DOCTORAATSPROEFSCHRIFT

2009 | School voor Levenswetenschappen

Topics in Analysis and Sensitivity Analysis for Incomplete Longitudinal Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

Cristina SOTTO

Promotor: prof. dr. Geert Molenberghs
Copromotor: dr. Caroline Beunckens

universiteit hasselt
I-BioStat
KATHOLIEKE UNIVERSITEIT
LEUVEN

D/2009/2451/42

Topics in Analysis and Sensitivity Analysis for Incomplete Longitudinal Data

Cristina Sotto

Promotor: Prof. dr. Geert Molenberghs

Co-Promotor: dr. Caroline Beunckens

Samenvatting

Wetenschappelijk onderzoek steunt in vele gevallen op empirisch, op gegevens gebaseerd onderzoek. Dergelijke gegevens zijn meer dan eens onvolledig. Experimenteel of klinisch onderzoek, bijvoorbeeld, is meestal gericht op het verzamelen van volledige gegevens, doch kan uitmonden in verlies of beschadiging van experimentele eenheden, uitval of overlijden van proefpersonen. Een ander voorbeeld wordt gevormd door respondenten in een survey die weigeren om bepaalde vragen te beantwoorden of zonder meer niet aan het onderzoek wensen deel te nemen. Wanneer vervanging van de ontbrekende gegevens niet mogelijk is, dient de analyse noodzakelijkerwijze te gebeuren op basis van onvolledige gegevens. Dit kan leiden tot de aanpassing van het oorspronkelijke statistisch analyseplan om er op die manier voor te zorgen dat de studie vooralsnog tot een geldig antwoord op de onderzoeksvraag leidt.

Observationale en experimentele studies nemen hetzij een cross-sectionele hetzij een longitudinale vorm aan. Waar de eerste vorm leidt tot het verzamelen van één enkele responsveranderlijke, geeft de tweede aanleiding tot het verzamelen van een reeks metingen, gespreid in de tijd. In beide types kunnen de gegevens onvolledig zijn. De structuur van de onvolledige gegevens neemt over het algemeen één van twee mogelijke vormen aan. In een longitudinale studie, waar een bepaalde responsveranderlijke herhaald in de tijd gemeten wordt, definieert de volgorde van de metingen op een natuurlijke manier de patronen. In voorkomend geval spreken we vaak van uitval (*dropout*), monotoon ontbreken, of attritie; dus, wanneer alle metingen volledig zijn tot op een bepaald punt in de tijd, waarna de metingen ontbreken. Monotone patronen kunnen ook in andere situaties voorkomen, ook al is het concept vrij natuurlijk geassocieerd aan herhaalde metingen. Daarnaast hebben we ook niet-monotone ontbrekende waarden, waarbij de patronen geen natuurlijke ordening toelaten. Bij longitudinale metingen komt het erop neer dat studiesubjecten ontbrekende metingen hebben op een bepaald moment in de tijd, doch eventueel nadien opnieuw geobserveerd kunnen worden. Nochtans, ook al kan het dus voorkomen in longitudinale studies, het is meer gebruikelijk dit soort patronen in multivariate situaties tegen te komen.

Het onvolledig zijn van de gegevens wordt meestal als een ongewenst neveneffect van empirisch onderzoek beschouwd. Het induceert een bijkomende graad van complexiteit in de data-analyse. Heel wat data analytische technieken waren oorspronkelijk niet ontworpen om dit aspect te behandelen. De keuze van hoe met onvolledigheid wordt omgegaan kan belangrijke gevolgen hebben voor de uit de analyse volgende conclusies. Bijvoorbeeld, wanneer in een longitudinale studie patiënten met een minder gunstige evolutie een hogere kans hebben op uitval en als bovendien dergelijke patiënten van analyse worden uitgesloten, dan kan de resulterende analyse te optimistisch uitvallen. Op dezelfde manier, wanneer respondenten vóór (of tegen) een bepaald gevoelig thema met een hogere kans weigeren van op een survey vraag te antwoorden, dan zullen de resulterende conclusies in de regel vertekening vertonen.

Het concept van een nonrespons mechanisme werd ingevoerd door Rubin (1976), minder uit nood om het te modeleren, dan wel om af te leiden onder welke aannames het mechanisme kan verwaarloosd worden zonder de geldigheid van de statistische conclusies te verstoren. Waar frequentistische methoden in principe de sterkere MCAR conditie vereisen, is dat voor likelihood en Bayesiaanse methodologie gelukkig slechts de zwakkere MAR assumptie. In elk geval kan een MNAR mechanisme niet verwaarloosd worden. Deze resultaten zijn zeer nuttig omdat ze aangeven hoe analyses kunnen vereenvoudigd worden. Het is nochtans niet mogelijk om op een onomstotelijke wijze vast te stellen in welke categorie een bepaald mechanisme precies valt. Er dienen altijd niet te verifiëren aannames gemaakt te worden. Dit is intuïtief duidelijk, omdat het, bijvoorbeeld, zeer moeilijk te achterhalen is waarom een patiënt precies uitvalt in een klinische studie. Tenzij het expliciet zou geweten zijn, kan de onderzoeker vaak niet verder gaan dan een verantwoorde gok.

Gedurende de jongste decennia is de methodologie om met ontbrekende gegevens om te gaan in belangrijke mate in volume en kracht toegenomen. Alles wijst erop dat deze trend zich zal verderzetten. In het verleden waren er zowel methodologische als computationale barrières om te beletten dat meer geavanceerde methodologie kon gebruikt worden. Dit leidde er dan toe dat men eenvoudige methoden gebruikte, zoals *complete case analysis*. Een dergelijke aanpak is meestal erg inefficiënt, in het bijzonder wanneer er een substantieel deel van de gegevens ontbreekt. Vaak is er ook belangrijke vertekening. Formeel komt het erop neer dat de sterke MCAR aanname vereist is. Heden ten dage is de reeks mogelijkheden, ter beschikking van de onderzoeker, sterk uitgebreid. Zulke meer algemeen geldige methoden maken het gebruik van de eenvoudige methoden overbodig; een aantal ervan wordt in deze thesis onderzocht. Hoogstens kunnen de eenvoudige methoden als eerste stap aangezien worden, of als aanvulling op meer gesofisticeerde methodologie. Ondanks uitgebreide onderzoeks- en onderwijsinspanningen op het gebied van onvolledige gegevens, toch blijft het omgaan met dit type data ingewikkeld. Dit is voornamelijk toe te schrijven aan het feit dat assumpties waarop de methodolgie rust zo goed als nooit helemaal verifieerbaar zijn. Een groot deel van deze thesis is erop gericht dit te onderstrepen en verder te verduidelijken. In dezelfde geest besteden we heel wat aandacht aan zogenaamde *sensitiviteitsanalyse*.

Het doel van deze thesis is om cruciale aspecten aan te halen van en inzicht te verschaffen in de methodologie voor de analyse van ontbrekende gegevens. De klemtoon ligt op herhaalde metingen. Nieuwe en/of aangepaste methodologie wordt voorgesteld en bestudeerd, en vaak in verband gebracht met bestaande methoden (hoofdstukken 5, 6, 8 en 9). Er worden ook verscheidene methoden voorgesteld voor sensitiviteitsanalyse (hoofdsukken 4, 7 en 8). Sensitiviteitsanalyse is waarschijnlijk de veiligste en voorzichtigste optie wanneer de gegevens onvolledig zijn. We zullen nu de verscheidene hoofdstukken kort voorstellen.

Hoofdstukken 2 en 3 omvatten inleidend materiaal, nodig als achtergrond bij de latere hoofdstukken. Motiverende *case studies* worden beschreven in hoofdstuk 2. Definities, algemene begrippen en fundamentele concepten worden aangereikt in hoofdstuk 3. In hetzelfde hoofdstuk wordt een overzicht gegeven van bestaande methodologie, zoals eenvoudige methoden en imputatie. Ook een aantal veelgebruikte modellen wordt geschetst, zoals het Diggle and Kenward (1994) model.

Het begrip van een MAR tegenhanger voor een gefit MNAR model (Molenberghs *et al.*, 2008) wordt geschetst in hoofdstuk 4. Een dergelijke tegenhanger is equivalent met het oorspronkelijke model voor wat betreft de fit aan de geobserveerde gegevens, maar verschilt in de predictie van de niet geobserveerde waarnemingen. Van hieruit kunnen we vertrekken om aan te geven dat de fit van een MNAR model *nooit* gebruikt kan worden om na te gaan dat een mechanisme MNAR of MAR is. We verwijzen hierbij ook naar Gill, van der Laan and Robins (1997) en Schafer and Graham (2002). Wanneer een parametrisch MAR model de geobserveerde gegevens niet goed beschrijft, kan een meer uitgebreid MNAR model gefit worden, waarna dan de MAR tegenhangen kan afgeleid worden. Uiteraard kan dit alles mooi ingepast worden in sensitiviteitsanalyse.

Methoden gebaseerd op MAR worden verder bestudeerd in hoofdstukken 5 en 6. In het eerste van beide hoofdstukken beschouwen we marginale MAR modellen voor onvolledige dichotome longitudinale gegevens, met bijzondere aandacht voor variaties op de semi-parametrische aanpak van Liang and Zeger (1986), d.w.z. *generalized estimating equations* (GEE). Het feit dat GEE een frequentistische methode is, betekent dat ze in beginsel slechts geldig is onder MCAR, zoals eerder aangegeven een sterke, onrealistische aanname. Om die reden worden twee aanpassingen bestudeerd: gewogen GEE (WGEE) (Robins, Rotnitzky and Zhao, 1995) aan de ene kant en GEE gecombineerd met *multiple imputation* (Rubin, 1987), of MI-GEE, aan de andere kant. Waar asymptotische simulaties de theoretische eigenschappen van beide methodes aantoonden, gaven simulaties met eindige steekproeven aan wat de praktische eigenschappen zijn, nuttig voor gebruik in concrete data analyse (Beunckens, Sotto and Molenberghs, 2008). De aantrekkelijke asymptotische eigenschappen van WGEE werden nauwelijks gereproduceerd in kleine steekproeven. In een aantal scenarios waren de schattingen gebaseerd op WGEE een stuk minder precies dan de schattingen gebaseerd op MI-GEE. Verder, in tegenstelling tot MI-GEE dat een zekere mate van robuustheid geniet bij verkeerd gespecifieerde modellen (zowel voor het meetmodel als voor het imputatiemodel), is WGEE vrij sensitief voor verkeerd gespecifieerde dropout- en meetmodellen. Samengevat illustreren de vergelijkingen gemaakt in dit hoofdstuk de relatieve sterkte van imputatie, eerder dan weging, in het bijzonder in de context van GEE.

Hoofdstuk 6 beschouwt het zogenaamde *pattern-mixture model*, PMM (Little, 1993, 1994a) voor de analyse van categorische gegevens met monotoon ontbrekende gegevens. Via het gebruik van gesimuleerde gegevens bestuderen we de methode van Jansen and Molenberghs (2007), die gebruik maakt van identificatie-restricties om PMM parameters te schatten. We leiden ook asymptotische varianties af voor gemarginaliseerde effectschattingen, in analogie met Sotto *et al.* (2009a). De resultaten geven aan dat de precisie van de schattingen sterk afhangt van de hoeveelheid ontbrekende gegevens in de verschillende dropout patronen. Dit impliceert op zijn beurt dat een identificatie schema best gebruikt maakt van behoorlijk goed gevulde patronen. Desalniettemin zijn de precisieschattingen bevredigend, ook als ze gebaseerd zijn op eerder sporadisch gevulde patronen. Het effect van behandeling op het moment van de laatste meting geeft meestal de beste precisie, onafhankelijk van het dropout mechanisme of de gebruikte restricties. Met betrekking tot de gemarginaliseerde effecten van behandeling zijn deze voorgesteld door Jansen and Molenberghs (2007) lichtjes minder vertekend dan de directe lineaire methode (Park and Lee, 1999) wanneer de gegevens binair zijn. De directe lineaire methode wordt dus best beperkt tot gebruik bij continue, Gaussisch verdeelde gegevens.

In hoofdstuk 7 wordt het gebruik van invloedsmaten als instrumenten voor sensitiviteitsanalyse bestudeerd. Veralgemeende lokale invloedsmaten worden afgeleid en variaties op de bestaande methodologie voor de vorm die perturbaties aannemen worden geïntroduceerd. In analogie met Beunckens *et al.* (2009) is de ontwikkeling van ideeën in dit hoofdstuk sterk geworteld in een omvattende sensitiviteitsanalyse van de zogenaamde *Slovenian Public Opinion Survey*. We combineren resultaten van bestaande strategieën voor sensitiviteitsanalyse, zoals niet-parametrische grenzen, een familie van identificeerbare modellen en intervallen afgeleid van niet-geïdentificeerde modellen, met deze gebaseerd op invloedsmaten.

De moeilijkheden bij het fitten van likelihood-gebaseerde non-ignorable modellen voor longitudinale gegevens ontstaan uit de noodzaak om de likelihood voor de volledige gegevens te integreren over de ontbrekende gegevens. Hoofdstuk 8 onderzoekt het gebruik van computationele algoritmen om de complexe en tijdsintensieve aspecten van dit numeriek integratieproces te vereenvoudigen. Sotto *et al.* (2009b) beschouwde het gebruik van het zogenaamde *stochastic expectation-maximization* of SEM algoritme (Celeux and Diebolt, 1985) om een likelihood-gebaseerd model voor longitudinale gegevens met non-ignorable missingness te fitten. Het is een variatie op het zeer populaire *expectation-maximization* (EM) algoritme. Op basis van een simulatiestudie werden stochastische en niet-stochastische methoden met elkaar vergeleken. De stochastische methode is zeer stabiel, terwijl de niet-stochastische methode tot een iets grotere precisie leidt. Hiermee hebben we niet alleen een computationeel aantrekkelijk alternatief voor bestaande methoden, doch SEM opent ook wegen voor sensitiviteitsanalyse.

De grote verspreiding van makkelijk te gebruiken doch flexibele software om imputatie te implementeren heeft ongetwijfeld bijgedragen aan de populariteit ervan. In hoofdstuk 9 passeren een aantal implementaties de revue en worden ze met elkaar vergeleken. Het praktische nut is dat gebruikers zich kunnen laten leiden door deze resultaten bij het maken van een keuze. De methoden worden vergeleken, op basis van simulaties, m.b.t. de resulterende parameterschattingen en m.b.t. maten voor individuele imputatie-vertekening, d.w.z. de mate van vertekening in de imputatie geëvalueerd op het niveau van het individu. Ondanks het feit dat de verscheidene methoden over het algemeen tot vergelijkbare schattingen en standaardfouten leiden, zijn er duidelijke verschillen op het niveau van de inviduele imputatie-vertekening. Verder blijkt de performantie van de verscheidene routines af te hangen van twee factoren, met name de fractie ontbrekende gegevens en de variabiliteit in de respons.

De ideeën ontwikkeld in deze thesis werden voorgesteld om aan te geven wat de opties zijn, beschikbaar voor de onderzoeker. Bestaande methoden werden ook verder bestudeerd. De keuze voor een bepaalde analysemethode is meestal ingegeven door een samengaan van verscheidene pragmatische beschouwingen, zoals de specifieke wetenschappelijke vraagstelling, computationele beschouwingen, en de hoeveelheid onvolledige gegevens. Omdat een dergelijke keuze vaak nogal overweldigend kan

zijn voor de gebruiker hopen we met ons werk bijgedragen te hebben aan het vergemakkelijken van de keuze. Tegelijk hebben we het belang aangegeven van sensitiviteitsanalyse. Verscheidene instrumenten voor sensitiviteitsanalyse werden aangereikt en bestudeerd. In het bijzonder wanneer er onzekerheid is over het onderliggende missing data mechanisme, wat bijna altijd het geval is, is het gebruik van verscheidene methoden tegelijkertijd een belangrijke bijdrage om na te gaan hoe onzekerheid over de ontbrekende gegevens de conclusies al dan niet kan verstoren.

Acknowledgements

Before I left Manila to come to Belgium for the Masters Program, the Dean of our College at my home university told me to immediately look for Ph.D. programs to apply for after the Masters program. Feeling that I did not have what it takes to complete a Ph.D., I totally ignored this advice. After the BioStat program, I found myself not wanting to leave Belgium just yet (no, it was not because of the weather), and so I applied for a 1-year consulting position here at CenStat. Given my earlier resistance to applying for a Ph.D., you can imagine my surprise when Noel (Veraverbeke) told me that I had been accepted, not for the consulting position, but for a 4-year Ph.D. grant. After my first published paper, most of my apprehension had all but disappeared. And now, at the culmination of my Ph.D., I look back with no regret but rather with gratitude that I had been given such an opportunity.

The completion of this thesis could not have been possible without the help and support of a number of people. First and foremost, I would like to thank Geert for his thorough guidance, unwavering support and extreme patience in supervising me over the last 4 years. He has been a great mentor and I have truly enjoyed collaborating with him. I have also been fortunate to have had a very supportive co-promotor, Lien Beunckens. Thanks Bella, not just for discussing mathematical equations and simulation codes with me, but for simply making the 4 years so much easier and a lot less boring.

I would also like to extend my sincere appreciation to all our co-authors, as well as to all the colleagues who have contributed, in some way or other, to this endeavor. Special thanks also to Saskia, Kaat and Annouschka, on whom I constantly relied for moral and logistic support and advice on administrative and procedural matters.

Being in a country so different from my own for 6 long years has not exactly been a walk in the park for me. Between the crazy Belgian weather and absolute absence of food delivery services and take-away coffee, it's a wonder I survived. And so I am especially grateful to my SATC-meets-DH BFFs. At this point I don't know if I should thank you for *not* making me go crazy or for making me go crazy once in a while. Our entertaining evenings of hilarious jokes combined with gluttonous feasts were exactly what I needed to get me through the tougher periods. Thanks, girls!

Finally, I would like to express my deepest gratitude to my parents. Mom, Dad, I know that you have always wanted me to become a doctor. And though it is not exactly in the field that you would have wanted, I know that you are nevertheless happy and proud. Thank you for your continuing love and support.

Cristina Sotto

Diepenbeek, 09 September 2009

"I'd like to know myself better. Unfortunately the key data are missing."

Contents

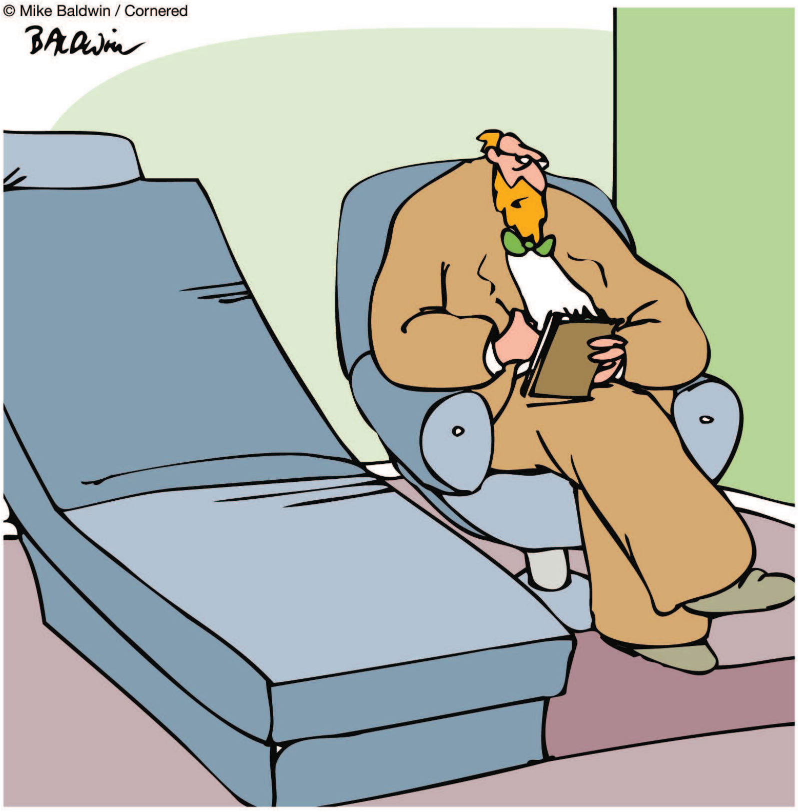
Appendices

List of Tables

List of Figures

List of Abbreviations

ACMV	Available Case Missing Values
ARMD	Age-Related Macular Degeneration
BDM	Bivariate Dale Model
BRD	Baker, Rosenberger and DerSimonian
CCA	Complete Case Analysis
CCMV	Complete Case Missing Values
GEE	Generalized Estimating Equations
LOCF	Last Observation Carried Forward
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMC	Markov Chain Monte Carlo
MDM	Multivariate Dale Model
MI	Multiple Imputation
MI-GEE	Multiple-Imputation Generalized Estimating Equations
MNAR	Missing Not At Random
MSE	Mean Squared Error
NCMV	Neighboring Case Missing Values
PMM	Pattern-Mixture Model
SEM	Stochastic Expectation-Maximization
SPO	Slovenian Public Opinion
WGEE	Weighted Generalized Estimating Equations



“This is the third session you’ve missed.
We need to look at that.”

1

Introduction

Scientific investigation frequently relies on data-based research, and, as such, is inevitably prone to incompleteness. Experimental or clinical studies, for instance, though probably designed to yield complete data, can result in the loss or damage of the experimental unit (or clinical subject) during the conduct of the experiment. When replacement of the lost unit is infeasible, the research investigator is faced with analysis of incomplete data. Incompleteness in survey-based studies is perhaps even more common, as respondents may refuse to provide an answer to one or several questions for a variety of reasons. It is also possible that the study design itself induces a lack of information for certain individuals of the study population. Regardless of the specific missing data scenario, modification of the original plan of analysis may be required to still be able to address the research question in a scientifically sound manner.

In a broad sense, observational or experimental studies can be of a longitudinal or cross-sectional nature. Whereas the latter entails data collection at a single point in time, the former involves the collection of information from the study units over periods of time. Both longitudinal and cross-sectional studies can result in data with missing values, and the structure of the incomplete data is generally classified into two forms. In a longitudinal study in which a particular response variable is observed repeatedly over time, the sequence of resulting measurements naturally defines a logical arrangement. In such cases, missingness often occurs in the form of the loss of a study unit at a specific point in time and thereafter, yielding what is conventionally termed *dropout* or attrition, or, more generally, *monotone missingness.* The latter, though most frequently encountered in longitudinal types of studies, can also occur in cross-sectional data, provided some ordering of the variables of interest can be considered. In contrast, one can speak of so-called *non-monotone missingness*, for which the missing values do not define any particular structure, but rather occur intermittently within the set of variables. Non-monotone missingness, though entirely conceivable in longitudinal data, is actually more commonly encountered in multivariate data.

Missingness, which is typically viewed as a nuisance, usually induces additional complexity in the analysis, because most data analysis procedures were not designed for them. The manner in which incompleteness in the data is addressed can have important implications on the resulting conclusions. In a longitudinal study assessing the efficacy of a particular treatment, for example, if patients with poorer clinical condition are the ones who drop out from the study and are excluded from the analysis, the efficacy of the treatment may be too optimistically estimated. Similarly, if respondents in favor (or against) a sensitive issue fail to supply responses and are discarded from the analysis a survey, resulting conclusions about the issue will most probably be biased accordingly. Though these and similar situations are somewhat extreme and perhaps unlikely, they nevertheless illustrate the consequences that the analysis approach can have on the conclusions derived from incomplete data.

When missing values occur for reasons unknown or beyond the control of the researcher, assumptions about the process that generates them need to be made. Rubin (1976) defined a classification scheme for the non-response process in terms of its statistical relation with the process generating the data itself (i.e., the response process). When the two processes are independent, possibly conditionally on covariates, the missing data or non-response process is described as being *missing completely at random* (MCAR). A *missing at random* (MAR) non-response mechanism, on the other hand, consists of a relation between the non-response process and the observed, but not the unobserved, responses. Dependence between the non-response and the missing responses characterizes a *missing not at random* (MNAR) missingness mechanism.

To help motivate these different missing data mechanisms, consider a typical longitudinal clinical trial to assess the efficacy of a given treatment for a particular disease or condition. A patient who fails to continue the study for reasons not related to his/her clinical condition (e.g., relocation to a different region from which continuation in the trial would not be possible) would have missing values that most probably fall within the category of MCAR. Alternatively, a patient who was improving over three visits, lapsed at the fourth visit, and, as a consequence of the regressing condition, dropped out at the fifth visit might most plausibly have missing-at-random values, since the nature of the missingness is related to the previously observed (worsening) outcome. An MNAR situation might best be depicted by a patient who improves to the third visit, deteriorates in between the third and the fourth visit, and, because of his/her aggravated condition, drops out at the fourth visit. In this case, the outcome for the fourth visit will be unobserved and the reason for the missingness is actually related to the yet-to-be observed deteriorating outcome at that visit.

The notion of a non-response process was introduced by Rubin (1976) not so much out of a need to model it, but rather to help clarify under what conditions it could possibly be ignored and yet have valid inferences. While frequentist statistical procedures generally admit ignoring the missing data process only when the data are MCAR, likelihood and Bayesian approaches allow ignoring the missing data process under MAR. In contrast, the missingness process cannot be ignored in the case of MNAR, which is also commonly referred to as the *non-ignorable* situation. These results, though already substantially providing some level of simplification of the missing data problem in the sense of narrowing down viable alternatives for analysis, rely on unverifiable assumptions about the underlying missing data mechanism. From the previous illustrations, it is clear that unless one knows, or at least has some idea about, the motivations for a subject's dropping out, it can be almost impossible to determine which underlying missing data mechanism is in play. Without such knowledge, a researcher can only attempt, at best, a reasonable guess. And unless strong conviction about the underlying mechanism driving the missingness can be had, more confidence in resulting conclusions might be gained by means of so-called *sensitivity analysis*. Under such, a researcher might consider analysis approaches under various missing data scenarios and compare resulting inferences.

Prior to advances in the theory, methodology, as well as computational capability, for handling missing data, analysis generally entailed the exclusion of cases with missing values - an approach referred to as *complete case analysis*. Such an approach is clearly inefficient, especially for data with large amounts of missingness, and potentially biased, as was already pointed out earlier, but more importantly, is valid only if the underlying missingness process is MCAR. Owing to the development of less naïve methodologies in the recent decades, many more options are now available to the researcher. Moreover, with the continuing growth in this area of research, missing data problems have become much more manageable and a lot less of a nuisance than in the past.

The aim of this thesis is to highlight and provide insight on existing methodology in the area of incomplete data, with primary interest falling on longitudinal outcome data. In addition, new or modified techniques shall be proposed and assessed, often in relation to existing approaches. Because the inferential validity of a method depends heavily on the unverifiable assumption regarding the underlying missing data mechanism, a considerable part of the thesis shall also be devoted to the use of various tools for *sensitivity analysis*, which provides a more prudent approach compared to "putting all one's eggs in a single basket." It is important to clarify at this point that incompleteness can arise in any variable in a study, be it a response variable or a covariate. The main concern here, however, shall be missingness in the response or outcome, rather than in covariates. Moreover, inasmuch as outcome variables of a continuous or of a discrete nature are both entirely conceivable in practice, methods for both types are explored. An overview of each of the chapters within the thesis now follows.

Chapters 2 and 3 comprise the introductory material necessary for the development of this thesis. In Chapter 2, various data sets used within the thesis are described briefly. Basic definitions, general notations and fundamental concepts to be used throughout the thesis are then outlined in Chapter 3. The latter will also already provide a very broad overview of some of the existing methodology, e.g., simple techniques and imputation, as well as an enumeration of various commonly used models for dealing with incomplete data, e.g., the Diggle and Kenward (1994) model.

As has been stressed previously, the choice of the assumed type of non-response mechanism is a critical one in the modeling of incomplete data, due to the unverifiable nature of such a choice. Moreover, as will be developed in Chapter 4, under certain situations, a model derived under one particular missing data mechanism may be equivalent, in some statistical sense, to another model derived under a different mechanism. In Chapter 4, such an idea is presented, relating missingness at random with missingness not at random (Molenberghs *et al.*, 2008). The result is of substantive interest as it not only further underscores the risk in placing one's belief entirely on a specific missing data mechanism, but also emphasizes the importance of performing sensitivity-based analyses.

MAR-based approaches in modeling incomplete data are the subject of Chapters 5 and 6. In the former, focusing on non-Gaussian longitudinal data with dropout, two versions of a semi-parametric method – one using a weighting scheme to address missingness and the other using imputation of the missing data – are compared. The relative merits of the different approaches are highlighted via a simulation study conducted under various misspecifications. Chapter 6, on the other hand, considers the *pattern-mixture modeling* framework (Little, 1993, 1994a). Building upon the ideas proposed in Jansen and Molenberghs (2007), regarding pattern-mixture models for categorical outcomes with missingness and their subsequent marginalization, simulations are conducted and asymptotic variances for the marginalized estimates are further derived in line with the results of Sotto *et al.* (2009a).

Chapters 7 and 8 focus on methodology for the analysis of longitudinal or multivariate data with non-ignorable missingness. In Chapter 7, the use of influence measures are explored as possible tools for sensitivity analysis. Parallel to Beunckens *et al.* (2009), the development of ideas in this chapter are rooted within a comprehensive sensitivity analysis on the Slovenian Public Opinion Survey (Chapter 2), combining results of existing strategies for sensitivity analyses with those based on influence measures. Chapter 8 explores the application of a computational algorithm to fit likelihood-based models to longitudinal data with missing data that are of an MNAR nature (Sotto *et al.*, 2009b). The algorithm is particularly useful in the latter setting as it provides an approach that precludes the often tedious and computationally demanding aspects of numerical integration that is required under non-ignorable models. Moreover, such a technique, when not taken as the primary course of analysis, can further serve as an additional tool within the framework of a sensitivity analysis.

The widespread availability of easy-to-use yet flexible software-based routines for implementing imputation - broadly covering all methods that entail filling-in missing observations with some form of imputed values - has undoubtedly boosted the popularity of imputation as a technique to deal with missing data. In Chapter 9, a number of these software routines or packages is reviewed under simulated settings, as well as within a real-life context, to help bring out features of the different routines relative to each other, thereby providing practitioners with bases for making a choice regarding which software package might best suit their needs.

Finally, in the last chapter (Chapter 10), ideas, results, issues and recommendations are presented regarding the different methods considered within the thesis. Technical details and derivations that are excluded from the main text are provided in the Appendix section.

2

Key Examples

In this chapter, three data sets that shall serve as key examples in the thesis are introduced. These data sets, which are either of a longitudinal or multivariate nature and contain missing observations, will be used in illustrating methodologies in subsequent chapters.

2.1 Age-Related Macular Degeneration Trial

The first data set considered arises from a randomized multicenter clinical trial comparing an experimental treatment (interferon-α) with a corresponding placebo in the treatment of patients with age-related macular degeneration, or ARMD, with particular focus on the comparison between the placebo and the highest dose (6 million units daily) of interferon-α (Z). Full results of this trial are reported in Pharmacological Therapy for Macular Degeneration Study Group (1997).

Patients with macular degeneration progressively lose vision. In the trial, visual acuity was assessed at different time points (4, 12, 24 and 52 weeks) through patients' ability to read lines of letters on standardized vision charts. These charts display lines of five letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The raw visual acuity is the total number of letters read correctly. Table 2.1 shows the average visual acuity by treatment group at baseline and at the four subsequent measurement occasions.

Table 2.1: *Age-Related Macular Degeneration Trial. Mean (standard error) visual acuity at baseline, at 4, 12, 24 and 52 weeks, by randomized treatment group.*

Time Point	Placebo	Interferon-α	Total
Baseline	55.3 (1.4)	54.6 (1.4)	55.0 (1.0)
4 weeks	54.0 (1.5)	50.9 (1.5)	52.5 (1.1)
12 weeks	52.9 (1.6)	48.7 (1.7)	50.8 (1.2)
24 weeks	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
52 weeks (1 year)	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

Table 2.2 presents an overview of the amount and nature of missingness, monotone and non-monotone, for the ARMD data. A total of 240 patients were observed, of which about 78% had complete observations. Roughly 16% of the cases represented dropout at either the 12$^{\text{th}}$, 24$^{\text{th}}$ or 52$^{\text{nd}}$ week, less than 5% exhibited intermittent missingness, and 6 patients (2.5%) had no observations.

In subsequent analyses of the ARMD data, whenever considering monotone missingness, patients with intermittent missingness, as well as patients with no recorded measurements, are excluded, yielding a final analysis set consisting of 226 patients. When intermittent missingness is of interest, on the other hand, all 240 patients are included in the analysis set.

Table 2.2: *Age-Related Macular Degeneration Trial. Overview of missingness patterns and corresponding frequencies. (O: observed, M: missing.)*

	Measurement Occasion				Number	%
	4 weeks	12 weeks	24 weeks	52 weeks		
Completers	O	O	O	O	188	78.33
Dropouts						
	O	O	O	M	24	10.00
	O	O	M	M	8	3.33
	O	M	M	M	6	2.50
	M	M	M	M	6	2.50
Non-Monotone						
	O	O	M	O	4	1.67
	O	M	M	O	1	0.42
	M	O	O	O	2	0.83
	M	O	M	M	1	0.42

2.2 Hepatitis B Virus Immunization Trial

The second case study involves a Hepatitis B virus vaccination trial to study the persistence of vaccine-induced anti-HBs antibodies in mentally retarded patients (Van Damme *et al.*, 2006). Hepatitis B is a liver infection caused by the Hepatitis B virus (HBV) for which the mentally retarded are at high risk. The trial assessed the level of anti-HBs antibodies in the blood 20 years after the original vaccinations. Patient anti-HBs antibody levels were then recorded at various times points within a period of 61 months. Information on the number of booster shots received by the patient during the immunization period was also recorded.

Of 275 subjects originally enrolled in the trial, only 214 were available for follow-up after the 20-year period. The log-scaled level of anti-HBs was the response of interest, and, for the purposes of this thesis, analysis of this response was restricted to the last 2 time points in the sequence (months 60 and 61). Table 2.3 shows the distribution of the subjects across the various non-monotone patterns. Of the 214 subjects, about 60% had complete information at the time points of interest, while a little more than 40% had incomplete data. More specifically, about 29% of the cases had both measurements missing, 10% had responses only for the first time point and 2% had responses only for the second time point.

Table 2.3: *Hepatitis B Virus Immunization Trial. Overview of missingness patterns and corresponding frequencies. (O: observed, M: missing.)*

	Measurement Occasion		Number	%
	60 months	61 months		
Completers	O	O	126	58.89
Dropouts				
	O	M	21	9.81
	M	M	62	28.97
Non-Monotone				
	M	O	5	2.34

2.3 The Slovenian Public Opinion Survey

In 1991, Slovenians voted for independence from former Yugoslavia in a plebiscite. To prepare for this result, the Slovenian government collected data on its possible outcome by inserting questions in the so-called Slovenian Public Opinion (SPO) Survey. The survey, administered to 2074 voting-age Slovenians, was conducted a month prior to the plebiscite using a three-stage sampling design (Barnett, 2002). The three fundamental questions about the ensuing plebiscite added to the survey were:

(1) Are you in favor of Slovenian independence?

(2) Are you in favor of Slovenia's secession from Yugoslavia?

(3) Will you attend the plebiscite?

In spite of their apparent equivalence, questions (1) and (2) are different since independence would have been possible in a transition from the existing federal structure to a looser, confederal, form as well, and therefore the secession question was added. Question (3) is highly relevant since the political decision was taken that not attending was treated as an effective NO to question (1). Thus, the primary estimand is the proportion θ of people that will be considered as voting YES, which, in

Table 2.4: *Slovenian Public Opinion Survey. Cross-tabulation of the responses on the three questions added to the survey. (M: Missing.)*

Secession	Attendance	Independence		
		Yes	No	M
Yes	Yes	1191	8	21
	No	8	0	4
	M	107	3	9
No	Yes	158	68	29
	No	7	14	3
	M	18	43	31
M	Yes	90	2	109
	No	1	2	25
	M	19	8	96

Table 2.5: *Slovenian Public Opinion Survey. Cross-tabulation of the responses on the attendance and independence questions, collapsed over the secession question. (M: Missing.)*

Attendance	Independence		
	Yes	No	M
Yes	1439	78	159
No	16	16	32
M	144	54	136

the context of the three questions, can be defined as the fraction of people answering YES on *both* the independence and attendance questions (1) and (3), respectively, regardless of their response to question (2). The raw data are presented in Table 2.4.

In subsequent chapters that use the SPO Survey data, unless otherwise indicated, analysis will be restricted to the results of the independence and attendance questions, i.e., questions (1) and (3), respectively. The raw data in Table 2.4, collapsed over the secession question, are summarized in Table 2.5.

3

Fundamental Concepts

In this chapter, basic terminology, notations and fundamental concepts that are used in the area of incomplete (longitudinal) data and that will be used throughout the thesis, are introduced. Some commonly used methodologies are also briefly described.

3.1 Incomplete Longitudinal Data

Suppose that for subject $i, i = 1, 2, \ldots, N$, a sequence of measurements Y_{ij} is designed to be measured at time points $t_{ij}, j = 1, 2, \ldots, n_i$. The resulting vector of planned measurements, $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$, is referred to as the *complete* data. Suppose further that, for each measurement in the series, a corresponding non-response indicator R_{ij} is defined as:

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed, and} \\ 0 & \text{otherwise,} \end{cases}$$

which are organized into a vector $\boldsymbol{R}_i$ of parallel structure to $\boldsymbol{Y}_i$. The set of measurements, along with the non-response indicators, $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$, comprise what is called the *full* data. Typically, $\boldsymbol{Y}_i$ can be partitioned into two sub-vectors: $\boldsymbol{Y}_i^o$ consisting of those Y_{ij} for which $R_{ij} = 1$, and $\boldsymbol{Y}_i^m$ consisting of the remaining components, respectively referred to as the *observed* and *missing* components. For longitudinal data with missingness, only $(\boldsymbol{Y}_i^o, \boldsymbol{R}_i)$ is available.

The structure of the non-response vector admits two basic types of missingness: monotone and non-monotone. When non-response is *monotone* or of a *dropout* nature, the unobserved measurements within the longitudinal series all occur after a particular measurement occasion, and in that sense, the subject is said to have "dropped out" of the study. In such cases, the non-response vector $\boldsymbol{R}_i$ consists of a very particular form, with all R_{ij} equal to one up to a particular time point j and zero thereafter. This structure allows the non-response indicators in $\boldsymbol{R}_i$ to be collapsed into a single variate, D_i, defined as

$$D_i = 1 + \sum_{j=1}^{n_i} R_{ij},$$

denoting the time point at which subject i drops out. *Non-monotone* missingness, on the other hand, occurs when missing values arise intermittently within the series, leading to no distinct configuration of the non-response indicators, and thus, simplification of $\boldsymbol{R}_i$ into a lower-dimensional form is not straightforward.

3.2 Modeling Frameworks

Owing to the presence of the two stochastic components, $\boldsymbol{Y}_i$ and $\boldsymbol{R}_i$, models for incomplete longitudinal data involve working with the joint distribution (suppressing, for the moment, dependence on covariates), $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the respective parameter vectors describing the response and non-response processes. The choice of factorization of this joint distribution characterizes various modeling frameworks, each of which naturally warranting a particular interpretation. Under a *selection model* (SeM) framework (Rubin, 1976; Little and Rubin, 1987), the joint distribution is factored into a marginal model for the measurements and a conditional model for non-response given the measurements, that is,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{\theta})\, f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}). \tag{3.1}$$

Selection models are an obvious choice for clinicians, for instance, who are often interested in the marginal effect, $\boldsymbol{\theta}$, of the independent variables (e.g., treatment) on the response.

The reverse factorization of $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})$ into a marginal model for the non-response and a model for the measurements conditional on the non-response,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{r}_i, \boldsymbol{\theta})\, f(\boldsymbol{r}_i|\boldsymbol{\psi}), \tag{3.2}$$

characterizes *pattern-mixture models* (Little, 1993, 1994a), or PMM. It is clear that the structure of (3.2) describes a mixture of pattern-specific models for the measurements via the mixing distribution $f(\boldsymbol{r}_i|\boldsymbol{\psi})$. And in contrast with selection models, the parameters $\boldsymbol{\theta}$ in a PMM denote pattern-specific effects of the independent variables on the response.

Finally, if the measurement and non-response processes are taken to be independent, conditional on a common set $\boldsymbol{b}_i$ of latent variables or random effects, via the factorization,

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta})\, f(\boldsymbol{r}_i|\boldsymbol{b}_i, \boldsymbol{\psi}), \tag{3.3}$$

then, a *shared-parameter model* (Wu and Carroll, 1988; Wu and Bailey, 1989), or SPM, is in effect. Here, $\boldsymbol{\theta}$ denotes the effects of the independent variables, conditional on the random effects.

It should be noted that whereas $\boldsymbol{\theta}$ in (3.1) represent marginal effects, $\boldsymbol{\theta}$ in both (3.2) and (3.3) describe conditional effects, conditional on the missingness pattern and on the random effects, respectively. As such, direct comparison of parameter estimates arising from an SeM, a PMM or an SPM is not immediately possible. Further marginalization, however, over either the non-response patterns or over the random effects, may be considered to yield fully marginal effects.

3.3 Missing Data Mechanisms

Within the selection model framework, Rubin (1976) developed a taxonomy to classify the missingness process based on its dependence (or lack thereof) on the measurement process. The classification is hinged on the structure of the second term in the right hand side of (3.1), which upon partitioning of the response vector into its observed and missing components, can be expressed as $f(\boldsymbol{r}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})$. Independence of the measurement and non-response processes, possibly conditionally on covariates, defines a missing data mechanism that is said to be *missing completely at random* (MCAR). A less rigid assumption would be one of *missing at random* (MAR), for which the missingness may depend on the observed outcomes and on covariates but, given these, not further on the unobserved outcomes. When, in addition to such dependencies, the unobserved data provide further information about the non-response process, the mechanism is referred to as being *missing not at random* (MNAR). These can be depicted in terms of the conditional distribution of $\boldsymbol{r}_i$ given $\boldsymbol{y}_i$ as follows:

Mechanism	$f(\boldsymbol{r}_i\|\boldsymbol{y}_i, \boldsymbol{\psi})$
MCAR	$f(\boldsymbol{r}_i\|\boldsymbol{\psi})$
MAR	$f(\boldsymbol{r}_i\|\boldsymbol{y}_i^o, \boldsymbol{\psi})$
MNAR	$f(\boldsymbol{r}_i\|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})$

Though these frameworks are frequently considered as distinct, technically speaking, the MNAR family can be viewed as encompassing both MCAR and MAR models as special cases. In addition, an MCAR model, under which no dependence between $\boldsymbol{r}_i$ and $\boldsymbol{y}_i^m$ exists, can actually be (more broadly) classified as MAR as well. These connections give rise to the following relations:

$$\text{MCAR} \quad \subset \quad \text{MAR} \quad \subset \quad \text{MNAR}.$$

MCAR, though leading to the simplest approaches in terms of the modeling task, since the missing observations can be altogether excluded from the analysis, requires quite a stringent assumption, which, in real-life contexts, is often not realistic. MNAR, on the other hand, is plausible in (longitudinal) clinical trials, where it is common to encounter subjects who do well until midway into the trial, relapse after the last observed visit, and are then lost to follow-up. If, in such cases, the missed visit is, to some extent, attributable to the subject's worsening clinical condition, as represented by the yet-to-be observed response, then the non-response process depends on the missing response (that would have been recorded had the patient been present for the visit). Analogously, if a subject consistently worsens over the previous visits and, as a consequence, decides to drop out at the next scheduled visit, then the missing response somehow depends on the previous observed responses and may most likely be of an MAR nature.

It is important to note at this point that MCAR, MAR and MNAR are assumptions made regarding the underlying non-response process. As such, absolute certainty about them can never be had. Incompleteness in the data induces a certain degree of unidentifiability, which has to be compensated for, ideally by a wise choice for the operating missingness mechanism. Though an experienced researcher may have some general idea on the nature of the missing data, this would be, at best, an educated guess. Fazed with unfamiliarity with clinical (and other) aspects of non-response, one can perhaps avoid the more implausible MCAR, or even go for some sort of sensitivity analysis, comparing results under different mechanisms, for instance. As will become clear shortly, validity of inferences made under different statistical methods depend primarily on the assumed missingness mechanism. In extreme cases, contradictory results can be obtained under two different mechanisms. Moreover, a model obtained under an MNAR missingness mechanism admits an equivalent MAR model, at least in terms of the models' fit to the observed data, but with contrasting inferences regarding the unobserved data. Such ideas will be discussed in greater detail in Chapter 4.

3.4 Likelihood and Ignorability

Traditionally, likelihood-based methods entail maximization of the full-data likelihood

$$L^* \equiv L^*(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{r}) = \prod_{i=1}^{N} f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi}). \tag{3.4}$$

In the presence of incomplete data, however, inference must be based on what is observed, and thus, the full-data likelihood L^* must be replaced by the *observed-data* likelihood L, for which the individual likelihood contributions need to be integrated over the missing values, i.e.,

$$L \equiv L(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{r}) = \prod_{i=1}^{N} \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i|\boldsymbol{\theta}, \boldsymbol{\psi})\, d\boldsymbol{y}_i^m. \tag{3.5}$$

Restricting focus on the observed-data likelihood contribution of subject i, under an SeM framework, the integral in (3.5) becomes

$$L_i \equiv L_i(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}_i, \boldsymbol{r}_i) = \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta})\, f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})\, d\boldsymbol{y}_i^m. \tag{3.6}$$

Under MCAR (or MAR), independence of all (or the unobserved) measurements and non-response leads to

$$L_i = f(\boldsymbol{r}_i|\boldsymbol{\psi}) \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta})\, d\boldsymbol{y}_i^m = f(\boldsymbol{r}_i|\boldsymbol{\psi})\, f(\boldsymbol{y}_i^o|\boldsymbol{\theta}), \text{ or} \tag{3.7}$$

$$L_i = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi}) \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta})\, d\boldsymbol{y}_i^m = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi})\, f(\boldsymbol{y}_i^o|\boldsymbol{\theta}). \tag{3.8}$$

If the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, i.e, the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is equivalent to the product of the parameter spaces of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, then the *separability* condition is satisfied. Under such a case, within the context of likelihood inference, MCAR and MAR are *ignorable*, while an MNAR process is non-ignorable. Frequentist inference, on the other hand, require the stronger condition of MCAR to ensure ignorability (Rubin, 1976). Modifications can allow for the possibility to model the data under MAR, provided the missingness is addressed by, for instance, a dropout or an imputation model, thus rendering the missing-data mechanism not ignorable.

The implication of ignorability within likelihood-based approaches is that $\boldsymbol{\theta}$ can be estimated directly from the observed data, by maximizing the logarithm of the component $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$ in the likelihood contribution L_i – a task easily done by standard software procedures that allow for missing observations. Unless the researcher is particularly interested in the missingness model, non-response can altogether be "ignored." For the non-ignorable situation (MNAR), the likelihood contributions do not admit further simplification and evaluation of the integral is necessary to compute the observed-data likelihood. Depending on the complexity of the model used, not only is the approximation/evaluation of the integral itself involved, but its subsequent maximization (for maximum likelihood, for instance) can be computationally demanding. This integration step is what often poses a challenge in fitting likelihood-based parametric models for non-ignorable missing data.

