4), e60650.

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

DerSimonian, R. and Kacker, R. (2007) Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, **28**(2), 105 – 114.

DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled clinical trials.*, **7**(3), 177–188. DOI: 10.1016/0197-2456(86)90046-2.

Easterbrook, P. J., Berlin, J. A., Gopalan, R. and Matthews, D. R. (1991) Publication bias in clinical research. *Lancet*, **337**(8746), 867–872.

Eddelbuettel, D. and Francois, R. (2011) Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, **40**(1), 1–18.

Fibrinogen Studies Collaboration (2009) Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine.*, **28**(8), 1218–1237. DOI: 10.1002/sim.3540.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014) *Bayesian data analysis*, volume 2. Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741.

Goldstein, H. (2011) *Multilevel Statistical Models.* Wiley. 4th Edition.

Goldstein, H. (2014) *Heteroscedasticity and Complex Variation.* John Wiley & Sons, Ltd.

Goldstein, H., Carpenter, J., Kenward, M. and Levin, K. (2009) Multilevel models with multivariate mixed response types. *Statistical Modelling.*, **9**(3), 173–197. DOI: 10.1177/1471082X0800900301.

Goldstein, H., Carpenter, J. and Browne, W. (2014) Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A.*, **177**(2), 553–564. DOI: 10.1111/rssa.12022.

Gueyffier, F., Boutitie, F., Boissel, J. P., Coope, J., Cutler, J., Ekbom, T., Fagard, R., Friedman, L., Perry, H. M. and Pocock, S. (1995) INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. *Therapie*, **50**(4), 353–362.

Hardy, R. J. and Thompson, S. G. (1996) A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, **15**(6), 619–629.

Hartung, J. and Knapp, G. (2001) A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, **20**(24), 3875–3889.

Higgins, J. P., Thompson, S. G., Deeks, J. J. and Altman, D. G. (2003) Measuring inconsistency in meta-analyses. *BMJ*, **327**(7414), 557–560.

Higgins, J. P. T. and Green, S. (eds) (2008) *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, fifth edition.

Higgins, J. P. T. and Thompson, S. G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**(11), 1539–1558.

Holford, T. R. (1980) The analysis of rates and of survivorship using log-linear models. *Biometrics*, **36**(2), pp. 299–305.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K. and Sterne, J. A. (2014) Joint modelling rationale for chained equations. *BMC Med Res Methodol*, **14**, 28.

Hunter, J. and Schmidt, F. (1990) *Methods of meta-analysis: correcting error and bias in research findings*. Sage Publications.

Jolani, S., Debray, T. P., Koffijberg, H., van Buuren, S. and Moons, K. G. (2015) Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*, **34**(11), 1841–1863.

Koopman, L., van der Heijden, G. J., Grobbee, D. E. and Rovers, M. M. (2008) Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. *Am. J. Epidemiol.*, **167**(5), 540–545.

Laird, N. and Olivier, D. (1981) Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, **76**(374), 231–240.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963–974.

Lee, K. J. and Thompson, S. G. (2008) Flexible parametric models for random-effects distributions. *Statistics in Medicine*, **27**(3), 418–434.

Liu, J., Gelman, A., Hill, J., Su, Y.-S. and Kropko, J. (2013) On the stationary distribution of iterative imputations. *Biometrika*.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(2), pp. 226–233.

Mason, A., Richardson, S. and Best, N. (2012) Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods. *Journal of Official Statistics.*, **28**(2), 279–302.

Mavridis, D. and Salanti, G. (2013) A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research*, **22**(2), 133–158.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), pp. 109–142.

Meinfelder, F. (2011) *BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and single imputation for discrete data*. R package version 0.1-6.

Meng, X.-L. (1994) Multiple Imputation inferences with uncongenial sources of input. *Statist. Sci.*, **9**(4), 566–573.

Molenberghs, G., Michiels, B., Kenward, M. G. and Diggle, P. J. (1998) Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, **52**(2), 153–161.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A. and Verbeke, G. (2014) *Handbook of Missing Data Methodology*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

Olkin, I. and Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, **32**(2), 448–465.

Pocock, S. J., Ariti, C. A., McMurray, J. J., Maggioni, A., Køber, L., Squire, I. B., Swedberg, K., Dobson, J., Poppe, K. K., Whalley, G. A. and Doughty, R. N. (2013) Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur. Heart J.*, **34**(19), 1404–1413.

Quartagno, M. and Carpenter, J. (2014) *jomo: A package for Multilevel Joint Modelling Multiple Imputation*.

Quartagno, M. and Carpenter, J. R. (2015) Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine, in press*.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team (2015) *R language.* R Foundation for Statistical Computing, Vienna, Austria.

Resche-Rigon, M., White, I., Bartlett, J., Peters, S. and Thompson, S. (2013) Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine.*, **32**(28), 4890–905. DOI: 10.1002/sim.5894.

Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L. and Boutitie, F. (2008) Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, **27**(11), 1870–1893.

Riley, R. D., Lambert, P. C. and Abo-Zaid, G. (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, **340**, c221.

Ripley, B. (2008) *Stochastic Simulation.* Wiley.

Rubin, D. (1976) Inference and missing data. *Biometrika.*, **63**(3), 581–592. DOI:10.1093/biomet/63.3.581.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schafer, J. (1997) *Analysis of Incomplete Multivariate Data.* Chapman and Hall.

Schafer, J. L. (2015) *mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data.* R package version 1.0-9.

Schafer, J. L. and Graham, J. W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, **7**, 147–177.

Seaman, S., Bartlett, J. and White, I. (2012) Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, **12**(1), 46.

Shrier, I., Boivin, J.-F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., Brophy, J. and Rossignol, M. (2007) Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? a critical examination of underlying principles. *American Journal of Epidemiology*, **166**(10), 1203–1209.

Sidik, K. and Jonkman, J. N. (2005) Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(2), 367–384.

Simmonds, M. C., Higgins, J. P., Stewart, L. A., Tierney, J. F., Clarke, M. J. and Thompson, S. G. (2005) Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials*, **2**(3), 209–217.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Spratt, M., Carpenter, J., Sterne, J. A. C., Carlin, J. B., Heron, J., Henderson, J. and Tilling, K. (2010) Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology*, **172**(4), 478–487.

Stanley, T. D. and Jarrell, S. B. (1989) Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys*, **3**(2), 161–170.

Stata (2013) *Stata Statistical Software: Release 13*. StataCorp., College Station, TX: StataCorp LP.

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C. and Stewart, L. A. (2012) Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS ONE*, **7**(10), e46042.

Stewart, L. A. and Parmar, M. K. (1993) Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*, **341**(8842), 418–422.

Ted Harding, F. T. and Schafer, J. L. (2012) *cat: Analysis of categorical-variable datasets with missing values*. R package version 0.0-6.5.

van Buuren, S. (2011) *The Handbook of Advanced Multilevel Analysis.*, chapter Multiple imputation of multilevel data, pp. 173–196. Milton Park,UK: Routledge.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Von Hippel, P. T. (2009) How to impute interactions, squared and other transformed variables. *Sociological Methodology*, **39**(1), 265–291.

White, I. R. and Royston, P. (2009) Imputing missing covariate values for the Cox model. *Stat Med*, **28**(15), 1982–1998.

White, I. R., Daniel, R. and Royston, P. (2010) Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal*, **54**(10), 2267–2275.

Yucel, R. (2011) Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling.*, **11**(4), 351–370. DOI: 10.1177/1471082X100110040.

Zhao, J. and Schafer, J. (2013) *pan: Multiple Imputation for multivariate panel or clustered data*. R Foundation for Statistical Computing. R Package, version 0.9.

Part V

Appendices

A

MCMC Algorithms for Multilevel Models

A.1 MCMC Algorithm for common covariance matrix

In this appendix we present the algorithm for joint modelling imputation of hierarchical data. We begin with the algorithm for fitting the MCMC with a model with a common covariance matrix across studies (clusters).

Henceforth, i will index patients and j studies (or general level 2 units). $\mathbf{Y}$ is a matrix whose p-dimensional row vectors are the set of partially observed variables for the different units in the study; $\mathbf{X}$ is a matrix whose r-dimensional rows are the observations on completely observed variables whose effects are fixed across clusters and $\mathbf{Z}$ is the design matrix for the random effects, i.e. it will generally, but not necessarily, be a subset of $\mathbf{X}$ and its rows will be q-dimensional.