Perhaps the most widely used methodology for continuous longitudinal data within the likelihood framework is the general *linear mixed-effects model* (Laird and Ware, 1982), which takes the form

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{b}_i \sim N(\boldsymbol{0}, D) \\ \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma_i) \\ \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \text{ independent} \end{cases} \tag{3.9}$$

where $\boldsymbol{Y}_i$ is the n_i-dimensional (longitudinal) response vector for subject i, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are, respectively, the $(n_i \times p)$ and $(n_i \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p-dimensional vector containing the fixed effects, $\boldsymbol{b}_i$ is the q-dimensional vector containing random effects, and $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional vector of residual components, combining measurement error and serial correlation. D and Σ_i are general covariance matrices of size $(q \times q)$ and $(n_i \times n_i)$, respectively. In the case of no serial correlation, Σ_i reduces to $\sigma^2\boldsymbol{I}_{n_i}$.

It follows from (3.9) that, conditional on $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ is normally distributed with mean vector $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ and with covariance matrix Σ_i. Upon integration over the random effects, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, the resulting marginal (i.e., averaged over the random effects) model for the response can be expressed as:

$$\boldsymbol{Y}_i \sim N\left(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i D \boldsymbol{Z}_i' + \Sigma_i\right). \tag{3.10}$$

Ample detail and examples can be found in Verbeke and Molenberghs (2000).

Extension of the linear mixed-effects model to non-Gaussian (e.g., binary) longitudinal responses defines the class of *generalized linear mixed models* (Molenberghs and Verbeke, 2005). GLMMs expand the generalized linear model framework (McCullagh and Nelder, 1989) to the case of correlated responses by (1) including subject-specific regression parameters $\boldsymbol{b}_i$ in the linear predictor to address correlations among the repeated measures, and (2) assuming that conditional on the random effects $\boldsymbol{b}_i$, the elements of $\boldsymbol{Y}_i$ are independent. A typical GLMM assumes that all Y_{ij} have densities of the form $f_i(y_{ij}) \in$ exponential family, and the mean μ_{ij} is modeled through a linear predictor containing fixed regression parameters $\boldsymbol{\beta}$, as well as subject-specific parameters $\boldsymbol{b}_i$, with some known link function $\eta(\cdot)$. It is further assumed that the random effects follow a normal distribution.

The general specification of a GLMM thus includes (Molenberghs and Verbeke, 2005):

- the conditional distribution of the response Y_{ij} given the random effects $\boldsymbol{b}_i$ is assumed to follow from the exponential family;

- the conditional expectation of the response is modeled through a linear predictor containing the fixed regression parameters as well as the subject-specific parameters $\boldsymbol{b}_i$, through some known link function η, i.e.,

$$E\left(Y_{ij}|\boldsymbol{b}_i\right) = \mu_{ij} \text{ and } \eta\left(\mu_{ij}\right) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i, \tag{3.11}$$

  where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors containing known covariate values, with conditional variance given by:

$$\text{Var}\left(Y_{ij}|\boldsymbol{b}_i\right) = \phi v\left(\mu_{ij}\right); \tag{3.12}$$

- conditional on the $\boldsymbol{b}_i$, the repeated measurements in $\boldsymbol{Y}_i$ are independent; and,

- the $\boldsymbol{b}_i$ are independent and identically distributed as $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$.

Note that if $\eta(\cdot)$ is the natural (identity) link function, (3.11) becomes

$$\mu_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i,$$

and the GLMM reduces to the linear mixed model.

3.5 Methodology for Incomplete Longitudinal Data

This section is devoted to providing an overview of the various strategies in handling missing data in longitudinal studies.

3.5.1 Simple Methods

One of the simplest, but not necessarily most efficient, ways to handle incompleteness in any data is to use only the complete cases. Such an analysis is referred to as *complete case analysis* (CCA). Not only is implementation of CCA straightforward, most standard software delete incomplete cases at the onset of the procedure, and thus, no special data management is often necessary. The simplicity of CCA comes with a price – loss of information, arising from discarding incomplete cases, and, as a consequence, inefficient estimators and a severe decrease in power. Moreover, exclusion of incomplete cases is valid only under MCAR, and bias is inevitable under MAR. From an intuitive point of view, if the completers in a longitudinal study are generally the "better" patients, CCA would lead to overly optimistic results for the general study population.

Last observation carried forward or LOCF does exactly what its name states – the last observed measurement is used for succeeding unobserved measurements. If missingness is of a dropout nature, then the last recorded value for the response is used for all succeeding time points up to the end of the study. For intermittent missingness types, the last observed response is substituted for all missing observations within 2 observed responses. The idea of LOCF is based on the assumption that a subject will stay at the same response level until the end of the study, or at least, until he/she returns for a visit. Given that these periods of dropout (or missingness) can be months or years in many clinical trials, it is difficult to conceive that a patient's clinical condition will not change over such a possibly lengthy period. Moreover, when a patient goes off trial, changes in his/her condition are almost always expected. Finally, when LOCF is employed as a means of handling incompleteness in the data, no attempt is made to distinguish the actual observed responses and those filled in by LOCF (e.g., a weighting scheme which downweights the replaced values).

Molenberghs and Kenward (2007) (Section 4.3) showed, using hypothetical data, that, even under the unrealistically strong assumption of MCAR, while CCA produces unbiased estimates, the bias in the LOCF estimator does not vanish, and can even induce an apparent treatment effect when there is none. Under MAR, they showed that both can be biased and bias can go in either direction. The same authors further examined the nature of the resulting missing data mechanism implied by using LOCF (Kenward and Molenberghs, 2009). They determined that LOCF effects a missing data mechanism that is forced to depend on future, unobserved measurements – a sharp contradiction and incompatibility with MCAR, under which LOCF has been thought, apparently incorrectly, to be valid.

3.5.2 Imputation

Imputation broadly refers to any method that addresses the missing data problem by substituting imputed values for the missing measurements. This can be done in a variety of ways. LOCF, for instance, can be viewed as an imputation strategy in the broad sense. Alternatively, more intuitive ways of arriving at imputed values might be considered. *Unconditional mean imputation* (Little and Rubin, 1987) uses the average of all observed values of the same variable (e.g., response at a particular time point) over the other subjects as a substitute for missing values of that variable. The term unconditional arises from the fact that the imputation does not condition on a subject's available information, but rather, uses information from other subjects. In contrast, when information on a subject's missing response is imputed using information from the same subject's observed responses, the imputation method is called *conditional mean imputation* or *Buck's method* (Buck, 1960; Little and Rubin, 1987).

LOCF, unconditional and conditional mean imputation, among others, each replace the missing value with a single imputed value. Several such replacements can also be considered for the missing observation, via so-called *multiple imputation* (Rubin, 1978, 1987). Under this approach, the completed data sets are separately analyzed and subsequently pooled for final inference. The rationale behind multiple imputation is to take into account variability in the imputation of the missing values by taking several copies of it. The key idea of the procedure is to first replace each missing value with a set of M plausible values drawn from the conditional distribution of the unobserved values, given the observed ones. This conditional distribution represents the uncertainty about the right value to impute. In this way, M imputed data sets are generated (imputation stage), which are then analyzed using standard complete-data methods (analysis stage). Finally, the results from the M analyses have to be combined into a single inference (pooling stage) by means of the method laid out in Rubin (1978). In its basic form, multiple imputation requires the missingness mechanism to be MAR, though versions under MNAR have been proposed (Rubin, 1987; Molenberghs, Kenward and Lesaffre, 1997).

In line with previous notation, suppose that the vector of repeated measures $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$ is described by the parameter vector $\boldsymbol{\theta}$. At the imputation stage, the objective is to fill in the missing data with draws from the conditional distribution $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is unknown, an estimate for it, say $\widehat{\boldsymbol{\theta}}$, is first obtained from the data, after which $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \widehat{\boldsymbol{\theta}})$ is used to impute the missing data. This implies generating draws from the distribution of $\widehat{\boldsymbol{\theta}}$, thereby taking sampling uncertainty into account. Alternatively, a Bayesian approach, in which uncertainty about $\boldsymbol{\theta}$ is incorporated by means of some prior distribution for $\boldsymbol{\theta}$, can also be taken. A random $\boldsymbol{\theta}^*$ is first drawn from this prior distribution, then a random $\boldsymbol{Y}_i^m$ is selected from $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}^*)$. The so-imputed missing data are next augmented to the observed data, yielding completed data $(\boldsymbol{Y}^o,\boldsymbol{Y}^{m^*})$, which are then used to obtain an estimate of $\boldsymbol{\theta}$ and its variance, $U=\widehat{\text{Var}}(\widehat{\boldsymbol{\theta}})$. These steps are repeated a number of times, say M times, producing $\widehat{\boldsymbol{\theta}}^m$ and U^m, for $m=1,\ldots,M$. In the last phase of multiple imputation, the results of the analyses for the M imputed data sets are pooled into a single inference. The combined point estimate for the parameter of interest $\boldsymbol{\theta}$ from the multiple imputation is simply the average of the M complete-data point estimates. That is, the estimate and its estimated variance are given by:

$$\overline{\widehat{\boldsymbol{\theta}}}=\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\theta}}^m \quad \text{and} \quad \boldsymbol{V}=\boldsymbol{W}+\left(\frac{M+1}{M}\right)\boldsymbol{B}, \tag{3.13}$$

where

$$\boldsymbol{W}=\sum_{m=1}^{M}\frac{U^m}{M} \quad \text{and} \quad \boldsymbol{B}=\sum_{m=1}^{M}\frac{(\widehat{\boldsymbol{\theta}}^m-\overline{\widehat{\boldsymbol{\theta}}})(\widehat{\boldsymbol{\theta}}^m-\overline{\widehat{\boldsymbol{\theta}}})'}{M-1}, \tag{3.14}$$

with $\boldsymbol{W}$ denoting the average *within* imputation variance and $\boldsymbol{B}$ the *between* imputation variance (Rubin, 1987).

3.5.3 Direct Likelihood

In general, likelihood-based methods can be applied to incomplete longitudinal data after deletion (e.g., CCA) or imputation of the missing observations, or to the incomplete data in its raw form (i.e., without any pre-processing or prior treatment of the missing values). In the former, since missing values are not longer present in the set of complete cases or in the imputed data set, likelihood approaches are based on the full-data likelihood (3.4) of the complete (or completed) data. In contrast, for incomplete longitudinal data in the "raw," so to speak, any method within a likelihood framework would require working with the observed-data likelihood (3.5). For the remainder of this thesis, the terminology *direct likelihood*, unless otherwise stated, shall be used to refer to the latter case, that is, likelihood methods applied to the data as they are, rather than on data subjected to any pre-processing of the missing observations.

In Section 3.4, it was noted that, under ignorability, the observed-data likelihood simplifies into either (3.7) for MCAR or (3.8) for MAR. Under such cases, the observed-data log-likelihood splits into two component parts, having, as a consequence of the separability condition, separable parameter spaces. For MAR, for instance,

$$\ell_i \equiv \log L_i = \log f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi}) + \log f(\boldsymbol{y}_i^o|\boldsymbol{\theta}). \qquad (3.15)$$

Maximum likelihood estimation would thus simply entail separate maximization of the component terms, which would translate, for instance, to fitting two maximum likelihood models: one for the observed responses and another for the non-responses conditional on the observed responses. Additional simplification arises for the case of MCAR,

$$\ell_i \equiv \log L_i = \log f(\boldsymbol{r}_i|\boldsymbol{\psi}) + \log f(\boldsymbol{y}_i^o|\boldsymbol{\theta}), \qquad (3.16)$$

where the model for the non-response need not be conditioned on the observed responses. Moreover, if the focus of inference lies on the response process parameters $\boldsymbol{\theta}$, estimation of the (conditional) non-response model (given the observed measurements) can altogether be bypassed. As such, the direct likelihood approach is no more complicated than fitting a likelihood-based model on the complete cases. Standard software procedures that allow for incomplete observations, ensuring that the correct form of the likelihood is manipulated, would be able to obtain such a solution.

Direct likelihood under non-ignorability (e.g., MNAR) is a lot less straightforward in comparison with the ignorable case. Unlike the latter, the former does not admit further simplification of the observed-data log-likelihood contributions, due to the dependence of non-response on the unobserved (and possibly also, observed) outcomes,

$$\ell_i \equiv \log L_i = \log \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\boldsymbol{\theta})\, f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\psi})\, d\boldsymbol{y}_i^m. \qquad (3.17)$$

The integration over the missing values, which vanishes under ignorability, brings about additional levels of complexity to the direct likelihood approach for non-ignorable missingness. In addition to evaluation/approximation of the integral to compute ℓ_i in (3.17), which in itself can already be intricate, especially for high dimensions of missingness, obtaining a maximum of such a complex log-likelihood can be extremely demanding computationally. In Chapter 8, a computational algorithm to help overcome such numerical complexities will be explored.

3.5.4 Semi-Parametric Approaches

Often, full maximum likelihood can be prohibitive when the form of the likelihood is complex; this can be particularly so in the case of longitudinal sequences that are of moderate to large lengths. In such situations, working with the full likelihood can be avoided by specifying the likelihood only partially, resulting in a *semi-parametric* method. Broadly speaking, the application of semi-parametric techniques is not exclusively confined to the area of longitudinal data, though such approaches have gained popularity, particularly for the case of categorical (e.g., binary) repeated measures. For the latter, fully specified marginal models, e.g., Bahadur (1961) and Molenberghs and Lesaffre (1994), exist and can be fitted; however, the intricacies can be restrictive. As an alternative, Liang and Zeger (1986) proposed *generalized estimating equations* (GEE), which can be used to obtain marginal models for non-Gaussian longitudinal data, but, at the same time, circumventing the computational complexity of full likelihood. It is primarily useful whenever interest is restricted to the mean parameters. The semi-parametric nature of GEE arises from the fact that the method requires only the correct specification of the univariate marginal distributions, provided one is willing to adopt so-called *working assumptions* about the association structure of the vector of repeated measurements. The moments that are specified under GEE coincide with those of the Bahadur model (Bahadur, 1961), so that the former can be seen as a non-likelihood version of the latter, and as such, is sometimes viewed as a moment-based version of the Bahadur model. Alternatively, GEE can also be loosely described as a “correlation-corrected version of logistic regression.”

In the classical (or standard) form of GEE (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005), the score equations for a non-Gaussian outcome are

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} (A_i^{1/2} C_i A_i^{1/2})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.18}$$

where $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$, $\boldsymbol{\beta}$ is the vector of regression parameters, A_i is a diagonal matrix with the marginal variances, and C_i is the marginal correlation matrix for the repeated measures. Although $A_i = A_i(\boldsymbol{\beta})$ follows directly from the marginal mean model, $\boldsymbol{\beta}$ commonly contains no information about C_i. Therefore, the correlation matrix C_i typically is written in terms of a vector $\boldsymbol{\alpha}$ of unknown parameters, i.e., $C_i = C_i(\boldsymbol{\alpha})$. Liang and Zeger (1986) dealt with this set of nuisance parameters $\boldsymbol{\alpha}$ by allowing for specification of an incorrect structure or so-called working correlation matrix. Some of the more popular choices for the working correlations include:

Structure	Correlation	
Independence	$\rho(Y_{ij}, Y_{ik}) = 0,$	$j \neq k$
Exchangeability	$\rho(Y_{ij}, Y_{ik}) = \alpha,$	$j \neq k$
Autoregressive (1)	$\rho(Y_{ij}, Y_{i,j+t}) = \alpha^t,$	$t = 0, 1, \ldots, n_i - j$
Unstructured	$\rho(Y_{ij}, Y_{ik}) = \alpha_{jk},$	$j \neq k$

Assuming that the marginal mean $\boldsymbol{\mu}_i$ has been correctly specified as $h(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, Liang and Zeger (1986) showed that, under mild regularity conditions, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from solving (3.18) is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and with covariance matrix

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = I_0^{-1} I_1 I_0^{-1}, \tag{3.19}$$

where

$$I_0 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right) \quad \text{and} \quad I_1 = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \text{Var}(\boldsymbol{y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right),$$

with $V_i = A_i^{1/2} C_i A_i^{1/2}$. When the working correlation structure differs strongly from the true underlying structure, there is no price to pay in terms of the consistency of the asymptotic normality of $\widehat{\boldsymbol{\beta}}$, but such a poor choice may result in loss of efficiency.

The frequentist nature of GEE renders it valid only under the most restrictive of assumptions about the missingness mechanism – MCAR. With incomplete data that are MAR or MNAR, an erroneously specified working correlation matrix may additionally lead to bias (Molenberghs and Kenward, 2007). Extensions of GEE for longitudinal data with dropout of an MAR nature were proposed by Robins, Rotnitzky and Zhao (1995) and these shall be the subject of Chapter 5.

3.6 Model Families

The manner of treatment of the correlated repeated measures present in any longitudinal (or multivariate) data defines three model families. Considering, for the moment, Gaussian outcomes, a *marginal model* is characterized by a marginal mean function of the form

$$E(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}, \tag{3.20}$$

where $\boldsymbol{x}_{ij}$ denotes a vector of covariates for subject i at time point j and $\boldsymbol{\beta}$ denotes the vector of regression coefficients for these. Excepting conditioning on covariates, responses are modeled marginally over all other responses, as opposed to being modeled conditionally on other outcomes, for instance. Often also referred to as population-averaged models, marginal models directly relate covariates to the marginal response expectation, and thus, $\boldsymbol{\beta}$ in (3.20) represent fully marginal covariate effects (e.g., marginal treatment effect).

Random-effects models, on the other hand, further condition on a vector of random effects $\boldsymbol{b}_i$, over and above conditioning on covariates, that is,

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i. \tag{3.21}$$

Under such models, correlations among the repeated measures are addressed by the inclusion of parameters that are specific to a subject, so that given the collection of these subject-specific parameters or random effects, the responses within the longitudinal series are assumed to be independent. In contrast with population-averaged covariate effects, $\boldsymbol{\beta}$ in (3.21) reflects covariate effects that are conditional on a subject with a particular level of the random effects, say $\boldsymbol{b}_i = \boldsymbol{b}_{i_o}$, further denoting somewhat of a sub-population average, for the sub-population of subjects having random effects equal to $\boldsymbol{b}_{i_o}$.

Finally, correlations within the longitudinal responses can be dealt with by considering a particular outcome in the series and modeling it conditionally on the other outcomes (or subsets thereof), thereby defining the family of *conditional models*. A first-order stationary transition model, for instance, involves expectations of the form

$$E(Y_{ij}|Y_{i,j-1}, \ldots, Y_{i1}, \boldsymbol{x}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha Y_{i,j-1}. \qquad (3.22)$$

Conditional models, in general, describe features of one of the responses in a set given others in the same set. While often meaningful in a multivariate (non-longitudinal) context, the use of conditional models for longitudinal data is infrequent since scientific interest in such studies usually lies in estimating the effect of some covariate(s) over time, rather than on its effect at one particular time point given others.

Owing to the elegant properties of the multivariate normal distribution, random-effects models neatly imply a simple marginal model in the linear mixed model case (Verbeke and Molenberghs, 2000). In particular, expectation (3.20) follows from (3.21) either by marginalizing over the random effects or by conditioning on the random-effects vector $\boldsymbol{b}_i = \mathbf{0}$. As was pointed out earlier, in Section 3.4, the linear mixed model (3.9), for instance, which belongs to the random-effects model family, implies marginal model (3.10). Thus, the fixed-effects parameters $\boldsymbol{\beta}$ have both a marginal and hierarchical model interpretation. Certain autoregressive models, in which later-time residuals are expressed in terms of earlier ones, can also lead to particular instances for which the general linear mixed-effects model implies some marginal function of the form (3.20).

Such relations between marginal and random-effects models are much less straightforward in the case of non-Gaussian (e.g., binary) responses, arising primarily from the fact that the mean response is usually modeled through a link function $\eta(\cdot)$. And though formulation of the model remains linear at the level of the link, direct relations with the raw (non-Gaussian) response itself, is frequently of a non-linear nature.

For illustration, consider a binary response, the success probability of which is formulated in terms of a time covariate, t_{ij}, via the logit link. A marginally specified model, i.e., logit $P(Y_{ij} = 1|t_{ij}) = \beta_0 + \beta_1 t_{ij}$, would imply the following marginal mean response:

$$E(Y_{ij}|t_{ij}) = P(Y_{ij} = 1|t_{ij}) \quad = \quad \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}, \tag{3.23}$$

whereas a random-effects model could be formulated as

$$\text{logit } P(Y_{ij} = 1|t_{ij}, b_i) = \beta_0 + \beta_1 t_{ij} + b_i,$$

leading to the following conditional mean response:

$$E(Y_{ij}|t_{ij}, b_i) = P(Y_{ij} = 1|t_{ij}, b_i) \quad = \quad \frac{\exp(\beta_0 + \beta_1 t_{ij} + b_i)}{1 + \exp(\beta_0 + \beta_1 t_{ij} + b_i)}. \tag{3.24}$$

From the latter model, the marginal average evolution over time would require averaging (3.24) over the the random effects:

$$E(Y_{ij}|t_{ij}) = E\left[E(Y_{ij}|t_{ij}, b_i)\right] \quad = \quad E\left[\frac{\exp(\beta_0 + \beta_1 t_{ij} + b_i)}{1 + \exp(\beta_0 + \beta_1 t_{ij} + b_i)}\right], \tag{3.25}$$

which will clearly not lead to (3.23) because of the non-linearity of expression (3.24) in b_i, thus demonstrating that the implied marginal model from a random effects specification does not necessarily reduce to the marginally specified model.

The consequence of this disparity between random-effects and marginally specified models in the non-Gaussian setting is an obvious distinction between the parameters under each. In contrast with the Gaussian case, for which $\boldsymbol{\beta}$, whether defined on a marginal or random-effects model, can be unambiguously interpreted as marginal covariate effects, for non-Gaussian responses, the choice of model family will dictate either marginal or hierarchical inference on the model parameters. In addition, there is no straightforward relation between the parameter vector $\boldsymbol{\beta}^{RE}$ in the random-effects model and the parameter vector $\boldsymbol{\beta}^{M}$ in the marginal model, except in a few special cases. Moreover, not only are there interpretational differences between both families, but when comparing parameter estimates across families, one often observes substantial differences. The key reason is that they estimate different "true" parameters.

The inherent differences across the model families, particularly for non-Gaussian responses, warrants careful consideration regarding the type of model to be employed. Understanding of the key differences and the implications of each type of model is essential in making the appropriate choice. Ultimately, the choice rests on the research question at hand. If fully marginal covariate effects are of interest, marginal models may be the most appropriate route. From a practical point of view, the computational requirements can also help in the choice. A partially specified marginal model for non-Gaussian responses, for instance, may be less computationally demanding than a fully specified one.

3.7 Additional Concepts

3.7.1 The Bahadur Model

Bahadur (1961) proposed a marginal model for binary outcomes, accounting for the association via marginal correlations. Let the marginal success probability be denoted as $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$ and define the standardized deviations as:

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \qquad \text{and} \qquad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}, \tag{3.26}$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further define the correlations $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$, $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3}), \ldots, \rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{i,n_i})$. Then, the general Bahadur model can be represented by the expression

$$f(\boldsymbol{y}_i) \;=\; f_1(\boldsymbol{y}_i)\,c(\boldsymbol{y}_i), \tag{3.27}$$

where

$$\begin{aligned} f_1(\boldsymbol{y}_i) &= \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}\,(1-\pi_{ij})^{1-y_{ij}} \qquad \text{and} \\ c(\boldsymbol{y}_i) &= 1 + \sum_{j_1<j_2} \rho_{ij_1j_2}e_{ij_1}e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3}e_{ij_1}e_{ij_2}e_{ij_3} \\ &\quad + \ldots + \rho_{i12\ldots n_i}e_{i1}e_{i2}\ldots e_{i,n_i}. \end{aligned}$$

Thus, the probability mass function $f(\boldsymbol{y}_i)$ in (3.27) is the product of the independence model $f_1(\boldsymbol{y}_i)$ and the correction factor $c(\boldsymbol{y}_i)$. One viewpoint is to consider the factor $c(\boldsymbol{y}_i)$ as a model for overdispersion.

Even though a variety of flexible full-likelihood models exist, maximum likelihood can be unattractive due to excessive computational requirements, especially when high-dimensional vectors of correlated data arise, as alluded to in the context of the Bahadur model.

3.7.2 The Multivariate Dale Model

The multivariate Dale model or MDM (Molenberghs and Lesaffre, 1994) extends the bivariate global-cross ratio model (Dale, 1986), which accounts for dependence across multiple ordinal outcomes as well as their dependence on continuous and/or discrete, possibly time-varying, covariates. The model decomposes the joint probabilities of the outcomes into main effects and interactions. Thus, for a series of n outcomes, for instance, the MDM involves describing the marginal distributions for each of the n outcomes, $n(n-1)/2$ pairs of two-way interactions and three-way or higher-ordered associations. The description is completed by specifying link functions and linear predictors for both the univariate margins and the association parameters. A convenient choice for the univariate marginal links is the logistic link function, while the association parameters are often assumed to be constant modeled on a log odds ratio scale. Full specification of the association is done in terms of marginal global odds ratios. A trivariate version of the model has been presented by McCullagh and Nelder (1989), but they did not provide a way to calculate joint probabilities as needed in the likelihood equations. Solutions for the multivariate case have been provided by Molenberghs and Lesaffre (1994), Lang and Agresti (1994) and Glonek and McCullagh (1995). A detailed review, as well as various applications, is given in Molenberghs and Verbeke (2005) (Chapter 7).

Though developed for categorical (ordinal) responses with 2 or more levels, the MDM as applied to binary variables, say $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$, yields a simplification since these indicators define a single cutpoint, thus obviating the need to construct tables with cumulative counts and probabilities. For $n = 3$, the following logistic-regression and odds-ratio formulations describe a trivariate Dale model (subject-specific indices i are omitted for ease of notation):

$$\ln\left(\frac{p_{0++}}{1-p_{0++}}\right) = \boldsymbol{X}_1\boldsymbol{\theta}, \tag{3.28a}$$

$$\ln\left(\frac{p_{+0+}}{1-p_{+0+}}\right) = \boldsymbol{X}_2\boldsymbol{\theta}, \tag{3.28b}$$

$$\ln\left(\frac{p_{++0}}{1-p_{++0}}\right) = \boldsymbol{X}_3\boldsymbol{\theta}, \tag{3.28c}$$

$$\ln\psi_{12} = \ln\left(\frac{p_{00+}(1-p_{0++}-p_{+0+}+p_{00+})}{(p_{0++}-p_{00+})(p_{+0+}-p_{00+})}\right) = \boldsymbol{X}_4\boldsymbol{\theta}, \tag{3.28d}$$

$$\ln\psi_{13} = \ln\left(\frac{p_{0+0}(1-p_{0++}-p_{++0}+p_{0+0})}{(p_{0++}-p_{0+0})(p_{++0}-p_{0+0})}\right) = \boldsymbol{X}_5\boldsymbol{\theta}, \tag{3.28e}$$

$$\ln\psi_{23} = \ln\left(\frac{p_{+00}(1-p_{+0+}-p_{++0}+p_{+00})}{(p_{+0+}-p_{+00})(p_{++0}-p_{+00})}\right) = \boldsymbol{X}_6\boldsymbol{\theta}, \tag{3.28f}$$

$$\ln\psi_{123} = \ln\left(\frac{p_{000}p_{011}p_{101}p_{110}}{p_{001}p_{010}p_{100}p_{111}}\right) = \boldsymbol{X}_7\boldsymbol{\theta}, \tag{3.28g}$$

with $p_{ijk} = P(Y_1 = i, Y_2 = j, Y_3 = k|\boldsymbol{x})$, for $i, j, k = 0, 1$, and a $+$ in lieu of a subscript indicating that the marginal probability over this index needs to be used. Extensions to cases where $n > 3$ follow similar formulations.

When $n = 2$, the MDM reduces to Dale (1986)'s bivariate global cross-ratio model, which will be referred to as the bivariate Dale model (BDM). The joint probabilities $p_{ij} = P(Y_1 = i, Y_2 = j|\boldsymbol{x})$, for $i, j = 0, 1$ can now be decomposed into two marginal distributions for the main effects and one log cross-ratio for the association between the two non-response indicators:

$$\begin{aligned} \ln\left(\frac{p_{0+}}{1-p_{0+}}\right) &= \boldsymbol{X}_1\boldsymbol{\theta}, \\ \ln\left(\frac{p_{+0}}{1-p_{+0}}\right) &= \boldsymbol{X}_2\boldsymbol{\theta}, \\ \ln\psi = \ln\left(\frac{p_{00}p_{11}}{p_{01}p_{10}}\right) &= \boldsymbol{X}_3\boldsymbol{\theta}, \end{aligned} \tag{3.29}$$

which yield the following probabilities $p_{ij}(\boldsymbol{x})$ (dependence on $\boldsymbol{x}$ is omitted for ease of notation):

$$\begin{aligned} p_{00} &= \begin{cases} \dfrac{1 + (p_{0+} + p_{+0})(\psi - 1) - S(p_{0+}, p_{+0}, \psi)}{2(\psi - 1)} & \text{if } \psi \neq 1 \\ p_{0+}p_{+0} & \text{if } \psi = 1 \end{cases}, \\ p_{01} &= p_{0+} - p_{00}, \\ p_{10} &= p_{+0} - p_{00}, \\ p_{11} &= 1 - p_{0+} - p_{+0} - p_{00}, \end{aligned} \tag{3.30}$$

where

$$S(\lambda_1, \lambda_2, \psi) = \sqrt{[1 + (\lambda_1 + \lambda_2)(\psi - 1)]^2 + 4\psi(1 - \psi)\lambda_1\lambda_2}.$$

Extensions or variations to the above model are possible, e.g., associations can be assumed constant, or intercepts and/or covariate parameters can be kept constant over time, or relations between the covariate parameters over time can be included.

Among the advantages of the MDM is the flexibility with which the marginal and association parameters can be modeled. In addition, because some investigators may find the odds ratio easier to interpret, this model also has some interpretative advantages over, say, the multivariate probit model. The MDM, however, does not support the analysis of nominal categorical data, though this can be addressed by modifying the model to cover marginal descriptions using generalized logits instead of cumulative logits. Finally, one of the less attractive features of the model is that it is only implicitly defined, i.e., only parts of the model are "visible" – the marginal probabilities and the association structure. The consequence is that some statistical properties of the model are "hidden."

3.7.3 The BRD Family of Models

Baker, Rosenberger and DerSimonian (BRD, 1992) proposed a log-linear based family of models for the four-way classification of two binary outcomes, together with their respective missingness indicators. The generic expressions for the cell counts and corresponding probabilities are $Z_{r_1r_2,j_1j_2}$ and $\pi_{r_1r_2,j_1j_2}$, respectively, where $r_\ell = 0(1)$ if the measurement for outcome ℓ is missing (observed) and j_ℓ (=1,2) denotes the value for binary outcome ℓ (e.g., for the SPO Survey, 1=yes, 2=no). For instance, $Z_{11,21}$ denotes the number of completers, i.e., $(r_1, r_2) = (1, 1)$, with a (2) on the first outcome and a (1) on the second outcome. The models can be written as:

$$\begin{aligned}
\nu_{11,j_1j_2} &= E(Z_{11,j_1j_2}) = Z_{++,++}\pi_{11,j_1j_2}, \\
\nu_{10,j_1j_2} &= E(Z_{10,j_1j_2}) = \nu_{11,j_1j_2}\beta_{j_1j_2}, \\
\nu_{01,j_1j_2} &= E(Z_{01,j_1j_2}) = \nu_{11,j_1j_2}\alpha_{j_1j_2}, \\
\nu_{00,j_1j_2} &= E(Z_{00,j_1j_2}) = \nu_{11,j_1j_2}\alpha_{j_1j_2}\beta_{j_1j_2}\gamma,
\end{aligned}$$

with

$$\alpha_{j_1j_2} = \frac{\phi_{01|j_1j_2}}{\phi_{11|j_1j_2}}, \qquad \beta_{j_1j_2} = \frac{\phi_{10|j_1j_2}}{\phi_{11|j_1j_2}}, \qquad \gamma = \frac{\phi_{11|j_1j_2}\phi_{00|j_1j_2}}{\phi_{10|j_1j_2}\phi_{01|j_1j_2}},$$

where $\phi_{r_1r_2|j_1j_2} = P(R_1 = r_1, R_2 = r_2|Y_1 = j_1, Y_2 = j_2)$.

The α (β) parameters describe missingness in the first (second) outcome as the proportion of subjects with a missing response on the first (second) outcome relative to the proportion of subjects with both responses present, given a particular response combination (j_1, j_2). The γ parameter, on the other hand, captures the interaction between the two missingness indicators via the (conditional) odds ratio for a given response combination. The subscripts are missing from γ since Baker, Rosenberger and DerSimonian (1992) have shown that this quantity is independent of j_1 and j_2 in every identifiable model. These authors considered nine models, based on setting $\alpha_{j_1j_2}$ and $\beta_{j_1j_2}$ constant in one or both indices, and, enumerated using the 'BRD' abbreviation, are:

$$\begin{array}{llllll}
\text{BRD1}: & (\alpha_{..}, \beta_{..}) & \text{BRD4}: & (\alpha_{..}, \beta_{.j_2}) & \text{BRD7}: & (\alpha_{.j_2}, \beta_{.j_2}) \\
\text{BRD2}: & (\alpha_{..}, \beta_{j_1.}) & \text{BRD5}: & (\alpha_{j_1.}, \beta_{..}) & \text{BRD8}: & (\alpha_{j_1.}, \beta_{.j_2}) \\
\text{BRD3}: & (\alpha_{.j_2}, \beta_{..}) & \text{BRD6}: & (\alpha_{j_1.}, \beta_{j_1.}) & \text{BRD9}: & (\alpha_{.j_2}, \beta_{j_1.}).
\end{array}$$

Interpretation is straightforward; for example, in BRD1, both missingness indicators do not depend on the responses, thereby characterizing an MCAR mechanism. In view, however, of the nesting relations among the different types of mechanisms described in Section 3.3, although BRD1 is more precisely classified as MCAR, it can

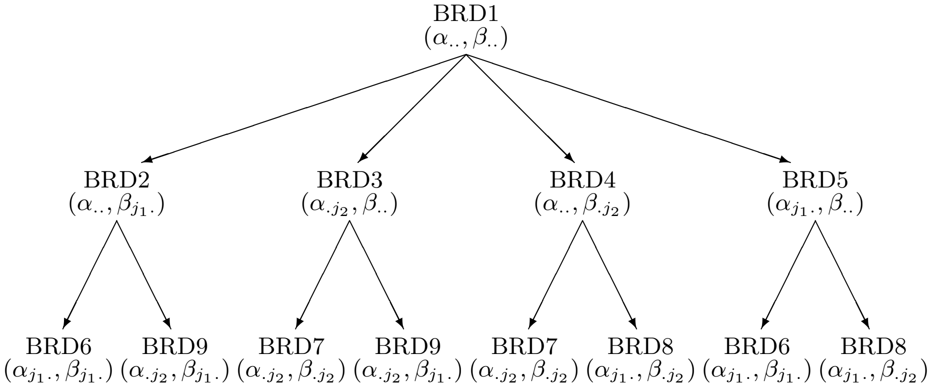

BRD1
$(\alpha_{..}, \beta_{..})$
BRD2
$(\alpha_{..}, \beta_{j_1 .})$
BRD3
$(\alpha_{.j_2}, \beta_{..})$
BRD4
$(\alpha_{..}, \beta_{.j_2})$
BRD5
$(\alpha_{j_1 .}, \beta_{..})$
BRD6
$(\alpha_{j_1 .}, \beta_{j_1 .})$
BRD9
$(\alpha_{.j_2}, \beta_{j_1 .})$
BRD7
$(\alpha_{.j_2}, \beta_{.j_2})$
BRD9
$(\alpha_{.j_2}, \beta_{j_1 .})$
BRD7
$(\alpha_{.j_2}, \beta_{.j_2})$
BRD8
$(\alpha_{j_1 .}, \beta_{.j_2})$
BRD6
$(\alpha_{j_1 .}, \beta_{j_1 .})$
BRD8
$(\alpha_{j_1 .}, \beta_{.j_2})$

Figure 3.1: *Graphical representation of the nesting structure within the BRD model family.*

also be considered a special case of MAR, or even MNAR. In BRD4, missingness in the first variable is constant, while missingness in the second variable depends on its value. Moreover, BRD6–BRD9 saturate the observed data degrees of freedom, while the lower numbered ones leave room for evaluating the goodness-of-fit of the model to the observed data. The nesting structure existing within the nine BRD models is displayed in Figure 3.1.

3.7.4 The Diggle-Kenward Model

Diggle and Kenward (1994) proposed a framework for modeling longitudinal data of a continuous nature with monotone missingness. Using a selection model formulation, as in (3.1), they combine a multivariate normal marginal model for the measurements $\boldsymbol{Y}_i$ with a logistic regression model for dropout D_i conditional on the measurements. The measurement model assumes that the vector $\boldsymbol{Y}_i$ satisfies the linear regression model $\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, V_i)$, where the matrix V_i can be left unstructured or assumed to be of a specific form, e.g., resulting from a linear mixed model, a factor-analytic structure, or spatial covariance structure (Verbeke and Molenberghs, 2000). The dropout model, on the other hand, takes the form of a logistic regression, for example,

$$\text{logit } P(D_i = j | D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi}) = \psi_0 + \psi_c\, y_{ij} + \psi_p\, y_{i,j-1}, \tag{3.31}$$

where $\boldsymbol{h}_{ij} = (y_{i1}, y_{i2}, \ldots, y_{i,j-1})$ represents the history of subject i up to time $t_{i,j-1}$. The latter implies that the conditional probability for dropout at occasion j, given that the subject was still observed at the previous occasion, is allowed to depend on the history $\boldsymbol{h}_{ij}$ of the subject up to time $t_{i,j-1}$, and on the possibly unobserved current outcome y_{ij}, but not on future outcomes $y_{ik}, k > j$. Clearly, the dependence of D_i on the possibly unobserved current outcome y_{ij} characterizes an MNAR mechanism, though special cases can be obtained by setting $\psi_c = 0$, implying MAR, or $\psi_p = \psi_c = 0$, which corresponds to MCAR.

Formulation (3.31) can also be extended to include the complete history, as well as external covariates. In fact, one could allow dropout at a specific occasion to be related to future responses as well, but this is usually counterintuitive. Moreover, including future outcomes seriously complicates the calculations, since computation of the likelihood in (3.5) will then require evaluation of a possibly high-dimensional integral.

From the conditional probabilities defined in (3.31), the (unconditional) probabilities of dropout at each occasion can then be calculated as follows:

$$P(D_i = j|\boldsymbol{y}_i, \boldsymbol{\psi}) = P(D_i = j|\boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi})$$

$$= \begin{cases} P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi}) & \text{for } j = 2, \\[2ex] P(D_i = j|D_i \geq j, \boldsymbol{h}_{ij}, y_{ij}, \boldsymbol{\psi}) & \\ \quad \times \displaystyle\prod_{k=2}^{j-1} [1 - P(D_i = k|D_i \geq k, \boldsymbol{h}_{ik}, y_{ik}, \boldsymbol{\psi})] & \text{for } j = 3, 4, \ldots, n_i, \\[2ex] \displaystyle\prod_{k=2}^{n_i} [1 - P(D_i = k|D_i \geq k, \boldsymbol{h}_{ik}, y_{ik}, \boldsymbol{\psi})] & \text{for } j = n_i + 1. \end{cases} \tag{3.32}$$

Diggle and Kenward (1994) obtained parameter and precision estimates by means of maximum likelihood. The likelihood necessitates marginalization over the unobserved outcomes, which, in practice, can involve relatively tedious and computationally demanding forms of numerical integration. This, combined with likelihood surfaces tending to be rather flat or otherwise awkward in shape, makes the model difficult to use – a feature shared by virtually all MNAR models. Application of the Diggle-Kenward model to various data sets can be found in Molenberghs and Kenward (2007) and Verbeke and Molenberghs (2000).

4

Missing At Random *versus* Missing Not At Random

The suitability of any methodology applied to incomplete longitudinal data, and as a consequence, the validity of resulting inferences, depends heavily on assumptions made regarding the nature of the missingness mechanism. While the more established statistical techniques are sufficiently equipped with means to assess the validity of underlying assumptions (e.g., diagnostic checking of residuals in regression analysis), incomplete data methods are much less developed in this respect. As it is close to impossible to ascertain the nature of the incompleteness, one needs to rely on the strength of one's belief that a particular mechanism is in effect. The implications of an incorrect choice on the resulting conclusions make the task even more crucial.

Traditionally, simple methods such as a CCA or simple forms of imputation (e.g., LOCF) have been in use, though such analyses have been shown to be prone to severe bias and/or losses of efficiency and should be avoided, as has been pointed out here (Section 3.5.1) and elsewhere (Kenward and Molenberghs, 2009; Molenberghs and Kenward, 2007; Molenberghs, Kenward and Lesaffre, 1997). Since a likelihood-based or Bayesian analysis is valid under MAR, for as long as all observed data are included into the analysis, direct likelihood approaches, their Bayesian counterparts, or multiple imputation, are often regarded as candidates for the primary analysis.

One can never exclude, however, the possibility that MNAR models may be operating. Even though a variety of statistical models have been proposed for the MNAR situation (Diggle and Kenward, 1994; Baker, 1995; Molenberghs, Kenward and Lesaffre, 1997; Troxel, Harrington and Lipsitz, 1998), and in spite of the dramatically increased computational power, such models are prone to considerable sensitivity. This was made clear by a variety of discussants to Diggle and Kenward (1994), such as Laird (1994), Little (1994b) and Rubin (1994). Several authors have laid bare such sensitivities and proposed methods for informal and formal sensitivity analysis (Kenward, 1998; Robins, Rotnitzky and Scharfstein, 1998; Molenberghs, Kenward and Goetghebeur, 2001; Van Steen *et al.*, 2001; Verbeke *et al.*, 2001; Thijs *et al.*, 2002; Jansen *et al.*, 2003). Overviews are provided in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005).

There is no quick and easy strategy to determine the missing data mechanism that is believed to be operating and that will eventually be assumed for the analysis. One view is that testing the MAR null hypothesis against an MNAR alternative is of a conventional nature. Although Diggle and Kenward (1994) have conducted such tests, it is very important to realize that such tests are conditional on the alternative model holding. Strictly speaking, the correctness of the alternative model can only be verified in as far as it fits the *observed* data. Thus, evidence for or against MNAR can only be provided within a particular, predefined parametric family, the plausibility of which cannot be verified in empirical terms alone.

In this chapter, an important result is presented regarding the impossibility of an overall (omnibus) assessment of MAR *versus* MNAR. This arises from the result of Molenberghs *et al.* (2008) that a uniquely defined MAR counterpart can be obtained for an MNAR model and this counterpart produces exactly the same fit as the original MNAR model, in the sense that it produces exactly the same predictions to the observed data (e.g., fitted counts in an incomplete contingency table) as the original MNAR model, and depends on exactly the same parameter vector. It will also be shown here that, while this so-called MAR counterpart generally does not belong to a conventional parametric family, its existence has important ramifications.

4.1 General Result

In this section, it will be shown that for every MNAR model fitted to a set of data, there is an MAR counterpart providing exactly the same fit to the data. Here, the concept of model fit should be understood as measured using such conventional methods as deviance measures and, of course, in as far as the observed data are concerned. The following steps are involved:

(1) fitting an MNAR model to the data;

(2) reformulating the fitted model in PMM form;

(3) replacing the density or distribution of the unobserved measurements given the observed ones and given a particular response pattern by its MAR counterpart; and,

(4) establishing that such an MAR counterpart uniquely exists.

Throughout this section, covariates $\boldsymbol{x}_i$ will be suppressed from notation, though they are assumed to be present.

In the first step, an MNAR model is fitted to the observed set of data. Recall, from Section 3.4, that the observed-data likelihood is:

$$L = \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) \, d\boldsymbol{y}_i^m. \tag{4.1}$$

Denoting the resulting parameter estimates, e.g., obtained by likelihood-based or Bayesian methods, as $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$, respectively, the fit to the hypothetical full data is

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \widehat{\boldsymbol{\theta}}) \, f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \widehat{\boldsymbol{\psi}}). \tag{4.2}$$

To undertake the second step, full density (4.2) is re-expressed in PMM form as:

$$f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \, f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \, f(\boldsymbol{y}_i^o | \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \, f(\boldsymbol{r}_i | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}). \tag{4.3}$$

A similar reformulation can be considered for an SPM. In a PMM, the model would have been expressed in this form to begin with. Note that, in line with PMM theory, the first term on the right hand side of (4.3),

$$f(\boldsymbol{y}_i^m | \boldsymbol{y}_i^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}), \tag{4.4}$$

is not identified from the observed data. In this case, it is determined solely from modeling assumptions. Within the PMM framework, identifying restrictions have to be considered (Little, 1994a; Molenberghs *et al.*, 1998; Kenward, Molenberghs and Thijs, 2003).

The third step requires replacing (4.4) by the appropriate MAR counterpart. To this end, the following lemma, formulating MAR equivalently within the PMM framework, is necessary.

Lemma 1 *In the PMM framework, the missing data mechanism is MAR if, and only if,*

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i,\boldsymbol{\theta}) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{\theta}).$$

This means that, in a given pattern, the conditional distribution of the unobserved components given the observed ones equals the corresponding distribution marginalized over the patterns. Lemma 1 further implies that MAR can be formulated in terms of $\boldsymbol{R}$ given $\boldsymbol{Y}$, but also in terms of $\boldsymbol{Y}$ given $\boldsymbol{R}$. The proof is rather straightforward and similar to what can be found in Molenberghs *et al.* (1998).

Proof of Lemma 1

Suppressing parameters and covariates from notation, the decomposition of the full data density, in both SeM and PMM fashion, whereby MAR is applied to the SeM version, produces:

$$f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m)\,f(\boldsymbol{r}_i|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m|\boldsymbol{r}_i)\,f(\boldsymbol{r}_i). \tag{4.5}$$

Further factoring the right hand side and moving the second factor on the left to the right as well gives:

$$\begin{aligned}
f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)\,\frac{f(\boldsymbol{y}_i^o|\boldsymbol{r}_i)\,f(\boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)}\\
f(\boldsymbol{y}_i^o,\boldsymbol{y}_i^m) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)\,\frac{f(\boldsymbol{y}_i^o,\boldsymbol{r}_i)}{f(\boldsymbol{r}_i|\boldsymbol{y}_i^o)}\\
f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o)f(\boldsymbol{y}_i^o) &= f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i)\,f(\boldsymbol{y}_i^o),
\end{aligned}$$

and hence

$$f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i).$$

☒

Using Lemma 1, it is clear that $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}})$ needs to be replaced with

$$h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i) = h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}}), \tag{4.6}$$

where the $h(\cdot)$ notation is used for shorthand purposes. Note that the density in (4.6) follows from the SeM-type marginal density of the complete data vector. Sometimes, therefore, it may be more convenient to replace the notation $\boldsymbol{y}_i^o$ and $\boldsymbol{y}_i^m$ by one that explicitly indicates which components are observed and missing in pattern $\boldsymbol{r}_i$ under consideration:

$$h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o,\boldsymbol{r}_i) = h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f[(y_{ij})_{r_j=0}|(y_{ij})_{r_j=1},\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}}]. \tag{4.7}$$

Thus, (4.7) provides a unique way of extending the model fit to the observed data, belonging to the MAR family. As stated before, the above construction does not lead to a member of a conventional parametric family. While this obviously implies limitations on its use, such is not dissimilar to the construction of some semi- and non-parametric estimators. Also, it helps to understand that an overall, definitive conclusion about the nature of the missing data mechanism is not possible, even though one can make progress if attention is confined to a given parametric family, in which one puts sufficiently strong prior belief. To show formally that the fit remains the same, consider the observed-data likelihood based on (4.1) and (4.3):

$$\begin{aligned}\widehat{L} &= \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m|\widehat{\boldsymbol{\theta}})\, f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \widehat{\boldsymbol{\psi}})\, d\boldsymbol{y}_i^m \\ &= \prod_i \int f(\boldsymbol{y}_i^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, d\boldsymbol{y}_i^m \\ &= \prod_i \quad f(\boldsymbol{y}_i^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) \\ &= \prod_i \int f(\boldsymbol{y}_i^o|\boldsymbol{r}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, f(\boldsymbol{r}_i|\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})\, h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o)\, d\boldsymbol{y}_i^m.\end{aligned}$$

The above results justify the following theorem.

Theorem 1 *Every fit to the observed data, obtained from fitting an MNAR model to a set of incomplete data, is exactly reproducible from an MAR decomposition.*

The key computational consequence is the need to compute $h(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o)$ in (4.6) or (4.7). This means that, for each pattern, the conditional density of the unobserved measurements given the observed ones needs to be extracted from the marginal distribution of the complete set of measurements. Molenberghs *et al.* (1998) have shown that, for the case of dropout, the so-called *available case missing value restrictions* (ACMV) provide a practical computational scheme. Precisely, ACMV states that $\forall t \geq 2$ and $\forall s < t$:

$$f(y_{it}|y_{i1}, \cdots, y_{i,t-1}, d_i = s) = f(y_{it}|y_{i1}, \cdots, y_{i,t-1}, d_i \geq t). \tag{4.8}$$

In other words, the density of a missing measurement, conditional on the measurement history, is determined from the corresponding density over all patterns for which all of these measurements are observed. For example, the density of the third measurement in a sequence, given the first and second ones, in patterns with only 1 or 2 measurements taken, is determined from the corresponding density over all patterns with 3 or more measurements. Thijs *et al.* (2002) and Verbeke and Molenberghs (2000) derived a practical computational method for the factors in (4.8), which, upon letting α_d denote the probability of belonging to pattern d, are given by:

$$\begin{aligned} f(y_{it}|y_{i1},\cdots,y_{i,t-1},d_i=s) & \\ &= \frac{\displaystyle\sum_{d=s}^{n} \alpha_d\, f_d(y_{i1},\ldots,y_{is})}{\displaystyle\sum_{d=s}^{n} \alpha_d\, f_d(y_{i1},\ldots,y_{i,s-1})} \qquad (4.9)\\ &= \sum_{d=s}^{n} \left[\frac{\alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}{\sum_{d=s}^{n_i} \alpha_d f_d(y_{i1},\ldots,y_{i,s-1})}\right] f_d(y_s|y_{i1},\ldots,y_{i,s-1}). \qquad (4.10)\end{aligned}$$

The above identifications for the monotone case are useful when an MNAR pattern-mixture model has been fitted to begin with, since then, the identifications under MAR can be calculated from the pattern-specific marginal distributions. When a selection model is fitted in the initial step, on the other hand, all conditional distributions needed in (4.7), can be derived from the estimated $f(y_{i1},\ldots,y_{in_i}|\widehat{\boldsymbol{\theta}})$. For an MNAR PMM initial model for non-monotone missing data patterns, it is necessary to first rewrite the PMM in SeM form, then subsequently derive the required conditional distributions for the so-obtained SeM measurement model. This essentially comes down to calculating a weighted average of the pattern-specific measurement models. In some cases, such as for contingency tables, this step can be done in an alternative way by fitting a saturated MAR SeM to the fit obtained from the PMM.

The following sections illustrate and contrast the monotone and non-monotone cases using a bivariate and trivariate outcome with dropout on the one hand, and a bivariate non-monotone outcome on the other. While the theorem applies to both the monotone and non-monotone settings, it is insightful to see that only for the former do relatively simple and intuitively appealing expressions arise, while the latter setting involves the need for iterative computation.

4.1.1 A Bivariate Outcome With Dropout

Considering monotone missing data patterns (dropout) for the case of two outcomes, the first of which is always observed and the second one partially missing, the SeM and PMM decompositions will be equated, enabling derivation of the expressions for the MAR counterparts. As it is likewise interesting and straightforward to derive results under the MCAR setting, these shall be presented as well.

Dropping covariates, parameters, and the subject index i from notation, the SeM-PMM equivalence for the case of the bivariate outcome with dropout is given by:

$$\begin{aligned} f(y_1,y_2)\,\widetilde{g}(d=2|y_1,y_2) &= f_2(y_1,y_2)\,\widetilde{\alpha}(d=2),\\ f(y_1,y_2)\,\widetilde{g}(d=1|y_1,y_2) &= f_1(y_1,y_2)\,\widetilde{\alpha}(d=1). \end{aligned}$$

Note that this is the setting considered by Glynn, Laird and Rubin (1986). Here, $\widetilde{g}(\cdot)$ is used for the SeM dropout model, with $\widetilde{\alpha}(\cdot)$ denoting the PMM probabilities to belong to one of the patterns. Since $\widetilde{\alpha}(d=1)+\widetilde{\alpha}(d=2)=1$ and a similar result holds for the $\widetilde{g}(\cdot)$ functions, it is convenient to write:

$$\begin{aligned} f(y_1,y_2)\,g(y_1,y_2) &= f_2(y_1,y_2)\,\alpha & (4.11)\\ f(y_1,y_2)\,[1-g(y_1,y_2)] &= f_1(y_1,y_2)\,[1-\alpha]. & (4.12) \end{aligned}$$

Assuming MCAR, it is clear that $\alpha = g(y_1,y_2)$, producing, without any difficulty,

$$f(y_1,y_2) = f_2(y_1,y_2) = f_1(y_1,y_2). \qquad (4.13)$$

Under MAR, y_2 has to be removed from $g(\cdot)$ for incomplete observations, but since a single parametric function is assumed for the missingness model, it follows that $g(y_1,y_2)=g(y_1)$ and hence (4.11) produces

$$f(y_1)\,f(y_2|y_1)\,g(y_1) = f_2(y_1)\,f_2(y_2|y_1)\,\alpha.$$

Upon reordering,

$$\frac{f(y_1)\,g(y_1)}{f_2(y_1)\,\alpha} = \frac{f_2(y_2|y_1)}{f(y_2|y_1)}. \qquad (4.14)$$

The same arguments can be applied to (4.12), from which the following can be derived:

$$f(y_2|y_1) = f_2(y_2|y_1) = f_1(y_2|y_1). \qquad (4.15)$$

Note that (4.15) is strictly weaker than (4.13). The last term in (4.15) is not identified by itself, and hence, needs to be set equal to its counterpart from the completers which, in turn, is equal to the marginal distribution. This is in agreement with (4.7) as well as with the specific identifications applicable in the monotone and hence ACMV setting.

4.1.2 A Trivariate Outcome With Dropout

In the bivariate case, it can be observed that identification (4.15) does not involve mixtures. This changes as soon as there are three or more outcomes. The equations corresponding to (4.11)–(4.12), specialized to the MAR case, are:

$$
\begin{aligned}
f(y_1,y_2,y_3)\,g_0 &= f_0(y_1,y_2,y_3)\,\alpha_0, &(4.16)\\
f(y_1,y_2,y_3)\,g_1(y_1) &= f_1(y_1,y_2,y_3)\,\alpha_1, &(4.17)\\
f(y_1,y_2,y_3)\,g_2(y_1,y_2) &= f_2(y_1,y_2,y_3)\,\alpha_2, &(4.18)\\
f(y_1,y_2,y_3)\,g_3(y_1,y_2) &= f_3(y_1,y_2,y_3)\,\alpha_3. &(4.19)
\end{aligned}
$$

Pattern 0, without any follow-up measurements, has been included in the above formulations, but discussion of this will be deferred for the moment. Also, it can be noted that $g_3(\cdot)$ can be written as a function of y_3 as well, but because the sum of the $g_d(\cdot)$ equals one, it is clear that $g_3(\cdot)$ ought to be independent of y_3.

With arguments similar to the ones developed in the case of two measurements, (4.19) can be rewritten as:

$$\frac{f(y_1,y_2)}{f_3(y_1,y_2)}\cdot\frac{g_3(y_1,y_2)}{\alpha_3}=\frac{f_3(y_3|y_1,y_2)}{f(y_3|y_1,y_2)}.$$

Exactly the same consideration can be made based on (4.18), and hence

$$f_3(y_3|y_1,y_2)=f(y_3|y_1,y_2)=f_2(y_3|y_1,y_2). \qquad (4.20)$$

The first factor identifies the second one, and hence also the third one. Equation (4.17) leads to

$$f_1(y_2,y_3|y_1)=f(y_2,y_3|y_1),$$

which produces, in fact, two separate identities:

$$
\begin{aligned}
f_1(y_2|y_1) &= f(y_2|y_1), &(4.21)\\
f_1(y_3|y_1,y_2) &= f(y_3|y_1,y_2)=f_3(y_3|y_1,y_2)=f_2(y_3|y_1,y_2). &(4.22)
\end{aligned}
$$

For the latter one, identity (4.20) has been used as well. The density $f(y_2|y_1)$, needed in (4.21), is determined from the general ACMV result (4.10) as

$$f(y_2|y_1)=\frac{\alpha_2 f_2(y_2|y_1)+\alpha_3 f_3(y_2|y_1)}{\alpha_2+\alpha_3}.$$

Finally, turning attention to (4.16), it is clear that $g_0=\alpha_0$ and hence also $f_0(y_1,y_2,y_3)=f(y_1,y_2,y_3)$. From the latter density, only $f(y_1)$ has not been determined yet, but this one follows again very easily from the general ACMV result, i.e.,

$$f(y_1)=\frac{\alpha_1 f_1(y_1)+\alpha_2 f_2(y_1)+\alpha_3 f_3(y_1)}{\alpha_1+\alpha_2+\alpha_3}.$$

In summary, the necessary MAR identifications easily follow from both the PMM and the SeM formulations of the model.

4.1.3 A Bivariate Outcome With Non-Monotone Missingness

For a bivariate outcome with non-monotone missingness, the corresponding equations to (4.11)–(4.12) and (4.16)–(4.19) are:

$$\begin{aligned} f(y_1,y_2)\,g_{00}(y_1,y_2) &= f_{00}(y_1,y_2)\,\alpha_{00}, \quad (4.23)\\ f(y_1,y_2)\,g_{10}(y_1,y_2) &= f_{10}(y_1,y_2)\,\alpha_{10}, \quad (4.24)\\ f(y_1,y_2)\,g_{01}(y_1,y_2) &= f_{01}(y_1,y_2)\,\alpha_{01}, \quad (4.25)\\ f(y_1,y_2)\,g_{11}(y_1,y_2) &= f_{11}(y_1,y_2)\,\alpha_{11}. \quad (4.26)\end{aligned}$$

Clearly, under MCAR, the $g_{r_1r_2}(\cdot)$ functions do not depend on the outcomes, and hence, $f_{r_1r_2}(y_1,y_2) = f(y_1,y_2)$ for all four patterns. For the MAR case, (4.23)–(4.26) simplify to

$$\begin{aligned} f(y_1,y_2)\,g_{00} &= f_{00}(y_1,y_2)\,\alpha_{00}, \quad (4.27)\\ f(y_1,y_2)\,g_{10}(y_1) &= f_{10}(y_1,y_2)\,\alpha_{10}, \quad (4.28)\\ f(y_1,y_2)\,g_{01}(y_2) &= f_{01}(y_1,y_2)\,\alpha_{01}, \quad (4.29)\\ f(y_1,y_2)\,g_{11}(y_1,y_2) &= f_{11}(y_1,y_2)\,\alpha_{11}. \quad (4.30)\end{aligned}$$

It can be observed that there are four identifications across the $g_{r_1r_2}(y_1,y_2)$ functions, i.e.,

$$g_{00} + g_{10}(y_1) + g_{01}(y_2) + g_{11}(y_1,y_2) = 1,$$

for each (y_1,y_2). Also, $\sum_{r_1,r_2}\alpha_{r_1,r_2} = 1$. Applying the usual algebra to (4.27)–(4.30) yields three identifications for the unobservable densities:

$$\begin{aligned} f_{00}(y_1,y_2) &= f(y_1,y_2), \quad (4.31)\\ f_{10}(y_1|y_2) &= f(y_1|y_2), \quad (4.32)\\ f_{01}(y_2|y_1) &= f(y_2|y_1). \quad (4.33)\end{aligned}$$

Using these in conjunction with the identifiable parts of the distributions yields the MAR counterpart.

4.2 The General Case of Incomplete Contingency Tables

In Sections 4.1.1–4.1.3, general identification schemes for an MAR extension of a fitted model to a binary or trivariate outcome with dropout, as well as to a bivariate outcome with non-monotone missingness, have been derived. Whereas the monotone cases provide explicit expressions in terms of the pattern-specific densities, (4.31)–(4.33) provide an identification only in terms of the marginal probability. This in itself is not a problem, since the marginal density is always available, either directly when an SeM is fitted, or through marginalization when a PMM or an SPM is fitted.

In the specific case of contingency tables, further progress can be made. As will be shown shortly, a saturated MAR model is indeed always available, for any incomplete contingency table setting. This implies that one can start from the fit of an MNAR model to the observed data, and then extend it, using this result, towards MAR. The general formulation of the result, as well as a discussion of its precise implications for practice, shall be presented in this section.

Consider a $\prod_{k=1}^{n} c_k$ contingency table with supplemental margins, where k indexes the n dimensions in the table and c_k is the number of alternatives the k^{th} categorical variable can take. The table of completers is indexed by $\boldsymbol{r} = \mathbf{1} = (1, \ldots, 1)$. A particular incomplete table is indexed by $\boldsymbol{r} \neq \mathbf{1}$. The full set of tables can but does not have to be present. The number of cells is:

$$\#\text{cells} = \sum_{\boldsymbol{r}} \prod_{k=1}^{n} c_k^{r_k}. \tag{4.34}$$

The measurement model probabilities can be denoted by $p_{\boldsymbol{j}} = p_{j_1 \ldots j_n}$, for $j_k = 1, \ldots, c_k$ and $k = 1, \ldots, n$, and these probabilities sum to one. The missingness probabilities, assuming MAR, are:

$$p(\boldsymbol{r}|\boldsymbol{j}) = \begin{cases} p(\boldsymbol{r}|j_k \text{ with } r_k = 1) & \text{if } \boldsymbol{r} \neq \mathbf{1}, \\ 1 - \sum_{\boldsymbol{r} \neq \mathbf{1}} p(\boldsymbol{r}|\boldsymbol{j}) & \text{if } \boldsymbol{r} = \mathbf{1}. \end{cases} \tag{4.35}$$

Summing over $\boldsymbol{r}$ implies summing over those patterns for which actual observations are available. The number of parameters in the saturated model can be computed as

$$\#\text{parameters} = \left(\prod_{k=1}^{n} c_k - 1 \right) + \sum_{\boldsymbol{r} \neq \mathbf{1}} \prod_{k=1}^{n} c_k^{r_k}. \tag{4.36}$$

The first term in (4.36) is for the measurement model, while the second one is for the missingness model. Clearly, the number of parameters equals one less than the number of cells, establishing the claim. The situation where covariates are present is covered automatically, merely by considering one extra dimension in the contingency table, $j = 0$ say, with c_0 referring to the total number of covariate levels in the set of data.

In what follows, the implications for the simple but important settings studied in Sections 4.1.1 and 4.1.3 are laid out.

4.2.1 A Bivariate Contingency Table With Dropout

In Section 4.1.1, identifications were derived for the general case of bivariate outcomes with monotone missingness. In the specific setting of contingency tables, these can be derived as well by further fitting the saturated MAR model, described in the previous section, to the fit obtained from the original MNAR model.

Suppose that $z_{2,jk}$ and $z_{1,j}$ denote, for the completers and dropouts, respectively, the counts obtained from the fit of the original model, p_{jk} denote the measurement model probabilities, and q_j the dropout probabilities. Then, due to ignorability, the likelihood factors into two components:

$$\ell_1 = \sum_{j,k} z_{2,jk} \ln p_{jk} + \sum_{j} z_{1,j} \ln p_{j+} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \tag{4.37}$$

$$\ell_2 = \sum_{j,k} z_{2,jk} \ln q_j + \sum_{j} z_{1,j} \ln(1 - q_j). \tag{4.38}$$

An undetermined Lagrange multiplier λ can be used to incorporate the sum constraint on the marginal probabilities. Solving the score equations for (4.37) and (4.38) produces, with simple and well-known algebra,

$$\widehat{p_{jk}} = \frac{1}{n} z_{2,jk} \left(\frac{z_{2,j+} + z_{1,j}}{z_{2,j+}} \right), \tag{4.39}$$

$$\widehat{q_j} = \frac{z_{2,j+}}{z_{2,j+} + z_{1,j}}, \tag{4.40}$$

where n is the total sample size. Combining parameter estimates leads to the new, MAR-based, fitted counts:

$$\widehat{z_{2,jk}} = n \widehat{p_{jk}} \widehat{q_j} = z_{2,jk}, \tag{4.41}$$

$$\widehat{z_{1,jk}} = n \widehat{p_{jk}} (1 - \widehat{q_j}) = z_{1,j} \frac{z_{2,jk}}{z_{2,j+}}, \tag{4.42}$$

$$\widehat{z_{1,j+}} = z_{1,j+}. \tag{4.43}$$

From (4.41) and (4.43) it is clear that the fit in terms of the observed data has not changed. The expansion of the incomplete data into a complete one is described by (4.42). Equations (4.41) and (4.42) can be used to produce the MAR counterpart to the original model, without any additional calculations. This, however, is not so simple for the non-monotone case, as will be shown next.

4.2.2 A Bivariate Contingency Table With Non-Monotone Missingness

For bivariate contingency tables with non-monotone missingness, the counterparts to (4.37)–(4.38) are:

$$\begin{aligned} \ell_1 &= \sum_{j,k} z_{11,jk} \ln p_{jk} + \sum_j z_{10,j} \ln p_{j+} + \sum_k z_{01,k} \ln p_{+k} \qquad (4.44) \\ &+ z_{00} \ln p_{++} - \lambda \left(\sum_{j,k} p_{jk} - 1 \right), \\ \ell_2 &= \sum_{j,k} z_{11,jk} \ln(1 - q_{10,j} - q_{01,k} - q_{00}) \qquad (4.45) \\ &+ \sum_j z_{10,j} \ln q_{10,j} + \sum_k z_{01,k} \ln q_{01,k} + z_{00} \ln g_{00}. \end{aligned}$$

Notation has been modified in accordance with the design. The q quantities correspond to the $g(\cdot)$ model in Section 4.1.3. While $p_{++} = 1$ and hence z_{00} does not contribute information to the measurement probabilities, it does add to the estimation of the missingness model.

Deriving the score equations from (4.44) and (4.45) is straightforward but, unlike in the previous section, no closed form exists. Chen and Fienberg (1974) derived an iterative scheme for the probabilities p_{jk}, based on setting the expected sufficient statistics equal to their *complete-data* counterparts, yielding

$$np_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}} + z_{00} \frac{p_{jk}}{p_{++}},$$

and hence, with $p_{++} = 1$,

$$(n - z_{00})\, p_{jk} = z_{11,jk} + z_{10,j} \frac{p_{jk}}{p_{j+}} + z_{01,k} \frac{p_{jk}}{p_{+k}}. \qquad (4.46)$$

The same equation is obtained from the first derivative of (4.44). The iterative scheme of Chen and Fienberg (1974) results from initiating the process with a set of starting values for the p_{jk}, e.g., from the completers, and then evaluating the right hand side of (4.46). Equating it to the left hand side provides an update for the parameters. The process is repeated until convergence.

While there are no closed-form counterparts to (4.39) and (4.40), the expressions equivalent to (4.41)–(4.43) are:

$$\begin{aligned}
\widehat{z_{11,jk}} &= z_{11,jk}, && (4.47)\\
\widehat{z_{10,jk}} &= z_{10,j}\,\frac{p_{jk}}{p_{j+}}, && (4.48)\\
\widehat{z_{01,jk}} &= z_{01,k}\,\frac{p_{jk}}{p_{+k}}, && (4.49)\\
\widehat{z_{00,jk}} &= z_{00}\,p_{jk}. && (4.50)
\end{aligned}$$

An important difference can be noted between (4.41)–(4.43) on the one hand and (4.47)–(4.50) on the other. In the monotone case, the expressions on the right hand side are in terms of the counts z only, whereas here the marginal probabilities p_{jk}, which have to be determined from a numerical fit, intervene.

4.3 Analysis of the Slovenian Public Opinion Survey

In this section, the practical use of the results described in the previous sections are illustrated on the Slovenian Public Opinion (SPO) Survey data (questions 1 and 3), on which derivation of the MAR counterpart to certain MNAR models will be shown. The ideas developed in this chapter can be illustrated easily by means of 4 models from the BRD family (Section 3.7.3), fitted to the independence and attendance outcomes (Table 2.5). Particular interest shall be paid on models BRD1, BRD2, BRD7 and BRD9. Whereas model BRD1 assumes missingness to be MCAR, all others are of the MNAR type. Model BRD2 has 7 free parameters, and hence does not saturate the observed data degrees of freedom, while models BRD7 and BRD9 saturate the 8 data degrees of freedom. The collapsed data, together with the model fits, are displayed in Table 4.1. Each of the four models is doubled up with its MAR counterpart.

Table 4.1 presents, apart from the raw data, for each of the models and its MAR counterpart, the fit to the observed and the hypothetical complete data. The fits of models BRD7, BRD9 and their MAR counterparts to the observed data, coincide with the observed data. As the theory states, every MNAR model and its MAR counterpart produce exactly the same fit to the observed data, which is therefore also seen for BRD1 and BRD2. However, while models BRD1 and BRD1(MAR) coincide in their fit to the hypothetical complete data, this is not the case for the other three models. The reason is clear: since model BRD1 is MCAR, which can be also more broadly classified as belonging to the MAR family, its counterpart BRD1(MAR) will not produce any difference, but merely copies the fit of BRD1 to the unobserved data,

Table 4.1: *The Slovenian Public Opinion Survey, analysis restricted to the independence and attendance questions. Observed data and the fit of models BRD1, BRD2, BRD7 and BRD9, and their MAR counterparts, to the observed data and to the hypothetical complete data. Rows (columns) of the contingency tables correspond to 'yes'* vs. *'no' on the attendance (independence) question. The four tables in each row correspond to people: (i) responding to both questions; (ii) responding to independence only; (iii) responding to attendance only; and, (iv) responding to neither question.*

Observed data & fit of BRD7, BRD7(MAR), BRD9 and BRD9(MAR) to incomplete data

1439	78	159		144	54	136
16	16	32				

Fit of BRD1 and BRD1(MAR) to incomplete data

1381.6	101.7	182.9		179.7	18.3	136.0
24.2	41.4	8.1				

Fit of BRD2 and BRD2(MAR) to incomplete data

1402.2	108.9	159.0		181.2	16.8	136.0
15.6	22.3	32.0				

Fit of BRD1 and BRD1(MAR) to complete data

1381.6	101.7	170.4	12.5	176.6	13.0	121.3	9.0
24.2	41.4	3.0	5.1	3.1	5.3	2.1	3.6

Fit of BRD2 to complete data

1402.2	108.9	147.5	11.5	179.2	13.9	105.0	8.2
15.6	22.3	13.2	18.8	2.0	2.9	9.4	13.4

Fit of BRD2(MAR) to complete data

1402.2	108.9	147.7	11.3	177.9	12.5	121.2	9.3
15.6	22.3	13.3	18.7	3.3	4.3	2.3	3.2

Fit of BRD7 to complete data

1439	78	3.2	155.8	142.4	44.8	0.4	112.5
16	16	0.0	32.0	1.6	9.2	0.0	23.1

Fit of BRD9 to complete data

1439	78	150.8	8.2	142.4	44.8	66.8	21.0
16	16	16.0	16.0	1.6	9.2	7.1	41.1

Fit of BRD7(MAR) and BRD9(MAR) to complete data

1439	78	148.1	10.9	141.5	38.4	121.3	9.0
16	18	11.8	20.2	2.5	15.6	2.1	3.6

given the observed ones. Finally, while BRD7 and BRD9 produce a different fit to the complete data, BRD7(MAR) and BRD9(MAR) coincide. This is because the fits of BRD7 and BRD9 coincide with respect to their fit to the observed data, and indeed, due to their saturation, coincide with the observed data as such. This fit is the sole basis for the models' MAR extensions. It is noteworthy that, while BRD7, BRD9 and BRD7(MAR)≡BRD9(MAR) all saturate the observed data degrees of freedom, their complete-data fits are dramatically different.

It is worthwhile to examine the implications of these results for the primary estimand θ, the proportion of people voting YES by simultaneously being in favor of independence and deciding to take part in the vote. Rubin, Stern and Vehovar (1995) considered, apart from simple models such as complete case analysis ($\widehat{\theta} = 0.928$) and available case analyses ($\widehat{\theta} = 0.929$), both ignorable models ($\widehat{\theta} = 0.892$ when based on the two main questions and $\widehat{\theta} = 0.883$ when using the secession question as an auxiliary variable) and a non-ignorable one ($\widehat{\theta} = 0.782$). Since the value of the plebiscite was $\theta_{\text{pleb}} = 0.885$, an important benchmark obtained four weeks after the SPO Survey, these authors concluded the MAR was preferable. Molenberghs, Kenward and Goetghebeur (2001) supplemented these analysis with a so-called pessimistic-optimistic interval, obtained from replacing the incomplete data with NO and YES, respectively, and obtained: $\theta \in [0.694, 0.904]$. Further, they considered all nine BRD models, producing a range for θ from 0.741 to 0.892. Ultimately, these authors devised a method to consider overspecified models, in which point estimates are replaced by interval estimates, giving rise to so-called *intervals of ignorance.*

Turning now to the results obtained from fitting each of the nine BRD models (Table 4.2), it can be observed that BRD1 produces $\widehat{\theta} = 0.892$, exactly the same estimate as the first MAR estimate obtained by Rubin, Stern and Vehovar (1995). This should not come as a surprise, since both BRD1 and Rubin's model assume MAR and use information from the two main questions. Before continuing with the models' interpretation, it is necessary to assess their fit. Conducting likelihood ratio tests for BRD1 *versus* the ones with 7 parameters, BRD2–BRD5, and then in turn for BRD2–BRD5 *versus* the saturated models BRD6–BRD9, suggests the lower numbered models do not fit well, leaving models BRD6–BRD9. The impression might be generated that the poor model fit of BRD1 might be seen as evidence for discarding the MAR-based value 0.892. However, studying the MAR values from each of the models BRD1(MAR)–BRD9(MAR), as displayed in the last column of Table 4.2, it is clear that this value is remarkably stable and hence a value of $\widehat{\theta} = 0.892$, based on the four counterparts BRD6(MAR)–BRD9(MAR), is a sensible choice after all. Thus, a main contribution resulting from considering the counterparts in this particular

Table 4.2: *The Slovenian Public Opinion Survey, analysis restricted to the independence and attendance questions. Summaries on each of the models BRD1–BRD9.*

Model	Structure	d.f.	loglik	$\widehat{\theta}$	C.I.	$\widehat{\theta}_{\text{MAR}}$
BRD1	$(\alpha_{..}, \beta_{..})$	6	-2495.29	0.892	[0.878;0.906]	0.8920
BRD2	$(\alpha_{..}, \beta_{j_1.})$	7	-2467.43	0.884	[0.869;0.900]	0.8915
BRD3	$(\alpha_{.j_2}, \beta_{..})$	7	-2463.10	0.881	[0.866;0.897]	0.8915
BRD4	$(\alpha_{..}, \beta_{.j_2})$	7	-2467.43	0.765	[0.674;0.856]	0.8915
BRD5	$(\alpha_{j_1.}, \beta_{..})$	7	-2463.10	0.844	[0.806;0.882]	0.8915
BRD6	$(\alpha_{j_1.}, \beta_{j_1.})$	8	-2431.06	0.819	[0.788;0.849]	0.8919
BRD7	$(\alpha_{.j_2}, \beta_{.j_2})$	8	-2431.06	0.764	[0.697;0.832]	0.8919
BRD8	$(\alpha_{j_1.}, \beta_{.j_2})$	8	-2431.06	0.741	[0.657;0.826]	0.8919
BRD9	$(\alpha_{.j_2}, \beta_{j_1.})$	8	-2431.06	0.867	[0.851;0.884]	0.8919

example, is the provision of a solid basis for the MAR-based estimate. Obviously, since models BRD6(MAR)–BRD9(MAR) are exactly the same and exhibit a perfect fit, the corresponding probabilities $\widehat{\theta}_{\text{MAR}}$ are exactly equal too. And here, even though BRD2(MAR)–BRD5(MAR) differ among each other, the probability of being in favor of independence and attending the plebiscite is constant across these four models. This is a mere coincidence, since all three other cell probabilities are different, but only slightly so. For example, the probability of being in favor of independence combined with not attending ranges over 0.066–0.0685 across these four models.

In the preceding analyses, a two-stage use of models BRD6(MAR)–BRD9(MAR) has been employed. At the first stage, in a conventional way, the fully saturated model is selected as the only adequate description of the observed data. At the second stage, these models are transformed into their MAR counterpart, from which inferences are drawn. The MAR counterpart usefully supplements the original models BRD6–BRD9 and provide one further, important scenario to model the incomplete data. In principle, the same exercise can be conducted when the additional secession variable would be used.

4.4 Discussion

In this chapter, it has been shown that every MNAR model, fitted to a set of incomplete data, can be replaced by an MAR version which produces exactly the same fit to the observed data. Several important implications follow from this. First, unless one puts strong *a priori* belief in the posited MNAR model, it is not possible to use the fit of an MNAR model for or against MAR – a message in line with Gill, van der Laan and Robins (1997) and Schafer and Graham (2002). Second, it sometimes happens that the obvious parametric MAR model does not fit the observed data well. This is the case for BRD1, fitted to the Slovenian Public Opinion Survey. It is then appealing to fit a sufficiently versatile MNAR model, to ensure a good fit to the observed data, and then to use the MAR version. Various forms of use can be given to this MAR counterpart. To begin with, the MAR counterpart of a single, well-fitting MNAR model can be used as the sole basis for inference. More realistically, one can consider a variety of well-fitting MNAR models, such as BRD6–BRD9, and then switch to the corresponding collection of MAR counterparts. For example, for the SPO data, BRD6(MAR)–BRD9(MAR) all provide the same answer, unlike the MNAR models they originate from. Finally, an MAR counterpart or several MAR counterparts, can be used as a component of a sensitivity analysis. In this respect, it is useful to recall that all MNAR models, saturating the observed data, produce the same counterpart.

It should also be pointed out that the collection of counterparts provides a way of constructing an entire collection of MAR models. This is a less than trivial matter, especially with non-monotone missing data, as opposed to the straightforward determination of the MAR version of an MNAR model in the case of dropout, since the ACMV restrictions, established by Molenberghs *et al.* (1998) and translated in a computational scheme by Thijs *et al.* (2002), provides a convenient algorithm. In the case of non-monotone missingness, the marginal density of the outcomes is needed, and this can be directly obtained when the model fitted is of the SeM type. When a PMM is fitted, the marginal density follows from a weighted sum over the pattern-specific measurement models, for which no explicit construction exists.

Another issue worth noting here is that the best possible MAR model can always be obtained through the construction proposed in this paper. One merely has to consider the best fitting, perhaps a saturated, model, and then construct its MAR counterpart, which, by definition, does not alter the fit.

The analyst might want to examine the differences between how the MNAR model, on the one hand, and its MAR counterpart, on the other, fit the hypothetical complete data, so as to better understand what individuals and/or which parts of the data make the missing data mechanism appear MNAR. The sensitivity analysis part of Molenberghs and Kenward (2007) provides a thorough discussion. Beunckens *et al.* (2009) have considered this issue in the context of the SPO Survey.

The preceding analyses of the SPO Survey data has shown that, while a set of

MNAR models produces a widely varying range of conclusions about the proportion of people who are jointly in favor of independence and plan to attend the plebiscite, the corresponding MAR models produce a very narrow range of estimates, which, in addition, all lie close to the outcome of the plebiscite. This provides evidence for the claim, also made in Rubin, Stern and Vehovar (1995), that choosing an MAR model as one's main route of analysis is a sensible one.

While the result of Theorem 1 is general, focus here remained on SeM and PMM formulations. It is worth re-emphasizing that the SPM modeling framework is also covered without any problem. In this case, the likelihood is expressed as

$$L = \prod_i \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{\theta}, \boldsymbol{b}_i)\, f(\boldsymbol{r}_i | \boldsymbol{\psi}, \boldsymbol{b}_i)\, d\boldsymbol{y}_i^m, \tag{4.51}$$

with $\boldsymbol{b}_i$ denoting the shared parameter and often taking the form of random effects. To apply the result, $f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \widehat{\boldsymbol{\theta}}, \boldsymbol{b}_i)$ needs to be integrated over the shared parameter. The model as a whole needs to be used to produce the fit to the observed data, and then (4.7) is used to extend the observed-data fit to complete-data MAR version.

5

Non-Gaussian Longitudinal Data and Missing At Random: Multiple Imputation *versus* Weighting

Methodology for longitudinal data have historically been developed and revolve primarily around continuous outcomes. Discrete (e.g., binary) responses, however, have become popular as well over the years, sparking, as a consequence, increased research interest in methods for such. Within the three model families (Section 3.6), a broad set of methods are available, though the marginal and random-effects models are perhaps the most frequently used in longitudinal non-Gaussian settings. Under a likelihood framework, ignorability can be invoked for MAR, and hence, likelihood-based marginal (e.g., Bahadur model) or random-effects (e.g., GLMM) models can be used. For non-likelihood marginal models, the semi-parametric approach based on so-called *generalized estimating equations* (GEE), proposed by Liang and Zeger (1986), provides a useful alternative. The frequentist nature of standard GEE, for which dropout need not be modeled, renders it valid only under the restrictive MCAR condition. Robins, Rotnitzky and Zhao (1995) extended the GEE approach for applicability under the weaker MAR case by the use of inverse probability weights, resulting in *weighted generalized estimating equations* (WGEE). An alternative route would be multiple imputation, developed by Rubin (1987), in which missing values are imputed several times, and the resulting completed data sets are analyzed using a standard method, such as GEE. The so-obtained multiple inferences are subsequently combined into a single one (MI-GEE). Standard multiple imputation requires MAR to hold, even though extensions exist. In both methods, missingness (e.g., dropout) needs to be addressed, either by means of a dropout model for WGEE or by an imputation model for multiple-imputation-based GEE, implying that the missing-data mechanism is then not ignorable.

In this chapter, the important setting of incomplete non-Gaussian longitudinal data is considered. Of particular interest is a comparison of the two aforementioned approaches – WGEE and MI-GEE – to obtain marginal models for discrete longitudinal outcomes with dropout that are of an MAR nature. Focusing on the case of incomplete binary measures, both methods are compared using so-called asymptotic, as well as small-sample, simulations, in a variety of correctly and incorrectly specified models. Beunckens, Sotto and Molenberghs (2008) investigated this comparison under modest proportions of incompleteness in the data, which will be complemented here by considering more substantial amounts of missing data.

5.1 Missing-At-Random-Based Non-Gaussian Longitudinal Data Methods

This section describes three specific approaches to modeling non-Gaussian longitudinal data with MAR-type of missingness, particularly, dropout. Though primary interest lies on methodology to obtain marginal models for the latter, e.g., WGEE and MI-GEE, conditional models are also explored.

5.1.1 Weighted Generalized Estimating Equations

Liang and Zeger (1986) pointed out that inferences based on their GEE-based approach are valid only under MCAR, due to the fact that they are based on frequentist considerations. An important exception, mentioned by these authors, is the situation where the working correlation structure happens to be correct, since then the estimates and model-based standard errors are valid under the weaker MAR. In general, the working correlation structure will not be correctly specified, and hence, Robins, Rotnitzky and Zhao (1995) proposed a class of weighted estimating equations to allow for MAR. The general idea behind this approach is to weight each subject's contribution in the GEE by the inverse probability of dropping out at the time that he/she dropped out. Thus, a subject staying in the study is considered representative of himself as well as of a number of similar subjects that did drop out from the study. The incorporation of these weights, reduces possible bias in the regression parameter estimates. Restricting focus on the specific case of monotone missingness, such a weight can be expressed as:

$$\begin{aligned}\nu_{ij} \equiv P[D_i = j] &= \prod_{k=2}^{j-1} (1 - P[R_{ik} = 0 | R_{i2} = \ldots = R_{i,k-1} = 1]) \times \\ &\qquad P[R_{ij} = 0 | R_{i2} = \ldots = R_{i,j-1} = 1]^{I\{j \le n_i\}},\end{aligned}$$

where $j = 2, 3, \ldots, n_i + 1$, or, in terms of the dropout indicator D_i,

$$\nu_{ij} = \begin{cases} P(D_i = j | D_i \ge j) & \text{for } j = 2, \\ P(D_i = j | D_i \ge j) \prod_{k=2}^{j-1} [1 - P(D_i = k | D_i \ge k)] & \text{for } j = 3, \ldots, n_i, \\ \prod_{k=2}^{n_i} [1 - P(D_i = k | D_i \ge k)] & \text{for } j = n_i + 1. \end{cases} \tag{5.1}$$

For the weighted GEE approach, the score equations to be solved are then given by:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{d=2}^{n_i+1} \frac{I(D_i = d)}{\nu_{id}} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}(d) \, (A_i^{1/2} R_i A_i^{1/2})^{-1}(d) \, [\boldsymbol{y}_i(d) - \boldsymbol{\mu}_i(d)] = \mathbf{0},$$

where $\boldsymbol{y}_i(d)$ and $\boldsymbol{\mu}_i(d)$ are the first $(d-1)$ elements of $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$, respectively. The terms $\dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}(d)$ and $(A_i^{1/2} R_i A_i^{1/2})^{-1}(d)$ are defined analogously, in line with the definitions of Robins, Rotnitzky and Zhao (1995).

WGEE falls within the broad class of schemes that employ inverse probability weighting of complete cases (IPWCC). As such, WGEE provides an appealing alternative for correcting for MAR missingness, since only a model for the missingness process is required, without necessitating assumptions about the conditional distribution of the unobserved given the observed outcomes. IPW methods, however, have, as their main criticism, inefficiency, in comparison with, for instance, likelihood-based approaches (Clayton *et al.*, 1998) or multiple imputation (Carpenter, Kenward and Vansteelandt, 2006). Moreover, when certain subsets of the population have low response probabilities, resulting estimates can be unstable (Little and Rubin, 1987).

5.1.2 Multiple-Imputation-Based Generalized Estimating Equations and Multiple Imputation-Transition

In Section 3.5.2, multiple imputation or MI (Rubin, 1978) was described to consist of three stages. At the imputation stage, M imputed data sets are generated by replacing each missing value with a set of M plausible values, drawn from the conditional distribution of the unobserved values, given the observed ones. Analysis of the M completed data sets, by means of any standard complete data method, then follows. At the final stage – the pooling stage – the results from the M analyses are combined into a single inference, e.g., (3.13) and (3.14), by means of the method laid out in Rubin (1978). In its basic form, multiple imputation requires the missingness mechanism to be MAR, even though versions under MNAR have been proposed (Rubin, 1987; Molenberghs, Kenward and Lesaffre, 1997).

At the analysis stage of MI, any standard complete data method can be used. One can consider MI together with GEE or with a transition model, hereinafter referred to as MI-GEE and MI-Transition, respectively. In essence, these methods come down to first using the predictive distribution of the unobserved outcomes given the observed ones and perhaps covariates. After this step, the missing-data mechanism can be further ignored, provided the missing-data mechanism is MAR. Under either case, a misspecification made in the imputation step will only affect the unobserved (i.e., imputed) but not the observed part of the data. Meng (1994) showed that as long as the imputation model is not grossly misspecified, these approaches will perform well.

5.2 Simulation Study

In the previous section, two approaches proposed to overcome the bias occurring in GEE under MAR have been introduced. WGEE is unbiased for a correctly specified dropout and mean structure of the measurement model. MI-GEE requires compatibility between the imputation and estimation model to be correctly specified. Therefore, it is of interest to quantify the bias and precision under various types of misspecification. To this end, an asymptotic simulation study, as well as small-sample simulations, were conducted on various underlying data-generating models. Whereas asymptotic simulations give a nice paradigm to explore the situation of "large" samples, small-sample simulations give insight into the behavior of the methods in real-life settings.

In the simulation study, it is important to distinguish between two stages: (1) the data-generating stage and (2) the analysis stage. In the first stage, a data-generating model is defined. Under the selection model framework, this generating model consists of a measurement model on the one hand, and a dropout model given the measurement model on the other. The analysis stage involves three distinct models: a measurement model, a dropout model and an imputation model. For the WGEE approach, only a marginal measurement model and a dropout model need to be specified. In contrast, the analysis stage for MI-GEE would entail the specification of an imputation model, rather than a dropout model, as well as a marginal measurement model. Finally, for MI-Transition, a conditional rather than marginal measurement model is needed, as well as an imputation model. For the MI-GEE and MI-Transition approaches, the predictors of dropout are included in the imputation model. These distinctions can be more clearly visualized in the following schematic:

MODEL	Data-Generating	Analysis/Fitting Stage		
	Stage	WGEE	MI-GEE	MI-Transition
Measurement	✓	✓	✓	✓
Dropout	✓	✓		
Imputation			✓	✓

To assess the distinctive and relative merits of the methods of interest, their performance for binary longitudinal outcomes with monotone missingness shall be considered, first in the case without any misspecification, then under various misspecifications. Since interest lies in comparing WGEE and MI-GEE as methods for dealing with missing data in a non-Gaussian longitudinal setting, the misspecification can be made either in the dropout model, in the imputation model, or in the measurement model. Misspecification in the missingness mechanism, however, e.g., using MCAR for an underlying MAR mechanism, is not further explored, as this is not the main focus here and has already been investigated extensively (Jansen *et al.*, 2006).

In this section, the various data-generating models employed for the simulations are first defined. A description of the design of the simulation study follows, after which the results of the simulation, under each of the various scenarios, are presented.

5.2.1 Data-Generating Models

For the simulation study, an outcome at 3 time points was generated using three different measurement models: first, three-dimensional binary outcomes were generated from a Bahadur model, as well as from a second-order autoregressive, AR(2), transition model; further, a three-dimensional continuous outcome (that was later dichotomized) was generated from a trivariate Gaussian distribution. Whereas the choice of the first two is obvious, since focus lies on binary repeated measures, the third case depicts real-life settings for which a continuous outcome is available, but the scientific question is based on a dichotomized version of it. For all three cases, the measurement model incorporated a binary treatment indicator, such as a treatment *versus* placebo classification. In addition, for the dropout model, an MAR mechanism was considered. Assuming that dropout can occur only after the first time point, there are three possible dropout patterns: (1) dropout at the second time point, (2) dropout at the third time point, or (3) no dropout. The combination of the various measurement models and the dropout model gives rise to three data-generating models, which will hereinafter be denoted as:

Generating Model	Measurement Model	Dropout Model
GM I	Bahadur	MAR logistic
GM II	AR(2)	MAR logistic
GM III	Gaussian	MAR logistic

GM I is based on a Bahadur (1961) model, which, upon denoting by t_j the time point at which measurement j is taken and by x_i the treatment indicator, follows general formulation (3.27) with

$$\text{logit}(\pi_{ij}) \;=\; \text{logit}\; P(Y_{ij}=1|x_i,t_j) \;=\; \beta_0 \;+\; \beta_x\, x_i \;+\; \beta_t\, t_j \;+\; \beta_{xt}\, x_i\, t_j, \tag{5.2}$$

where $\boldsymbol{\beta}_0 = -0.25$, $\beta_x = 0.5$, $\beta_t = 0.2$ and $\beta_{xt} = -0.8$, with two- and three-way correlation coefficients equal to $\rho_{ij_1j_2} = 0.2$ and $\rho_{ij_1j_2j_3} = 0$, respectively. The latter define an exchangeable correlation structure. The missingness process for GM I is assumed to be MAR, and the probability of dropout at time point j given x_i and the measurement at the previous time point, $y_{i,j-1}$, is modeled by a logistic regression of the form

$$\text{logit}\; P(D_i=j|x_i,y_{i,j-1},D_i\geq j) \;=\; \psi_0 \;+\; \psi_x\, x_i \;+\; \psi_{prev}\, y_{i,j-1}, \tag{5.3}$$

where $j = 2,3,4$. To be able to explore the effects of the amount of missingness as well, the following two sets of parameters, giving rise to varying proportions of missing data patterns were considered:

Setting	ψ_0	ψ_x	ψ_{prev}
1	-0.50	-0.60	-3.50
2	-0.10	-0.29	-1.20

The resulting proportions of the 3 dropout patterns under each of these settings are enumerated in Table 5.1. While setting 1 reflects a situation with only a moderate amount of missingness in the data, a somewhat more extreme case, with more than half of the data incomplete, is represented by dropout setting 2.

Table 5.1: *Generating Model I. Proportions of missing data patterns, by treatment level and overall, under each dropout setting. (O: observed, M: missing.)*

Pattern	Setting 1			Setting 2		
	$x = 0$	$x = 1$	Overall	$x = 0$	$x = 1$	Overall
OOO	32.86	34.83	67.69	21.86	23.05	44.91
OOM	7.03	7.64	14.67	10.75	11.61	22.35
OMM	10.11	7.53	17.64	17.39	15.34	32.74

GM II consists of an autoregressive transition model of order 2, AR(2), defined by the following:

$$\begin{aligned} P(x_i) &= \mu_x, \\ \text{logit } P(Y_{i1} = 1|x_i) &= \alpha_0 + \alpha_x x_i, \\ \text{logit } P(Y_{i2} = 1|x_i, y_{i1}) &= \phi_0 + \phi_x x_i + \phi_1 y_{i1}, \quad \text{and} \\ \text{logit } P(Y_{i3} = 1|x_i, y_{i1}, y_{i2}) &= \gamma_0 + \gamma_x x_i + \gamma_1 y_{i1} + \gamma_2 y_{i2}, \end{aligned}$$

where $\mu_x = 0.5$, $\alpha_0 = -0.2$, $\alpha_x = 0.3$, $\phi_0 = -0.1$, $\phi_x = 0.5$, $\phi_1 = 0.7$, $\gamma_0 = -0.25$, $\gamma_x = 0.35$, $\gamma_1 = 0.4$ and $\gamma_2 = 0.6$. This is combined with the same dropout models considered for GM I, which lead to the missing data pattern proportions summarized in Table 5.2.