The general formulation of the joint multivariate normal model for the partially observed data is:

$$\begin{aligned} \mathbf{Y}_{i,j} &= (\mathbf{I}_p \otimes \mathbf{X}_{i,j})\boldsymbol{\beta} + (\mathbf{I}_p \otimes \mathbf{Z}_{i,j})\mathbf{u}_j + \mathbf{e}_{i,j} \\ \mathbf{u}_j &\sim N(0, \Omega_u) \\ \mathbf{e}_{i,j} &\sim N(0, \Omega_e), \end{aligned} \tag{A.1.1}$$

where $\otimes$ is the Kronecker product, defined as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix},$$

for an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$.

In the above model we have the following parameters:

- the vector of fixed effects $\boldsymbol{\beta}$;
- the J vectors of random effects $\mathbf{u}_j$;
- the level 1 residuals $\mathbf{e}_{i,j}$;
- the level 1 covariance matrix Ω_e;
- the level 2 covariance matrix Ω_u.

Furthermore, in the Gibbs sampler we treat missing data as additional parameters.

We then need to choose the priors for all these parameters. We decided to take flat priors for all the parameters in the model apart from the two covariance matrices; for these, we chose an inverse-Wishart prior instead, mainly because it is the conjugate distribution.

Starting values for all the parameters were chosen as well: matrices of zeros for fixed and random effects parameters, identity matrices for covariance matrices and random draws from a normal distribution with mean set to the mean of the observed cases for the missing data.

We are now ready to outline the general algorithm to fit the above model. At each iteration of the sampler we have to update the value of each parameter from the appropriate conditional distribution, conditioning on all the observed data and the current values for all the other parameters:

1. Update the fixed effect parameters vector $\boldsymbol{\beta}$ drawing from the following pr-variate normal distribution:

$$\boldsymbol{\beta} \sim N(\mu_\beta, \boldsymbol{\Omega}_\beta)$$

$$\mu_\beta = \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right) \right)^{-1} \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{Y}_{i,j} - \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right) \mathbf{u}_j \right) \right)$$

$$\boldsymbol{\Omega}_\beta = \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right) \right)^{-1}$$

2. Update the random effects parameters $\mathbf{u}_j$, for $j = 1, \ldots, J$, drawing each from the following pq-variate normal distribution:

$$\mathbf{u}_j \sim N(\mu_{uj}, \boldsymbol{\Omega}_{uj})$$

$$\mu_{uj} = \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right) + \boldsymbol{\Omega}_u^{-1} \right)^{-1} \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{Y}_{i,j} - \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right) \boldsymbol{\beta} \right) \right)$$

$$\boldsymbol{\Omega}_{uj} = \left(\sum_{i,j} \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right)^T \boldsymbol{\Omega}_e^{-1} \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right) + \boldsymbol{\Omega}_u^{-1} \right)^{-1}$$

3. Calculate level 1 residuals $\mathbf{e}_{i,j}$:

$$e_{i,j} = Y_{i,j} - \left(\mathbf{I}_p \otimes \mathbf{X}_{i,j} \right) \boldsymbol{\beta} - \left(\mathbf{I}_p \otimes \mathbf{Z}_{i,j} \right) \mathbf{u}_j$$

4. Draw the new level 2 covariance matrix Ω_u from:

$$\mathbf{\Omega}_u \sim W^{-1}(\nu_u, \mathbf{S}_u)$$

$$\nu_u = J + \nu_{u,prior}$$

$$\mathbf{S}_u = \left(\sum_j \mathbf{u}_j \mathbf{u}_j^T + \mathbf{S}_{u,prior} \right)^{-1}$$

5. Draw the new level 1 covariance matrix Ω_e from:

$$\mathbf{\Omega}_e \sim W^{-1}(\nu_e, \mathbf{S}_e)$$

$$\nu_e = I + \nu_{e,prior}$$

$$\mathbf{S}_e = \left(\sum_{i,j} \mathbf{e}_{i,j} \mathbf{e}_{i,j}^T + \mathbf{S}_{e,prior} \right)^{-1}$$

6. Draw the new imputed values for all the missing data. This is done by drawing for each observation with n_{miss} missing data from the appropriate conditional n_{miss}-variate normal distribution.

When handling binary or categorical data with latent normal variables, a further step is the rejection sampling procedure to draw new values for the latent normals. These are simply drawn from the appropriate conditional multivariate normal distribution, but then they are accepted or rejected whether or not they satisfy the conditions presented in Subsection 3.3.1, i.e. for an N-level variable, accept a draw where all the $N-1$ latent normals are negative if the observed category is N, accept a draw where the maximum is for latent normal K and is greater than zero if the observed category is K.