As WGEE and MI-GEE both involve marginal models, so as to allow comparison, the conditional AR(2) transition model needs to be further marginalized to obtain so-called marginalized "true" parameters, which then approximately describe a marginal logistic function. For this simulation study, it is assumed that the underlying marginal model for GM II is of the form given in (5.2). Inasmuch as the underlying measurement model is in fact conditional, rather than marginal, there is no way to verify whether this assumed underlying marginal model is "true". The marginalization is done by computing the marginal probabilities from the underlying conditional AR(2) transition model probabilities, i.e., for a given outcome vector and treatment level, $(y_{i1}, y_{i2}, y_{i3}, x_i)$,

$$P(y_{i1}, y_{i2}, y_{i3}, x_i) = P(y_{i3}|x_i, y_{i1}, y_{i2})\, P(y_{i2}|x_i, y_{i1})\, P(y_{i1}|x_i)\, P(x_i). \tag{5.4}$$

Then, on a hypothetical data set consisting of all 16 possible combinations of the form $(y_{i1}, y_{i2}, y_{i3}, x_i)$, with corresponding marginal probabilities given in (5.4) as weights, a GEE model of the form (5.2) is fitted. The resulting marginalized "true" parameters of GM II are $\beta_0 = -0.3658$, $\beta_x = 0.2673$, $\beta_t = 0.2265$ and $\beta_{xt} = 0.0790$.

Table 5.2: *Generating Model II. Proportions of missing data patterns, by treatment level and overall, under each dropout setting. (O: observed, M: missing.)*

Pattern	Setting 1			Setting 2		
	$x = 0$	$x = 1$	Overall	$x = 0$	$x = 1$	Overall
OOO	32.46	40.10	72.56	21.63	27.27	48.90
OOM	6.75	3.71	10.46	10.49	8.70	19.19
OMM	10.78	6.19	16.98	17.88	14.03	31.91

Finally, GM III arises from a Gaussian outcome, W_{ij}, at three time points, where:

$$\mu_{ij} \;=\; E(W_{ij}|x_i, t_j) \;=\; \eta_0 \;+\; \eta_x\, x_i \;+\; \eta_t\, t_j \;+\; \eta_{xt}\, x_i\, t_j,$$

for $i = 0, 1$ and $j = 1, 2, 3$, with $\eta_0 = 3.5$, $\eta_x = 0$, $\eta_t = 1.75$ and $\eta_{xt} = 0.5$. That is,

$$\boldsymbol{\mu} \;=\; \begin{bmatrix} \boldsymbol{\mu}_0 \\ \boldsymbol{\mu}_1 \end{bmatrix} = \begin{bmatrix} (\mu_{01}, \mu_{02}, \mu_{03})' \\ (\mu_{11}, \mu_{12}, \mu_{13})' \end{bmatrix} = \begin{bmatrix} (5.75, 8.00, 10.25)' \\ (5.25, 7.00, 8.75)' \end{bmatrix},$$

for which the following unstructured covariance matrix is assumed:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.80 & 0.35 \\ 0.80 & 1 & 0.50 \\ 0.35 & 0.50 & 1 \end{pmatrix}.$$

The missingness process for this GM is given by:

$$\text{logit}\; P(D_i = j|x_i, w_{i,j-1}, D_i \geq j) \;=\; \delta_0 \;+\; \delta_x\, x_i \;+\; \delta_{prev}\, w_{i,j-1}, \qquad (5.5)$$

where $j = 2, 3, 4$. As in each of the two previous GMs, two dropout settings are again considered, the first with $\delta_0 = -0.15$, $\delta_x = 0.8$ and $\delta_{prev} = -0.35$, and for the second dropout setting, only the latter parameter was modified to $\delta_{prev} = -0.15$. Combining these dropout models with the measurement model yields, on average, over the 500 generated samples, percentages of missing data patterns that are presented in Table 5.3.

The binary outcome Y_{ij} was then obtained from the continuous outcome W_{ij} by defining a cut-off value of 6.5, i.e., $Y_{ij} = 1$, if $W_{ij} \geq 6.5$, and 0, otherwise. Although the generated outcomes are continuous in nature, the focus here is on the analysis of the binary version Y_{ij}. For this reason, "true" parameters corresponding to this dichotomized response need to be obtained by fitting a GEE model of the form (5.2). Note, however, that this model is again not necessarily the unknown underlying marginal model for the binary outcomes. Using the complete data, the resulting parameters are $\beta_0 = -3.0373$, $\beta_x = 0.0095$, $\beta_t = 1.7812$ and $\beta_{xt} = 0.4828$.

Table 5.3: *Generating Model III. Proportions of missing data patterns, by treatment level and overall, averaged over the 500 generated samples, under each dropout setting. (O: observed, M: missing.)*

Pattern	Setting 1			Setting 2		
	$x = 0$	$x = 1$	Overall	$x = 0$	$x = 1$	Overall
OOO	40.31	35.38	75.69	27.50	17.52	45.02
OOM	3.18	4.19	7.37	8.11	10.14	18.25
OMM	6.50	10.44	16.94	14.39	22.34	36.73

The choice for linear time evolutions, at the scale of the linear predictor and within each of the treatment arms, allows distinguishing between misspecification effects on cross-sectional parameters (β_0 and β_x), longitudinal parameters (β_t), and parameters combining aspects of both (β_{xt}). In practice, for example in a clinical trial, it might be advisable to allow for an unstructured, saturated treatment-by-time model, reducing the risk of model misspecification and in line with recommendations made by Molenberghs *et al.* (2004) and several references listed therein.

5.2.2 Design of the Simulation Study

Given that the sequence of outcomes and the missing data process for GM I and GM II are discrete, quantification of bias under specific assumptions about the non-response process can be done via an algorithm first proposed by Rotnitzky and Wypij (1994). This so-called asymptotic simulation method entails first creating a hypothetical data set consisting of all possible outcome sequences for each level of the covariate(s) and for each of the possible missingness patterns. The probability mass with which each of these outcome combinations occurs can be computed based on the assumed data-generating model (measurement and dropout models). The asymptotic simulation then entails fitting the prescribed model on the hypothetical data set, using the probability mass values as weights, thereby resulting in the asymptotic solution. Asymptotic simulations, though enabling computation of asymptotic quantities (e.g., bias and variance), have only theoretical value and may provide guidance as to what happens in large to very large samples, but are of no meaningful use with conventional data analysis. As such, these are best supplemented with small-sample simulations.

For the three-dimensional binary outcome $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})'$ and the binary treatment indicator x_i considered here, there are $2^3 = 8$ possible outcome sequences at each level of x_i, or a total of 16 possibilities over both covariate (treatment) levels. From the assumed measurement model, the probability masses, $P(\boldsymbol{y}_i, x_i)$, for each of these 16 sequences can be computed. Now, for each such case, there are 3 possible dropout patterns – dropout at second time point, dropout at the third time point, and no dropout – yielding a total of 48 possibilities. The probabilities $P(\boldsymbol{y}_i, x_i)$ are thus further split among the 3 missingness patterns according to the dropout probabilities. Specifically, denoting by $P(D_i = 2|D_i \geq 2), P(D_i = 3|D_i \geq 3)$ and $P(D_i = 4|D_i \geq 4)$ the probabilities of dropout at time points 2, 3 and 4, respectively, and using (5.1), the probabilities for the different outcome combinations are given by:

$$\begin{aligned} P(\boldsymbol{y}_i, x_i, D_i = 4|D_i \geq 4) &= P(\boldsymbol{y}_i, x_i) \prod_{j=2}^{4} [1 - P(D_i = j|D_i \geq j)], \\ P(\boldsymbol{y}_i, x_i, D_i = 3|D_i \geq 3) &= P(\boldsymbol{y}_i, x_i) \prod_{j=2}^{3} [1 - P(D_i = j|D_i \geq j)]\ P(D_i = 4|D_i \geq 4), \\ P(\boldsymbol{y}_i, x_i, D_i = 2|D_i \geq 2) &= P(\boldsymbol{y}_i, x_i)\ [1 - P(D_i = 2|D_i \geq 2)] \prod_{j=3}^{4} P(D_i = j|D_i \geq j). \end{aligned}$$

The estimating equations are then applied to this hypothetical data set with the application of the resulting probability weighting. The solutions obtained are the limiting (i.e., asymptotic) solutions, which can then be compared with the known parameters of the simulation model, so as to conveniently derive the asymptotic bias of the estimators.

For the small-sample simulations, a sample of size of $N = 100$ subjects, equally divided between the two treatment groups, was considered. Such a choice is directly applicable to practitioners, since many bio-pharmaceutical trials employ about 50 to 100 patients per treatment arm. Based on the underlying probabilities from GM I or GM II, 50 observations were generated randomly for each treatment group, and $S = 500$ such samples were then generated. Similarly, for GM III, $S = 500$ samples were generated, each with $n_0 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $n_1 = 50$ observations from $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. While asymptotic simulations were conducted only for GM I and GM II, small-sample simulations were done for all three generation models. When using a GEE approach, the same working correlation structure as assumed during data generation was employed for the analysis.

5.2.3 Results

For the ensuing discussion, various properties are quantified to help assess and compare WGEE and MI-GEE. Firstly, bias for some generic parameter θ is defined here as the difference between its estimate and its true value, i.e., $\mathrm{Bias}(\widehat{\theta}) = \widehat{\theta} - \theta$. For the asymptotic version, probability weights, computed from the underlying GM, are applied in solving the estimating equations, giving the limiting solutions, from which the asymptotic bias (Bias_∞) can be computed, as well as asymptotic variances (Var_∞) of the parameter estimators.

For the small-sample simulations, the average ($\overline{\text{Est}}$) of the estimators over all $S = 500$ samples, its estimated variance for a sample of size N ($\widehat{\text{Var}}_N$) and MSE are computed as:

$$\begin{aligned} \overline{\text{Est}} &\equiv \overline{\overline{\widehat{\theta}}} &&= \sum_{i=1}^{S} \frac{\widehat{\theta}_i}{S}, \\ \widehat{\text{Var}}_N &\equiv \widehat{\text{Var}}_N(\overline{\text{Est}}) &&= \sum_{i=1}^{S} \frac{(\widehat{\theta}_i - \overline{\overline{\widehat{\theta}}})^2}{S-1}, \qquad \text{and} \\ \text{MSE} &\equiv \text{MSE}(\overline{\text{Est}}) &&= \text{Bias}_N^2(\overline{\text{Est}}) + \widehat{\text{Var}}_N(\overline{\text{Est}}). \end{aligned}$$

5.2.3.1 Everything Correctly Specified

The individual merits of each method are first investigated when every one of its aspects is correctly specified. It can be recalled that GM I is based on a Bahadur measurement model and a logistic model for dropout that is reflective of an MAR mechanism, i.e., depending on the previous measurement as well as the treatment indicator. An appropriate analysis model would therefore consist of a measurement model and a dropout model that match those of this GM. Since GEE methods can be viewed as moment-based versions of the Bahadur model (Section 3.5.4), a GEE-based solution using the same structure as that of the underlying measurement model would be suitable. To address the MAR nature of the missingness, the GEE approach is supplemented with a weighting scheme, obtained from a model of the same form as that of the underlying dropout model, resulting now in WGEE. Thus, WGEE was fitted for GM I, using weights taken from fitting a logistic dropout model with the treatment indicator and the previous measurement as predictors. It should be noted that under WGEE the imputation model is not relevant since the missingness is addressed, not by imputation, but rather, by means of the dropout model. The results for both the asymptotic and small-sample simulations for both dropout settings are shown in Table 5.4.

The asymptotic unbiasedness of the WGEE estimators under a correctly specified mean structure is demonstrated by the results of the asymptotic simulation. The same cannot be said, however, for the small-sample simulations, under which a substantial amount of bias is observed under setting 1, though this improves considerably under setting 2. Similarly, though MSEs are still relatively large in setting 1, demonstrating the inefficiency of WGEE for small samples, improvement is observed for the case with more missing data. These observations are indicative that, for a sample of size

Table 5.4: *Generating Model I. Asymptotic bias* (Bias_∞)*, small-sample bias* (Bias_N) *and mean squared error* (MSE) *of the parameter estimates, for WGEE analysis with everything correctly specified, for two dropout settings and for* $N = 100$.

Parameter	Setting 1			Setting 2		
	Bias_∞	Bias_N	MSE	Bias_∞	Bias_N	MSE
β_0	-1.87E-06	-0.3956	1.2346	-1.95E-07	-0.0272	0.2964
β_x	1.99E-07	0.1226	2.1254	-1.37E-07	0.0681	0.6361
β_t	2.02E-07	0.1018	0.2491	-0.15E-07	0.0143	0.0848
β_{xt}	-1.66E-07	-0.1357	0.4625	-0.08E-07	-0.0470	0.1944

$N = 100$, the consistency of the WGEE estimators does not seem to be achieved, at least not for this particular generating model. The improvement of the precision under increased incompleteness, though seemingly counterintuitive, might suggest that the benefits of weighting are better realized under situations with more missing data.

For GM II, which uses an AR(2) transition model for the mean structure and a conditional logistic model for dropout, after multiple imputation of the missing values, an AR(2) transition model (MI-Transition), which is consistent with the underlying measurement model, was fitted. The multiple imputations were carried out with the SAS procedure PROC MI, which employs a conditional logistic imputation model for binary outcomes, a model in line with the underlying measurement model of GM II and fully parametric, admitting valid inferences under MAR (Schafer, 2003). Thus, the MI-Transition analysis, both the imputation as well as the measurement models, are correctly specified in the sense that they are compatible with the underlying measurement model. Note also that a dropout model need not be defined for this mode of analysis, since imputations, rather than dropout weights, are used to cope with the missingness. For the asymptotic simulation, $M = 500$ data sets were imputed, while for the small-sample simulations, since efficient results can be obtained even under a small number of imputations (Rubin, 1987), a more practically relevant value of $M = 5$ was chosen. Table 5.5 gives the results for both types of simulations for the two dropout settings considered.

The first panel, reflecting the results for the model for the first outcome Y_1, shows asymptotically unbiased parameter estimates for both settings, which is expected for this outcome, since data for all subjects are available and are thus not imputed. The small-sample simulations for this outcome, however, show small amounts of bias, which can most probably be attributed to finite sampling. For the second and third

Table 5.5: *Generating Model II. Asymptotic bias* (Bias_∞)*, small-sample bias* (Bias_N) *and mean squared error* (MSE) *of the parameter estimates, for MI-Transition analysis with everything correctly specified, for two dropout settings and for $N = 100$.*

Parameter	Setting 1			Setting 2		
	Bias_∞	Bias_N	MSE	Bias_∞	Bias_N	MSE
α_0	-0.0000	-0.0313	0.0935	0.0000	-0.0313	0.0935
α_x	-0.0000	0.0369	0.1805	-0.0000	0.0369	0.1805
ϕ_0	-0.0096	0.0317	0.2056	-0.0049	0.0326	0.2860
ϕ_x	-0.0666	0.0041	0.2635	-0.1348	0.0119	0.3688
ϕ_1	0.0343	0.0241	0.2698	-0.0838	0.0413	0.3360
γ_0	0.0236	0.0798	0.3535	0.0650	0.1136	0.5918
γ_x	-0.0568	0.0090	0.3024	-0.1016	0.1033	1.9510
γ_1	-0.0594	-0.0971	0.2448	-0.1078	0.0464	0.6039
γ_2	0.0072	0.0382	0.3264	-0.2052	-0.1822	0.3168

panels, respectively representing the models for Y_2 and Y_3, some bias is observed, asymptotically and for small samples, but the amounts are generally of small magnitudes. It is worthwhile to point out the increase not only in both types of bias, but also in MSE, for most of the parameters when going from dropout setting 1 to 2. As the latter consists of more missingness in the data than the former, this observation seems to suggest that the merits of MI tend to diminish under increased incompleteness. This is not surprising since more missing values would require more imputations from less information, which would naturally inflate the variability of the MI estimates, subsequently decreasing their efficiency.

Since the correctly specified MI-Transition model fitted is a conditional model, resulting estimates warrant further marginalization to obtain so-called “marginal” parameters. It can be recalled from Section 5.2.1 that the MI-Transition model defines three sets of conditional probabilities, from which marginal probabilities can be derived as in (5.4). These marginal probabilities can be estimated using now the parameter estimates, rather than the true parameter values, and these estimated marginal probabilities can be subsequently used as weights in fitting a GEE model of the form (5.2) on a data set consisting of all possible combinations of outcome sequences and treatment level. This yields estimates for the “marginal” parameters

Table 5.6: *Generating Model II. Asymptotic bias* (Bias$_\infty$)*, small-sample bias* (Bias$_N$) *and mean squared error* (MSE) *of the parameter estimates, for the marginalized MI-Transition analysis with everything correctly specified, for two dropout settings and for* $N = 100$*.*

Parameter	Setting 1			Setting 2		
	Bias$_\infty$	Bias$_N$	MSE	Bias$_\infty$	Bias$_N$	MSE
β_0	0.0058	-0.0616	1.1239	0.0501	-0.0350	1.1192
β_x	0.0270	0.0470	2.4650	0.0571	-0.0113	2.4931
β_t	-0.0031	0.0394	0.2064	-0.0475	0.0214	0.2045
β_{xt}	-0.0352	-0.0153	0.4499	-0.0741	0.0308	0.4548

of the MI-Transition model, which can then be compared with the corresponding "marginal" parameters in (5.2); the results are shown in Table 5.6. Asymptotic bias for the parameter estimates is generally small, while its small-sample counterpart is larger for setting 1. The reverse is observed under the setting of increased missingness, which is also observed to inflate the asymptotic bias. With respect to MSE, however, only very slight differences are observed across the two dropout settings.

Finally, recall that GM III is based on a Gaussian measurement model and a logistic dropout model. The analysis model used for this GM was MI-GEE, which requires an imputation model and a measurement model, but not a dropout model. Multiple imputations of the missing Gaussian outcomes were first obtained using a Gaussian imputation model, thereby ensuring a correctly specified imputation model, that is, one that uses the underlying measurement process to generate the imputations for the missing observations. The Gaussian outcome was then dichotomized based on the previously defined cutoff value, after which standard GEE, using a probit link, was applied to the dichotomized outcome of the imputed data sets. Since the underlying distribution for the outcomes is not discrete, only small-sample simulations are possible. Although initially $S = 500$ samples were generated, after dichotomization of the Gaussian outcome, there were 51 (and 108) samples under dropout setting 1 (and 2) for which convergence was not attained. Inspection of these samples showed that the treatment-by-time interaction could not be estimated because, at one time point, all dichotomized outcomes belonged to only one treatment group, and thus, these samples had to be excluded.

Table 5.7 gives the results of the simulation only for the $S' = 449$ and $S' = 392$ convergent samples for dropout settings 1 and 2, respectively. The "true" parameter

Table 5.7: *Generating Model III. Small-sample bias* (Bias_N) *and mean squared error* (MSE) *of the parameter estimates, for MI-GEE analysis with everything correctly specified, for dropout setting 1 (using* $S' = 449$*) and for dropout setting 2 (using* $S' = 392$*), for* $N = 100$.

Parameter	Setting 1		Setting 2	
	Bias_N	MSE	Bias_N	MSE
β_0	0.0016	0.1978	0.0208	0.1986
β_x	0.0056	0.3968	0.0077	0.3792
β_t	-0.0004	0.0601	-0.0135	0.0579
β_{xt}	-0.0060	0.1481	-0.0145	0.1297

values used to compute the bias were obtained by fitting the same measurement model using the complete (binary) data from the $S' = 449$ or $S' = 392$ samples. Consistent with the theory on MI, only very small bias for the estimates was observed and might be expected to decrease even further under larger samples. In addition, MSEs generally decrease slightly under setting 2.

5.2.3.2 Dropout and Measurement Models Correct, Imputation Model Incorrect

A comparison is now considered between WGEE and MI-GEE, both having a correctly specified measurement model, but with an incorrectly specified imputation model for the latter and a correctly specified dropout model for the former. In both cases, the measurement model used is consistent with the underlying Bahadur measurement model of GM I. In the previous section, it was discussed that fitting WGEE for GM I, using the same mean structure as that of the underlying measurement model and with weights obtained from a logistic dropout model with the treatment indicator and the previous measurement as predictors, ensures every aspect is correctly specified. For MI-GEE, imputations are done using a conditional logistic imputation model for binary outcomes – a model that is *not* consistent with the marginal nature of the underlying Bahadur measurement model and is, therefore, incorrectly specified. Thus, the said comparison, of WGEE with correctly specified dropout and measurement models against MI-GEE with correctly specified measurement model but incorrectly specified imputation model, is possible under GM I. The results of the correct and incorrect analyses are compared in Table 5.8.

As was already noted in the previous section, WGEE does not yield unbiased and

Table 5.8: *Generating Model I. Asymptotic bias* (Bias$_\infty$)*, small-sample bias* (Bias$_N$) *and mean squared error* (MSE) *of the parameter estimates, for WGEE, with correctly specified dropout, and MI-GEE, with incorrectly specified imputation model, for two dropout settings and for* $N = 100$*.*

Parameter	WGEE (correct)			MI-GEE (incorrect)		
	Bias$_\infty$	Bias$_N$	MSE	Bias$_\infty$	Bias$_N$	MSE
Setting 1						
β_0	-0.0000	-0.3956	1.2346	0.0101	-0.0182	0.2322
β_x	0.0000	0.1226	2.1254	-0.3515	0.0181	0.4723
β_t	0.0000	0.1018	0.2491	0.0007	0.0082	0.0558
β_{xt}	-0.0000	-0.1357	0.4625	0.3056	-0.0040	0.1171
Setting 2						
β_0	-0.0000	-0.0272	0.2964	0.0836	-0.0039	0.2906
β_x	-0.0000	0.0681	0.6361	-0.6024	-0.0111	0.5779
β_t	-0.0000	0.0143	0.0848	-0.0647	-0.0011	0.0826
β_{xt}	-0.0000	-0.0470	0.1944	0.5388	0.0183	0.1701

consistent estimators for the particular sample size used, and the small-sample bias is considerably less for MI-GEE, particularly under dropout setting 1. Under setting 2, however, differences between the small-sample bias for the two methods are much smaller. It can also be observed that the precision of the estimators from MI-GEE tends to decrease, though only slightly, under increased proportions of missing data. But, despite this decreased precision, MI-GEE still yields more efficient estimators than those obtained for WGEE, as evidenced by smaller MSEs, notwithstanding the fact that the WGEE analysis was entirely correctly specified. These observations suggest a certain amount of robustness in MI-GEE with respect to misspecifications in the imputation model.

5.2.3.3 Imputation and Measurement Models Correct, Dropout Model Incorrect

Whereas in the previous section the relative performances of WGEE with a correctly specified dropout model and MI-GEE with an incorrectly specified imputation model were compared, in this section the reverse is investigated, that is, a comparison of WGEE with an incorrectly specified dropout model against MI-GEE with a correctly specified imputation model. For this assessment, GM III is considered. Ear-

Table 5.9: *Generating Model III. Small-sample bias* (Bias_N) *and mean squared error* (MSE) *of the parameter estimates, for WGEE, with incorrectly specified dropout, and MI-GEE, with correctly specified imputation model, for two dropout settings and for* $N = 100$.

Parameter	MI-GEE (correct)		WGEE (incorrect)	
	Bias_N	MSE	Bias_N	MSE
Setting 1				
β_0	0.0016	0.1978	-0.2961	0.9656
β_x	0.0056	0.3968	0.1295	1.4962
β_t	-0.0004	0.0601	0.1833	0.2975
β_{xt}	-0.0060	0.1481	-0.0444	0.5448
Setting 2				
β_0	0.0208	0.1986	-0.0737	0.3607
β_x	0.0077	0.3792	0.0734	0.6127
β_t	-0.0135	0.0579	0.0595	0.1262
β_{xt}	-0.0145	0.1297	-0.0654	0.2511

lier, it was noted that under GM III, a Gaussian imputation model combined with a subsequent standard GEE analysis on the dichotomized outcomes of the completed data sets results in MI-GEE with everything correctly specified. To enable comparison with WGEE using an incorrectly specified dropout model, weights are obtained from a logistic dropout model with the treatment indicator and the *binary* version of the previous measurement as predictors. The latter is a clear misspecification in the dropout model, since the underlying dropout model uses the *continuous* previous measurement as predictor. In both cases, the measurement model corresponds to the assumed underlying measurement model for the dichotomized version of the continuous response. The results of this comparison are given in Table 5.9. Only small-sample simulations are possible since the underlying GM does not consist of a discrete set of outcomes.

Bias is much smaller for MI-GEE, which can be expected as this is a correctly specified analysis model, with discrepancies much more pronounced for modest amounts of missingness (setting 1). As was already previously observed, increased missingness, though worsening the MI-GEE results slightly, tends to improve the results of WGEE. With respect to MSE, the estimators obtained from MI-GEE are superior to those from WGEE, for both dropout settings. These observations seem to highlight the

Table 5.10: *Marginalized Generating Model II. Asymptotic bias* (Bias_∞)*, small-sample bias* (Bias_N) *and mean squared error* (MSE) *of the parameter estimates, for WGEE, with correctly specified dropout and incorrectly specified measurement model, and MI-GEE, with correctly specified imputation model and incorrectly specified measurement model, for two dropout settings and for* $N = 100$*.*

Parameter	WGEE (incorrect)			MI-GEE (incorrect)		
	Bias_∞	Bias_N	MSE	Bias_∞	Bias_N	MSE
Setting 1						
β_0	-0.0633	-0.4193	1.3114	0.0034	-0.0576	0.2523
β_x	-0.0556	-0.1487	3.0037	0.0332	0.0532	0.4914
β_t	0.0380	0.1222	0.2311	-0.0013	0.0352	0.0611
β_{xt}	0.0352	0.0812	0.5945	-0.0409	-0.0237	0.1168
Setting 2						
β_0	-0.0633	-0.1281	0.3301	0.0497	-0.0464	0.2957
β_x	-0.0556	-0.0150	0.7424	0.0665	0.0427	0.5870
β_t	0.0380	0.0775	0.0985	-0.0479	0.0276	0.0856
β_{xt}	0.0352	0.0298	0.2225	-0.0833	-0.0146	0.1645

sensitivity of WGEE to misspecifications in the dropout model, in contrast to MI-GEE, which was noted to be somewhat robust to misspecifications in the imputation model.

5.2.3.4 Imputation and Dropout Models Correct, Measurement Model Incorrect

As a final comparison, the case of an incorrectly specified measurement model is considered, particularly under the marginalized version of GM II. For WGEE, weights are obtained from a dropout model consistent with the underlying dropout model of GM II, while for MI-GEE, imputations are generated from a conditional AR(2) transition model, which is in line with the underlying measurement model of GM II. In this way, both the dropout and imputation models are correctly specified. However, the fitted measurement models for both WGEE and MI-GEE are clearly misspecified, in the sense that the outcomes are modeled marginally (i.e., GEE), rather than conditionally (i.e., AR(2)). The resulting estimates for each case were compared to the "marginal" parameters of GM II. The results of this comparison (Table 5.10) indicate generally less small-sample bias and consistently smaller MSEs for MI-GEE for both settings. It is particularly interesting to note that quite an improvement is observed in the MSE (and on a lesser degree, in the small-sample bias) of the WGEE estimates for setting 2 compared to setting 1, while the MSE for MI-GEE increases slightly for more missing values in the data.

5.3 Discussion

In this chapter, the analysis of binary longitudinal data with missingness of an MAR nature was investigated. Apart from likelihood-based methods, such as generalized linear mixed-effects models (Molenberghs and Verbeke, 2005), non-likelihood methods are attractive, especially when a so-called marginal model is of interest. Since standard generalized estimating equations (Liang and Zeger, 1986) are unbiased only under MCAR, a variety of modifications and alternatives to GEE have been proposed. Undoubtedly the most popular route is through weighted estimating equations, proposed by Robins, Rotnitzky and Zhao (1995), and a number of later extensions. Also of attraction is a combination of GEE and multiple imputation (Rubin, 1987) methods, i.e., MI-GEE. Once multiple imputation is considered an option, it has the merit of allowing for a variety of imputation techniques, whereafter several analysis methods can be considered. Two such routes considered in this chapter are MI-GEE and MI-Transition.

The simulation study presented here has provided quantitative evidence, based on asymptotic, as well as small-sample, simulations, that can be usefully applied in the decision-making process. WGEE and MI-GEE, as well as MI-Transition, have been compared under a variety of scenarios, shedding light on their respective merits. Although, asymptotically, WGEE exhibits the desirable properties it is theoretically known to possess, these are barely reproduced for small samples, even when every aspect of the analysis is correctly specified. Moreover, the observed sensitivity of WGEE to misspecification in either the dropout or measurement model renders these asymptotic properties meaningless. Misspecifications are common in practice and it is seldom the case that one would have an entirely correctly specified analysis model. This, along with the fact that the desirable properties of WGEE are not attained for modest sample sizes, which is common in typical clinical trials, discourages its recommendation. Some amount of improvement, however, was observed under increased proportions of missing data. MI-GEE and MI-Transition, on the other hand, demonstrate a certain degree of robustness to misspecification in either the imputation or measurement model, yielding more precise estimates than WGEE, even despite a further marginalization for the MI-Transition case. Furthermore, while application of MI to the case with missingness in the covariates is relatively easy, WGEE's applicability under such is much less straightforward. Moreover, one can do MI under MAR with intermittent missing data. Though the results of the simulation study provide insight about the methods under consideration, whenever inference is critical, it is always wise to try a couple of different methods, by way of sensitivity analysis.

Previous work on the merits of inverse probability weighting methods *versus* multiple imputation can be found in the discussion of Scharfstein, Rotnitzky and Robins (1999), as well as in Clayton *et al.* (1998) and in Carpenter, Kenward and Vansteelandt (2006). Clayton *et al.* (1998) investigated the use of inverse probability weighting (IPW) and multiple imputation, among others, in the context of longitudinal binary data in a multi-phase sampling setting. They found that, while IPW was inefficient for such a (2×2) phase design, MI showed remarkable efficiency. Moreover, this, along with possible extension to data arising from other designs, indicates the substantial strengths of MI. Carpenter, Kenward and Vansteelandt (2006), on the other hand, used simulations to study a so-called doubly-robust IPW estimator, introduced by Scharfstein, Rotnitzky and Robins (1999), in comparison with standard IPW, maximum likelihood and MI. The doubly-robust IPW estimator is a modified version of the usual IPW, proposed to improve the efficiency of IPW estimators. IPW estimators were again found to be inefficient and sensitive to the choice of the weight model, but the doubly-robust version proves to be as efficient as MI and robust to misspecification. Although applied to continuous Gaussian data, they expect the results to generalize to the discrete case. Whereas Clayton *et al.* (1998) used actual data and Carpenter, Kenward and Vansteelandt (2006) used simulations of a small-sample nature, this chapter presents a small-sample simulation study complemented with asymptotic simulations. Results reinforce the strength of MI over IPW, specifically in application to GEE. WGEE, a type of IPW scheme that uses as weights the inverse of the probability of dropout (taken from some dropout model), was found to be inefficient for small-samples with modest amounts of missing data, in line with the findings of these two papers regarding the inefficiency of such IPW schemes. This (lack of) efficiency, however, might well be addressed by adopting the doubly-robust IPW version Scharfstein, Rotnitzky and Robins (1999) in obtaining the WGEE solutions. And finally, though WGEE showed some amount of improvement under the setting with more missingness in the data, in most of the comparisons made, MI-GEE still outperformed WGEE.

6

Multiple-Imputation-Based Pattern-Mixture Modeling

Clinical studies that give rise to, often incomplete, longitudinal data frequently utilize, as mode of analysis, the selection modeling framework (Rubin, 1976; Little and Rubin, 1987). The choice is natural and obvious, particularly when scientific interest lies on the marginal effects of the independent variables on the response, since resulting estimates from selection models immediately provide such marginal effects. Sometimes, interest can also very well be on the effects of the covariates on a particular group of subjects, e.g., on completers, and on how these effects vary across different response groups. In such cases, pattern-mixture models (Little, 1993, 1994a) could provide a suitable course of analysis, since parameter estimates in a PMM denote pattern-specific effects of the covariates on the response.

References to pattern-mixture models include Rubin (1977); Glynn, Laird and Rubin (1986); Little and Rubin (1987); Hogan and Laird (1997); Molenberghs and Kenward (2007). Several authors have contrasted selection models and pattern-mixture models, to either compare the answer to the same scientific question, such as a marginal treatment effect or time evolution, as a form of sensitivity analysis, or to gain additional insight by supplementing the results from a selection model with those of a pattern-mixture approach. Examples can be found in Verbeke, Lesaffre and

Spiessens (2001) and Michiels *et al.* (2002) for continuous outcomes, while categorical outcomes have been treated by Michiels, Molenberghs and Lipsitz (1999a,b). On a more directed slant, i.e., not necessarily *vis a vis* the selection modeling approach, Thijs *et al.* (2002) discussed various strategies in fitting pattern-mixture models for continuous outcomes via model simplification, while Jansen and Molenberghs (2007) explored strategies for the categorical case using identifying restrictions.

In this chapter, pattern-mixture models are applied to categorical data with monotone missingness using the approach described by Jansen and Molenberghs (2007), which will be investigated over simulated settings, so as to be able to assess the performance of the method. Moreover, while Jansen and Molenberghs (2007) derived formulae for marginalized point estimates, asymptotic variance expressions derived by Sotto *et al.* (2009a) will be introduced here to supplement these.

6.1 Pattern-Mixture Models

Pattern-mixture models (Little, 1993, 1994a, 1995; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007) employ a different response model for each pattern of the missing values, the observed data being a mixture of these, weighted by the probability of each missing value or dropout pattern. The family of pattern-mixture models is based on factorization (3.2), where the conditional density of the measurements given the missingness process is combined with the marginal density describing the missingness mechanism. Note that the latter can depend on covariates, but not on the random outcomes, and, in addition, it is possible to have different covariate dependencies in either component of the factorization.

A key issue in this modeling framework is that pattern-mixture models are, by construction, under-identified, that is, over-specified. Little (1993, 1994a) addresses this problem with the use of so-called identifying restrictions, whereby inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. Under *complete case missing values* (CCMV), information that is unavailable is always borrowed from the completers. Alternatively, the nearest identified pattern can be used as a donor (*neighboring case missing values*, NCMV). Using such identification schemes, within a specific pattern, the conditional distribution of the unobserved measurements, given the observed ones becomes identified. Yet a third possibility is to borrow information for an unidentified distribution from all patterns for which it is identified (*available case missing values*, ACMV). A further set of restrictions is presented in Kenward, Molenberghs and Thijs

(2003). The concept of restrictions will be elaborated upon in the following subsection.

Alternative to the use of identifying restrictions, model simplification can be considered as a means of addressing the under-identification in pattern-mixture models. This approach might, for instance, restrict trends to functional forms supported by the observed data within a pattern, e.g., linear or quadratic time trend. One can also take the route of allowing the parameters to vary in a controlled parametric way, and, for example, assume that the time evolution within each pattern is unstructured but parallel across patterns. Thijs *et al.* (2002) discussed these sub-strategies in detail, in the context of continuous longitudinal outcomes. Examples can be found in Molenberghs and Kenward (2007).

6.1.1 Identification Schemes

Restricting attention to the case of monotone missingness or dropout, assume that all subjects are to be observed at time points $t = 1, 2, \ldots, n$, and thus, $n_i = n, \forall i = 1, 2, \ldots, N$. This results in n patterns, for which the complete data density is given by

$$f_t(y_1, y_2, \ldots, y_n) = f_t(y_1, y_2, \ldots, y_t)\, f_t(y_{t+1}, y_{t+2}, \ldots, y_n | y_1, y_2, \ldots, y_t). \tag{6.1}$$

Owing to the monotone nature of the incompleteness, the number of repeated measurements is equal to the number of potential patterns, even though some of the latter may turn out to be empty. Density (6.1) describes all n outcomes, conditional on pattern t being observed. Although the first factor on the right hand side is clearly identified from the observed data, and can hence be modeled using the observed data, the second is not, necessitating the application of identifying restrictions.

While, in principle, completely arbitrary restrictions can be used by means of any valid density function over the appropriate support, strategies which relate back to the observed data deserve privileged interest. One can base identification on all patterns for which a given component y_s is identified. A general expression for this is

$$f_t(y_s | y_1, y_2, \ldots, y_{s-1}) = \sum_{j=s}^{n} \omega_{sj}\, f_j(y_s | y_1, y_2, \ldots, y_{s-1}), \quad s = t+1, t+2, \ldots, n. \tag{6.2}$$

Let $\boldsymbol{\omega}_s = (\omega_{ss}, \omega_{s,s+1}, \ldots, \omega_{sn})'$. Every $\boldsymbol{\omega}_s$ with components summing to one provides a valid identification scheme. Three special and important cases are considered. Little (1993) proposed *complete case missing values* (CCMV), which uses the following identification:

$$f_t(y_s | y_1, y_2, \ldots, y_{s-1}) = f_n(y_s | y_1, y_2, \ldots, y_{s-1}), \quad s = t+1, t+2, \ldots, n. \tag{6.3}$$

In other words, the conditional distribution beyond time t is always borrowed from the corresponding conditional distribution of the completers, i.e., $\omega_{sn} = 1$ and all other ω_{sj} are set to zero. This identification scheme is perhaps most reasonable and applicable when a large bulk of the subjects have complete data and only small proportions fall within the various dropout patterns. Extension of this approach to non-monotone patterns is also particularly easy. Alternatively, the nearest identified pattern can be used to identify missing components, i.e.,

$$f_t(y_s|y_1, y_2, \dots, y_{s-1}) = f_s(y_s|y_1, y_2, \dots, y_{s-1}), \quad s = t+1, t+2, \dots, n. \tag{6.4}$$

Such restrictions are referred to as *neighboring case missing values* (NCMV), with $\omega_{ss} = 1$ and the other ω_{sj} set to zero. The third special case of (6.2) is *available case missing values* (ACMV), under which derivation of the corresponding $\boldsymbol{\omega}_s$ vectors is straightforward and results in

$$\omega_{sj} = \frac{\pi_j \, f_j(y_1, y_2, \dots, y_{s-1})}{\displaystyle\sum_{\ell=s}^{n} \pi_\ell \, f_\ell(y_1, y_2, \dots, y_{s-1})}, \tag{6.5}$$

where π_j is the fraction of observations in pattern j (Molenberghs *et al.*, 1998). Clearly, $\boldsymbol{\omega}_s$ defined by (6.5) consists of components which are nonnegative and sum to one, i.e., a valid density function is obtained. Molenberghs *et al.* (1998) showed that, for monotone missing data, ACMV in the pattern-mixture framework is the natural counterpart of MAR in the selection model framework.

It should be noted that identifying restrictions are unverifiable assumptions, which also follows from the unidentified nature of the ω parameters, affording sensitivity analysis rather than providing an unambiguous answer to the incompleteness issue. One might, for example, prefer CCMV in cases where the bulk of the data is complete, or perhaps ACMV since, in the monotone case, it is the counterpart of MAR (Molenberghs *et al.*, 1998). However, it is wise not to put too much emphasis on one particular set of restrictions and to consider several. It is also worthwhile to point out that family (6.2) is termed “interior” by Kenward, Molenberghs and Thijs (2003), since identification is done based on distributions following from the data itself. One could, of course, envisage restrictions coming from external sources, such as expert opinions, historical studies, etc. These authors also described identifying restrictions that ensure that dropout does not depend on future, unobserved values, termed *missing non future dependent* (MNFD).

6.1.2 Pattern Mixture Models for Categorical Outcomes with Monotone Missingness

In this section, a procedure for fitting PMMs to categorical data with monotone missingness, proposed by Jansen and Molenberghs (2007), with an application to actual data, will be described. In general, the method can be broken down into three stages. In the first stage, initial models are fitted to the observed data: a univariate model for subjects with one measurement; a bivariate model for subjects with two measurements; and so on. Identification occurs at the second stage, in which a particular identifying restriction is chosen. Using the initial models obtained in stage 1, the required probabilities for the chosen identifying restriction are computed and subsequently used to perform multiple imputation of the missing data to complete each of the patterns. At the last stage, analysis is conducted by fitting full-vector (n-variate) models to the completed data of each pattern and pooling these analyses over the multiple imputations. The mixture of these pattern-specific models comprises the final pattern-mixture model.

In line with Jansen and Molenberghs (2007), the procedure is outlined in detail for the case of three binary outcomes. Extension to more outcomes and/or to more than two outcome categories is straightforward. The multivariate Dale model (Molenberghs and Lesaffre, 1994) will be used to estimate the parameters of the identified densities. For completers (pattern 3), a trivariate Dale model of the form (3.28) will be used, for pattern 2, a bivariate Dale model as given by (3.29), and for pattern 1, a univariate Dale model, which is equivalent to conventional logistic regression. This is referred to as the *minimal approach* (Jansen and Molenberghs, 2007), which will be described shortly.

It is important to point out that, using the MDM formulations given by (3.28), the incomplete patterns provide information neither about the unobserved outcomes nor about the associations involving those unobserved outcomes. Thus, for pattern 2, only an analogous model involving (3.28a), (3.28b) and (3.28d) can be obtained from the data, while for pattern 1 only (3.28a) will be available. Of course, the functions (3.28a)-(3.28g) are specific to a particular pattern and therefore the design matrices may also change from pattern to pattern. This is necessary, among others, when different patterns correspond to different sets of parameters.

Also in this setting, one is interested in model parameters for the full set of repeated outcomes, and thus identifying restrictions are necessary to determine the unknown probabilities by equating them to functions of known probabilities. In the normal case, restrictions are very natural to apply, because marginal as well as conditional distributions can be expressed as simple functions of the mean vector and the covariance matrix components. For categorical data in general and for the Dale model in particular, there is no easy transition from marginal to conditional distributions in terms of the model parameters.

Going back now to the minimal approach, a trivariate Dale model for the complete pattern is combined with a bivariate and univariate Dale model for the incomplete patterns, resulting in densities $f_3(y_1, y_2, y_3)$, $f_2(y_1, y_2)$ and $f_1(y_1)$, respectively. From this approach, the following underlying probabilities can be estimated:

$$\begin{aligned} p_{y_1,y_2,y_3|3} &= P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3|t = 3), \\ p_{y_1,y_2|2} &= P(Y_1 = y_1, Y_2 = y_2|t = 2), \\ p_{y_1|1} &= P(Y_1 = y_1|t = 1). \end{aligned}$$

For pattern 2, there is only one possibility to impute the missing cell counts, since information on the third measurement can only be borrowed from pattern 3. So, the partial counts $Z_{y_1,y_2|2}$ and the conditional probabilities $p_{y_3|y_1,y_2,3} = P(Y_3 = y_3|Y_1 = y_1, Y_2 = y_2, t = 3)$ have to be used to identify $Z^*_{y_1,y_2,y_3|2}$ as $Z_{y_1,y_2|2} \times p_{y_3|y_1,y_2,3}$. The asterisk refers to a completed count, while corresponding counts are indicated by the symbol Z. For pattern 1, several possibilities are available to impute the missing cell counts, since information on the second measurement can be borrowed from pattern 2 as well as from pattern 3. Using (6.2), the joint probability of y_1, y_2 and y_3 in pattern 1 can be written as:

$$p_{y_1,y_2,y_3|1} = p_{y_1|1} \left[\omega \, p_{y_2|y_1,2} + (1-\omega) \, p_{y_2|y_1,3}\right] p_{y_3|y_1,y_2,3},$$

where specific choices of ω lead to the previously defined identifying restrictions, i.e.,

$$\begin{aligned} \text{CCMV} &: \; p_{y_1|1} \, p_{y_2|y_1,3} \, p_{y_3|y_1,y_2,3}, \\ \text{NCMV} &: \; p_{y_1|1} \, p_{y_2|y_1,2} \, p_{y_3|y_1,y_2,3}, \\ \text{ACMV} &: \; \omega = \frac{\pi_2 p_{y_1|2}}{\pi_2 p_{y_1|2} + \pi_3 p_{y_1|3}}, \end{aligned}$$

such that the missing cell counts can be identified as follows:

$$\text{CCMV} : \; \widehat{Z}^*_{y_1,y_2,y_3|1} = Z_{y_1|1} \, \widehat{p}_{y_2|y_1,3} \, \widehat{p}_{y_3|y_1,y_2,3}, \tag{6.6}$$

$$\text{NCMV} : \; \widehat{Z}^*_{y_1,y_2,y_3|1} = Z_{y_1|1} \, \widehat{p}_{y_2|y_1,2} \, \widehat{p}_{y_3|y_1,y_2,3}, \tag{6.7}$$

$$\text{ACMV} : \; \widehat{Z}^*_{y_1,y_2,y_3|1} = Z_{y_1|1} \left[\frac{\widehat{\pi}_2 \widehat{p}_{y_1,y_2|2} + \widehat{\pi}_3 \widehat{p}_{y_1,y_2|3}}{\widehat{\pi}_2 \widehat{p}_{y_1|2} + \widehat{\pi}_3 \widehat{p}_{y_1|3}}\right] \widehat{p}_{y_3|y_1,y_2,3}. \tag{6.8}$$

Multiple imputations are then generated by drawing uniformly from Bernoulli variables with probabilities embedded in (6.6)–(6.8). As stated earlier, once the imputations have been generated, the so-called analysis task model, i.e., the final model, can be fitted and multiple-imputation inference conducted. Using conventional multiple-imputation machinery (Little and Rubin, 1987), obtaining parameter and precision estimates is straightforward via expressions (3.13) and (3.14).

Although Jansen and Molenberghs (2007) described an extension of the above procedure to the case of non-monotone missingness, their analyses were confined to the monotone case, owing to an almost negligible proportion of intermittent missingness in the data they considered. They further mention that generalization of the procedure to the non-monotone case would result in a proliferation of parameters, leading to a trade-off between clarity and parsimony.

6.1.3 Marginalization Across Patterns

By way of its factorization of the joint distribution of $\boldsymbol{y}_i$ and $\boldsymbol{r}_i$, as specified in (3.2), a PMM gives rise to pattern-specific estimates. In cases where the scientific interest is on such, a PMM would, of course, be a natural choice. If, however, interest lies on marginal effects, one might opt for a selection model instead. It is also possible though that one wants to obtain, from a fitted PMM, so-called "marginalized" effects, i.e., pattern-specific effects marginalized across all patterns. This might be the case, for example, when one wants to compare PMM results with their selection model counterparts. In the case of continuous data, the overall (or marginalized) effect is simply a weighted average of the pattern-specific effects. For categorical data, however, the marginalization is less straightforward. Jansen and Molenberghs (2007) propose an approximation for the marginalized effects of a PMM for a logistic model.

Suppose that to model the data from pattern t, one considers a logistic regression of a binary response Y_{ij} on some treatment of interest, X_i, of the form

$$P(Y_{ij} = 1|t) = \frac{e^{\alpha_t + \beta_t X_i}}{1 + e^{\alpha_t + \beta_t X_i}}. \tag{6.9}$$

In general, $t = 1, 2, \ldots, K$, where $K = n$ for the monotone case (as in Section 6.1.1), while $K = 2^n$ for non-monotone missingness, in which case, for instance, the multivariate Dale model might be used to model the non-response. The parameters α_t and β_t can depend on j, but this index is suppressed from notation. Assuming that interest is on one particular treatment effect X, e.g., treatment effect at the last occasion, and that π_t denote the pattern probabilities as defined previously, the marginal success probability over all patterns, is then equal to

$$P(Y_{ij} = 1) = \sum_{t=1}^{K} \pi_t \, \frac{e^{\alpha_t + \beta_t X}}{1 + e^{\alpha_t + \beta_t X}}. \tag{6.10}$$

From this formulation, there are three ways to calculate the marginal effects (e.g., the intercept A and treatment effect B) at the last occasion. First, using a direct linear approach (Park and Lee, 1999), where a weighted average over all patterns is taken, i.e.,

$$A_{PL} = \sum_t \pi_t \alpha_t \quad \text{and} \quad B_{PL} = \sum_t \pi_t \beta_t. \tag{6.11}$$

Though seemingly logical, this marginalization may not be entirely appropriate for binary responses, as will be shown shortly. Second, the marginal probability can be approximated via a logistic regression, a probit model, or by fully using the longitudinal nature of the design, through a Dale model, a generalized linear mixed model, etc. Third, classical averaging can be performed. To this effect, function (6.10) is kept as is and is then computed, graphed, or sampled from. Note that averaging in this last way will be similar to the marginalization of generalized linear mixed-effects model (Molenberghs and Verbeke, 2005). Here, the marginalization is over the patterns, rather than over random effects. When a GLMM is used in each pattern, then a double marginalization is necessary, one over the random effects and another over the patterns. Focus here will be on the second approach – using a Dale model – but comparison will also be made with the less appropriate but sometimes used direct linear approach.

Jansen and Molenberghs (2007) approximate (6.10) by a logistic regression, i.e., letting A_{JM} and B_{JM} denote the marginalized effects,

$$f(X) \;=\; \sum_t \pi_t \frac{e^{\alpha_t+\beta_t X}}{1+e^{\alpha_t+\beta_t X}} \;\cong\; \frac{e^{A_{JM}+B_{JM}X}}{1+e^{A_{JM}+B_{JM}X}}, \tag{6.12}$$

and derived the following expressions for A_{JM} and B_{JM}:

$$\begin{aligned} A_{JM} &= \text{logit}\left(\sum_t \pi_t \frac{e^{\alpha_t}}{1+e^{\alpha_t}}\right) \qquad \text{and} \\ B_{JM} &= \frac{\displaystyle\sum_t \pi_t \, \beta_t \frac{e^{\alpha_t}}{1+e^{\alpha_t}} \frac{1}{1+e^{\alpha_t}}}{\left(\displaystyle\sum_t \pi_t \frac{e^{\alpha_t}}{1+e^{\alpha_t}}\right)\left(\displaystyle\sum_t \pi_t \frac{1}{1+e^{\alpha_t}}\right)}. \end{aligned} \tag{6.13}$$

They further showed that when the treatment effects are equal across patterns, the marginalized treatment effect at the last occasion obtained through approximation (6.12), will not be larger in absolute value than that obtained using the direct linear approach (6.11). Moreover, marginalization may both increase or decrease the effect in absolute value when the treatment effects differ across patterns.

Considering now the direct linear approach, Park and Lee (1999) assume that the pattern-specific success probability (6.9) is approximately linear, i.e.,

$$P(Y_{ij} = 1|t) \;=\; \frac{e^{\alpha_t+\beta_t X_i}}{1+e^{\alpha_t+\beta_t X_i}} \;\cong\; \alpha_t + \beta_t X_i, \tag{6.14}$$

whereby averaging over all patterns yields the following expression for the marginal success probability (6.10):

$$f(X) \;=\; \sum_t \pi_t \, \frac{e^{\alpha_t+\beta_t X}}{1+e^{\alpha_t+\beta_t X}} \;\cong\; A_{PL} + B_{PL} X, \tag{6.15}$$

with A_{PL} and B_{PL} as defined in (6.11). Clearly, the assumption that the pattern-specific success probability is linear, though probable in certain cases, is generally not realistic for most scenarios. Moreover, approximation (6.14) essentially ignores the presence of the link function in modeling the binary response. Though this approach to obtain marginal effects is entirely appropriate for Gaussian outcomes, for which the response is modeled directly, i.e., with an identity link function, such an approximation can easily fail when modeling is done via a link, especially when the link entails a highly non-linear transformation of the response.

Suppose that $\widehat{\alpha}_t$ and $\widehat{\beta}_t$ respectively denote the maximum likelihood estimates of the pattern-specific parameters, α_t and β_t, of the trivariate Dale model for pattern t, for $t = 1, 2, 3$. Suppose further that $\widehat{\pi}_t$ denotes the sample proportion for pattern t, for $t = 1, 2, 3$. Note that sample proportions are the maximum likelihood estimates for the true pattern proportions π_t in a multinomial model. Substituting now $\widehat{\alpha}_t, \widehat{\beta}_t$ and $\widehat{\pi}_t$ into expressions (6.11) and (6.13), the estimates for the marginalized effects are given by:

$$\widehat{A}_{PL} \;=\; \sum_t \widehat{\pi}_t \, \widehat{\alpha}_t \qquad \text{and} \qquad \widehat{B}_{PL} = \sum_t \widehat{\pi}_t \, \widehat{\beta}_t, \tag{6.16}$$

$$\begin{aligned} \widehat{A}_{JM} \;&=\; \text{logit}\left(\sum_t \widehat{\pi}_t \, \frac{e^{\widehat{\alpha}_t}}{1+e^{\widehat{\alpha}_t}}\right) \qquad \text{and} \\ \widehat{B}_{JM} \;&=\; \frac{\displaystyle\sum_t \widehat{\pi}_t \, \widehat{\beta}_t \, \frac{e^{\widehat{\alpha}_t}}{1+e^{\widehat{\alpha}_t}} \, \frac{1}{1+e^{\widehat{\alpha}_t}}}{\left(\displaystyle\sum_t \widehat{\pi}_t \, \frac{e^{\widehat{\alpha}_t}}{1+e^{\widehat{\alpha}_t}}\right)\left(\displaystyle\sum_t \widehat{\pi}_t \, \frac{1}{1+e^{\widehat{\alpha}_t}}\right)}. \end{aligned} \tag{6.17}$$

The asymptotic variances of these estimators, derived in Sotto *et al.* (2009a), can be obtained using the delta method, for which independence of the pattern-specific estimates across patterns is assumed. Such an assumption of independence of the pattern-specific estimates across patterns is entirely justified in a PMM since, by (3.2), the outcome vector is modeled conditionally on a given dropout pattern and thus, the pattern-specific parameters for one particular pattern can be viewed as being estimated independently of those in other patterns. In fact, the model for the outcomes given a pattern can differ across patterns. It is further assumed that the pattern probabilities and the pattern-specific parameters are estimated independently of each other. Again, this is plausible in a PMM since, by way of factorization (3.2), the dropout process, from which the pattern probabilities arise, is modeled independently of the outcomes, thereby rendering the estimates of the former independent of the pattern-specific estimates that characterize the latter. Finally, because the maximum likelihood estimates of α_t, β_t and π_t are used, the above estimates (6.16) and (6.17) are consistent for (6.11) and (6.13), respectively, and this follows directly from standard maximum likelihood principles, under the usual regularity conditions (Welsh, 1996).

There is a need to be concerned, a priori, with potential dependencies between observations after the imputation process has taken place, because, for example, information on the completers is used to impute sequences that are incomplete. However, application of multiple imputation here follows the general theory (Little and Rubin, 1987), where the observed data are used to estimate the parameters from the imputation distribution. Observations would definitely become dependent if this parameter were used as if it were known. To alleviate this problem, first, draws are made from the parameter's posterior distribution, separately for each of the multiple imputations, and only thereafter are imputations generated. While not straightforward to establish independence in a particular situation, general theory (Rubin, 1987) states that this procedure removes or at least alleviates this problem. Moreover, it is sensible to assume that the correlations are mild to begin with. In the particular case considered here, this rests on the following assumptions:

- Independence of pattern-specific estimates across patterns implies that $\forall k \neq \ell$,

$$\text{Cov}(\widehat{\alpha}_k, \widehat{\alpha}_\ell) = 0, \quad \text{Cov}(\widehat{\beta}_k, \widehat{\beta}_\ell) = 0 \quad \text{and} \quad \text{Cov}(\widehat{\alpha}_k, \widehat{\beta}_\ell) = 0.$$

- Independence of estimates for pattern probabilities and (measurement model) pattern-specific estimates further implies:

$$\begin{aligned} \text{Cov}(\widehat{\alpha}_k, \widehat{\pi}_k) = 0 \qquad &\text{and} \qquad \text{Cov}(\widehat{\beta}_k, \widehat{\pi}_k) = 0, \quad \forall k = 1, 2, 3, \\ \text{Cov}(\widehat{\alpha}_k, \widehat{\pi}_\ell) = 0 \qquad &\text{and} \qquad \text{Cov}(\widehat{\beta}_k, \widehat{\pi}_\ell) = 0, \quad \forall k \neq \ell. \end{aligned}$$

With the aforementioned assumptions, for the specific case of three outcomes considered here, derivation of the asymptotic variances for the direct linear approach (6.11) is straightforward, yielding

$$\begin{aligned}
\mathrm{Var}(\widehat{A}_{PL}) &= \sum_{t=1}^{3} {\pi_t}^2 \mathrm{Var}(\widehat{\alpha}_t) + \sum_{t=1}^{3} {\alpha_t}^2 \mathrm{Var}(\widehat{\pi}_t) + 2\sum_{t<\ell} \alpha_t \alpha_\ell \mathrm{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell), \\
\mathrm{Var}(\widehat{B}_{PL}) &= \sum_{t=1}^{3} {\pi_t}^2 \mathrm{Var}(\widehat{\beta}_t) + \sum_{t=1}^{3} {\beta_t}^2 \mathrm{Var}(\widehat{\pi}_t) + 2\sum_{t<\ell} \beta_t \beta_\ell \mathrm{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell).
\end{aligned} \tag{6.18}$$

For the approach of Jansen and Molenberghs (2007), the marginalized effects in (6.13) can be expressed in terms of the following functions:

$$\begin{aligned}
U \equiv f_1(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}}, \\
V \equiv f_2(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \frac{1}{1+e^{\alpha_k}}, \qquad \text{and} \\
W \equiv f_3(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \, \beta_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}} \frac{1}{1+e^{\alpha_k}},
\end{aligned}$$

with corresponding estimators given by:

$$\begin{aligned}
\widehat{U} \equiv f_1(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}}, \\
\widehat{V} \equiv f_2(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \frac{1}{1+e^{\widehat{\alpha}_k}}, \qquad \text{and} \\
\widehat{W} \equiv f_3(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \, \widehat{\beta}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}} \frac{1}{1+e^{\widehat{\alpha}_k}}.
\end{aligned}$$

Application of the delta method on (6.13), as expressed in terms of these functions, yields the following asymptotic variances:

$$\begin{aligned}
\mathrm{Var}(\widehat{A}_{JM}) &= \frac{1}{U^2(1-U)^2} \mathrm{Var}(\widehat{U}), \\
\mathrm{Var}(\widehat{B}_{JM}) &= \frac{W^2}{U^4V^2} \mathrm{Var}(\widehat{U}) + \frac{W^2}{U^2V^4} \mathrm{Var}(\widehat{V}) + \frac{1}{U^2V^2} \mathrm{Var}(\widehat{W}) + \\
&\quad \frac{2W^2}{U^3V^3} \mathrm{Cov}(\widehat{U}, \widehat{V}) + \frac{2W}{U^3V^2} \mathrm{Cov}(\widehat{U}, \widehat{W}) + \frac{2W}{U^2V^3} \mathrm{Cov}(\widehat{V}, \widehat{W}),
\end{aligned} \tag{6.19}$$

where:

$$
\begin{aligned}
\operatorname{Var}(\widehat{U}) &= \sum_{t=1}^{3} {\pi_t}^2 \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^2} \operatorname{Var}(\widehat{\pi}_t) \;+ \\
&\quad 2 \sum_{t<\ell} \frac{e^{\alpha_t} e^{\alpha_\ell}}{(1+e^{\alpha_t})(1+e^{\alpha_\ell})} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Var}(\widehat{V}) &= \sum_{t=1}^{3} {\pi_t}^2 \frac{(-e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^2} \operatorname{Var}(\widehat{\pi}_t) \;+ \\
&\quad 2 \sum_{t<\ell} \frac{1}{(1+e^{\alpha_t})(1+e^{\alpha_\ell})} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Var}(\widehat{W}) &= \sum_{t=1}^{3} {\pi_t}^2 {\beta_t}^2 \frac{(e^{\alpha_t})^2(1-e^{\alpha_t})^2}{(1+e^{\alpha_t})^6} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} {\pi_t}^2 \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Var}(\widehat{\beta}_t) \;+ \\
&\quad \sum_{t=1}^{3} {\beta_t}^2 \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Var}(\widehat{\pi}_t) \;+\; 2 \sum_{t=1}^{3} {\pi_t}^2 \beta_t \frac{(e^{\alpha_t})^2(1-e^{\alpha_t})}{(1+e^{\alpha_t})^5} \operatorname{Cov}(\widehat{\alpha}_t, \widehat{\beta}_t) \;+ \\
&\quad 2 \sum_{t<\ell} \beta_t \beta_\ell \frac{e^{\alpha_t} e^{\alpha_\ell}}{(1+e^{\alpha_t})^2(1+e^{\alpha_\ell})^2} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{U}, \widehat{V}) &= \sum_{t=1}^{3} {\pi_t}^2 (-1) \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} \frac{e^{\alpha_t}}{(1+e^{\alpha_t})^2} \operatorname{Var}(\widehat{\pi}_t) \;+ \\
&\quad \sum_{t<\ell} \frac{e^{\alpha_t} + e^{\alpha_\ell}}{(1+e^{\alpha_t})(1+e^{\alpha_\ell})} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{U}, \widehat{W}) &= \sum_{t=1}^{3} {\pi_t}^2 \beta_t \frac{(e^{\alpha_t})^2(1-e^{\alpha_t})}{(1+e^{\alpha_t})^5} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} \beta_t \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^3} \operatorname{Var}(\widehat{\pi}_t) \;+ \\
&\quad \sum_{t=1}^{3} {\pi_t}^2 \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Cov}(\widehat{\alpha}_t, \widehat{\beta}_t) \;+ \\
&\quad \sum_{t<\ell} e^{\alpha_t} e^{\alpha_\ell} \frac{\beta_t(1+e^{\alpha_\ell}) + \beta_\ell(1+e^{\alpha_t})}{(1+e^{\alpha_t})^2(1+e^{\alpha_\ell})^2} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{V}, \widehat{W}) &= \sum_{t=1}^{3} {\pi_t}^2 (-1) \beta_t \frac{(e^{\alpha_t})^2(1-e^{\alpha_t})}{(1+e^{\alpha_t})^5} \operatorname{Var}(\widehat{\alpha}_t) \;+\; \sum_{t=1}^{3} \beta_t \frac{e^{\alpha_t}}{(1+e^{\alpha_t})^3} \operatorname{Var}(\widehat{\pi}_t) \;+ \\
&\quad \sum_{t=1}^{3} {\pi_t}^2 (-1) \frac{(e^{\alpha_t})^2}{(1+e^{\alpha_t})^4} \operatorname{Cov}(\widehat{\alpha}_t, \widehat{\beta}_t) \;+ \\
&\quad \sum_{t<\ell} \frac{(1+e^{\alpha_t})\beta_\ell e^{\alpha_\ell} + (1+e^{\alpha_\ell})\beta_t e^{\alpha_t}}{(1+e^{\alpha_t})^2(1+e^{\alpha_\ell})^2} \operatorname{Cov}(\widehat{\pi}_t, \widehat{\pi}_\ell)
\end{aligned}
$$

From (6.12) and (6.15) it is easy to see that (A_{JM}, B_{JM}) and (A_{PL}, B_{PL}) are two approximations for some true underlying marginal effects, say (A, B), which are unknown but might be estimated by fitting a selection model. For the ensuing simulation study, the marginal effects were obtained by factoring the underlying data generating mechanism in a selection model (SeM) formulation, via a procedure that will be defined shortly. The estimators $(\widehat{A}_{JM}, \widehat{B}_{JM})$ and $(\widehat{A}_{PL}, \widehat{B}_{PL})$ can thus be used to estimate these underlying SeM-type marginal effects. In the simulation, these two approximations are examined with respect to how they fare in estimating the true SeM-type marginal parameters.

The procedure to obtain underlying marginal parameters from a PMM, with which the above-defined marginalized effects estimates, $(\widehat{A}_{PL}, \widehat{B}_{PL})$ and $(\widehat{A}_{JM}, \widehat{B}_{JM})$, can be subsequently compared, will be described for the specific case of 3 time points, though generalization to more time points is straightforward. Recall that the general formulation of a PMM given in (3.2) allows the computation of the joint distribution of the binary responses, $\boldsymbol{Y}$, and the dropout pattern, t, conditional on the treatment indicator, X, as

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}, t|X) &= P(\boldsymbol{Y} = \boldsymbol{y}|t, X)\, P(t|X) \\ &= P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3|t, X)\, P(t|X) \\ &= P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3|t, X)\, \pi_{t|X}, \end{aligned}$$

where $\pi_{t|X}$ denotes the pattern proportions conditional on the level of the treatment indicator. The marginal success probability, say, at the last time point, given treatment X, can then be obtained by summing over all response values at all other time points and over all patterns, i.e.,

$$\begin{aligned} P(Y_3 = 1|X) &= \sum_{\forall y_1, y_2, t} P(\boldsymbol{Y} = \boldsymbol{y}, t|X) \\ &= \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} \sum_{t=1}^{3} P(Y_1 = y_1, Y_2 = y_2, Y_3 = 1|t, X)\, \pi_{t|X} \\ &= \sum_{t=1}^{3} \pi_{t|X}\, P(Y_3 = 1|t, X). \end{aligned} \tag{6.20}$$

Assuming now a logistic regression of Y_3 on the treatment indicator, X,

$$\text{logit } P(Y_3 = 1|X) = A_3 + B_3 X,$$

implying

$$\begin{aligned} \text{logit } P(Y_3 = 1|X = 0) &= A_3 \qquad \text{and} \\ \text{logit } P(Y_3 = 1|X = 1) &= A_3 + B_3, \end{aligned}$$

which can then be solved for the true underlying parameters, A_3 and B_3, as

$$\begin{aligned} A_3 &= \text{logit } P(Y_3 = 1|X = 0) \qquad \text{and} \\ B_3 &= \text{logit } P(Y_3 = 1|X = 1) - \text{logit } P(Y_3 = 1|X = 0). \end{aligned}$$

Though interest is usually on the effects at the last time point, effects at intermediate time points can also be obtained analogously.

Note that $P(Y_3 = 1|t, X)$ in (6.20) is modeled logistically as specified by (3.28c) in the Dale model, i.e.,

$$\text{logit } P(Y_3 = 1|t, X) \quad = \quad \alpha_{t,3} + \beta_{t,3}X,$$

which is equivalent to

$$P(Y_3 = 1|t, X) \quad = \quad \frac{e^{\alpha_{t,3}+\beta_{t,3}X}}{1 + e^{\alpha_{t,3}+\beta_{t,3}X}},$$

which, upon substitution into (6.20) leads to

$$P(Y_3 = 1|X) \quad = \quad \sum_{t=1}^{3} \pi_{t|X} \frac{e^{\alpha_{t,3}+\beta_{t,3}X}}{1 + e^{\alpha_{t,3}+\beta_{t,3}X}}. \qquad (6.21)$$

Clearly, this is not equivalent to (6.10), which uses the unconditional pattern proportions, π_t, as weights. Since $\widehat{A}_{JM}$ and $\widehat{B}_{JM}$, as well as $\widehat{A}_{PL}$ and $\widehat{B}_{PL}$, are both based on (6.10), it would be natural to expect bias for these when estimating the underlying SeM-type marginal effects. Thus, in presenting the results of the simulation, MSE, rather than variances, shall be used to assess the precision of the said estimates. Of course in the case that the conditional pattern proportions are equal to the unconditional ones, $\pi_{t|X} = \pi_t, \forall X$, e.g., under an MCAR mechanism, then the estimates of Jansen and Molenberghs (2007) are unbiased for the SeM-type marginal parameters.

6.2 Simulation Study

In this section, the design of the simulation study is first presented, followed by the primary results. The last subsection includes results of additional simulations conducted to investigate convergence properties, as well as to gain more insight into the performance of the different identifying restrictions.

6.2.1 Design of the Simulation Study

For the simulation study, 3 binary outcomes, (Y_1, Y_2, Y_3), and a single two-level covariate, X, possibly denoting a treatment indicator, were considered. To formulate an underlying PMM for the outcomes, a distinct trivariate Dale model of the general form (3.28) was defined for each pattern. Here, for each outcome, a simple logistic model was specified, consisting of an intercept and treatment effect, respectively denoted as α_t and β_t, $t = 1, 2, 3$. Further, the association parameters (ψ) were set to be constant. This implies a 10-dimensional full parameter vector, i.e.,

$$\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \ln\psi_{12}, \ln\psi_{13}, \ln\psi_{23}, \ln\psi_{123})'.$$

Letting $\boldsymbol{C} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, the design matrices in (3.28) can be defined as:

$$\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{C} & {}_2\mathbf{0}_8 \\ {}_8\mathbf{0}_2 & {}_8\mathbf{0}_8 \end{pmatrix}, \quad \boldsymbol{X}_2 = \begin{pmatrix} {}_2\mathbf{0}_2 & {}_2\mathbf{0}_2 & {}_2\mathbf{0}_6 \\ {}_2\mathbf{0}_2 & \boldsymbol{C} & {}_2\mathbf{0}_6 \\ {}_6\mathbf{0}_2 & {}_6\mathbf{0}_2 & {}_6\mathbf{0}_6 \end{pmatrix}, \quad \boldsymbol{X}_3 = \begin{pmatrix} {}_4\mathbf{0}_4 & {}_4\mathbf{0}_2 & {}_4\mathbf{0}_4 \\ {}_2\mathbf{0}_4 & \boldsymbol{C} & {}_2\mathbf{0}_4 \\ {}_4\mathbf{0}_4 & {}_4\mathbf{0}_2 & {}_4\mathbf{0}_4 \end{pmatrix},$$

while for $i = 4, 5, 6$ and 7,

$$\begin{aligned} \boldsymbol{X}_4 &= [x_4]_{ij}\,, \text{ where } x_4 = \begin{cases} 1, & \text{for } i = 7 \text{ and } j = 7 \\ 0, & \text{otherwise} \end{cases}, \\ \boldsymbol{X}_5 &= [x_5]_{ij}\,, \text{ where } x_5 = \begin{cases} 1, & \text{for } i = 8 \text{ and } j = 8 \\ 0, & \text{otherwise} \end{cases}, \\ \boldsymbol{X}_6 &= [x_6]_{ij}\,, \text{ where } x_6 = \begin{cases} 1, & \text{for } i = 9 \text{ and } j = 9 \\ 0, & \text{otherwise} \end{cases}, \\ \boldsymbol{X}_7 &= [x_7]_{ij}\,, \text{ where } x_7 = \begin{cases} 1, & \text{for } i = 10 \text{ and } j = 10 \\ 0, & \text{otherwise} \end{cases}, \end{aligned}$$

which further implies

$$\begin{aligned} \boldsymbol{X}_1\boldsymbol{\theta} &= \boldsymbol{C}\boldsymbol{\theta}_1 = \boldsymbol{C}(\alpha_1, \beta_1)' = (\alpha_1, \alpha_1 + \beta_1)', \\ \boldsymbol{X}_2\boldsymbol{\theta} &= \boldsymbol{C}\boldsymbol{\theta}_2 = \boldsymbol{C}(\alpha_2, \beta_2)' = (\alpha_2, \alpha_2 + \beta_2)', \\ \boldsymbol{X}_3\boldsymbol{\theta} &= \boldsymbol{C}\boldsymbol{\theta}_3 = \boldsymbol{C}(\alpha_3, \beta_3)' = (\alpha_3, \alpha_3 + \beta_3)', \\ \boldsymbol{X}_4\boldsymbol{\theta} &= 1\theta_4 = \ln\psi_{12}, \\ \boldsymbol{X}_5\boldsymbol{\theta} &= 1\theta_5 = \ln\psi_{13}, \\ \boldsymbol{X}_6\boldsymbol{\theta} &= 1\theta_6 = \ln\psi_{23}, \\ \boldsymbol{X}_7\boldsymbol{\theta} &= 1\theta_7 = \ln\psi_{123}. \end{aligned}$$

The values chosen for $\boldsymbol{\theta}$, for each pattern, are given in Table 6.1. The mixture of the 3 trivariate Dale models defined by these sets of parameters gives rise to the underlying PMM used for the simulations.

Monotone missingness was then generated by formulating a logistic dropout model of the form

$$\text{logit } P(D = j | D \geq j, X) = \nu_0 + \nu_1 X,$$

Table 6.1: *Trivariate Dale model parameter values specified for each of the three dropout patterns. (O: observed, M: missing.)*

Pattern	α_1	β_1	α_2	β_2	α_3	β_3	ψ_{12}	ψ_{13}	ψ_{23}	ψ_{123}
1 (OMM)	0.214	0.096	0.182	0.067	0.142	0.090	1.4	1.1	1.6	1.2
2 (OOM)	0.155	0.135	0.130	0.084	0.100	0.083	1.6	1.2	1.5	1.7
3 (OOO)	0.109	0.033	0.155	0.028	0.142	0.056	2.2	1.8	2.3	1.5

with X denoting the treatment indicator. Two dropout settings were considered for the above model, using the following sets of parameters: $(\nu_0, \nu_1) = (-2.2, 0.8)$ and $(\nu_0, \nu_1) = (-1.5, 0.8)$. The resulting percentages of subjects per pattern, as well as per covariate (treatment) level, are given in Table 6.2. Combining these mixing weights from each of these dropout models with the three trivariate Dale models defines two underlying PMMs on which the simulations will be based. The underlying SeM-type marginal parameters corresponding to these two PMMs were computed using the approach described at the end of Section 6.1.3 and are further shown in Table 6.3.

On each of the two defined underlying PMMs, $S = 500$ samples, each of size $N = 1000$, were generated, and the procedure described in Section 6.1.2 was applied to each of the generated samples. Marginalized parameter estimates using both (6.13) and the direct linear method (6.11) were also computed. The choice of parameter values (Table 6.1), as well as the choice of sample size $N = 1000$, was dictated by limitations in fitting both the initial, as well as the final, analysis models. Fitting a trivariate or a bivariate Dale model, or even an ordinary logistic model for that matter, requires that all combinations of the outcome vector are present per level of the covariate. In generating each of the $S = 500$ samples, it was therefore necessary

Table 6.2: *Proportions of missing data patterns, by treatment level and overall, under each dropout setting. (O: observed, M: missing.)*

Pattern	Setting 1			Setting 2		
	$x = 0$	$x = 1$	Overall	$x = 0$	$x = 1$	Overall
1 (OMM)	4.99	9.89	14.88	9.12	16.59	25.71
2 (OOM)	4.49	7.93	12.42	7.46	11.09	18.54
3 (OOO)	40.52	32.17	72.70	33.42	22.32	55.75

Table 6.3: *True marginal parameter values for the two underlying pattern-mixture models. (The marginal intercept and marginal treatment effect for the j^{th} outcome, $j = 1, 2, 3$, are denoted A_j and B_j, respectively.)*

PMM	Y_1		Y_2		Y_3	
	A_1	B_1	A_2	B_2	A_3	B_3
1	0.1236	0.0746	0.1551	0.0457	0.1381	0.0640
2	0.1350	0.0950	0.1559	0.0558	0.1356	0.0700

to ensure that all combinations were "observed" within each pattern. The values for $\boldsymbol{\theta}$ in Table 6.1 were thus chosen so that the probabilities for the trivariate Dale models (per pattern), when combined with the dropout probabilities, would result in joint probabilities that are still large enough to generate all combinations within one pattern for a sample of $N = 1000$.

6.2.2 Primary Results

The discussion of the results are organized into those pertaining to the initial estimates, the pattern-specific estimates, and the marginalized effects estimates, respectively.

6.2.2.1 Initial Estimates

For the first stage, a trivariate Dale model for cases in pattern 3, a bivariate Dale model for the second pattern, and a logistic model for pattern 1 were fitted using the observed data of each sample. At this stage, no attempt is made to address the missingness yet; appropriate models are simply fit to the observed data. The resulting estimates, averaged over the $S = 500$ samples, were then compared with the true parameter values from the underlying trivariate Dale models, so as to obtain the bias and MSE, both of which are presented in Table 6.4.

In both settings, parameter estimates exhibit very small bias and reasonably small MSEs. The results for pattern 3 under setting 1 are the most precise, as might be expected, since it is under this particular setting and within this particular pattern that a large proportion of the subjects fall. For the same pattern, but under setting 2, bias and MSEs are slightly larger, since setting 2 consists of more missingness and thus fewer subjects within pattern 3 (see Table 6.2). For patterns 1 and 2, under setting 1, although biases are quite acceptable, the treatment effects, as well as the association, seem to be less precisely estimated than the intercepts, but this improves in setting 2.

Table 6.4: *Bias and MSE of the parameter estimates for the initial models fitted, under each dropout setting, to the observed data: trivariate Dale model for pattern 3, bivariate Dale model for pattern 2 and logistic model for pattern 1.*

Parameter	Pattern 1		Pattern 2		Pattern 3	
	Bias	MSE	Bias	MSE	Bias	MSE
Setting 1						
α_1	0.0044	0.0913	0.0099	0.0908	0.0045	0.0112
β_1	0.0151	0.1284	-0.0302	0.1413	-0.0014	0.0236
α_2	——	——	0.0030	0.0917	0.0049	0.0101
β_2	——	——	-0.0123	0.1386	0.0001	0.0205
α_3	——	——	——	——	0.0012	0.0104
β_3	——	——	——	——	-0.0025	0.0223
$\ln\psi_{12}$	——	——	0.0121	0.1354	-0.0042	0.0252
$\ln\psi_{13}$	——	——	——	——	-0.0057	0.0222
$\ln\psi_{23}$	——	——	——	——	0.0039	0.0217
$\ln\psi_{123}$	——	——	——	——	0.0008	0.0994
Setting 2						
α_1	0.0021	0.0476	0.0054	0.0543	0.0122	0.0139
β_1	0.0005	0.0745	0.0006	0.0840	-0.0115	0.0345
α_2	——	——	-0.0014	0.0570	-0.0005	0.0125
β_2	——	——	0.0039	0.0858	-0.0027	0.0298
α_3	——	——	——	——	0.0070	0.0128
β_3	——	——	——	——	-0.0110	0.0311
$\ln\psi_{12}$	——	——	0.0286	0.0976	0.0006	0.0310
$\ln\psi_{13}$	——	——	——	——	-0.0034	0.0298
$\ln\psi_{23}$	——	——	——	——	0.0047	0.0326
$\ln\psi_{123}$	——	——	——	——	-0.0126	0.1166

Finally, in contrast with the results for pattern 3, the incomplete patterns 1 and 2 yield better results under setting 2, since there are more subjects within these patterns than there are under setting 1.

6.2.2.2 Pattern-Specific Estimates

The results from the initial models fitted to the observed data were then used to apply various identifying restrictions, under which multiple imputations ($M = 5$) were obtained to fill in the missing data. Under each setting and for each identifying restriction considered, three trivariate Dale models were then fitted to the completed

Table 6.5: *Bias and MSE of the pattern-specific parameter estimates for the trivariate Dale models fitted to the completed data using various identifying restrictions, for dropout setting 1.*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
Pattern 1						
α_1	0.0037	0.0911	0.0049	0.0906	0.0047	0.0909
β_1	0.0161	0.1279	0.0144	0.1270	0.0147	0.1281
α_2	0.3139	0.1190	0.2645	0.0946	-0.0464	0.1190
β_2	-0.0423	0.0311	-0.0454	0.0392	0.0162	0.1713
α_3	0.4004	0.1830	0.3514	0.1421	0.2940	0.1087
β_3	-0.0676	0.0387	-0.0739	0.0334	-0.0583	0.0354
$\ln\psi_{12}$	0.4387	0.2346	0.4116	0.2115	0.1578	0.1899
$\ln\psi_{13}$	0.4085	0.2268	0.4752	0.2733	0.4193	0.2286
$\ln\psi_{23}$	0.2775	0.1349	0.3634	0.1818	0.3350	0.1642
$\ln\psi_{123}$	0.3877	0.4634	0.2283	0.3062	0.2464	0.3099
Pattern 2						
α_1	0.0105	0.0901	0.0099	0.0907	0.0098	0.0903
β_1	-0.0310	0.1397	-0.0302	0.1412	-0.0299	0.1406
α_2	0.0030	0.0921	0.0037	0.0915	0.0030	0.0914
β_2	-0.0123	0.1392	-0.0133	0.1375	-0.0124	0.1375
α_3	0.3241	0.1288	0.3312	0.1350	0.3269	0.1318
β_3	-0.0436	0.0416	-0.0578	0.0430	-0.0513	0.0394
$\ln\psi_{12}$	0.0132	0.1355	0.0132	0.1355	0.0131	0.1352
$\ln\psi_{13}$	0.3225	0.1626	0.3308	0.1716	0.3236	0.1631
$\ln\psi_{23}$	0.4009	0.2148	0.3986	0.2167	0.3963	0.2110
$\ln\psi_{123}$	-0.1062	0.2881	-0.0678	0.2934	-0.0876	0.2828

data – one for each pattern. The bias and MSE of the pattern-specific parameter estimates under dropout settings 1 and 2 are given in Tables 6.5 and 6.6, respectively. Only the results for patterns 1 and 2 are presented, since the results for pattern 3 remain the same as in Table 6.4.

Considering first the results for pattern 2, in general, under both dropout settings, substantial bias can be observed for the intercept of the third outcome, α_3, as well as the two-way association parameters involving this outcome, i.e., $\ln\psi_{13}$ and $\ln\psi_{23}$. This seems reasonable to expect since, for pattern 2, it is this outcome that is imputed, and thus more susceptible to bias. All other parameters for this pattern exhibit

Table 6.6: *Bias and MSE of the pattern-specific parameter estimates for the trivariate Dale models fitted to the completed data using various identifying restrictions, for dropout setting 2.*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
Pattern 1						
α_1	0.0023	0.0472	0.0025	0.0474	0.0025	0.0475
β_1	0.0002	0.0740	-0.0001	0.0741	-0.0001	0.0744
α_2	0.3107	0.1081	0.2072	0.0574	-0.0436	0.0731
β_2	-0.0360	0.0195	-0.0225	0.0228	0.0103	0.1073
α_3	0.3978	0.1719	0.3409	0.1272	0.2918	0.0983
β_3	-0.0628	0.0228	-0.0619	0.0210	-0.0593	0.0259
$\ln\psi_{12}$	0.4298	0.2166	0.4070	0.2025	0.1617	0.1381
$\ln\psi_{13}$	0.3973	0.2196	0.5010	0.2942	0.4192	0.2242
$\ln\psi_{23}$	0.2726	0.1412	0.3469	0.1681	0.3282	0.1560
$\ln\psi_{123}$	0.3557	0.3809	0.2063	0.2328	0.2059	0.2291
Pattern 2						
α_1	0.0054	0.0542	0.0051	0.0545	0.0054	0.0542
β_1	0.0006	0.0839	0.0010	0.0843	0.0006	0.0837
α_2	-0.0019	0.0566	-0.0011	0.0569	-0.0017	0.0569
β_2	0.0048	0.0847	0.0034	0.0857	0.0044	0.0857
α_3	0.3239	0.1182	0.3228	0.1174	0.3223	0.1181
β_3	-0.0405	0.0262	-0.0391	0.0250	-0.0459	0.0256
$\ln\psi_{12}$	0.0284	0.0977	0.0286	0.0976	0.0287	0.0976
$\ln\psi_{13}$	0.3360	0.1665	0.3354	0.1694	0.3390	0.1657
$\ln\psi_{23}$	0.4087	0.2208	0.3986	0.2159	0.3969	0.2095
$\ln\psi_{123}$	-0.1231	0.2246	-0.1334	0.2282	-0.1467	0.2555

small bias. The MSEs, on the other hand, indicate that the intercepts α_1 and α_2 for outcomes 1 and 2, respectively, are slightly more precisely estimated than the treatment effects β_1 and β_2. For the association parameters, larger MSEs are obtained for those involving the third outcome, especially so for the three-way association $\ln\psi_{123}$. Finally, it seems that the treatment effect for the third outcome, β_3, is the most precisely estimated of all parameters, as indicated by the smaller MSE for this parameter under both settings.

For pattern 1, for both settings, it can be generally observed that the intercept parameters for the second and third outcomes, α_2 and α_3, respectively, as well as all association parameters, have fairly large bias. A particular exception can be seen for the NCMV case, under which α_2, exhibits much smaller bias compared to the CCMV and ACMV cases. On the other hand, the corresponding treatment effects for the last two outcomes, β_2 and β_3, have small bias. With respect to precision, all association parameters have relatively high MSE. In addition, the MSEs for the intercepts α_2 and α_3 are similar in magnitude, whereas those for the treatment effects β_2 and β_3 are considerably smaller. Finally, as was observed for pattern 2, the treatment effect for the third outcome, β_3, is also the most precisely estimated parameter in pattern 1.

Comparing now the results across the various identifying restrictions, for a particular parameter, all three identification schemes yield very similar values for bias and MSE for pattern 2, under both settings, as might be expected, since the three restrictions are in fact equivalent for this pattern. That is, information about the last outcome can only be borrowed from the completers' pattern 3, which is the neighboring pattern, as well as the only pattern for which such information is available. For pattern 1, however, the situation is quite different. Discrepancies in bias can be seen across the various identification schemes for a particular parameter, and these are larger in value than those observed under pattern 2. Looking further into the parameters related to the missing outcomes, under both settings, the smallest bias is observed under the NCMV case for $\alpha_2, \beta_2, \alpha_3, \beta_3$ and $\ln \psi_{12}$, while for $\ln \psi_{13}$ and $\ln \psi_{23}$, the CCMV restriction yields the smallest bias. For $\ln \psi_{123}$, it is the ACMV (NCMV) case that results in the least bias for setting 1 (2). In terms of precision, on the other hand, the parameters $\alpha_2, \beta_3, \ln \psi_{13}$ and $\ln \psi_{23}$ seem to be estimated fairly comparably across the identifying restrictions in both settings. The parameter β_2 is most precisely estimated under CCMV in setting 1, and under NCMV in setting 2, while α_3 is most precise under NCMV (both settings). Finally, $\ln \psi_{123}$ is most precise under ACMV (setting 1) and NCMV (setting 2).

To evaluate the effect of the amount of missingness on the fitting procedure, results for the two dropout settings within a particular pattern are now compared. For pattern 1, the bias and MSEs of the estimates are generally smaller under setting 2, with more missingness, i.e., more subjects within this pattern. For pattern 2, the setting with more missingness shows smaller bias for the α and β parameters, larger bias for the association parameters, and generally smaller MSEs for all parameters.

These results require careful qualification. With respect to how the results relate to the identifying restrictions, it is clear from the previous observations that no particular identifying restriction consistently led to the most precise estimates. This is understandable as the underlying parameters (Table 6.1) do not reflect any particular identification scheme. In Section 6.2.3, the performance of the various identifying restrictions are investigated further by specifically choosing underlying parameters that represent an NCMV setting. Whereas it is entirely possible to explore this situation within the context of a simulation study, the choice of which restriction leads to superior results is more difficult when dealing with actual data, since the true underlying identification pattern is usually unverifiable.

Secondly, particular attention needs to be accorded to the results insofar as the degree of incompleteness in the data is concerned. Observations seem to indicate that, within reason, the amount of missingness does not really pose additional difficulty in applying the proposed method, at least up to cases with moderate incompleteness (e.g., around 50%). Results showed that better precision is obtained when the data contains more missingness. Initially, this may seem contrary to what might be expected, since dropout setting 2 consists of more missing data, and might therefore be expected to yield worse results. However, it is also necessary to consider the proportions of subjects within each pattern. It is useful to recall that the conditional probabilities used for the imputations are initially estimated from the observed data. When a pattern has very few subjects, as would be the case when there are few dropouts and most subjects are completers, the conditional probabilities for the incomplete patterns may be poorly determined, thus yielding imputations that are probably less reliable. Hogan and Laird (1997) suggested that each pattern needs to be sufficiently "filled," requiring large numbers of dropouts. This is further validated by inspection of the results for the completers (Table 6.4), which indicate better precision for setting 1, under which pattern 3 has more subjects. Hence, in assessing the effects of the amount of missingness on the pattern-specific estimates, it is essential that the particular pattern is considered, since more missingness in the data may imply less subjects in one pattern (e.g., completers), but more subjects in the other (incomplete) patterns.

6.2.2.3 Marginalized Effects Estimates

The estimates for the marginalized effects (6.13) proposed by Jansen and Molenberghs (2007) were also computed from the pattern-specific parameter estimates and these were compared with the true marginalized parameter values of the underlying PMM (Table 6.3). The results for each identifying restriction and for the two dropout settings are summarized in Table 6.7. Although scientifically speaking, interest might really be placed on the marginalized effects estimates for the last outcome, i.e., at the end of the series, at which point the treatment presumably does or does not take its desired effect, values for the other outcomes are nevertheless presented here for a more

Table 6.7: *Bias and MSE of the marginalized parameter estimates (6.17) of Jansen and Molenberghs (2007) for the pattern-mixture model fitted to the completed data using various identifying restrictions, for both dropout settings. (The marginalized intercept and treatment effect for the* j^{th} *outcome,* $j = 1, 2, 3$*, are denoted* A_{JM_j} *and* B_{JM_j}*, respectively.)*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
Setting 1						
A_{JM_1}	0.0109	0.0088	0.0110	0.0088	0.0110	0.0088
B_{JM_1}	-0.0209	0.0163	-0.0211	0.0164	-0.0210	0.0164
A_{JM_2}	0.0486	0.0099	0.0414	0.0096	-0.0039	0.0122
B_{JM_2}	-0.0096	0.0143	-0.0101	0.0154	-0.0023	0.0209
A_{JM_3}	0.0950	0.0164	0.0891	0.0149	0.0806	0.0141
B_{JM_3}	-0.0119	0.0156	-0.0150	0.0150	-0.0123	0.0153
Setting 2						
A_{JM_1}	0.0170	0.0091	0.0170	0.0091	0.0170	0.0091
B_{JM_1}	-0.0322	0.0182	-0.0322	0.0181	-0.0323	0.0182
A_{JM_2}	0.0773	0.0128	0.0519	0.0113	-0.0117	0.0154
B_{JM_2}	-0.0150	0.0147	-0.0121	0.0177	-0.0045	0.0269
A_{JM_3}	0.1603	0.0325	0.1462	0.0281	0.1338	0.0247
B_{JM_3}	-0.0261	0.0164	-0.0259	0.0158	-0.0265	0.0171

concise evaluation of the marginalization procedure. In general, the magnitudes of the bias are small, and MSE values seem to indicate that the procedure for marginalization works pleasingly stably.

Comparing across identifying restrictions, under both settings, the bias and MSE are very similar for the marginalized parameters A_{JM_1} and B_{JM_1}, which is reasonable since this parameter relates to the first outcome, which is always observed and never imputed. For A_{JM_2} and B_{JM_2}, within both settings, although NCMV exhibits the least bias for these two parameters, their MSEs are largest under this restriction. For A_{JM_3} and B_{JM_3}, bias and MSE values are comparable across the identifying restrictions and no particular identifying restriction seems to show superiority. In addition, for setting 2, it can be observed that MSE values for A_{JM_3} are somewhat larger than those of the other parameters, implying that this parameter seems to be the least precisely estimated one. With respect to the direction of bias of a particular param-

Table 6.8: *Bias and MSE of the marginalized parameter estimates (6.16) using the direct linear approach of Park and Lee (1999) for the pattern-mixture model fitted to the completed data using various identifying restrictions, for both dropout settings. (The marginalized intercept and treatment effect for the j^{th} outcome, $j = 1, 2, 3$, are denoted A_{PL_j} and B_{PL_j}, respectively.)*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
Setting 1						
A_{PL_1}	0.0122	0.0090	0.0123	0.0090	0.0122	0.0090
B_{PL_1}	-0.0225	0.0168	-0.0227	0.0169	-0.0226	0.0169
A_{PL_2}	0.0513	0.0102	0.0438	0.0100	-0.0029	0.0126
B_{PL_2}	-0.0127	0.0147	-0.0131	0.0159	-0.0035	0.0222
A_{PL_3}	0.0995	0.0174	0.0930	0.0157	0.0840	0.0148
B_{PL_3}	-0.0169	0.0160	-0.0197	0.0153	-0.0167	0.0156
Setting 2						
A_{PL_1}	0.0180	0.0092	0.0180	0.0092	0.0180	0.0092
B_{PL_1}	-0.0333	0.0186	-0.0334	0.0185	-0.0334	0.0185
A_{PL_2}	0.0803	0.0134	0.0538	0.0116	-0.0110	0.0158
B_{PL_2}	-0.0173	0.0151	-0.0139	0.0182	-0.0052	0.0285
A_{PL_3}	0.1650	0.0341	0.1499	0.0293	0.1372	0.0257
B_{PL_3}	-0.0302	0.0168	-0.0296	0.0161	-0.0302	0.0175

eter, although generally consistent across the two settings within a given identifying restriction, the NCMV case differs from CCMV and ACMV for the parameter A_{JM_2}. Finally, it generally seems that the intercept parameters are overestimated, while the treatment effect parameters are underestimated.

Looking now across the two dropout settings, within a particular identifying restriction, bias and MSE values are smaller under setting 1, which has less missingness. Whereas the amount of missingness affects the pattern-specific estimates differently across patterns, its effects on the marginalized estimates are in line with what one would normally expect – that situations with less missing data will tend to show more accurate results.

Table 6.8 shows the bias and MSE for the marginalized effects estimates using the direct linear approach (6.11). Magnitudes of bias for almost all parameters are larger for this approach compared to those using the approximation proposed by Jansen and

Table 6.9: *Trivariate Dale model parameter values specified for each of the three dropout patterns for the underlying pattern-mixture model having an NCMV structure. (O: observed, M: missing.)*

Pattern	α_1	β_1	α_2	β_2	α_3	β_3	ψ_{12}	ψ_{13}	ψ_{23}	ψ_{123}
1 (OMM)	0.190	0.096	0.155	0.067	0.142	0.090	1.4	1.1	1.6	1.2
2 (OOM)	0.214	0.115	0.130	0.084	0.142	0.083	1.6	1.2	1.5	1.5
3 (OOO)	0.220	0.150	0.110	0.125	0.170	0.065	2.2	1.8	2.3	1.7

Molenberghs (2007) (in Table 6.7), as might be predicted, since the former is probably a less appropriate way of marginalizing the pattern-specific estimates. However, for the parameters A_{PL_2} and B_{PL_2}, under the NCMV case in both settings, the direct linear approach actually gives slightly smaller magnitudes of bias. For MSE, on the other hand, in all cases, the Jansen and Molenberghs (2007) approximation (6.13) gives more precise estimates.