Again, as discussed in Subsection 3.3.1 and throughout Chapter 8, the step for updating level 1 covariance matrix is slightly different because of the constraints imposed and will require an element-wise update of the covariance matrix with a Metropolis-Hastings step.

A.2 MCMC algorithm for random study-specific covariance matrices

In this appendix, we describe the MCMC algorithm for random level 1 covariance matrices introduced by (Yucel, 2011), and slightly modified to correct minor errors.

The basic assumption behind this method is that level 1 precision matrices for each level 2 unit are drawn from a Wishart distribution and thus the inverse of study-specific level 1 covariance matrices $\Omega_{1,j}$ have marginal distribution:

$$(\Omega_{1,j})^{-1} \sim W(a, A)$$

where a is the degrees of freedom and A the scale matrix.

Now, we consider the general algorithm for MCMC multilevel imputation presented in the previous section and we see the differences introduced in this algorithm by this further assumption. Firstly, the common covariance matrix Ω_1 needs to be substituted in the algorithm with the proper study-specific $\Omega_{1,j}$. Secondly, we have two more parameters to deal with, a

and A. So, we need to specify priors for these parameters; a simple choice might be:

$$a \sim \chi^2_\eta$$

$$A^{-1} \sim W(\gamma, \Gamma)$$

where η, γ and Γ need to be chosen by the analyst. Following Yucel, we choose in our analyses to set $\eta = p_1$, where $p_1 = dim(\Omega_{1,j})$, while for γ and Γ we choose the least informative values possible, that is $\gamma = p_1$ and $\Gamma = I$.

Then, we need to draw a new value for A^{-1}. We do know that the conditional distribution of A^{-1} given all the $\Omega_{1,j}$ and a is Wishart:

$$A^{-1} \sim W\left(\gamma + aJ, \left(\Gamma^{-1} + \sum_j \Omega_{1,j}^{-1}\right)^{-1}\right)$$

where J represents the number of level 2 units. Unfortunately, we don't know the form of the distribution of a given $\Omega_{1,j}$ and therefore we are forced to use a Metropolis Hastings step in order to draw the new value of a at each step of the MCMC.

We know that the density of a is proportional to:

$$f(a) \propto f_1(a)\Gamma_p(\frac{a}{2})^{-J}\left(\sum_{i=1}^{J}|\Omega_j|\right)^{-\frac{a+p+1}{2}}$$

$$\Gamma_{\frac{\gamma+aJ}{2}}|\Gamma^{-1} + \Omega_1^{-1} + \ldots + \Omega_J^{-1}|^{-\frac{aJ+\gamma}{2}}$$

where $f_1(a)$ is the prior distribution for a, i.e. χ^2_η, and Γ_p is the p-variate gamma function. Since a has a skewed distribution, then Yucel suggested to use a transformed variable u such that:

$$u = \log(a + p_1)$$

$$f_U(u) = f\left(\exp(u) - p_1\right) \left| \frac{\partial a}{\partial u} \right|$$

We then choose as proposal density a t_4 distribution centred at u_m, i.e. the mode of $f_U(u)$, that we calculate using Newton-Raphson search. The density of such distribution is:

$$h(u) \propto \left(1 + \frac{(u - u_m)^2}{4\lambda^2}\right)^{-\frac{5}{2}}$$

where we choose λ in order to match the curvature of f_U:

$$\lambda = \sqrt{-\frac{5}{4\frac{\partial^2 f_U(u)}{\partial_u^2}\big|_{u=u_m}}}$$

Then, our Metropolis-Hastings sampler accepts a proposed u^* with probability:

$$\min\left(1, \frac{f(u^*)h(u)}{f(u)h(u^*)}\right).$$

Lastly, given the new draws of a and A, we update the J covariance matrices from the correspondent distributions:

$$\Omega_{1,j}^{-1} \sim W\left(a + I_j, W_j^{-1}\right),$$

where:

$$W_j = A^{-1} + \sum_{i=1}^{I} (\mathbf{Y}_{i,j}^{(1)} - \mathbf{X}_{i,j}^{(1)}\beta^{(1)} - \mathbf{Z}_{i,j}^{(1)}\mathbf{u}_j^{(1)})^T (\mathbf{Y}_{i,j}^{(1)} - \mathbf{X}_{i,j}^{(1)}\beta^{(1)} - \mathbf{Z}_{i,j}^{(1)}\mathbf{u}_j^{(1)}).$$

Where I_j is the number of level 1 units within each level 2 cluster, $\mathbf{Y}_{i,j}^{(1)}$ is the vector of the outcomes for subject i in cluster j, $\mathbf{X}_{i,j}^{(1)}$ and $\mathbf{Z}_{i,j}^{(1)}$ are the covariates for the fixed and random effects and $\boldsymbol{\beta}^{(1)}$ and $\mathbf{u}_j^{(1)}$ are the vectors of fixed and random effects coefficients.

B

Further Results of Youth Cohort Study Analysis

B.1 Further Results of Youth Cohort Study Analysis

Below we present some further results comparing REALCOM and our first mex program using the YCS dataset as described in Chapter 6.

	REALCOM	Mex
$\hat{\beta}_1$	39.683	39.686
$\hat{\beta}_2$	0.103	0.101
$\hat{\beta}_3$	-0.065	-0.0626

Table B.1: Posterior Means for the three coefficients β in model 1 of Table 2.4

	REALCOM	Mex
$\hat{\beta}_1$	39.716	39.715
$\hat{\beta}_2$	1.907	1.901
$\hat{\beta}_3$	-0.899	-0.901
$\hat{\beta}_4$	-0.675	-0.676
$\hat{\beta}_5$	-0.840	-0.841
$\hat{\beta}_6$	-1.244	-1.249
$\hat{\beta}_7$	-1.053	-1.056

Table B.2: Posterior Means for the seven coefficients β in model 2 of Table 2.4

	REALCOM	Mex
$\hat{\beta}_1$	39.681	39.685
$\hat{\beta}_2$	0.099	0.101
$\hat{\beta}_3$	-0.066	-0.061
$\hat{\beta}_4$	1.906	1.899
$\hat{\beta}_5$	-0.903	-0.908
$\hat{\beta}_6$	-0.675	-0.680
$\hat{\beta}_7$	-0.843	-0.847
$\hat{\beta}_6$	-1.248	-1.258
$\hat{\beta}_7$	-1.053	-1.054

Table B.3: Posterior Means for the nine coefficients β in model 3 of Table 2.4

		REALCOM	Mex
$\hat{\beta_{i,1}}$	$\hat{\beta_{0,1}}$	32.621	32.621
	$\hat{\beta_{1,1}}$	2.612	2.621
	$\hat{\beta_{2,1}}$	2.933	2.933
$\hat{\beta_{i,2}}$	$\hat{\beta_{0,2}}$	0.058	0.060
	$\hat{\beta_{1,2}}$	-0.058	-0.058
	$\hat{\beta_{2,2}}$	0.036	0.036
$\hat{\beta_{i,3}}$	$\hat{\beta_{0,3}}$	-0.078	-0.074
	$\hat{\beta_{1,3}}$	0.018	0.018
	$\hat{\beta_{2,3}}$	0.001	0.001
$\hat{\beta_{i,4}}$	$\hat{\beta_{0,4}}$	2.141	2.142
	$\hat{\beta_{1,4}}$	-0.057	-0.053
	$\hat{\beta_{2,4}}$	-0.097	-0.100
$\hat{\beta_{i,5}}$	$\hat{\beta_{0,5}}$	-0.894	-0.894
	$\hat{\beta_{1,5}}$	0.062	0.065
	$\hat{\beta_{2,5}}$	-0.017	-0.019
$\hat{\beta_{i,6}}$	$\hat{\beta_{0,6}}$	-0.564	-0.556
	$\hat{\beta_{1,6}}$	-0.064	-0.064
	$\hat{\beta_{2,6}}$	-0.034	-0.038
$\hat{\beta_{i,7}}$	$\hat{\beta_{0,7}}$	-0.705	-0.707
	$\hat{\beta_{1,7}}$	-0.095	-0.092
	$\hat{\beta_{2,7}}$	-0.038	-0.039
$\hat{\beta_{i,8}}$	$\hat{\beta_{0,8}}$	-1.111	-1.119
	$\hat{\beta_{1,8}}$	-0.028	-0.014
	$\hat{\beta_{2,8}}$	-0.056	-0.058
$\hat{\beta_{i,9}}$	$\hat{\beta_{0,9}}$	-0.917	-0.914
	$\hat{\beta_{1,9}}$	-0.111	-0.101
	$\hat{\beta_{2,9}}$	-0.034	-0.038

Table B.4: Posterior Means for the twenty-seven coefficients β in model 4 of Table 2.4