6.2.3 Additional Results

In order to get deeper insight into the performance of the proposed method, additional simulations were performed, based on a new underlying PMM that was clearly of an NCMV type. It can be recalled that for the case of 3 outcomes, the identifying restrictions are equivalent for pattern 2, and therefore, the choice of parameters is dictated primarily with reference to pattern 1. Hence, in order to reflect an NCMV setting, pattern 1 parameters were chosen to be more similar to those of pattern 2 than to pattern 3. The values of these chosen parameters are shown in Table 6.9. Furthermore, an increased sample size of $N = 4000$ was used to be able to assess consistency of the previously defined estimates, as well as to be able to freely choose parameters in an NCMV way, without having to encounter samples with incomplete combination levels. With respect to missingness, the same dropout settings as in the primary simulation (Table 6.2) were considered. The underlying SeM-type marginal effects from these newly-defined PMMs are given in Table 6.10. Finally, as in the previous simulations, under these settings, $S = 500$ samples were again generated.

For conciseness, only partial results from these new simulations will be presented here. Inasmuch as the additional simulations exhibited the same results as the primary simulations did, at least as far as the effects of the different dropout settings are concerned, the presentation of results for these new simulations will be restricted to those under setting 2, having more missingness. In addition, with respect to com-

Table 6.10: *True marginal parameter values for the two underlying pattern-mixture models having an NCMV structure. (The marginal intercept and marginal treatment effect for the j^{th} outcome, $j = 1, 2, 3$, are denoted A_j and B_j, respectively.)*

PMM	Y_1		Y_2		Y_3	
	A_1	B_1	A_2	B_2	A_3	B_3
1	0.2165	0.1303	0.1163	0.1128	0.1647	0.0681
2	0.2136	0.1193	0.1212	0.1048	0.1607	0.0711

parison of the identifying restrictions, only the results for pattern 1 will be discussed, since the choice of scheme is immaterial in pattern 2, under which they are equivalent. As with the primary simulations, results will be sub-grouped into those for the initial estimates, for the pattern-specific estimates, and then for the marginalized effects estimates. Finally, asymptotic variances of the marginalized effects estimates with the variances obtained from the simulation study will also be compared.

Table 6.11 summarizes the results for the initial estimates for the NCMV-type PMM under setting 2. Bias for all parameters are quite small and the MSEs are generally smaller than those for the previous simulation (Table 6.4), demonstrating the consistency of these estimates for the pattern-specific parameters of the initial models. Somewhat larger MSEs are observed for the incomplete patterns, as these consist of fewer subjects than the group of completers (pattern 3). Finally, though the MSEs indicate fairly precise estimation of the initial parameters, it seems that the treatment effects and the associations are less accurately estimated than are the intercepts.

The results for the pattern-specific estimates after imputation of the missing values are given in Table 6.12. As expected, for any given parameter, bias and MSE are more or less equal across the identification schemes in pattern 2. Moreover, as observed under the previous simulations, parameters involving the missing outcome, namely $\alpha_3, \psi_{13}, \psi_{23}$ and ψ_{123}, exhibit higher bias and MSEs. However, the treatment effect at the last time point, β_3, is quite precisely estimated in comparison with all the other estimates. Moving on the the results for pattern 1, it can be observed that both bias and MSE are smallest under the NCMV restriction for most parameters, except β_3 and ψ_{23}, both of which were most precisely estimated under the CCMV case. For pattern 1, NCMV borrows from pattern 2, but the latter pattern contains no information about these two parameters and hence the actual information propagates from pattern 3 instead. As a result, uncertainty increases and performance worsens. The differences, however, in the bias and MSE between the NCMV and

Table 6.11: *Bias and MSE of the parameter estimates for the initial models fitted, under dropout setting 2, to the observed data: trivariate Dale model for pattern 3, bivariate Dale model for pattern 2 and logistic model for pattern 1, for the underlying pattern-mixture model having an NCMV structure.*

Parameter	Pattern 1		Pattern 2		Pattern 3	
	Bias	MSE	Bias	MSE	Bias	MSE
α_1	0.0005	0.0114	-0.0060	0.0138	0.0027	0.0031
β_1	-0.0056	0.0169	0.0150	0.0245	-0.0063	0.0077
α_2	——	——	0.0039	0.0136	-0.0008	0.0031
β_2	——	——	-0.0082	0.0227	-0.0017	0.0083
α_3	——	——	——	——	-0.0024	0.0028
β_3	——	——	——	——	0.0003	0.0068
$\ln\psi_{12}$	——	——	0.0068	0.0211	-0.0028	0.0073
$\ln\psi_{13}$	——	——	——	——	-0.0067	0.0083
$\ln\psi_{23}$	——	——	——	——	-0.0011	0.0075
$\ln\psi_{123}$	——	——	——	——	-0.0018	0.0333

CCMV case for the latter two parameters are much less pronounced than those for the other parameters. Noticeable too are the higher bias and MSEs for the association parameters in comparison with the α and β parameters, but the smallest values for these are generally still obtained under the NCMV case.

Moving on to the results for the marginalized effects estimates, which are tabulated in Table 6.13, a comparison of these results with those for setting 2 in Tables 6.7 and 6.8 is first examined. For both sets of estimates, substantially smaller MSEs are consistently observed for all parameters and for all identification schemes under this simulation using $N = 4000$. With respect to bias, under the NCMV case, all parameters exhibit smaller bias compared to the previous simulations with smaller sample size. Thus, at least under the NCMV strategy, both bias and MSE are considerably reduced for all parameters, which is compatible with these estimates' consistency. Under the CCMV and ACMV schemes, almost all parameter estimates yield smaller bias for this setting of increased sample size, with the exception of the estimates for the marginalized intercept for outcome 2, A_{JM_2} or A_{PL_2}. For this parameter, bias actually increased slightly for $N = 4000$ under CCMV and ACMV. Though this might seem contrary to intuition, as an increased sample size usually leads to a less biased estimate, the slight inflation of the bias might be attributed to the use of the incorrect identifying strategy.

Table 6.12: *Bias and MSE of the pattern-specific parameter estimates for the trivariate Dale models fitted, under dropout setting 2, to the completed data using various identifying restrictions, for the underlying pattern-mixture model having an NCMV structure.*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
Pattern 1						
α_1	-0.0000	0.0113	0.0007	0.0113	0.0007	0.0113
β_1	-0.0048	0.0169	-0.0059	0.0168	-0.0060	0.0168
α_2	0.3086	0.0984	0.2222	0.0528	-0.0239	0.0159
β_2	-0.0255	0.0055	-0.0178	0.0058	-0.0006	0.0250
α_3	0.3734	0.1424	0.3330	0.1136	0.2831	0.0832
β_3	-0.0550	0.0073	-0.0634	0.0085	-0.0669	0.0093
$\ln\psi_{12}$	0.4351	0.1979	0.3792	0.1523	0.1455	0.0450
$\ln\psi_{13}$	0.4223	0.1929	0.4653	0.2277	0.4056	0.1761
$\ln\psi_{23}$	0.2512	0.0782	0.3187	0.1136	0.3039	0.1034
$\ln\psi_{123}$	0.4761	0.2872	0.3549	0.1728	0.3380	0.1637
Pattern 2						
α_1	-0.0058	0.0137	-0.0057	0.0138	-0.0058	0.0137
β_1	0.0146	0.0243	0.0145	0.0243	0.0147	0.0243
α_2	0.0041	0.0135	0.0039	0.0135	0.0040	0.0136
β_2	-0.0085	0.0225	-0.0082	0.0224	-0.0084	0.0227
α_3	0.2857	0.0854	0.2853	0.0846	0.2882	0.0868
β_3	-0.0485	0.0084	-0.0473	0.0081	-0.0507	0.0086
$\ln\psi_{12}$	0.0068	0.0211	0.0068	0.0211	0.0068	0.0211
$\ln\psi_{13}$	0.3144	0.1122	0.3113	0.1109	0.3129	0.1108
$\ln\psi_{23}$	0.3733	0.1523	0.3726	0.1513	0.3713	0.1494
$\ln\psi_{123}$	0.1187	0.0619	0.1206	0.0642	0.1278	0.0689

In terms of the precision of the marginalized effects estimates relative to each other, it can be observed that the treatment effects generally seem to be more precisely estimated than the intercepts, for outcomes 2 and 3, under most schemes and for both sets of estimates – a result that is quite practical since interest is usually on the treatment effect rather than on the intercept, particulary so on that of the last outcome. The intercept for the last outcome, however, seems to be beset with a substantive amount of bias, and hence, a larger MSE, than the other parameters, for both the Jansen and Molenberghs (2007) and Park and Lee (1999) estimates. In comparison with the corresponding estimates under the simulations with smaller

Table 6.13: *Bias and MSE of the marginalized parameter estimates (6.17) of Jansen and Molenberghs (2007) and (6.16) of Park and Lee (1999) for the pattern-mixture model fitted, under dropout setting 2, to the completed data using various identifying restrictions, for the underlying pattern-mixture model having an NCMV structure. (The additional subscript $j, j = 1, 2, 3$, is used to identify the outcome number.)*

Parameter	CCMV		ACMV		NCMV	
	Bias	MSE	Bias	MSE	Bias	MSE
A_{JM_1}	-0.0023	0.0021	-0.0021	0.0021	-0.0021	0.0021
B_{JM_1}	0.0086	0.0044	0.0083	0.0044	0.0083	0.0044
A_{JM_2}	0.0820	0.0085	0.0606	0.0057	-0.0019	0.0036
B_{JM_2}	-0.0108	0.0042	-0.0088	0.0045	-0.0047	0.0067
A_{JM_3}	0.1423	0.0217	0.1322	0.0190	0.1203	0.0160
B_{JM_3}	-0.0182	0.0038	-0.0201	0.0040	-0.0217	0.0039
A_{PL_1}	-0.0020	0.0021	-0.0018	0.0021	-0.0018	0.0021
B_{PL_1}	0.0082	0.0044	0.0079	0.0044	0.0080	0.0044
A_{PL_2}	0.0839	0.0088	0.0616	0.0059	-0.0017	0.0036
B_{PL_2}	-0.0116	0.0042	-0.0095	0.0046	-0.0051	0.0068
A_{PL_3}	0.1448	0.0225	0.1343	0.0195	0.1221	0.0164
B_{PL_3}	-0.0193	0.0039	-0.0212	0.0041	-0.0228	0.0040

sample size, bias, though still substantial, has already improved in this setting with the use of a larger sample. However, since this improvement is not much, there would be no reason to believe that it will decrease much further under an even larger sample size. This might be a reflection of the earlier qualification that the marginalized effects estimates may be biased for the SeM-type marginal parameters, particularly when the unconditional pattern proportions are not equal to the conditional ones. On the one hand, one might try to improve the bias by using the conditional pattern probabilities as weights in the proposed estimates (6.17) of Jansen and Molenberghs (2007). In contrast, if focus does not lie on the marginalized intercept but on the marginalized treatment effect, one may proceed to use the estimates as they are, since the above results seem to indicate that the treatment effects are fairly well estimated by the proposed procedure. Finally, with respect to comparisons across the two sets of estimates, both bias and MSE are slightly higher for the approximation by Park and Lee (1999).

Finally, to evaluate the efficiency of the marginalized effects estimates, asymp-

Table 6.14: *Asymptotic variance (Asy) and simulation variance (Sim) of the marginalized parameter estimates (6.17) of Jansen and Molenberghs (2007) and (6.16) of Park and Lee (1999) for the pattern-mixture model fitted, under dropout setting 2, to the completed data using various identifying restrictions, for the underlying pattern-mixture model having an NCMV structure. (The additional subscript* $j, j = 1, 2, 3,$ *is used to identify the outcome number.)*

Parameter	CCMV		ACMV		NCMV	
	Asy	Sim	Asy	Sim	Asy	Sim
A_{JM_1}	0.0022	0.0021	0.0021	0.0021	0.0022	0.0021
B_{JM_1}	0.0043	0.0043	0.0043	0.0043	0.0043	0.0043
A_{JM_2}	0.0016	0.0018	0.0016	0.0021	0.0024	0.0036
B_{JM_2}	0.0037	0.0041	0.0037	0.0044	0.0050	0.0067
A_{JM_3}	0.0012	0.0015	0.0012	0.0015	0.0012	0.0015
B_{JM_3}	0.0026	0.0035	0.0026	0.0036	0.0026	0.0035
A_{PL_1}	0.0022	0.0021	0.0022	0.0021	0.0022	0.0021
B_{PL_1}	0.0043	0.0044	0.0043	0.0043	0.0043	0.0043
A_{PL_2}	0.0016	0.0018	0.0016	0.0021	0.0024	0.0036
B_{PL_2}	0.0037	0.0041	0.0037	0.0045	0.0050	0.0068
A_{PL_3}	0.0012	0.0015	0.0012	0.0015	0.0012	0.0014
B_{PL_3}	0.0026	0.0035	0.0026	0.0036	0.0026	0.0034

totic variances, as given in (6.18) and (6.19) were computed using the underlying parameter values. Variances and/or covariances of the pattern-specific estimates are obtained from the simulation runs, while variances and/or covariances of the pattern proportions are obtained using standard results for the multinomial and/or binomial distributions. For instance, $Var(\widehat{\pi}_t) = \pi_t(1-\pi_t)/n$. This was then compared with the simulation variance for the corresponding marginalized effects estimate. The results under each identification scheme are shown in Table 6.14. The very small differences, generally of order $(1 \times e^{-04})$, between the asymptotic and simulation variance demonstrate efficiency of the marginalized effects estimates.

6.3 Discussion

In this chapter, under simulated settings, a procedure for fitting pattern-mixture models to categorical data with monotone missingness via the use of identifying restrictions was applied. Asymptotic variances for marginalized effects estimates, as proposed by Jansen and Molenberghs (2007) and Park and Lee (1999), were derived and the performance of these marginalized effects was subsequently investigated.

It was observed that although the (maximum likelihood) estimates for the parameters of the pattern-specific initial models were consistent, precision was contingent on the amount of missingness within the pattern. This sparseness in some patterns may have further consequences on the resulting imputations required for the method, since within these patterns, conditional probabilities on which the imputations are based will probably be less reliable. In such situations, identifying strategies that are based on the more amply filled patterns may prove to be more effective. Similarly, the pattern-specific estimates of the trivariate Dale model parameters were also influenced by the amount of missing data in the particular pattern. However, even for the least filled pattern, the precision for the main parameters (e.g., intercepts and treatment effects) seemed to be quite reasonable. The association parameters, on the other hand, were generally seen to be poorly estimated, but as these are usually regarded as nuisance parameters, such a result is not too alarming. Moreover, these were specified in a simplistic manner – as constants – in the underlying Dale models, but one could always reformulate the Dale model in such a way that these associations are more meaningfully modeled in terms of some known covariates, thereby possibly improving their estimation. Interestingly, the results of the simulations showed that the treatment effect at the last time point, on which scientific interest usually lies, was the most precisely estimated parameter, under any dropout setting or any identification scheme.

With respect to the marginalized effects estimates, missingness again played a fairly important role, with decreased precision in the case of more incomplete data. In addition, the direct linear approach by Park and Lee (1999) yielded slightly more bias and less precision than the proposed estimates of Jansen and Molenberghs (2007). Yet again the primary parameter of interest, i.e., the marginalized treatment effect at the last outcome, was seen to be remarkably stable. The additional simulations allowed a more objective assessment of the behavior of these marginalized effects estimates in terms of the various identifying restrictions. The most favorable results were observed under the restriction that was consistent with the underlying identification scheme, demonstrating that the proposed method and its accompanying marginalization actually can be considered successful, provided the appropriate identification is used. However, the use of an identification scheme that is inconsistent with the underlying one could induce bias in some of the less important parameters (e.g., intercept at the last time point). This, of course, poses a more difficult issue to deal with, in the sense that the underlying identification pattern is almost always unverifiable. At best, the bias might be reduced by using the conditional pattern proportions in the estimation of the marginalized effects, as opposed to the unconditional ones, whenever obtaining SeM-type marginal effects from a fitted PMM is the goal.

The assumption of independence of the pattern-specific estimates across patterns raises a worthwhile point of contention. It would be natural to expect that employing the CCMV restriction, for instance, would induce some degree of dependence between the pattern-specific estimates of the incomplete patterns and those of the completers, thereby contradicting the said assumption. Though drawing from the parameters' posterior distribution, separately for each multiple imputation, and only subsequently generating imputations – in accordance with general MI theory (Little and Rubin, 1987) – might alleviate this problem, one might be inclined to believe that dependencies in the pattern-specific estimates across patterns still remain. In such cases, variance estimates for the marginalized parameters, as derived by the delta method, can nevertheless be adjusted to account for such dependencies. Formula (C.5) in Appendix C would still hold and would reduce to a more general formula than (C.6), and also (C.10), in which pertinent terms, which would not zero out, would have to be incorporated accordingly.

Regarding the choice of simulation settings, two points deserve mention here. The first is with regard to the sample size, $N = 1000$, which is undoubtedly large for a typical clinical trial. Although not quite realistic except for very large clinical trials, such a sample size is not unusual for survey-type data, in which case the outcome vector would be a truly multivariate one, rather than a single outcome measured longitudinally. The applicability therefore extends beyond longitudinal data. As was already pointed out earlier, the large sample size was necessary to ensure that all combinations were present within each pattern, to allow fitting of the initial and final models; this is a feature, and sometimes a drawback, of the PMM framework, which is indeed a pragmatic concern. In the simulation study, though this issue was circumvented through the use of a large sample size, one ought not forget that for real data analysis settings, sparsely filled or even empty levels are bound to occur. It is thus relevant to address rather than circumvent this issue. A second point that arises has to do with the type of missingness: monotone. One might argue that such type of missingness is more difficult to define for a multivariate outcome vector (e.g., survey data) than in the longitudinal case, where one can rely on the natural time-ordering of the measurement sequence. However, when the multivariate outcomes within the vector can somehow be ranked, e.g., by order of importance of the variables in the survey, then defining monotone missingness is sometimes an option.

Jansen and Molenberghs (2007) pointed out that the final analysis model can, but does not have to be, equal to the initial model fitted to the observed data. That is, although one fits a pattern-mixture model to the observed data at the first stage, it is entirely possible to fit a selection model after the identification/imputation stage, i.e., after the data have been completed. This, however, raises the issue of "proper" imputation, which Rubin (1987) defines as one where the analysis model is in agreement with the imputation model. This means, broadly, that the imputation model should ideally be a super-model of the analysis model. Fitting a selection model at the final stage would therefore pose a conflict in this regard. Thus, to avoid so-called improper imputation, final models considered here were of the pattern-mixture type. The nature of the scientific question, however, would, of course, be an important consideration as well. For instance, if marginal effects are of primary interest, then a selection model would be the natural choice for final analysis model. Hence, the final choice would have to be a balance between the desire for proper imputation and the nature of the scientific question.

Finally, regarding application of the procedure to the non-monotone case, several concerns arise. The first issue has to do with parsimony. In the case of non-monotone missingness, a lot more patterns arise, leading to a proliferation of parameters. In this study, which considered longitudinal sequences of three time points, only three patterns arise in the monotone case, each employing 10 parameters, and this is only assuming a very simple structure such as, for example, a single treatment indicator and/or association parameters that are not allowed to vary over covariate levels. For the non-monotone case, there will be 8 patterns, and, under the same simple structure considered here, the PMM will consist of a full set of 80 parameters! It is easy to see that increasing the number of time points, even just slightly, will necessitate estimation of an even much larger numbers of parameters. In addition, these patterns should be more or less sufficiently filled in order to get any useful information from them. And, unless the study employs a very large sample size, it is quite difficult to foresee such a situation. In line with this, complete combinations would also be necessary within each pattern, to fit the proposed models without any computational challenges, again requiring large numbers of subjects. On a somewhat different note, difficulties also arise in considering which outcome to impute first. Jansen and Molenberghs (2007) discussed that the formulations they presented can be seen merely as a few points in a vast, continuous, design space. To compound the issue, in the non-monotone setting, there are no explicit expressions for ACMV, which might be the more popular choice of identifying restriction as it is the equivalent of MAR a pattern-mixture framework (Molenberghs *et al.*, 1998). For these reasons, focus and discussion of the non-monotone case has been deferred, perhaps to a separate study on its own, so that such issues can be adequately dealt with and properly discussed therein.

7

Influence Analysis as a Means of Sensitivity

In a broad sense, sensitivity analysis can be defined as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools, such as diagnostic measures. Such an informal definition encompasses a wide variety of useful approaches. Among the simplest of procedures would be to fit a selected number of (non-random) models, which are all deemed plausible or in which a preferred (primary) analysis is supplemented with a number of variations. The extent to which conclusions (inferences) are stable across such a range provides an indication about the degree to which they are robust to inherently untestable assumptions about the missingness mechanism. Variations to a basic model can be constructed in different ways. The most obvious strategy, cast within the selection model paradigm, is to consider various dependencies of the missing data process on the outcomes and/or the covariates. Alternatively, the distributional assumptions of the models can be changed (as will be illustrated in Section 7.3.1.3). Related to this, one can assess how an MNAR model, or a collection of MNAR models, differ from the set of models with equal fit to the observed data but that are of a MAR nature (as was seen in Chapter 4). Additionally, sensitivity analysis can also be performed at the level of the individual observations rather than at the level of the models. In such cases, interest is directed towards finding those individuals that drive the conclusions towards one or more MNAR models, thereby warranting exploration of the influence of every individual separately. Two techniques exist: namely, global influence and local influence (Cook, 1986). The global influence methodology, also known as the case-deletion approach (Cook and Weisberg, 1986), was introduced by Cook (1979, 1986) in linear regression. Thijs, Molenberghs and Verbeke (2000), Verbeke *et al.* (2001) and Molenberghs *et al.* (2003) applied the local influence approach within the context of linear mixed models. Later, Van Steen *et al.* (2001) adapted these ideas to the model of Molenberghs, Kenward and Lesaffre (1997), for monotone repeated ordinal data. Beunckens *et al.* (2009) integrated global and local influence techniques within a comprehensive sensitivity analysis of the Slovenian Public Opinion Survey.

In this chapter, various tools for sensitivity analysis are examined. Of particular interest is the use of influence measures, both on a global as well as on a local scale, as means by which to assess sensitivity. Both existing and proposed methods, as well as ideas, are developed using the SPO Survey data – in particular, on responses arising from the independence and attendance questions – via a comprehensive sensitivity analysis. Parallel to a review of simple analyses of the SPO Survey, three main strands of sensitivity analysis are considered, i.e., the intervals of ignorance, influence analysis and the computation of a so-called MAR counterpart to the model considered (Chapter 4), thereby bringing together and contrasting existing sensitivity assessments with new, local-influence-based analyses that have never been applied to the SPO Survey data. Moreover, the local influence technology based on cell counts rather than parameters is new, as well as the approach of perturbing the cell probabilities rather than the model parameters. A subsection is devoted to the integration of all these analyses and their implications on the the primary estimand, i.e., the proportion of people voting for independence, further demonstrating the value of consolidated sensitivity analysis.

7.1 Global Influence

The idea of global influence originates from the concept of case deletion, whereby the so-called influence of a single subject is assessed by examining the resulting model differences when the particular subject is excluded from the analysis. With $\boldsymbol{\phi}$ denoting the unknown parameter vector of interest, and $\ell_i(\boldsymbol{\phi})$ denoting the log-likelihood contribution of individual i, the log-likelihood for the entire data set can be obtained as:

$$\ell(\boldsymbol{\phi}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\phi}).$$

Global influence approaches compare the log-likelihoods of the model of interest fitted to the entire data set, $\ell(\boldsymbol{\phi})$, on the one hand, with that excluding or doubling one subject, $\ell_{(\pm i)}(\boldsymbol{\phi})$, on the other. Cook's distances (CD) are based on measuring the discrepancy, induced by deletion or doubling of a single case, in either the log-likelihood

$$CD_{1i} = 2\left[\widehat{\ell}(\boldsymbol{\phi}) - \widehat{\ell}_{(\pm i)}(\boldsymbol{\phi})\right], \tag{7.1}$$

or in the parameter vector

$$CD_{2i} = 2(\widehat{\boldsymbol{\phi}} - \widehat{\boldsymbol{\phi}}_{(\pm i)})'\, \ddot{L}^{-1}\, (\widehat{\boldsymbol{\phi}} - \widehat{\boldsymbol{\phi}}_{(\pm i)}), \tag{7.2}$$

with $\ddot{L}$ the matrix of second derivatives of $\ell(\boldsymbol{\phi})$, with respect to $\boldsymbol{\phi}$, evaluated at $\widehat{\boldsymbol{\phi}}$. Both measures can be constructed for the entire parameter vector or for subvectors thereof; this includes, of course, a single parameter. The quantities CD_{1i} and CD_{2i} respectively represent the amount of displacement either in the log-likelihood or in the parameter values themselves, and larger values would be indicative of some amount of influence of the deleted or doubled case. Performing a global influence analysis on data with categorical outcomes is less time consuming than on data with continuous outcomes, since the data can then be organized into cells and case deletion (or doubling) need only be done for one case within each cell, rather than for all cases within that cell.

7.2 Local Influence

One of the drawbacks of global influence is that the specific cause of the influence is hard to retrieve because, by deleting or adding a subject, all types of influence stemming from it are lumped together. Local influence (Cook, 1986; Verbeke *et al.*, 2001) is presumably more suitable in this respect, since influence, in the form of a perturbation, is introduced into a specific area in the model. The method also often leads to closed forms.

Denote the log-likelihood corresponding to a particular model by

$$\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\phi}|\omega_i), \tag{7.3}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\psi})'$ is the s-dimensional vector, grouping, respectively, the parameters of the measurement and dropout models. The vector $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)'$, belonging to an open subset Ω of $\mathbb{R}^N$, denotes a vector defining infinitesimal perturbations around the model studied. Obviously, the original model would correspond to $\boldsymbol{\omega}_o = (0, 0, \ldots, 0)'$.

Let $\widehat{\boldsymbol{\phi}}$ be the maximum likelihood estimator for $\boldsymbol{\phi}$, obtained by maximizing $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$, and let $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ denote the maximum likelihood estimator for $\boldsymbol{\phi}$ under $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$. The local influence approach compares $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ with $\widehat{\boldsymbol{\phi}}$, and similar values would indicate that the parameter estimates are robust with respect to perturbations in the direction of the extended model. Cook (1986) quantified differences between $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ with $\widehat{\boldsymbol{\phi}}$ through the likelihood displacement:

$$LD(\boldsymbol{\omega}) = 2\,\left[\,\ell(\widehat{\boldsymbol{\phi}}|\boldsymbol{\omega}_o) - \ell(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}|\boldsymbol{\omega}_o)\,\right],$$

which takes into account the variability of $\widehat{\boldsymbol{\phi}}$. LD($\boldsymbol{\omega}$) will be large if $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$ is strongly curved at $\widehat{\boldsymbol{\phi}}$, implying that $\boldsymbol{\phi}$ is estimated with high precision. A graph of $LD(\boldsymbol{\omega})$ against $\boldsymbol{\omega}$, i.e., the geometric surface formed by values of the $(N+1)$-dimensional vector $\zeta(\boldsymbol{\omega}) = (\boldsymbol{\omega}', LD(\boldsymbol{\omega}))'$, depicts the influence of perturbations. Because this so-called *influence graph* (Lesaffre and Verbeke, 1998) can only be depicted when $N = 2$, Cook (1986) proposed considering local influence, i.e., the normal curvatures $C_{\boldsymbol{h}}$ of $\zeta(\boldsymbol{\omega})$ in $\boldsymbol{\omega}_o$, in the direction of some N-dimensional vector $\boldsymbol{h}$ of unit length. A general expression is given by

$$C_{\boldsymbol{h}} = 2\left|\boldsymbol{h}'\,\boldsymbol{\Delta}'\,(\ddot{L})^{-1}\,\boldsymbol{\Delta}\,\boldsymbol{h}\right|, \tag{7.4}$$

where $\ddot{L}$ is the $(s \times s)$ matrix of second-order derivatives of $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$ with respect to $\boldsymbol{\phi}$, evaluated at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$, i.e., for $i, \ell = 1, 2, \ldots, s$, the elements of $\ddot{L}$ are:

$$[\ddot{L}]_{i\ell} \quad = \quad \left.\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)}{\partial\phi_i \partial\phi_\ell}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}},$$

and $\boldsymbol{\Delta}$ is the $(s \times N)$ matrix having, as its i^{th} column, the s-dimensional vector

$$\boldsymbol{\Delta}_i = \left.\frac{\partial^2 \ell_i(\boldsymbol{\phi}|\omega_i)}{\partial\omega_i \partial\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}},\,\omega_i=0}.$$

A sensible choice for $\boldsymbol{h}$ is the vector with a one in the i^{th} position and zero elsewhere, giving rise to say $\boldsymbol{h}_i$ and C_i, corresponding to the perturbation of subject i only, or in the case of contingency tables, perturbation of cell i only. Another important direction is the direction $\boldsymbol{h}_{\max}$ of maximal normal curvature $C_{\max}$. It shows how to perturb the model to obtain the largest local changes in the likelihood displacement. Details can be found in Verbeke and Molenberghs (2000).

The above development is geared towards studying the influence on the likelihood function. Other choices are possible as well. One might assess, for instance, the influence of perturbations on a particular function(s) of the parameters, rather than on the log-likelihood itself. In such cases, letting $Z(\boldsymbol{\phi})$ denote some generic function of the model parameters, local influence of perturbations around the posited null model can be evaluated in terms of the difference between $Z(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})$ and $Z(\widehat{\boldsymbol{\phi}})$. Analogous to Cook (1986), the following generalized expression for $C_{\boldsymbol{h}}$ is proposed:

$$C_{\boldsymbol{h}} = 2 \left| \boldsymbol{h}' \, \boldsymbol{\Delta}' \, (\ddot{L})^{-1} \, \ddot{Z} \, (\ddot{L})^{-1} \, \boldsymbol{\Delta} \, \boldsymbol{h} \right|, \tag{7.5}$$

with $\boldsymbol{\Delta}$ and $\ddot{L}$ as previously defined, $\|\boldsymbol{h}\| = 1$, and $\ddot{Z}$ is the $(s \times s)$ matrix of second-order derivatives of $Z(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$, evaluated at $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}$, i.e., for $i, \ell = 1, 2, \ldots, s$, the elements of $\ddot{Z}$ are given by

$$[\ddot{Z}]_{il} \quad = \quad \left. \frac{\partial^2 Z(\boldsymbol{\phi})}{\partial \phi_i \partial \phi_\ell} \right|_{\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}} .$$

It can be easily seen that expression (7.5) reduces to (7.4) when the function of interest, $Z(\boldsymbol{\phi})$, is the log-likelihood $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$ itself. Moreover, whereas (7.4) quantifies influence in terms of the displacement in the log-likelihood, (7.5) describes influence through the displacement in the particular function of interest, $Z(\boldsymbol{\phi})$.

It is worthwhile to note that the local influence measures described here are not calibrated, because both (7.4) and (7.5) take the form of a squared second derivative, owing to the double occurrence of $\boldsymbol{\Delta}$, ‘divided by’ another second derivative, $\ddot{L}$. Changing units, therefore, changes scale, unlike in the mixed-models application of Lesaffre and Verbeke (1998) and Verbeke and Molenberghs (2000), where the influence measures approximately sum to twice the sample size. As a consequence, when applying local influence as presented here, interpretation ought to be relative rather than absolute. The important issue remains then as to ‘how large is large?’ It is extremely hard to provide firm guidelines, but it may be wise, as one possible rule of thumb, to scrutinize the subjects with the largest 5% of influence values. The issue has been studied in detail in Jansen *et al.* (2006).

Local influence is useful when assessing which (groups of) observations are most influential in driving the conclusions about the nature of the missing data mechanism in the direction of the more elaborate MNAR model. As Jansen *et al.* (2006) indicate, such a phenomenon should not be seen as evidence, let alone proof, that some observations are genuinely influenced by a complex MNAR mechanism rather than, for example, by a simpler MAR mechanism. Indeed, this would conflict with the MAR-counterpart results (Chapter 4). Rather, such influence graphs are instructive when assessing which observations have the power to drive the conclusions towards a more complex mechanism. Often, the issue is that other outlying features, such as

Table 7.1: *Theoretical distribution of the probability mass over the full and observed cells, respectively, for a bivariate binary outcome with missingness in none, one, or both responses.*

(a) Complete cells

$\pi_{11,11}$	$\pi_{11,12}$
$\pi_{11,21}$	$\pi_{11,22}$

$\pi_{10,11}$	$\pi_{10,12}$
$\pi_{10,21}$	$\pi_{10,22}$

$\pi_{01,11}$	$\pi_{01,12}$
$\pi_{01,21}$	$\pi_{01,22}$

$\pi_{00,11}$	$\pi_{00,12}$
$\pi_{00,21}$	$\pi_{00,22}$

(b) Observed cells

$\pi_{11,11}$	$\pi_{11,12}$
$\pi_{11,21}$	$\pi_{11,22}$

$\pi_{10,1+}$
$\pi_{10,2+}$

$\pi_{01,+1}$	$\pi_{01,+2}$

$\pi_{00,++}$

unusual values or unusual slopes in longitudinal observations, are responsible for the apparent conclusion about the missing data mechanism. Like the MAR-counterpart result, this type of sensitivity analysis issues a cautionary warning against excessive confidence regarding the nature of the missing data mechanism. A limiting feature, unsurprising in view of the foregoing discussion, is the absence of a 'yardstick,' or a threshold, demarcating influential subjects; arguably, influence graphs will blow a whistle over subjects that need further scrutiny.

7.3 Sensitivity Analysis of the Slovenian Public Opinion Survey

For the ensuing analyses, focus lies on the binary outcomes representing the responses for the independence and attendance questions of the SPO Survey. The general expression for the cell probabilities can be expressed as

$$\pi_{r_1 r_2, j_1 j_2} = P(R_1 = r_1, R_2 = r_2, J_1 = j_1, J_2 = j_2),$$

where R_1 and R_2 denote the missingness indicators for the attendance and independence questions, respectively, while J_1 and J_2 denote the actual responses for the same. Further, $r_1 = 0$ (or 1) if the answer to the attendance question is missing (or observed) and $j_1 = 1$ (or 2) if the answer to the attendance question is YES (or NO). The indices r_2 and j_2 are similarly defined for the independence question.

The four-way joint probabilities of the two missingness indicators and the two binary responses gives rise to a total of 16 theoretical *full* cell probabilities, which are depicted in Table 7.1a and which produce 15 full data degrees of freedom. The

Table 7.2: *The Slovenian Public Opinion Survey. Observed cells, collapsed over the secession question. (A simplified cell indexing system is shown in relation to the original cell indexing system. The columns refer to 'independence' whereas the rows refer to 'attendance.')*

Cell 1 $Z_{11,11} = 1439$	Cell 2 $Z_{11,12} = 78$
Cell 3 $Z_{11,21} = 16$	Cell 4 $Z_{11,22} = 16$

Cell 5 $Z_{10,1+} = 159$
Cell 6 $Z_{10,2+} = 32$

Cell 7 $Z_{01,+1} = 144$	Cell 8 $Z_{01,+2} = 54$

Cell 9 $Z_{00,++} = 136$

probabilities for the 9 *observed* cells are shown in Table 7.1b, in which the four tables represent, respectively, the cell probabilities for the bivariate binary outcome with missingness in none, the second, the first, or both responses. The cell counts, instead of probabilities, for the 9 observed cells for the SPO Survey data are shown in Table 7.2, which also presents the cell indexing system that will be used for the remainder of this chapter.

The parameter of interest, θ, i.e., the proportion of people answering YES to *both* the independence and attendance questions, is the marginal probability that the two responses are both equal to 1 (marginalized over the missingness indicators), and can be expressed as

$$\theta = \sum_{r_1=0}^{1} \sum_{r_2=0}^{1} \pi_{r_1 r_2,11} \;=\; \sum_{r_1=0}^{1} \sum_{r_2=0}^{1} P(R_1 = r_1, R_2 = r_2, J_1 = 1, J_2 = 1).$$

It is clear from the above expression that when the data are complete, evaluation of θ would simply entail summing the upper-left cells from each the four (2×2) tables in Table 7.1a. However, when the data are incomplete, as in Table 7.1b, the relation is not straightforward as the collapsed cells have to be split in order to obtain values for the cells pertinent to the estimation of θ.

7.3.1 Review of Existing Analyses

In this section, an overview of results from previous analyses of the SPO Survey is presented, in order to complement and compare these with new approaches considered here, thereby rendering a more comprehensive and integrated sensitivity analysis.

7.3.1.1 Simple Analyses

The SPO Survey data were used by Molenberghs, Kenward and Goetghebeur (2001) to illustrate their proposed sensitivity analysis tool – the interval of ignorance. Molenberghs *et al.* (2008) used the data to exemplify results about the relationship between

Table 7.3: *The Slovenian Public Opinion Survey. Some estimates of the proportion θ attending the plebiscite and voting for independence, as presented in Rubin, Stern and Vehovar (1995) and Molenberghs, Kenward and Goetghebeur (2001), as well as sensitivity analysis results.*

Estimation method	Voting in favor of independence: $\widehat{\theta}$
Standard analyses	
Non-parametric bounds	[0.694;0.905]
Complete cases	0.928
Available cases	0.929
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782
Sensitivity analyses	
BRD1–BRD9	[0.741;0.892]
Well-fitting BRD6–BRD9	[0.741;0.867]
BRD1(MAR)–BRD9(MAR)	[0.892;0.892]
Interval of Ignorance: Model 10	[0.762;0.893]
Interval of Ignorance: Model 11	[0.766;0.883]
Interval of Ignorance: Model 12	[0.694;0.905]
Plebiscite	0.885

MAR and MNAR models (Section 4.3). An overview of various analyses can be found in Molenberghs and Kenward (2007). These authors used the models proposed by Baker, Rosenberger and DerSimonian (1992) for the setting of two-way contingency tables, subject to missingness in either none, one, or both responses. Rubin, Stern and Vehovar (1995) conducted several analyses of the data. Their main emphasis was on determining the proportion θ of the population that would attend the plebiscite and vote for independence. Their estimates are reproduced in Table 7.3.

The pessimistic (optimistic) bounds, or non-parametric bounds, are obtained by setting all incomplete data that can be considered a YES (NO), as YES (NO). The complete case estimate for θ is based on the subjects answering all three questions and the available case estimate is based on the subjects answering the two questions of interest here. It is noteworthy that both of these estimates are out of bounds. This is not a mistake: it should be recalled that the political decision was made to treat a NO on the attendance question as an effective NO vote in the plebiscite. Disregarding incomplete cases ignores this aspect and thus discards available information, thereby causing the estimate to exceed the bounds. Note that the bounds apply only to those estimators making use of all available data, not to estimators based on subsets.

This underscores the growing conviction that such estimates should be disregarded in favor of more routine use of MAR (Molenberghs and Kenward, 2007). Nevertheless, it may sometimes be worthwhile to consider the complete case (CC) analysis as simply one valid under MCAR. Ideally, it should be set against the background of bounds, as is done here, or at least against a number of competing models. Rubin, Stern and Vehovar (1995) considered two MAR models, also reported in Table 7.3, with the first one based solely on the two questions of direct interest and the second one using all three. Finally, they considered a single MNAR model, based on the assumption that missingness on a question depends on the answer to that question but not on the other questions. Rubin, Stern and Vehovar (1995) concluded, owing to the proximity of the MAR analysis to the plebiscite value, that MAR in this and similar cases may be considered a plausible assumption. As argued before (Kenward, Goetghebeur and Molenberghs, 2001), one has to be careful with this conclusion, however. Arguments to support this position will be provided in the next section.

7.3.1.2 Missing At Random Counterpart

Molenberghs, Kenward and Goetghebeur (2001) fitted the BRD models (Section 3.7.3) to the independence and attendance questions of the SPO Survey. These were supplemented by Molenberghs *et al.* (2008) with estimates obtained by fitting the corresponding MAR counterpart model to a particular (MNAR) BRD model (Section 4.3). Table 7.4 summarizes the results. It can be recalled from Section 4.3 that the lower-numbered models BRD1-BRD5 have poor fit, which might be seen as evidence to disregard the MAR-based estimate of $\widehat{\theta} = 0.892$ from BRD1. Inspection, however, of the MAR values, $\widehat{\theta}_{\text{MAR}}$, from each of the counterpart models BRD1(MAR)–BRD9(MAR) reveals a remarkably stable estimate. Hence, a value of $\widehat{\theta} = 0.892$, based on the four counterparts BRD6(MAR)–BRD9(MAR) of the well-fitting models BRD6-BRD9, seems to be a sensible choice. Thus, considering the MAR counterparts in this particular example, provides solid basis for the MAR-based estimate, which, without additional insight from the MAR counterpart models, might have been discarded on the grounds of poor model fit.

Table 7.4: *The Slovenian Public Opinion Survey, analysis restricted to the independence and attendance questions. Summaries on each of the models BRD1–BRD9, with obvious column labels. The last column, labeled* $\widehat{\theta}_{MAR}$*, refers to the estimate from the MAR counterpart model corresponding to the given one.*

Model	Structure	d.f.	loglik	$\widehat{\theta}$	C.I.	$\widehat{\theta}_{\mathrm{MAR}}$
BRD1	$(\alpha_{..},\beta_{..})$	6	-2495.29	0.892	[0.878;0.906]	0.8920
BRD2	$(\alpha_{..},\beta_{j_1.})$	7	-2467.43	0.884	[0.869;0.900]	0.8915
BRD3	$(\alpha_{.j_2},\beta_{..})$	7	-2463.10	0.881	[0.866;0.897]	0.8915
BRD4	$(\alpha_{..},\beta_{.j_2})$	7	-2467.43	0.765	[0.674;0.856]	0.8915
BRD5	$(\alpha_{j_1.},\beta_{..})$	7	-2463.10	0.844	[0.806;0.882]	0.8915
BRD6	$(\alpha_{j_1.},\beta_{j_1.})$	8	-2431.06	0.819	[0.788;0.849]	0.8919
BRD7	$(\alpha_{.j_2},\beta_{.j_2})$	8	-2431.06	0.764	[0.697;0.832]	0.8919
BRD8	$(\alpha_{j_1.},\beta_{.j_2})$	8	-2431.06	0.741	[0.657;0.826]	0.8919
BRD9	$(\alpha_{.j_2},\beta_{j_1.})$	8	-2431.06	0.867	[0.851;0.884]	0.8919

7.3.1.3 Intervals of Ignorance

A sample from Table 7.1 produces empirical proportions representing the πs with error. This imprecision disappears asymptotically. What remains is ignorance regarding the redistribution of all but the first four πs over the missing values. This leaves ignorance regarding any probability in which at least one of the first or second indices is equal to 0, and hence regarding any derived parameter of scientific interest. For such a parameter, say θ, a region of possible values which is consistent with Table 7.1 is called a region of ignorance. Evidently, such a region will depend, not only on the data and the way it is incomplete, but also on the model for which it is constructed. Analogously, an observed incomplete table leaves ignorance regarding the would-be observed full table, which leaves imprecision regarding the true full probabilities. The region of estimators for θ that is consistent with the observed data provides an estimated region of ignorance. For a single parameter, the region becomes the *interval of ignorance*. Various ways of constructing regions of ignorance are conceivable. Typically, one selects the largest possible set of identifiable parameters. The remaining ones are then termed sensitivity parameters. For every value chosen for the latter, the former can be estimated by means of, for example, maximum likelihood. Repeating this for all values of the sensitivity parameters or, practically speaking, a sufficiently

Table 7.5: *The Slovenian Public Opinion Survey. Intervals of ignorance* (II) *and intervals of uncertainty* (IU) *for the proportion θ (confidence interval) of YES votes, following from fitting overspecified Models 10, 11 and 12.*

Model	d.f.	loglik	$\widehat{\theta}$	
			II	IU
10	9	-2431.06	[0.762;0.893]	[0.744;0.907]
11	9	-2431.06	[0.766;0.883]	[0.715;0.920]
12	10	-2431.06	[0.694;0.905]	

refined grid, one effectively obtains a region or, in the univariate case, an interval of estimates. The $(1-\alpha)100\%$ *region of uncertainty* is a larger region, encompassing the region of ignorance, in the spirit of a confidence region, designed to capture the combined effects of imprecision and ignorance. In essence, this amounts to constructing, for every point in the region of ignorance, a confidence region, the union of which then produces the interval of uncertainty. Details regarding construction and asymptotic properties can be found in Molenberghs, Kenward and Goetghebeur (2001), Kenward, Goetghebeur and Molenberghs (2001), and Vansteelandt *et al.* (2006).

In addition to the 9 BRD models, Molenberghs, Kenward and Goetghebeur (2001) considered 3 models on which they derived intervals of ignorance and intervals of uncertainty. Model 10 is defined as $(\alpha_{.j_2}, \beta_{j_1 j_2})$ with

$$\beta_{j_1 j_2} = \beta_0 + \beta_{j_1 .} + \beta_{.j_2}, \tag{7.6}$$

while Model 11 assumes $(\alpha_{j_1 j_2}, \beta_{j_1 .})$ and uses

$$\alpha_{j_1 j_2} = \alpha_0 + \alpha_{j_1 .} + \alpha_{.j_2}. \tag{7.7}$$

Finally, Model 12 is defined as $(\alpha_{j_1 j_2}, \beta_{j_1 j_2})$, a combination of both (7.6) and (7.7). The resulting intervals of ignorance and intervals of uncertainty for each of these models are shown in Table 7.5, while a graphical representation the YES votes is given in Figure 7.1, along with the results in Tables 7.3 and 7.4. Model 10 shows an interval of ignorance which is very close to $[0.741, 0.892]$, the range produced by the models BRD1–BRD9, while Model 11 is somewhat sharper and just fails to cover the plebiscite value. However, the corresponding intervals of uncertainty contain the true value.

Interestingly, Model 12 virtually coincides with the non-parametric range even though it does not saturate the complete data degrees of freedom. To do so, not 2

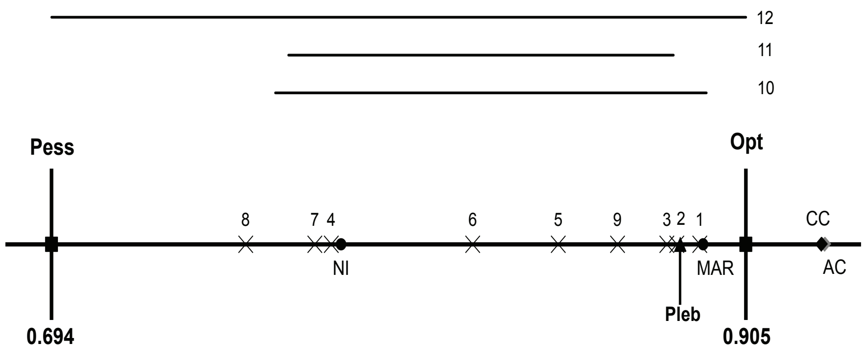


Figure 7.1: *The Slovenian Public Opinion Survey. Relative positions of the estimates of the proportion of YES votes, based on the BRD Models, on the nonparametric pessimistic-optimistic bounds, and on the models considered in Rubin, Stern and Vehovar (1995) and in Molenberghs, Kenward and Goetghebeur (2001), along with the actual plebiscite result. (Pess(■): pessimistic boundary; Opt(■): optimistic boundary; MAR(●): Rubin* et al.*'s MAR model; NI(●): Rubin* et al.*'s MNAR model; AC(♦): available cases; CC(♦): complete cases; Pleb(▲): plebiscite outcome. Crosses (×) and the numbers above them refer to the BRD models. Intervals of ignorance (Models 10–12) are represented by horizontal bars.)*

but in fact 7 sensitivity parameters would have to be included. Thus, it appears that a relatively simple sensitivity analysis is adequate to increase insight in the information provided by the incomplete data about the proportion of valid YES votes, in the study under consideration here.

7.3.2 Global Influence

Figure 7.2 shows a selection of the results for the global influence analysis conducted on the SPO Survey data. Inasmuch as Cook's distance measure CD_{2i} was approximately zero for all cells, indicating no substantial influence when adding or removing a single case from a particular cell, for all other models, only the results for BRD4, BRD7, and BRD8 are presented. For BRD4, it can be observed that addition of a single observation to cell 3 has a large influence on the parameters, as well as deletion from either cells 3 or 5. Cell 3 represents subjects with a NO on the attendance question and a YES on the independence question. An addition or removal of one such respondent can largely affect the parameters of BRD4. Similarly, exclusion of a single respondent with a YES on the attendance question but a missing response on the independence

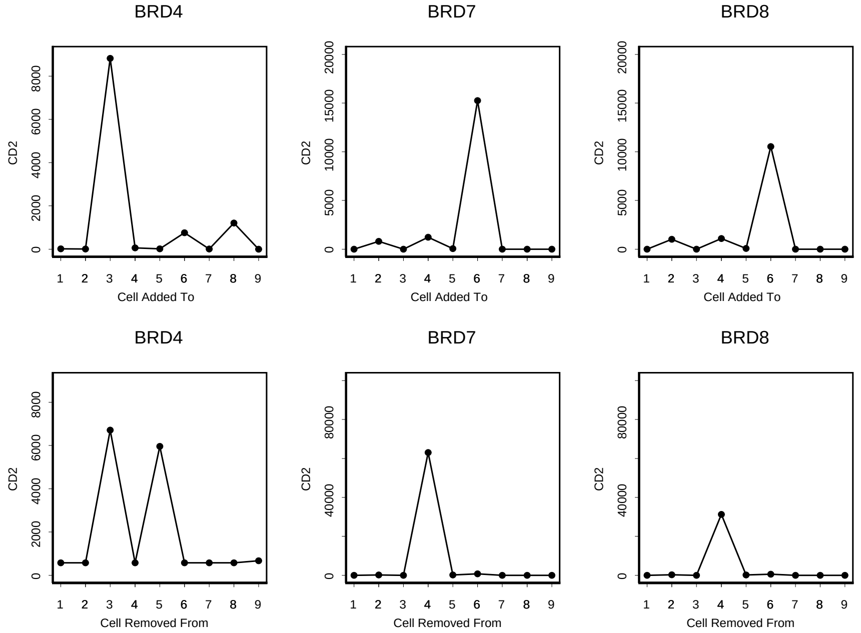


Figure 7.2: *The Slovenian Public Opinion Survey. Global influence analysis for BRD4, BRD7 and BRD8, using Cook's distance measure, CD_{2i}, evaluated when an observation is added to (first row) or deleted from (second row) the i^{th} observed cell.*

question (cell 5), also influences BRD4's model parameters, though to a lesser extent. For models BRD7 and BRD8, an additional observation in cell 6 or a deletion from cell 4 leads to significant influence on these models' parameters. Thus, adding a subject with a NO for attendance and a missing independence response, or excluding a respondent with NO on both questions, yields changes in the model parameters of BRD7 and BRD8. These findings hint on the influential nature of subjects with a NO on the attendance question, which is likely related with this group's sparseness.

7.3.3 Local Influence

In this section, two forms of local influence analysis are described and illustrated using the BRD models as applied to the SPO Survey data. Whereas the first approach, developed by Jansen *et al.* (2003), considers a perturbation in the model parameters, the second type introduces perturbations in the cell probabilities (Beunckens *et al.*, 2009). For both cases, local influence measures are obtained not only in terms of the displacement in the log-likelihood, but also in terms of the displacement in a function of the log-likelihood, namely, the predicted cell counts.

7.3.3.1 Perturbing Parameters: One BRD Model *vs.* Another

The nesting structure within the BRD family of models (Figure 3.1) motivates the consideration of local influence approaches rooted in perturbations of a given BRD model in the direction of a more elaborate one (Jansen *et al.*, 2003), i.e., one containing an additional parameter. For example, BRD4 includes the parameter $\beta_{.j_2}, (j_2 = 1, 2)$, whereas BRD1 only includes $\beta_{..}$. For this type of influence analysis, ω_i in (7.3) can be viewed, not as a parameter, but as an infinitesimal perturbation of the simpler model towards the more complex one, confined to a single subject. For the perturbation of BRD1 in the direction of BRD4, for instance, one considers $\beta_{..}$ and $(\beta_{..} + \omega_i)$. The vector of all ω_is defines the direction in which such a perturbation is considered. Consider the BRD log-likelihood, which is given by:

$$\begin{aligned} \ell(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{j_1,j_2} Z_{11,j_1j_2} \ln \pi_{11,j_1j_2} + \sum_{j_1} Z_{10,j_1+} \ln \pi_{10,j_1+} \\ &+ \sum_{j_2} Z_{01,+j_2} \ln \pi_{01,+j_2} + Z_{00,++} \ln \pi_{00,++}, \end{aligned} \tag{7.8}$$

where $\pi_{r_1r_2,j_1j_2} = p_{j_1j_2}\, q_{r_1r_2|j_1j_2}$ and

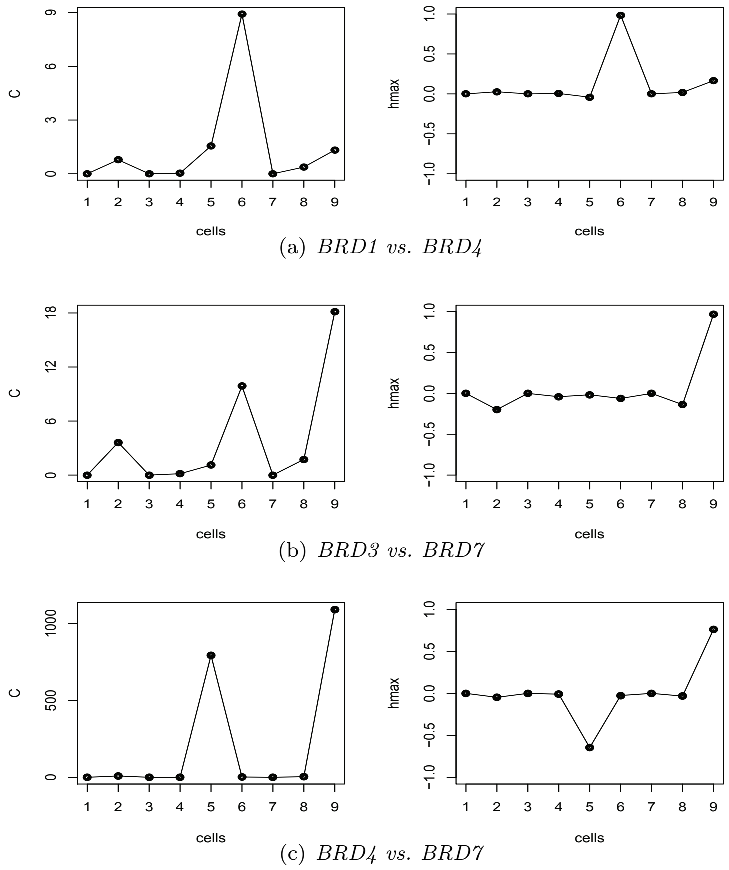
$$q_{r_1r_2|j_1j_2} = \frac{\exp\{\alpha_{j_1j_2}(1-r_1) + \beta_{j_1j_2}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{j_1j_2}) + \exp(\beta_{j_1j_2}) + \exp(\alpha_{j_1j_2} + \beta_{j_1j_2} + \gamma)}. \tag{7.9}$$

Consider for the moment the perturbation of BRD1 in the direction of BRD4. Letting $\beta_{.1} = \beta_{..}$ and $\beta_{.2} = (\beta_{..} + \omega_i)$ in expression (7.9) yields, for BRD4:

$$\begin{aligned} q_{r_1r_2|j_11} &= \frac{\exp\{\alpha_{..}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{..}) + \exp(\alpha_{..} + \beta_{..} + \gamma)}, \\ q_{r_1r_2|j_12} &= \frac{\exp\{\alpha_{..}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{..} + \beta_{..} + \omega_i + \gamma)}. \end{aligned}$$

Note that the perturbation ω_i defines a difference between the above 2 dropout probabilities, while under the simpler (null) model BRD1, the two expressions reduce to a single dropout probability. Local influence measures now follow from the general logic described in Section 7.2.

For the SPO Survey data, the model pairs BRD1 *vs.* BRD4, BRD3 *vs.* BRD7, and BRD4 *vs.* BRD7 are of particular interest, since it is precisely these pairs having high influence on the likelihood displacement. Figure 7.3 shows the influence measures C_i, as well as $\boldsymbol{h}_{\max}$, plotted against the i^{th} observed cell. For the comparison of BRD1 *vs.* BRD4, a peak is observed at cell 6, for both C_i and $\boldsymbol{h}_{\max}$, implying that respondents in this cell drive the data more towards BRD4 $(\alpha_{..}, \beta_{.j_2})$ rather than BRD1 $(\alpha_{..}, \beta_{..})$. That is, subjects with a NO on the attendance question and a missing value



(a) *BRD1 vs. BRD4*

(b) *BRD3 vs. BRD7*

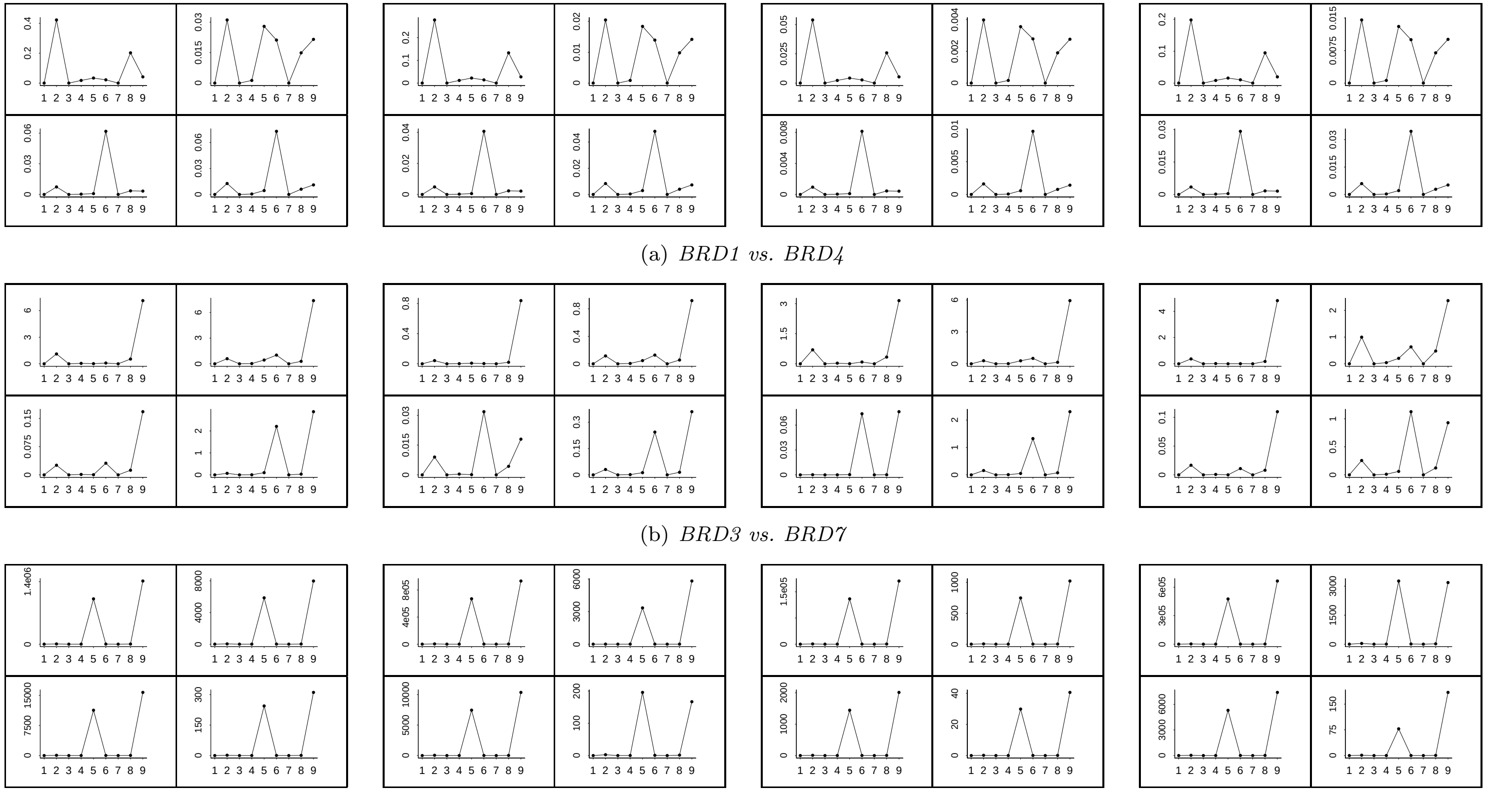
(c) *BRD4 vs. BRD7*

Figure 7.3: *The Slovenian Public Opinion Survey. Local influence analysis, via perturbation in parameters, for model pairs (a) BRD1 vs. BRD4, (b) BRD3 vs. BRD7, and (c) BRD4 vs. BRD7, using local influence measures C_i (first column) and $\boldsymbol{h}_{max}$ (second column) at the i^{th} observed cell, evaluated in terms of the displacement in the log-likelihood.*

on the independence question are influential when perturbing the model such that missingness in the independence question depends on the corresponding unobserved answer (BRD4) rather than being constant (BRD1). For BRD3 *vs.* BRD7, a considerably large value is observed at 'fully missing' cell 9. This means that missingness in the independence question is driven to depend on the corresponding unobserved answer by subjects with missing responses on both questions, and slightly by those with a NO on attendance and a missing value on independence (cell 6). Finally, it is primarily subjects with missing responses on both questions (cell 9) that seem to push the data from BRD4 $(\alpha_{..}, \beta_{.j_2})$ towards BRD7 $(\alpha_{.j_2}, \beta_{.j_2})$. These subjects, along with those that have a YES on attendance and a missing value on independence (cell 5), make the missingness in the attendance question depend on the response of the independence question.

In a contingency-table setting as the SPO Survey, it is instructive to study influence, not only on the parameters, but also on the predicted cell counts. In this case, $Z(\boldsymbol{\phi})$ in (7.5) is taken as a particular cell count, $Z_{r_1r_2,j_1j_2}$, of interest. The results of the local influence analysis on the 16 fitted cell counts are presented in Figure 7.4. The first panel, Figure 7.4(a), for model BRD1 *vs.* BRD4, shows similar shapes for the influence graphs, albeit with differing magnitudes, for a particular cell (j_1, j_2), across the four missingness patterns. For $(j_1, j_2) = (1, 1)$, it is cell 2 that shows influence, and also slightly cell 8. Respondents with either a YES or a missing value on attendance and a NO on independence thus drive the predicted cell count $Z_{r_1r_2,11}$ towards a model in which the missingness in the independence question depends on its value (BRD4). For $(j_1, j_2) = (1, 2)$, cells 2 and 5, as well as 6 and 9, stand out. These respondents make the predicted cell count $Z_{r_1r_2,11}$ seemingly come from BRD4 rather than BRD1. For $(j_1, j_2) = (2, 1)$ and $(j_1, j_2) = (2, 2)$, similar curves are obtained across the four missingness patterns, with a clear peak at cell 6, implying that the "NO-on-attendance/missing-on-independence" responses perturb predicted cell counts $Z_{r_1r_2,21}$ and $Z_{r_1r_2,22}$ in the direction of a model in which the missingness in the independence question is dependent on its value, rather than on one in which missingness in the independence question is constant.

The resulting patterns for the comparison of BRD3 against BRD7, seen in Figure 7.4(b), differs from what was observed for BRD1 *vs.* BRD4. Whereas for the latter, influence curves for a particular cell (j_1, j_2) remained the same across the missingness patterns, for BRD3 *vs.* BRD7, variations now arise across these missingness patterns, leading to a less clear-cut overall picture. This is further complicated by what can be observed for $(j_1, j_2) = (2, 1)$ and $(j_1, j_2) = (2, 2)$, i.e., the bottom row of the tables, for which curve shapes vary across the missingness patterns. Consider $(j_1, j_2) = (1, 1)$ and $(j_1, j_2) = (1, 2)$. Across missingness patterns, the predicted cell counts $Z_{r_1r_2,11}$ and $Z_{r_1r_2,12}$ are primarily influenced by subjects with both responses missing, and slightly by those having a YES on attendance/NO on independence. For cell $(j_1, j_2) = (2, 1)$, similar graphs are obtained for $(r_1, r_2) = (1, 1)$ and $(r_1, r_2) = (0, 0)$, i.e., the completers and double non-responders, respectively, with a peak at cell 9. It is therefore subjects with both responses missing that influence cell counts $Z_{11,21}$ and $Z_{00,21}$, in the direction of a model in which missingness in the independence question depends on its value. For the other two missingness patterns,



(a) *BRD1 vs. BRD4*

(b) *BRD3 vs. BRD7*

(c) *BRD4 vs. BRD7*

Figure 7.4: *The Slovenian Public Opinion Survey. Local influence analysis, via perturbation in parameters, for model pairs (a) BRD1 vs. BRD4, (b) BRD3 vs. BRD7, and (c) BRD4 vs. BRD7, using local influence measure C_i at the i^{th} observed cell, evaluated in terms of the displacement in each of the 16 predicted cell counts (in their respective positions as in Table 7.1).*

$(r_1, r_2) = (1, 0)$ and $(r_1, r_2) = (0, 1)$, referring to subjects with a single non-response, peaks occur at cells 6 and 9. Thus, subjects with a NO on attendance/missing on independence and those with both responses missing have an influence on predicted cell counts $Z_{10,21}$ and $Z_{01,21}$. These same subjects also influence the predicted cell counts $Z_{r_1 r_2,22}$.

Whereas the comparison of BRD3 *vs.* BRD7 presents the most variable influence graphs, BRD4 *vs.* BRD7 shows the most consistent ones. All 16 influence curves exhibit a single shape, although of varying magnitudes, implying that influence on any predicted cell count is coming from a common source, regardless of the missingness pattern. Here, a clear peak occurs at cells 9 and 5, similar to the likelihood-displacement results. Subjects with missing responses on both questions and those with YES on attendance/missingness on independence, have an influence that drives any predicted cell count towards a model where the missingness in the attendance question depends on the response of the independence question.

7.3.3.2 Perturbing Cell Probabilities

Alternative to perturbations in the parameters, which, in the case of the BRD family, implies perturbing one BRD model in the direction of another, one can consider introducing, within a given BRD model, infinitesimal perturbations in the cell probabilities, or consequently, perturbations in the cell counts. The perturbed likelihood is given by

$$\begin{aligned}
\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{j_1,j_2} (Z_{11,j_1 j_2} + N\omega_{11,j_1 j_2}) \ln \pi_{11,j_1 j_2} \\
&+ \sum_{j_1} (Z_{10,j_1+} + N\omega_{10,j_1+}) \ln \pi_{10,j_1+} \\
&+ \sum_{j_2} (Z_{01,+j_2} + N\omega_{01,+j_2}) \ln \pi_{01,+j_2} \\
&+ (Z_{00,++} + N\omega_{00,++}) \ln \pi_{00,++},
\end{aligned} \tag{7.10}$$

with all notation as before. Under the null model, $\boldsymbol{\omega} = \boldsymbol{\omega}_o = \mathbf{0}$, and the above log-likelihood reduces to the standard multinomial one. Because focus is now on cell probabilities rather than observations, interpretations will naturally be different, despite similarities in computation. A peak in the influence curve now represents the particular observed cell at which a probability perturbation causes substantial displacement in either the log-likelihood or some function of the parameters of a specific BRD model.

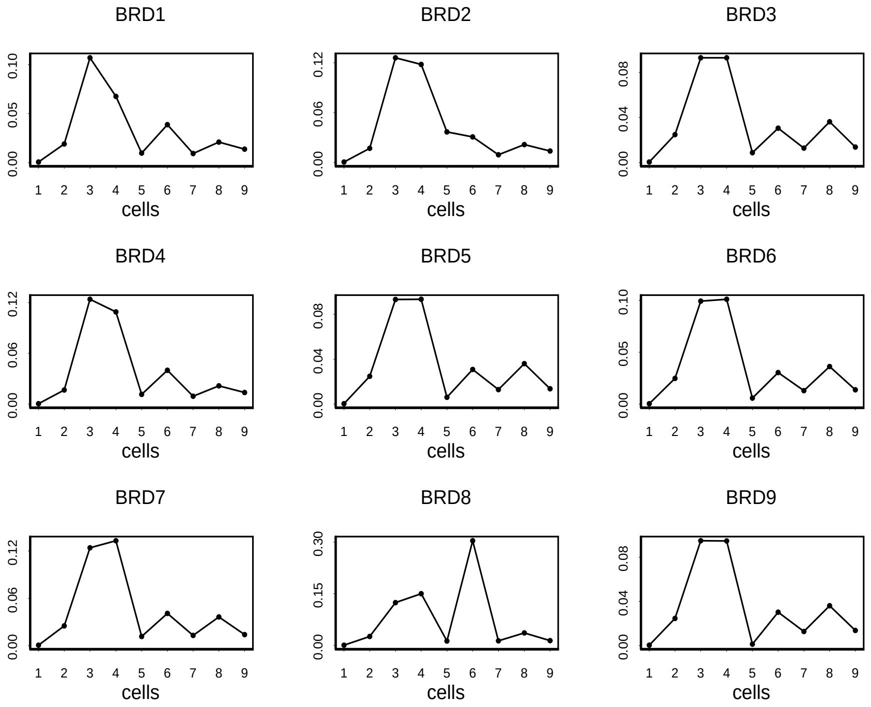


Figure 7.5: *The Slovenian Public Opinion Survey. Local influence analysis, via perturbation in cell probabilities, for each of the nine BRD models, using local influence measure C_i at the i^{th} observed cell, evaluated in terms of the displacement in the log-likelihood.*

For the SPO Survey data, consider first the influence on the likelihood displacement (Figure 7.5). For most models, the probabilities of cells 3 and/or 4 are influential. Also notable is the influence of changes in cell 6. Thus, the most influential cells for almost all models are the completers answering NO on attendance, likely attributable to the small counts in these cells, while for BRD8, those answering NO on attendance but with a missing response on independence are influential.

Results of local influence analysis on the predicted cell counts upon perturbing a particular cell probability are summarized in Table 7.6. Cells 1 and 2 are generally without influence, owing to large counts. While small perturbations in cell 3 seem to affect only the predicted cell counts in the top row ($j_1 = 1$, YES on attendance) under BRD5 and/or BRD6, such changes impact the cell position $(j_1, j_2) = (1, 2)$ (YES on attendance/NO on independence) under BRDs 1, 2, 3, 6, and/or 9. Perhaps the most striking observation that can be made from Table 7.6 is that a perturbation in cell

Table 7.6: *The Slovenian Public Opinion Survey. Local influence analysis, via perturbation in cell probabilities, for each of the nine BRD models, using local influence measure C_i at the i^{th} observed cell, evaluated in terms of the displacement in the each of the 16 predicted cell counts (in their respective positions as in Table 7.1). Entries in boxes denote the BRD model number for which influence is largest when the particular cell probability is perturbed.*

Adding ω to cell		Z_{11,j_1j_2}		Z_{10,j_1j_2}		Z_{01,j_1j_2}		Z_{00,j_1j_2}	
1		---	---	---	---	---	---	---	---
		---	---	---	---	---	---	---	---
	2	---	---	---	---	---	---	---	3
		---	---	---	---	---	---	---	---
		5,6	5,6	5,6	5,6	5,6	5,6	5	5
3		1-3,9	---	2,3,9	---	1-3,6	---	1-3,6	6
		4,6,7,9	4-7	4-7,9	4-8	4-7	4-8	4-7,9	3-9
	4	4-7,9	2-9	4-7	2-9	4-7,9	2-8	4-7,9	2-5,7,9
5		2	---	1-3	1,2	---	---	---	---
		---	---	---	---	---	---	---	---
		8	8	8	8	8	---	8	---
6		8	1,8	1,8	1	8	---	8	8
7		3,9	---	9	---	1-3,9	---	---	---
		---	---	---	---	---	---	---	---
	8	---	1-3,9	9	3,9	---	1-3,9	9	---
		---	---	---	---	---	1,9	---	---
9		1,2	---	---	---	---	---	1-3	1,2
		---	---	---	---	---	---	1	1

probability 4 (NO/NO respondents) yields influence on *all* 16 predicted cell counts in most of the higher-numbered BRD4–BRD9. Also, the results for perturbations in cell 6, indicate that it is primarily under BRD8 where a large influence is observed in most of the predicted cell counts. Finally, changes in the probability of the doubly missing cell 9 affect only the predicted cell counts of this missingness pattern and only under BRDs 1, 2, and/or 3.

7.3.4 How Sensitive is the Proportion of 'Yes' Voters?

The inferential target of the Slovenian Public Opinion Survey analyses is estimating the proportion of people voting in favor of independence – a goal hampered by incompleteness. It has been argued that putting blind belief in a single model may be too strong; it is not possible, from a purely statistical point of view, to unambiguously validate a single model, motivating the consideration of sensitivity analyses. The previous sections featured a variety of these, going beyond conventional sensitivity analysis applications, which are often confined to a single sensitivity assessment tool. Table 7.3 summarizes various methods. The simplest analysis considers the non-parametric bounds and deduces that even the least supportive scenario for independence would still produce a, roughly, 70% majority. Alternatively, one can fit a discrete class of models, such as the nine BRD models, or merely the well-fitting models BRD6–BRD9, and then construct the resulting interval; this narrows the non-parametric interval and even excludes the plebiscite value. It is here that the MAR counterparts (Table 7.4) prove to be quite useful, given that they essentially produce a single point estimate of 89.2% – a value very close to the actual plebiscite result. The concept of an interval naturally leads to the consideration of intervals of ignorance, nicely interpolating between the non-parametric interval and a single, identified model. From Table 7.5, it can be observed that Model 12 reproduces the non-parametric interval, while Models 10 and 11 further narrow it.

Turning to substantive considerations, the estimates for θ (Table 7.4), can be split into two:

Optimistic: MAR counterparts, BRD1, BRD3, (BRD5), (BRD6) and BRD9
Pessimistic: BRD4, BRD7 and BRD8

The parenthetical ones are slightly less pronounced than the others. The assumptions regarding the missingness mechanism, underpinning all twelve models, can be read from the second column in Table 7.4; they are also spelled out in Table 7.7. It is striking that the three fully identified, pessimistic estimates allow missingness in the

Table 7.7: *The Slovenian Public Opinion Survey. Meaning of the missingness mechanism in the nine identified models BRD1–BRD9 and its three overspecified extensions Models 10–12.*

Missingness in ⟶	attendance		independence	
Depends on ⟶	attendance	independence	attendance	independence
BRD1	—	—	—	—
BRD2	—	—	√	—
BRD3	—	√	—	—
BRD4	—	—	—	√
BRD5	√	—	—	—
BRD6	√	—	√	—
BRD7	—	√	—	√
BRD8	√	—	—	√
BRD9	—	√	√	—
Model 10	—	√	√	√
Model 11	√	√	√	—
Model 12	√	√	√	√

independence question to depend on independence, whereas the other six do not: these three models support the thesis that there is a large group of people, in favor of independence, that would not partake in the plebiscite, as can be seen from Table 4.1. Focusing on BRD7, BRD9, and their MAR counterparts, the pessimistic fit of BRD7 asserts that the proportion of people in favor of independence, yet that would not attend the plebiscite, amounts to $(78 + 155.8 + 44.8 + 112.5)/2074$, i.e., 18.9%. The optimistic BRD9 predicts this fraction to be as low as 7.3%, while for the counterparts it goes down further to 6.6%. It can thus be inferred, with reasonable confidence, that, in the actual plebiscite, people expressed their opinion, regardless of real or perceived pressure, and that an overwhelming majority favored independence, supporting the optimistic scenario.

There are more nuances in the overspecified Models 10–12. While Models 10 and 12 also include the "pessimistic relationship" of missingness in independence on the independence answer, the interval is not pessimistic. Rather, the three intervals encompass both very pessimistic and very optimistic scenarios. This is because, unlike the fully identified models, there is no need to sacrifice one type of dependence to maintain identifiability. Precisely, while none of BRD4, BRD7 and BRD8 allow missingness in the independence question to depend on the attendance response, the three intervals of ignorance do allow for such a dependence – a feature shared with BRD2, BRD6 and BRD9. It ought not to go unnoticed that, among the optimistic ones, BRD5 and BRD6 are somewhat less pronounced; these models allow missingness in the attendance question to depend on the respondent's attendance position. The effect of this is similar, but less sharp, than in the 'independence on independence' scenario sketched above. It is therefore no longer a surprise that BRD8, where these two effects play together, produces the most pessimistic estimate.

The question remains why the impact of changing the assumptions is rather spectacular, in the sense that the most pessimistic scenarios are very close to the pessimistic bound, whereas the most optimistic scenarios are virtually at the optimistic bound. It is instructive to return to the global and local influence analyses. First and foremost, the influence analysis results indicate that there is relatively little influence altogether, *except in BRD4, BRD7 and BRD8*, i.e., the entire pessimistic group. This is clear from the discussion in Sections 7.3.2 and 7.3.3, and from Figures 7.2, 7.3 and 7.5. Global influence results indicate a strong impact of the (relatively small) NO-on-independence cells in the pessimistic models: changes in these counts can dramatically alter the way in which the complete cells are split over the hypothetical complete cells, rendering $\widehat{\theta}$ unstable. Local influence focuses on a different aspect: which observations/cells drive the conclusions away from a given null model? For BRD1 *vs.* BRD4, this is Cell 6 – the not-in-favor-of-independence respondents without declared attendance status. Thus, a small but influential count is simultaneously responsible for a move towards a pessimistic scenario and the extent of pessimism.

The combined sensitivity analyses and substantive considerations demonstrate that the optimistic scenarios, whether from the MAR counterparts or the optimistic group of BRD models, are plausible descriptions of the mechanisms operating during the plebiscite exercise. The influence analyses show that the pessimistic ones are rather different from their optimistic counterparts and constitute plausible scenarios owing primarily to the presence of one or a few influential cells.

7.4 Discussion

In this chapter, a variety of existing and expanded sensitivity analyses have been presented to gain insight into the impact of missing data generating mechanisms governing the Slovenian Public Opinion Survey data. A first family of sensitivity analyses is based on considering a collection of models, (1) through non-parametric bounds, (2) a family of identified models, or (3) intervals resulting from over-specified models. Next to these, observation-varying analyses: (4) global influence, (5) local influence in terms of infinitesimally varying particular parameters, and (6) local influence through perturbing cell probabilities, were conducted. The combined conclusions of these sensitivity analyses, along with substantive considerations, lead to a coherent picture that points to increased confidence in the more optimistic scenarios, and provides a plausible explanation for these.

How does one then proceed, in similar, related or different data-analysis problems with incompleteness? First, many of the methods have been developed for contingency tables and beyond. This is true for local and global influence technology that also exists for continuous outcomes, possibly with covariates, and intervals of ignorance that have been developed for logistic regression. For a review, see Molenberghs and Kenward (2007). Other methods, not considered here, can be considered as well. For instance, pattern-mixture models (Thijs *et al.*, 2002) are worthy of attention, either for their own sake because they might appropriately address a particular scientific question, or as a useful contrast to selection models either: (1) to answer the same scientific question, based on these different modeling strategies; or (2) to gain additional insight by supplementing the selection model results with those from a pattern-mixture approach. Pattern-mixture models also have a special role in some multiple-imputation-based sensitivity analyses (Molenberghs and Kenward, 2007). Alternatively, one could also turn towards the shared-parameter framework, as one of its main advantages is the ability to easily handle non-monotone missingness, but these models are based on very strong parametric assumptions, such as normality for the shared random effect(s). Sensitivities abound in the selection and pattern-mixture frameworks as well, but the assumption of unobserved, random, or latent effects, further complicates the issue (Tsonaka, Verbeke and Lesaffre, 2009; Rizopoulos, Verbeke and Molenberghs, 2008). Yet a third approach brings together features of the selection, pattern-mixture and shared-parameter frameworks through the use of so-called *latent-class mixture models* (Beunckens *et al.*, 2008). Fourth, one can take a semi-parametric standpoint, where weighted generalized estimating equations (WGEE), proposed by Robins, Rotnitzky and Zhao (1994) play a central role. Rather than jointly modeling the outcome and missingness processes, the centerpiece is inverse probability weighting of a subject's contribution, where the weights are specified in terms of factors influencing missingness, such as covariates and observed outcomes (Robins, Rotnitzky and Scharfstein, 1998; Scharfstein, Rotnitzky and Robins, 1999). Robins, Rotnitzky and Scharfstein (2000) and Rotnitzky *et al.*

(2001) use this framework to conduct sensitivity analysis.

The field of sensitivity analysis towards the impact of incomplete data is vibrantly active and methodology as well as insight surrounding it is bound to emerge in time to come. The modeler and practitioner have an ever expanding toolkit of sensitivity analysis machinery at their disposal. Carefully selected sensitivity analysis equipment, supplemented with substantive arguments, can produce valuable insight and perhaps even confidence, above and beyond what is obtainable from a single analysis, be it of the MAR or MNAR type.

8

Gaussian Longitudinal Data and Missing Not At Random: Stochastic *versus* Non-Stochastic Solutions

In the setting of incomplete longitudinal data arising from clinical studies, the MCAR case is often regarded as perhaps the most stringent and unrealistic of beliefs regarding the missingness, requiring the simplest methodologies. At the opposite end of the spectrum lies the considerably reasonable and realistic situation of MNAR, entailing methods that are beset with numerical complexities. MNAR is plausible in (longitudinal) clinical trials, where it is common to encounter subjects who do well until midway into the trial, relapse after the last observed visit, and are then lost to follow-up. If, in such cases, the missed visit is, to some extent, attributable to the subject's 'worsening' clinical condition, as represented by the yet-to-be observed response, then the non-response process depends on the missing response (that would have been recorded had the patient been present for the visit), and consequently, the missing values are said to be non-randomly missing.

Under ignorability (MCAR or MAR), the observed-data likelihood (3.5) simplifies, obviating the need for any integration, and standard maximum likelihood procedures can be applied to the reduced form using only the observed data, provided the method's implementation can handle unbalanced data, in the sense of allowing for varying number of measurements per subject. The MNAR case, however, does not admit simplification and estimation of (3.5) would require evaluation of the integral numerically. This integration step is what often poses a challenge in fitting likelihood-based parametric models for non-ignorable missing data. Depending on the complexity of the model used, not only is the approximation/evaluation of the integral itself involved, but its subsequent maximization (for maximum likelihood, for instance) can be computationally demanding.

Broadly speaking, the evaluation and subsequent maximization of the observed-data likelihood (3.5) under non-ignorable missingness may be approached using either a stochastic or a non-stochastic solution. Non-stochastic solutions involve maximization of either a direct evaluation or a numerical approximation of the integral in (3.5). Diggle and Kenward (1994) provide such a solution for continuous Gaussian data with monotone missingness, while Troxel, Harrington and Lipsitz (1998) propose a method for the non-monotone case. Stochastic methods, on the other hand, involve (iteratively) simulating random draws from the underlying non-standard distribution via Markov chains to fill in the missing data, and subsequently maximizing the likelihood using the completed data. For this reason, the stochastic approach offers a less computationally demanding alternative because it precludes any numerical integration.

In this chapter, the use of the stochastic expectation-maximization (EM) algorithm (Celeux and Diebolt, 1985) for fitting selection models for continuous longitudinal data with MNAR missingness, be it of a monotone or non-monotone type, is explored and compared with non-stochastic solutions. Within the context of a simulation study, the merits of the stochastic approach over non-stochastic methods are highlighted, thereby emphasizing its value as a practical alternative. Ideas are further illustrated in applications on specific case studies.

8.1 Estimation Methodology

8.1.1 The Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977), which is a general-purpose iterative algorithm for calculating maximum likelihood estimates, has become quite popular for handling incomplete data problems, as well as for other data-augmented settings, such as latent-variable and random effects models. The fundamental idea behind the EM algorithm is to associate with the given incomplete data problem, a complete data problem for which maximum likelihood estimation is computationally more tractable. Each iteration of the algorithm consists of two steps: an expectation (E) step and a maximization (M) step. Given initial parameter values, the E-step calculates the conditional expectation of the complete-data log-likelihood given the observed data and the parameter estimates. Given then the complete-data log-likelihood, the M-step finds the parameter estimates to maximize the complete-data log-likelihood from the E-step. These steps are iterated till convergence.

More concretely, suppose $\boldsymbol{\theta}^{(t)}$ is the vector of current values for the parameter of interest $\boldsymbol{\theta}$. The E-step involves computing the objective function, which, in the case of a missing data problem, is equal to the expected value of the complete-data log-likelihood, given the observed data and the current parameters,

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = \int \ell(\boldsymbol{\theta},\boldsymbol{y}) f\left(\boldsymbol{y}^m\middle|\boldsymbol{y}^o,\boldsymbol{\theta}^{(t)}\right) d\boldsymbol{y}^m = E\left[\ell(\boldsymbol{\theta}|\boldsymbol{y})\middle|\boldsymbol{y}^o,\boldsymbol{\theta}^{(t)}\right], \tag{8.1}$$

that is, substituting the expected value of $\boldsymbol{Y}^m$, given $\boldsymbol{Y}^o$ and $\boldsymbol{\theta}^{(t)}$. At the M-step, the parameter vector that maximizes the log-likelihood of the imputed data, $\boldsymbol{\theta}^{(t+1)}$, is calculated. Thus, $\boldsymbol{\theta}^{(t+1)}$ satisfies

$$Q\left(\boldsymbol{\theta}^{(t+1)}\middle|\boldsymbol{\theta}^{(t)}\right) \geq Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) \quad , \quad \text{for all } \boldsymbol{\theta}.$$

Among the reasons for the widespread popularity of the EM algorithm is the guarantee that the likelihood function increases at every iteration, a property not shared by other optimization methods (e.g., Newton-Raphson), according it considerable numerical stability. Among the shortcomings of the EM algorithm, however, are a painfully slow rate of convergence, and in some cases, intractability of the E-step.

8.1.2 The Stochastic Expectation-Maximization Algorithm

Several stochastic variants of the EM algorithm have been introduced to overcome possible complexity in the E-step. Stochastic EM (Celeux and Diebolt, 1985), or SEM, replaces the E-step with an S-step, in which the missing data are imputed with plausible values, given the observed values and the current parameter estimates. A subsequent M-step then maximizes the complete-data likelihood based on the pseudo-complete data. Alternatively, Wei and Tanner (1990) proposed a Monte Carlo implementation of the E-step, resulting in Monte Carlo EM (MCEM), which generalizes SEM in the sense that in the S-step, m independent samples are drawn, in contrast to just one for SEM. Finally, Stochastic Approximation EM (Celeux and Diebolt, 1992), or SAEM, provides a hybrid of the EM and SEM algorithms, since, at the current estimate, a standard EM iteration is performed in addition to the SEM iteration. A review of the EM algorithm and its extensions can be found in McLachlan and Krishnan (1997) and Celeux, Chauveau and Diebolt (1995).

Though SEM has been around for some time, there have been few applications in the area of incomplete longitudinal data. Jolani and Ganjali (2007) applied SEM to fit a generalized Heckman selection model (Crouchley and Ganjali, 2002), which extends a sample selection model for missing data to the case of repeated responses with dropout. In contrast, Gad and Ahmed (2006) used the SEM algorithm for continuous longitudinal with intermittent missingness. They formulated a Gaussian model for $\boldsymbol{Y}_i$, with a logistic model for $\boldsymbol{R}_i$, conditional on the previous and current responses, but without any specification for the association among the response indicators, thereby implicitly assuming independence among these. Sotto *et al.* (2009b) proposed combining two fully multivariate models – a multivariate Gaussian model for $\boldsymbol{Y}_i$ and a multivariate Dale model for $\boldsymbol{R}_i$ – for the joint modeling of continuous longitudinal data with non-monotone missingness, and obtained solutions using the stochastic EM algorithm.

The main idea behind stochastic EM (SEM) is to "fill in," in what is referred to as the S-step, the missing data with a single draw from the conditional density of the missing values given the observed values at the current estimate of the parameter vector, thereby providing a plausible pseudo-complete sample. In the succeeding M-step, the resulting pseudo-complete sample is used to directly maximize the log-likelihood to obtain an updated estimate. The algorithm iterates these two steps until the resulting Markov-Chain converges.

In line with previous notation, let $f(\boldsymbol{y}_i, \boldsymbol{r}_i|\boldsymbol{\theta}) = f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i|\boldsymbol{\theta})$ denote the joint density function for the subject i. Each iteration t of the SEM algorithm consists of the following two steps:

- **S-step.** At the current estimate of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(t-1)}$, the missing values for the i^{th} subject are imputed using a single draw from the conditional distribution $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{r}_i, \boldsymbol{\theta}^{(t-1)})$. As the latter does not have a standard form, direct simulation from it is not possible, but an accept-reject procedure can be used:

  1. Simulate a candidate value $\boldsymbol{y}_i^*$ from the conditional distribution function $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\theta}^{(t-1)})$.
  2. Calculate the probability distribution of the missingness indicators, say $P(\boldsymbol{R}_i = \boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^*, \boldsymbol{\theta}^{(t-1)})$, according to the assumed non-response model.

These define a multinomial distribution for the different patterns of missingness, from which cumulative probabilities and ranges thereof can be obtained at each missingness pattern.

3. Generate a random variable u from the uniform distribution on the interval [0,1]. If u falls within the range of cumulative probabilities for the missingness pattern of the i^{th} subject, then the candidate value $\boldsymbol{y}_i^*$ is accepted as the imputation for $\boldsymbol{y}_i^m$. Otherwise, steps 1–3 are repeated until a suitable imputation is obtained.

- **M-step.** With the pseudo-complete data, denoted $\boldsymbol{y}^{ps}$, a likelihood maximization routine is then used to obtain updated parameters $\boldsymbol{\theta}^{(t)}$. The likelihood of the pseudo-complete data for subject i can be written as

$$f(\boldsymbol{y}_i^{ps}, \boldsymbol{r}_i|\boldsymbol{\theta}^{(t-1)}) = f(\boldsymbol{y}_i^{ps}|\boldsymbol{\theta}^{(t-1)})\, P(\boldsymbol{r}_i|\boldsymbol{y}_i^{ps}, \boldsymbol{\theta}^{(t-1)}).$$

Alternately imputing pseudo-complete data and performing maximization generates a Markov chain $\{\boldsymbol{\theta}^{(m)}\}$, which converges, after a sufficient burn-in period, to a stationary distribution $\pi(\cdot)$ under mild conditions (Ip, 1994). Unlike the deterministic EM algorithm, the final output from stochastic EM is a sample from this stationary distribution, which is approximately centered at the MLE of $\boldsymbol{\theta}$ and its mean can be considered as an estimate for $\boldsymbol{\theta}$. Diebolt and Ip (1996) refer to this mean as the *stochastic EM estimate* $\widetilde{\boldsymbol{\theta}}_n$, which does not, in general, coincide with the MLE.

The variance of $\widetilde{\boldsymbol{\theta}}_n$ can be estimated by the inverse of the observed information matrix, evaluated at $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_n$. Direct computation of this, however, can be just as inconvenient, because integration over the missing values will be necessary. Louis (1982) derives a useful identity relating the observed-data log-likelihood and the complete-data log-likelihood:

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{r})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\middle| \boldsymbol{y}^o, \boldsymbol{r}\right) - \text{Cov}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{r})}{\partial\boldsymbol{\theta}}\middle| \boldsymbol{y}^o, \boldsymbol{r}\right), \tag{8.2}$$

Efron (1994) proposes a Monte Carlo method to estimate the expectations and covariances in (8.2) by their empirical versions, by simulating, for the missing values $\boldsymbol{y}^m$, a set of, say G, independent and identically distributed samples $\boldsymbol{y}^{m_1}, \boldsymbol{y}^{m_2}, \ldots, \boldsymbol{y}^{m_G}$, using the conditional distribution $f(\boldsymbol{y}^m|\boldsymbol{y}^o, \boldsymbol{r}, \boldsymbol{\theta})$, and subsequently approximating (8.2) using:

$$\begin{aligned}
E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{r})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\middle| \boldsymbol{y}^o, \boldsymbol{r}\right) &\approx \frac{1}{G}\sum_{j=1}^{G} \frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y}^o, \boldsymbol{y}^{m_j}, \boldsymbol{r})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}, \\
\text{Cov}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{r})}{\partial\boldsymbol{\theta}}\middle| \boldsymbol{y}^o, \boldsymbol{r}\right) &\approx \frac{1}{G}\sum_{j=1}^{G}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y}^o, \boldsymbol{y}^{m_j}, \boldsymbol{r})}{\partial\boldsymbol{\theta}}\right)^2 - \left(\frac{1}{G}\sum_{j=1}^{G} \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y}^o, \boldsymbol{y}^{m_j}, \boldsymbol{r})}{\partial\boldsymbol{\theta}}\right)^2,
\end{aligned}$$

where $\boldsymbol{\theta}$ is fixed at $\widetilde{\boldsymbol{\theta}}$. To simulate values from $f(\boldsymbol{y}^m|\boldsymbol{y}^o, \boldsymbol{r}, \boldsymbol{\theta})$, the same accept-reject procedure, described in the S-step above, can be used.

The SEM algorithm is expected to converge faster than the EM algorithm, because calculation of the expected values of the missing data can be computationally demanding for high-dimensional integration. In addition, the SEM algorithm can also avoid, owing to its stochastic aspect, the saddle points or irrelevant local maxima (Diebolt and Ip, 1996). To determine convergence of a Markov chain, the approach proposed by Gelman and Rubin (1992), which is based on generating multiple, $m \geq 2$, chains in parallel, each of length $2n$ iterations, can be considered. Convergence of each scalar parameter is monitored by evaluating the so-called Potential Scale Reduction Factor (PSRF), which can be calculated as

$$\sqrt{\widehat{R}} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{m\,n}\,\frac{B}{W}\right)\frac{df}{df-2}}, \tag{8.3}$$

where B/n is the variance between the m sequence means, W is the average of the m within-sequence variances, both based only on the last n iterations of each sequence, and the term $df/(df-2)$ adjusts for sampling variability and can be ignored. The PSRF is the factor by which the scale of the current distribution might be reduced if the simulations were continued in the limit $n \rightarrow \infty$. Convergence is achieved if the PSRF is close to 1, implying that the parallel Markov chains are essentially overlapping. In such a case, the results are summarized using the simulated values from last halves of the m chains, rather than a single one. When the PSRF is large, further simulations may be necessary to improve inferences about the target distribution.

8.2 Simulation Study

To evaluate the performance of the SEM algorithm in fitting non-ignorable models for incomplete Gaussian longitudinal data, simulations for both the monotone and non-monotone situations were conducted. Data generating models, simulation settings and subsequent results under each case are described in the following subsections. These SEM results were compared with those obtained using Newton-Raphson with numerical integration, hereinafter referred to as the direct likelihood approach. The comparison of the two methods was done in terms of the mean squared error (MSE) of the parameter estimates. Although the true parameters used in the data generation, as well as the parameters obtained from the complete data, are both available, bias was computed (for the calculation of MSE) with respect to the latter. This was due to the fact that some sampling error was observed for some parameters. Moreover, since the focus here is to compare how the methods fare in handling the missing data, the emphasis lies more on bias due to missingness, rather than on bias due to finite sampling.

8.2.1 Monotone Missingness

The model framework considered for Gaussian longitudinal data with monotone missingness or dropout is that proposed by Diggle and Kenward (1994), i.e., a selection model consisting of a marginal Gaussian model for the responses and a logistic regression for dropout conditional on the possibly unobserved outcomes.

8.2.1.1 Data Generation

For the simulation study, trivariate Gaussian outcomes $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$, a treatment indicator $x_i(x_i = 0, 1)$, and 3 time periods $t_{ij}(t_{ij} = 1, 2, 3)$, were assumed to be related according to the following:

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_0 \;+\; \beta_x\, x_i \;+\; \beta_t\, t_{ij} \;+\; \beta_{xt}\, x_i\, t_{ij},$$

where $\beta_0 = 2.0, \beta_x = 0, \beta_t = 0.5$ and $\beta_{xt} = 1.0$. This gives rise to the following mean vector

$$E(\boldsymbol{Y}_i|x_i) \;\equiv\; \boldsymbol{\mu}_i = \begin{cases} (2.5, 3.0, 3.5)' & \text{for } x_i = 0 \\ (3.5, 5.0, 6.5)' & \text{for } x_i = 1 \end{cases}, \tag{8.4a}$$

with a pre-defined unstructured covariance matrix for $\boldsymbol{Y}_i$ given by:

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 0.7 & 0.4 & 0.2 \\ 0.4 & 0.6 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}. \tag{8.4b}$$

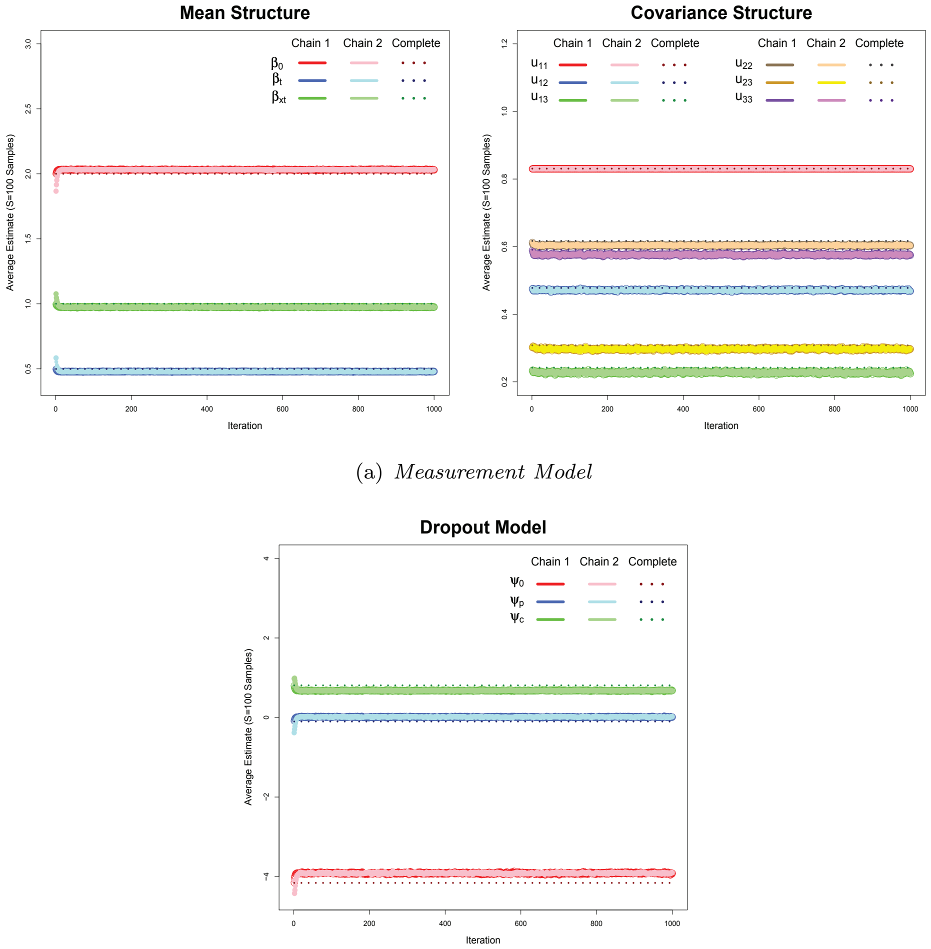
A total of $S = 100$ samples, each of size $N = 300$ (equally distributed between the two treatment arms), was generated from $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

To induce monotone missingness, in line with the Diggle and Kenward (1994) model, a logistic dropout model of the form (3.31) was used, with $\psi_0 = -4, \psi_p = -0.5$ and $\psi_c = 1$, the latter two denoting the coefficients for the previous and the current measurements, respectively. Over the $S = 100$ samples, this dropout model yields about 55% completers, and about 22% each for the other two dropout patterns.

Denoting by $u_{11}, u_{12}, \ldots, u_{33}$ the elements of the Cholesky decomposition of $\boldsymbol{\Sigma}$, and excluding β_x, since $\beta_x = 0$, as would be assumed in a typical randomized clinical trial, this data generating mechanism gives rise to the following full parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\psi}')' = (\beta_0, \beta_t, \beta_{xt}, u_{11}, u_{12}, u_{13}, u_{22}, u_{23}, u_{33}, \psi_0, \psi_p, \psi_c)'$.

8.2.1.2 Results

A Diggle and Kenward (1994) model, consisting of (8.4) and (3.31), was fitted to each of $S = 100$ generated samples with monotone missingness using the stochastic EM algorithm. For each sample, two parallel chains, each with $2n = 1000$ iterations, were simulated. Figure 8.1 depicts, at each iteration, the average value of the estimate over the $S = 100$ samples for each of the two chains, as well as for the complete data.



(a) *Measurement Model*

(b) *Dropout Model*

Figure 8.1: *Simulated Data (Monotone). Average parameter estimate per iteration, over the* $S = 100$ *samples, for each of the two simulated stochastic EM algorithm chains and for the complete data, for the Diggle-Kenward model.*

Table 8.1: *Simulated Data (Monotone). True parameter values, average parameter estimate for $S = 100$ complete samples, for stochastic EM* (SEM) *and for direct likelihood* (DL), *along with mean squared errors* (MSE), *for the Diggle-Kenward model.*

Effect	Parameter	True	Complete	Estimate		MSE × 10^2	
				SEM	DL	SEM	DL
Measurement model							
Intercept	β_0	2.00	2.00	2.03	2.00	0.55	0.52
Time	β_t	0.50	0.50	0.48	0.50	0.11	0.11
Treatment	β_{xt}	1.00	1.00	0.98	1.00	0.12	0.13
Variance 1	u_{11}	0.84	0.83	0.83	0.83	0.12	0.12
Covariance 1,2	u_{12}	0.48	0.48	0.47	0.48	0.14	0.21
Covariance 1,3	u_{13}	0.24	0.24	0.23	0.23	0.13	0.27
Variance 2	u_{22}	0.61	0.62	0.60	0.63	0.07	0.14
Covariance 2,3	u_{23}	0.30	0.31	0.30	0.32	0.10	0.32
Variance 3	u_{33}	0.59	0.59	0.58	0.59	0.06	0.12
Dropout model							
Intercept	ψ_0	-4.00	-4.16	-3.91	-4.08	17.99	25.77
Previous	ψ_p	-0.50	-0.10	0.01	-0.50	3.19	21.11
Current	ψ_c	1.00	0.81	0.68	1.01	2.93	8.75

The plots indicate convergence of the two chains to a common value, which was quite similar to the average value using the complete data. To further assess convergence, the PSRF (8.3) was calculated for each sample, and all $S = 100$ samples yielded a PSRF value that was sufficiently close to 1 for all parameters.

Given the convergence of the two chains, as assessed thru the PSRF, a pooled estimate for each sample was then obtained by averaging the last 500 iterations from both chains, while the covariance matrix was calculated using (8.2). Finally, the pooled estimates and covariance matrices were averaged over the $S = 100$ samples. The SEM results (estimates and MSEs), as well as those for direct likelihood, are presented in Table 8.1. The true parameters used in data generation, as well as the parameters computed from the complete data for the $S = 100$ samples are provided.

The two methods generally yield comparable estimates for both the measurement and dropout models, and these are sufficiently close to the parameters obtained using the complete data. This is not surprising in view of the fact that both approaches are simply different computational algorithms to obtain maximum likelihood estimates, and are naturally expected to produce consistent results. Larger discrepancies between the methods, however, are observed for the dropout model, but this is often the case and has been observed elsewhere (Molenberghs and Kenward, 2007).

Comparison of the errors (MSE), however, indicate consistently better estimation for SEM, except for the parameters β_0 and u_{11}, which show slightly smaller MSEs under direct likelihood. It can be noted that computation of the latter two involve only the first measurement, which is always observed, requiring no special treatment for addressing missingness, and as such, are not expected to differ beyond random variability under the two methods. With regard to computing time, using the same initial values, the solution for the first sample of the simulation was obtained much more quickly with SEM (45 min.) in comparison with direct likelihood (roughly 3 hours).

8.2.2 Non-Monotone Missingness

The Diggle and Kenward (1994) model lends applicability to continuous longitudinal data with dropout, and thus for intermittent or non-monotone missingness, an alternative model framework is necessary. Troxel, Harrington and Lipsitz (1998) proposed a full likelihood method for analyzing continuous longitudinal data with non-ignorable, non-monotone missing values, using a selection model consisting of a multivariate normal model for $\boldsymbol{Y}_i$ having a first-order antedependence structure. Owing to this assumption, the full joint distribution of the responses reduces to the product of conditional densities of the form $f(y_{ij}|y_{i,j-1})$, and, multiple integrals over the missing values also factor out as products of single integrals – thereby simplifying considerably the form of the observed-data likelihood. For the missingness model, each response indicator was formulated using a logistic model, for which the missingness probabilities depend on the current, possibly unobserved, values of the response. They further assumed independence of the response indicators, not only enabling the specification of the full joint distribution of $\boldsymbol{R}_i$ via the marginal distributions of the R_{ij}, but also precluding the need to model any association between them. These assumptions, though yielding much more tractable expressions for the observed-data likelihood, may be somewhat restrictive, especially when the outcomes require a more general dependence structure and/or when the response indicators necessitate some dependence among them. And yet despite the simplifications considered, Troxel, Harrington and Lipsitz (1998) noted that computations can become quite complicated and intractable for more than three or four repeated measurements.

For this simulation study, for the case of intermittent missingness, the selection model in Sotto *et al.* (2009b), which combines a multivariate normal model for the measurements $\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}_i, V_i)$ with a multivariate Dale model (Molenberghs and Lesaffre, 1994), described in Section 3.7.2, for the non-responses $\boldsymbol{R}_i$ given the measurements, is considered. As no particular assumptions are made on the covariance structure within the responses, nor on the independence of the non-response indicators, the proposed model provides a more general modeling framework than that presented by Troxel, Harrington and Lipsitz (1998).

8.2.2.1 Data Generation

For the simulation for the case of non-monotone missingness, bivariate Gaussian outcomes $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})'$, a treatment indicator $x_i(x_i = 0, 1)$, and two time periods $t_{ij}(t_{ij} = 1, 2)$, were assumed to be related according to the following:

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_0 \;+\; \beta_x\, x_i \;+\; \beta_t\, t_{ij} \;+\; \beta_{xt}\, x_i\, t_{ij},$$

where $\beta_0 = 2.0, \beta_x = 0, \beta_t = 0.5$ and $\beta_{xt} = 1.0$. A total of $S = 100$ samples, each of size $N = 300$ (equally distributed between the two treatment arms), was generated from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the mean vector and pre-defined unstructured covariance matrix for $\boldsymbol{Y}_i$ are:

$$E(\boldsymbol{Y}_i|x_i) \;\equiv\; \boldsymbol{\mu}_i \;=\; \begin{cases} (2.5, 3.0)' & \text{for } x_i = 0 \\ (3.5, 5.0)' & \text{for } x_i = 1 \end{cases} \qquad \text{and} \quad \boldsymbol{\Sigma}_i \;=\; \begin{pmatrix} 0.7 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}. \tag{8.5}$$

To induce non-monotone missingness, a bivariate Dale model was defined for the non-response vector $\boldsymbol{R} = (R_1, R_2)'$, i.e.,

$$\begin{aligned} \text{logit } p_{0+} \;&=\; \text{logit } P(R_{i1} = 0|x_i, y_{i1}, \boldsymbol{\psi}) \;=\; \psi_{a0} \;+\; \psi_{ax}\, x_i \;+\; \psi_{a1}\, y_{i1}, \\ \text{logit } p_{+0} \;&=\; \text{logit } P(R_{i2} = 0|x_i, y_{i2}, \boldsymbol{\psi}) \;=\; \psi_{b0} \;+\; \psi_{bx}\, x_i \;+\; \psi_{b2}\, y_{i2}, \\ \ln \psi \;&=\; \ln\left(\frac{p_{00}p_{11}}{p_{01}p_{10}}\right), \end{aligned} \tag{8.6}$$

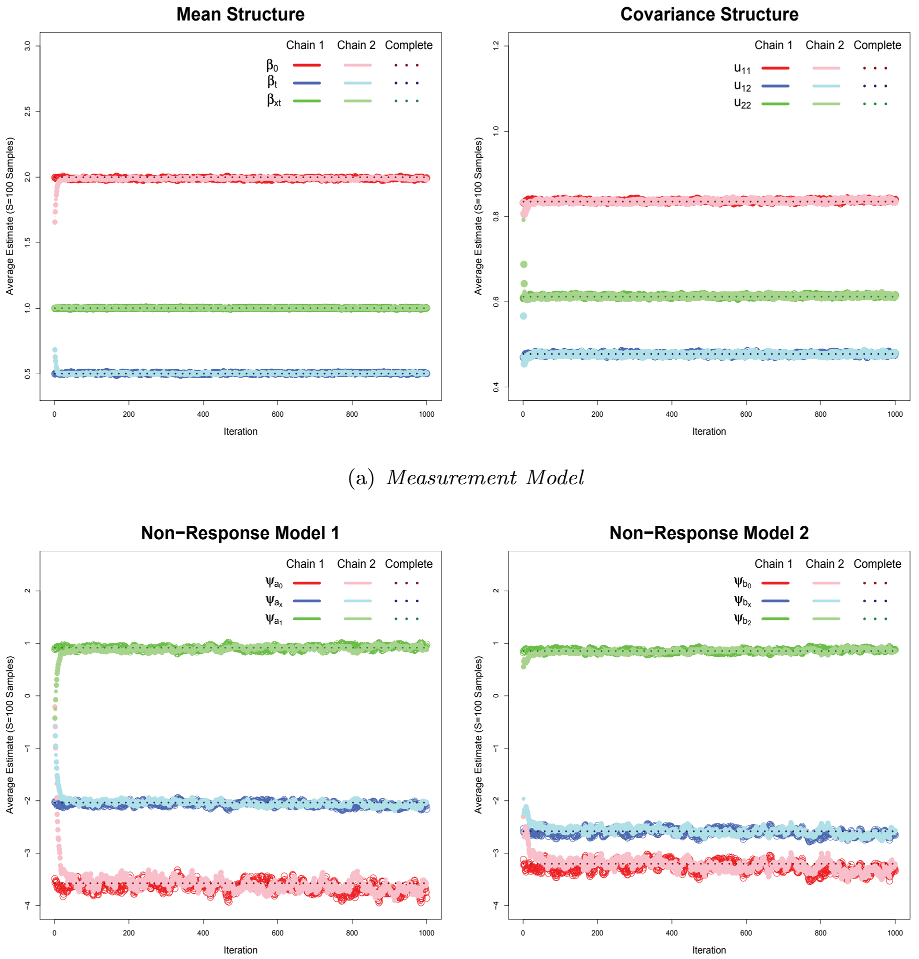
where $\psi_{a0} = -3.5, \psi_{ax} = -2.0, \psi_{a1} = 0.9, \psi_{b0} = -3.0, \psi_{bx} = -2.5, \psi_{b2} = 0.8$ and $\psi = 3$. Across the $S = 100$ samples, this missingness model led to about 63% completers, 19% with the second response missing, 8% with the first response missing and 9% with both responses missing.

With u_{11}, u_{12}, and u_{22} denoting the elements of the Cholesky decomposition of $\boldsymbol{\Sigma}$, and excluding β_x ($\beta_x = 0$) as in the monotone case, the full parameter vector for this Gaussian-Bivariate Dale model is given by:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\psi}')' = (\beta_0, \beta_t, \beta_{xt}, u_{11}, u_{12}, u_{22}, \psi_{a0}, \psi_{ax}, \psi_{a1}, \psi_{b0}, \psi_{bx}, \psi_{b2}, \ln\psi)'\,.$$

8.2.2.2 Results

For each of the $S = 100$ generated samples with non-monotone missingness, models (8.5) and (8.6) were fitted using the stochastic EM algorithm. Two parallel chains, each with $2n = 1000$ iterations, were simulated for each sample. For each of the simulated chains, the average estimate per iteration was plotted against the iteration number (Figure 8.2) and these plots indicated fairly acceptable convergence for almost



(a) *Measurement Model*

(b) *Non-Response Model*

Figure 8.2: *Simulated Data (Non-Monotone). Average parameter estimate per iteration, over the* $S = 100$ *samples, for each of the two simulated stochastic EM algorithm chains and for the complete data, for the Gaussian-Bivariate Dale model.*

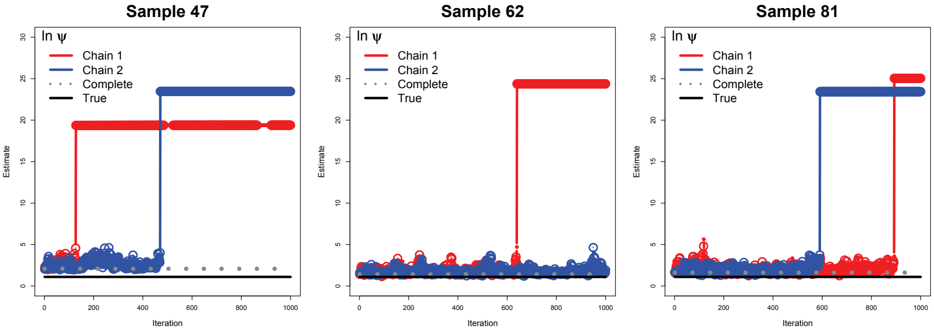


Figure 8.3: *Simulated Data (Non-Monotone). Estimates for the association parameter,* $\ln \psi$*, for each of the two simulated stochastic EM algorithm chains and for the complete data, for the three divergent samples, along with the true parameter value, for the Bivariate Dale Model.*

all the parameters. Computation of Gelman and Rubin (1992)'s PSRF (8.3) for all samples confirmed this for all parameters, with the exception of the association parameter, $\ln \psi$, which yielded a PSRF > 1.2 for 3 samples, possibly indicating non-convergence. The parallel chains for these 3 samples are shown in Figure 8.3, which clearly illustrates divergence of the association parameter, $\ln \psi$, for one or both chains.

A comparison of the average of the two simulated chains was done using all $S = 100$ samples and excluding the 3 divergent samples (i.e., using only $S' = 97$ samples). For each chain, at each of the 1000 iterations, the 100 (or 97) estimates for $\ln \psi$ were averaged to get the mean estimate per iteration. The latter was then plotted against the iteration number for each of the two chains using $S = 100$ and $S' = 97$ samples. The results are depicted in Figure 8.4. For Chain 1, when considering all samples (top left graph), it can be observed that the chain starts out relatively stable and close to the complete (and true) parameter values. At around iteration 150, however, the mean estimate moves to a somewhat larger value. The same happens again at iterations 650 and 900. In contrast, when the 3 divergent samples are excluded (bottom left graph), the chain remains stable all throughout. The same was observed for Chain 2. Closer inspection of the said samples did not seem to indicate any peculiarities, for instance, with regard to sparseness in some of the non-monotone missingness patterns. In an attempt to resolve the observed divergence, the chains were extended to 3000 iterations for these 3 identified samples, but no improvement was attained. Because fairly stable estimates for $\ln \psi$ were obtained for the remaining $S' = 97$ samples, exclusion of the 3 divergent samples was considered. Moreover, for some other parameters of the bivariate Dale missingness model, specifically the intercepts ψ_{a0} and ψ_{b0}, stability of the estimates also improved with the exclusion

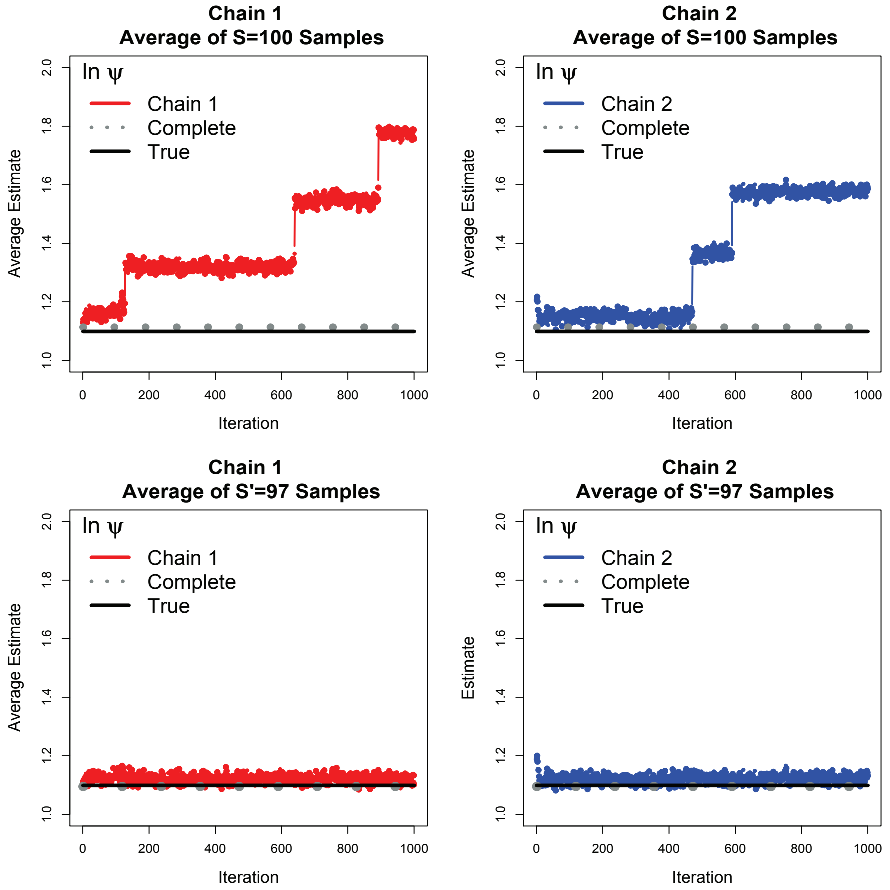


Figure 8.4: *Simulated Data (Non-Monotone). Average estimate of the association parameter,* $\ln\psi$*, per iteration, over* $S = 100$ *and over* $S' = 97$ *samples, for each of the two simulated stochastic EM algorithm chains and for the complete data, along with the true parameter value, for the Bivariate Dale Model.*

of the 3 divergent samples. The measurement model parameter estimates, however, remained more or less stable.

For each of the remaining $S' = 97$ samples, a pooled estimate was obtained by averaging the last 500 iterations from both chains, and the covariance matrix was calculated as described in Section 8.1.2. Finally, the estimates and covariance matrices for the $S' = 97$ samples were averaged to get the final estimates and quadratic errors (MSE), presented in Table 8.2. Both the true parameters used in data generation, as well as the parameters computed from the complete data of the $S' = 97$ samples, are provided.

Table 8.2: *Simulated Data (Non-Monotone). True parameter values, average parameter estimate for $S' = 97$ complete samples, for stochastic EM* (SEM) *and for direct likelihood* (DL), *along with mean squared errors* (MSE), *for the Gaussian-Bivariate Dale model.*

Effect	Parameter	True	Complete	Estimate		MSE $\times 10^2$	
				SEM	DL	SEM	DL
Measurement model							
Intercept	β_0	2.00	2.00	1.99	2.01	0.55	1.04
Time	β_t	0.50	0.50	0.50	0.52	0.18	0.55
Treatment	β_{xt}	1.00	1.00	1.00	0.99	0.16	0.30
Variance 1	u_{11}	0.84	0.84	0.84	0.85	0.10	0.20
Covariance 1,2	u_{12}	0.48	0.48	0.48	0.49	0.13	0.29
Variance 2	u_{22}	0.61	0.61	0.61	0.64	0.05	0.24
Missingness model							
Intercept for R_1	ψ_{a0}	-3.50	-3.57	-3.52	-5.51	27.56	584.31
Treatment for R_1	ψ_{ax}	-2.00	-2.04	-2.00	-2.69	16.73	87.69
Current for R_1	ψ_{a1}	0.90	0.92	0.88	1.50	3.32	53.37
Intercept for R_2	ψ_{b0}	-3.00	-3.20	-3.19	-4.32	28.15	309.96
Treatment for R_2	ψ_{bx}	-2.50	-2.58	-2.55	-3.24	20.03	121.09
Current for R_2	ψ_{b2}	0.80	0.86	0.84	1.18	2.64	25.58
Association	$\ln \psi$	1.10	1.10	1.12	0.77	11.96	33.50

Both approaches give again very similar estimates for the parameters of the measurement model and these values show only slight bias with respect to the complete parameters. In contrast, quite a bit of discrepancy between the two methods can be observed for the parameter estimates of the non-response model. Moreover, as was observed for the monotone case, quadratic errors are also consistently smaller for SEM. Given that both methods are consistent, the latter observations seem to indicate that the SEM algorithm is better able to overcome small-sample bias (with respect to the complete parameters) compared to direct likelihood. Computationally, with the same starting values, whereas estimates for the first generated sample were obtained using SEM in about 4 hours, the direct likelihood approach took only 1 hour.

8.3 Case Studies

In this section, application of the stochastic EM algorithm is illustrated for two longitudinal data sets, one with dropout and another with non-monotone missingness. The results are further compared with those obtained using direct likelihood.

8.3.1 Monotone: Age-Related Macular Degeneration Trial

A subset of the ARMD data, consisting of 226 subjects having either a complete set of responses or dropout type of missingness, was considered for analysis. For the 4 repeated measurements, a saturated means model was formulated, with the treatment (interferon-α or placebo) received by the patient, denoted x_i, as the primary covariate of interest, and assuming an unstructured covariance for the repeated measures, i.e.,

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_{0j} + \beta_{1j}\, x_i\, I_{t_{ij}}(j) \qquad \text{and} \qquad \Sigma_i \;=\; [\sigma_{rc}]\,,$$

for $i = 1, 2, \ldots, 226$ and $j, r, c = 1, 2, 3, 4$. For dropout, in line with the Diggle and Kenward (1994) model (Section 3.7.4), a logistic dropout model of the form (3.31), was used. The Diggle and Kenward (1994) measurement and dropout models were fitted using the SEM algorithm, generating two chains, each consisting of $M = 3000$ iterations. Convergence of the chains was first checked using the PSRF (Gelman and Rubin, 1992), all of which indicated fairly stable estimates. The SEM estimates were then computed using the last halves of these 2 chains and are presented in Table 8.3. Direct-likelihood estimates and standard errors are also shown.

For both models, all SEM estimates, as well as corresponding standard errors, were quite comparable with their direct likelihood counterparts, with a slight discrepancy observed for the intercept of the dropout model. The similarity in the resulting estimates under the two approaches might be attributed to the fact that the data contained only a very modest amount (17%) of missingness. Computationally, the SEM approach took a lot less time to converge than direct likelihood – 6 hours *vs.* 18 hours, respectively – given the same initial values.

8.3.2 Non-Monotone: Hepatitis B Virus Immunization Trial

For the analysis of the HBV data, the (log-scaled) anti-HBs antibody levels of 214 patients at months 60 and 61 was considered. Along with completers, all non-monotone patterns for these two time points were included in the analysis. For the bivariate measurements, a saturated means model was again formulated, with the number of booster shots received by the patient during the immunization period, denoted x_i,

Table 8.3: *Age-Related Macular Degeneration Trial (Monotone). Parameter estimates (standard errors) for stochastic EM* (SEM) *and for direct likelihood* (DL)*, for the Diggle-Kenward model.*

Effect	Parameter	SEM Estimate (s.e.)	DL Estimate (s.e.)
Measurement model			
Intercept 4	β_{01}	54.00 (1.47)	53.97 (1.47)
Intercept 12	β_{02}	52.98 (1.59)	52.96 (1.59)
Intercept 24	β_{03}	49.08 (1.72)	49.06 (1.73)
Intercept 52	β_{04}	43.58 (1.73)	43.58 (1.82)
Treatment 4	β_{11}	-3.11 (2.10)	-3.13 (2.09)
Treatment 12	β_{12}	-4.66 (2.26)	-4.68 (2.28)
Treatment 24	β_{13}	-3.79 (2.43)	-3.80 (2.49)
Treatment 52	β_{14}	-5.73 (2.43)	-5.66 (2.63)
Dropout model			
Intercept	ψ_0	-1.42 (0.39)	-1.80 (0.47)
Previous	ψ_p	0.01 (0.01)	0.01 (0.02)
Current	ψ_c	-0.04 (0.01)	-0.04 (0.02)

as covariate, and further assuming an unstructured covariance for the two repeated measures, i.e.,

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_{0j} \,+\, \beta_{1j}\, x_i\, I_{t_{ij}}(j) \qquad \text{and} \qquad \Sigma_i \;=\; [\sigma_{rc}]\,,$$

for $i = 1, 2, \ldots, 214$ and $j, r, c = 1, 2$. The non-monotone missingness in the non-response vector, $\boldsymbol{R} = (R_1, R_2)'$, was modeled using a bivariate Dale model of the form:

$$\begin{aligned} \text{logit}\, p_{0+} &= \text{logit}\, P(R_{i1} = 0|y_{i1}, \boldsymbol{\psi}) \;=\; \psi_0 \,+\, \psi_c\, y_{i1}, \\ \text{logit}\, p_{+0} &= \text{logit}\, P(R_{i2} = 0|y_{i2}, \boldsymbol{\psi}) \;=\; \psi_0 \,+\, \psi_c\, y_{i2}, \\ \ln \psi &= \ln\left(\frac{p_{00}p_{11}}{p_{01}p_{10}}\right). \end{aligned} \tag{8.7}$$

The SEM algorithm was applied, again using two chains of $M = 3000$ iterations each, and the (Gelman and Rubin, 1992) approach suggested convergence for all the parameters. Table 8.4 shows the results for both SEM and direct likelihood. As with

Table 8.4: *Hepatitis B Virus Immunization Trial (Non-Monotone). Parameter estimates (standard errors) for stochastic EM* (SEM) *and for direct likelihood* (DL), *for the Gaussian-Bivariate Dale model.*

Effect	Parameter	SEM Estimate (s.e.)	DL Estimate (s.e.)
Measurement model			
Intercept 60	β_{01}	4.69 (0.17)	4.70 (0.27)
Intercept 61	β_{02}	10.98 (0.23)	10.98 (0.44)
No. of Booster Shots 60	β_{11}	-0.37 (0.10)	-0.37 (0.13)
No. of Booster Shots 61	β_{12}	-0.87 (0.13)	-0.86 (0.24)
Missingness model			
Intercept	ψ_0	-1.08 (0.19)	-1.08 (0.20)
Current	ψ_c	0.07 (0.02)	0.07 (0.02)
Association	$\ln \psi$	4.23 (0.51)	4.22 (0.52)

the ARMD case study, both methods again yield estimates that are generally comparable, for both the response and non-response models, but the standard errors under SEM are consistently smaller. Larger discrepancies between the two methods might have been expected, as was observed for the non-monotone simulations. It should be noted, however, that the missingness model formulated here (8.7) is considerably simpler than that used in the simulation study (8.6), in the sense that a common intercept and common slope for the current measurement was used in the logistic models of both response indicators (a choice arising from some preliminary model building exercises on the data). As a result, this simplified model is probably unable to bring out any strengths and/or weaknesses of either method. At the same initial parameters, running time was roughly 3 hours for SEM as opposed to about 1 hour for direct likelihood.

8.4 Discussion

In this chapter, the performance of the stochastic EM algorithm – a modification of the EM algorithm whereby the E-step of the latter is replaced by a simulation step for the missing data – in fitting non-ignorable selection models for continuous longitudinal data with monotone or non-monotone missingness was assessed. Stochastic EM is a convenient alternative to the deterministic EM algorithm in the sense that integration of the complete-data likelihood over the missing values is altogether avoided since the S(imulation) step renders the data "pseudo"-complete. Via a simulation study, the SEM approach was shown to yield estimates with relatively smaller MSEs compared to direct likelihood, which exhibited substantial small-sample bias for the parameters of the non-response model. Though no categorical statement can be made regarding the superiority of either method in terms of computing time, it might be expected that what was observed for the monotone case (SEM converging much more quickly) will probably also be true for the non-monotone case with more repeated measures (i.e., $n > 2$), since, in the latter case, the form of the observed-data likelihood would require more integral evaluations than what has been considered here. Moreover, in light of the large errors observed under direct likelihood for the non-monotone case, one might be more inclined to advocate the use of the stochastic approach, despite its slower convergence.

The simulations and the case studies both suggested that the SEM algorithm yielded fairly stable results for the measurement model, but less so for the non-response model. In some situations, despite non-convergence of some missingness model parameter estimates, those for the measurement model remained close to the true values, implying some amount of robustness for the measurement model parameters. While it is indeed true that in the MAR case, the ignorability assumption necessarily implies that the measurement process and non-response process parameters are independent, one would not expect the same under MNAR. Apparently, however, the dependence does not appear to be very strong.

Though, at first sight, the instability of the parameter estimates for the missingness model might be perceived as an obstacle to the use of the SEM algorithm, such parameters are typically not of scientific interest, often being considered nuisance parameters. They are, however, necessary, within the context of incomplete longitudinal data, to avoid bias in the analysis. So although non-response is modeled explicitly, it is uncommon that a researcher is particularly interested in further examination of this model. Moreover, as researchers are primarily interested in the results for the measurement model, stochastic EM can most probably provide relatively stable results. And, in addition, whenever divergence does not present itself to be an issue, the non-response model parameters are, in fact, quite well estimated using the stochastic EM approach.

It is also important to note at this point some pragmatic considerations when using the stochastic EM algorithm. Regarding the number of iterations, one might consider initially running a chain to some moderately large value, say $2n = 1000$, and assessing its convergence (or lack thereof). If the chain has not seemed to achieve some stationarity, additional iterations can be made. Also, since it is virtually impossible to detect convergence using a single chain, it is strongly recommended to run the algorithm using at least 2 chains, assessing convergence using the approach of Gelman and Rubin (1992). With 3 or more chains, when divergence of one chain is observed, the SEM estimate can still be computed based on the other convergent chains. In addition, the burn-in period can be reduced by a wise choice of starting parameters. A practical approach would be to use, as starting values for the MNAR model, the resulting parameter estimates from an MAR model, which can be easily implemented using standard likelihood techniques, and some small arbitrary values for the additional non-ignorable parameters. Finally, with respect to the modeling framework, a Gaussian model for the repeated measures was combined here with a logistic model for dropout for the monotone case (Diggle and Kenward, 1994), whereas for the non-monotone case, the use the use of a Gaussian model for the longitudinal responses and a multivariate Dale model for the non-responses was proposed. Such a joint model is quite general, in comparison with, for instance, that proposed by Troxel, Harrington and Lipsitz (1998). A possible drawback, however, of a Gaussian-multivariate Dale type model would be a likely proliferation of parameters. For the simulations conducted here, considering only 2 time points, the missingness model alone consisted of 7 parameters. One can easily anticipate much more parameters for longer longitudinal sequences.

On a final note, it is important to emphasize that models for continuous longitudinal data with non-ignorable monotone or non-monotone missingness can be approached using either stochastic or non-stochastic methods. While the latter is frequently viewed as being the convention, the former provides a more computationally efficient approach to obtaining likelihood-based estimates, with, as our simulations showed, smaller magnitudes of error. In addition, the stochastic approach allows greater flexibility in the choice of joint model for $\boldsymbol{Y}$ and $\boldsymbol{R}$, since one need not be restricted by numerical complexities that would arise in the integration of likelihoods arising from more general, and consequently, more complicated, joint models. A stochastic solution can be obtained without such restrictions, especially so for the measurement model, and barring divergence, also for the non-response model. In general, whether applied to Gaussian longitudinal data with non-ignorable monotone or non-monotone missingness, the SEM algorithm appears to be not only a practical alternative to direct likelihood approaches, but also a useful tool within the context of a sensitivity analysis.

9

Comparison of Various Software Tools for Imputation

Statistical methods that address missingness have been extensively studied in recent years. In general, incompleteness in the data can be addressed either by pre-processing the data in order to render them complete prior to any analysis, i.e., by imputation of the missing values, or by employing methods that deal with the missingness directly in the analysis (e.g, via the use of a weighting scheme). Imputation, in its broadest sense, has perhaps become the more widely chosen route for dealing with missing data. The popularity of imputation methods can be attributed not only to its relatively easy to comprehend rationale, but also to its rather straightforward implementation. In addition, because imputation "solves" the missing data problem prior to the analysis, complete-data techniques can subsequently be applied without much further treatment. Moreover, to less technically equipped end users, more complex approaches for handling missing data directly in the analysis may seem much less appealing than imputation.

The widespread popularity of imputation has spurred the development of software-based implementation routines, procedures, or packages capable of generating imputations for incomplete data. Thus far, evaluation of existing implementations have frequently centered on the resulting parameter estimates of the prescribed model of interest after imputing the missing data. In some situations, however, interest may very well be on the quality of the imputed values at the level of the individual – an issue that has received relatively little attention.

In this chapter, in the particular setting of longitudinal data with dropout, the performance of different routines, procedures, or packages for imputing missing data is investigated. Evaluation of these existing implementations shall be done not only in terms of the resulting estimates, after imputing the missing data, for the model of interest, but also with respect to the quality of the generated imputations at the level of the individual. Via a simulation study, conducted under two different missingness mechanisms (missing at random and missing not at random), resulting imputations from the different routines are assessed at the level of the parameter estimates as well as at the level of the individual. The simulation study is further complemented with a comparison of the various imputation routines as applied to a case study.

9.1 Overview of Software Implementation Routines for Imputation

Imputation, which can be done singly or multiply, encompasses an entire scope of techniques that have been developed to make inferences about incomplete data, ranging from very simple strategies (e.g., mean imputation) to more advanced approaches that require, for instance, estimation of posterior distributions using MCMC methods. Additional complexity arises when the number of missingness patterns increases and/or when both categorical and continuous random variables are involved. Over the recent years, software-based implementations for these various imputation strategies have become widely available. An overview of some of the existing routines are presented in Table 9.1.

Horton and Kleinman (2007) evaluated a number of software routines in the context of a simple logistic regression analysis with missing values in the covariates. A slightly more extensive review of various software packages in Horton and Lipsitz (2001) investigated missingness in covariates as well as missingness in a single univariate outcome, albeit separately. For the review considered here, the more general situation of longitudinal outcomes is explored, with particular attention on missingness in the outcomes, rather than in the covariates. Moreover, among the procedures listed in Table 9.1, only those in boldface type shall be investigated, and for each of

Table 9.1: *Overview of some statistical software implementations for imputing missing data.*

Routine Name	Software package	Provider/Author	Approaches Implemented
Amelia II	**R**	**Honaker, King and Blackwell**	**Hybrid of EM with Bootstrap (`amelia`)**
Hmisc	**R and S-Plus**	**Frank Harrell**	**Multiple Imputation using Additive Regression, Bootstrapping and Predictive Mean Matching (`aregImpute`)**
ICE	Stata	Patrick Royston	Chained equation, imputing missing values by sampling from the posterior predictive distribution
IVEware	SAS or Standalone	Univ. of Michigan	Sequence of regression models, incorporating restriction to a relevant subpopulation and logical bounds for the imputation
LogXact	LogXact 7	Cytel	Maximum likelihood
MICE	R and S-Plus	van Buuren *et al.*	Chained equation (and potential MNAR models)
PROC MI	**SAS v9.1**	**SAS Institute**	**MCMC for Gaussian, Predictive Mean Matching, regression, logistic, polytomous and discriminant models**
Missing Data library	S-Plus 7	Insightful	Maximum likelihood or conditional Gaussian imputation model
randomForest	**R**	**Liaw and Wiener**	**Imputation of missing values using proximities from a Random Forest (`rfImpute`)**
SEM imputation	**R**	**Sotto *et al.***	**MCMC for non-standard dist'n., imputing missing values by sampling from the posterior predictive distribution**
SOLAS	Solas	Statistical Solutions	Supports predictive mean model (closest observed value to the predicted value), propensity score models and discriminant models for missing binary and categorical variables
yaimpute	R	Crookston and Finley	Supports nearest neighbor (NN) imputation, nearness based on distance; also provides random forest imputation (`yai`)

these, a general introduction is provided along with discussions regarding any particular issues related to its implementation.

It is important to mention that, despite the existence of numerous methods and software implementations for imputing incomplete data (Table 9.1), there is evidence that their use in applied settings remains limited. Burton and Altman (2004) reviewed the reporting (and handling) of missing data in 100 cancer prognostic studies published in the year 2002. From this total, missing data was reported in 81 articles, of which only 32 mentioned the use of some method to deal with the incomplete observations. Among these, the complete case approach, despite its potential for biased results, among other issues, and the available case method were the most commonly used tools for addressing missingness (12 articles each). Another 6 articles simply omitted variables with missing data, 4 included cases with missing values as a separate category within the modeling procedure, 3 used an *ad hoc* single-imputation procedure based on a surrogate variable or on median values from the non-missing observations, and only 1 paper applied multiple imputation.

Another article (Horton and Switzer, 2005) reviewed the use of statistical methods in research articles published between 2004 and 2005 in *The New England Journal of Medicine*. The authors reviewed a total of 331 papers, of which only 26 (8%) reported some way of handling missing data in their analysis. From the 26 manuscripts, 12 used a version of last observation carried forward (LOCF), 13 used an *ad hoc* imputation strategy (e.g., mean imputation), and 2 performed a kind of sensitivity analysis, by replacing missing values with worst case values.

Both of these reviews highlight the fact that, even when statistical methodology to tackle missing data problems exists, these are not frequently employed in practice. Of even greater concern is the use of incorrect statistical analyses, in spite of the clear messages in a suite of research articles arguing against complete case analysis or LOCF, for instance.

9.1.1 R Package: Amelia II

Amelia II (Honaker, King and Blackwell, 2008) provides features to "multiply impute" missing data in a single cross-sectional study, from time series data, or from a time-series-cross-sectional data set. It implements a bootstrapping-based algorithm (King *et al.*, 2001) that provides essentially the same answers as the standard Imputation Posterior (IP) algorithm (i.e., drawing random simulations from the multivariate normal observed data posterior using MCMC methods) or *EMis* approaches. The algorithm first bootstraps a sample data set with the same dimensions as the original data, estimates the sufficient statistics (with priors, if specified) by EM, and then imputes the missing values in the sample. It repeats this process m times to produce the m complete data sets, where the observed values are the same and the unobserved values are drawn from their posterior distributions. Amelia II is usually considerably faster than existing approaches and can handle a large number of variables. The program also generalizes existing approaches by allowing for trends in time series across observations within a cross-sectional unit, as well as priors that allow experts to incorporate beliefs they have about the values of missing cells in their data. Furthermore, the Amelia II package also includes useful diagnostics to help assess the fit of the imputation models.

9.1.2 R Package: Hmisc

The Hmisc package (Harrell, 2009) contains the function `aregImpute`, which can be used to perform multiple imputation for missing values. The function uses bootstrapping to approximate the process of drawing predicted values from a full Bayesian predictive distribution. Different bootstrap resamples are used for each of the multiple imputations. In other words, for the i^{th} imputation of a possibly incomplete variable, $i = 1, 2, \ldots, m$, a flexible additive model is fitted on a sample with replacement from the original data and this model is used to predict all of the original missing and non-missing values for the target variable. The sequence of steps used by the `aregImpute` algorithm consists of the following:

- For each incomplete variable, missing values are initialized to values obtained from a random sample, of the non-missing values (without replacement, if sufficient number of non-missing values exists), of size corresponding to the number of missing values for that variable.
- The next steps need to be repeated for a total of $(n_{\text{burn-in}} + m)$ iterations and only imputations from the last m iterations are saved.
  - For each incomplete variable, a sample with replacement is drawn from the observations in the entire data set for which the current variable being imputed is non-missing. A flexible additive model is then fitted and subsequently used to predict the target variable in all of the original observations. Imputation can be done using different ways of matching.
  - After the imputations are obtained, these randomly drawn imputations are used in the next iteration, this time taking the current target variable as a predictor of other sometimes-missing variables.

Some limitations of the procedure arise from the fact that predictive mean matching may not work well when fewer than three variables are used to predict the target variable, because many of the multiple imputations for an observation will be identical. When the missingness mechanism for a variable is so systematic that the distribution of observed values is truncated, predictive mean matching does not work. It will only yield imputed values that are near observed values, so intervals in which no values are observed will not be populated by imputed values.

9.1.3 R Package: randomForest

A *random forest* is an ensemble of many identically distributed trees generated from bootstrap samples of the original data. Each tree is constructed via a tree classification algorithm. The simplest random forest with random features is formed by selecting randomly, at each node, a small group of input variables to split on. The size of the group is fixed throughout the process of growing the forest. Each tree is grown using the CART (classification and regression trees) methodology without pruning (Breiman *et al.*, 1984). Using the trees, a proximity matrix is constructed, with proximities based on the number of times a pair of observations ends up in the same node, divided by two times the number of trees grown. This proximity measure is then used to update the missing values. In the case of continuous variables, missing values are imputed using a weighted average of non-missing observations, with proximities as weights. For categorical variables, the category with the largest average proximity is used to impute the missing values. It is recommended to repeat this step and update the proximity measures based on the new imputation.

The randomForest package in R (Breiman and Cutler, 2009) can be used to impute missing observations using the function `rfImpute` or `na.roughfix`. The latter simply replaces the missing values by their variable medians or most frequent level and is already incorporated as a first step in the function `rfImpute`. In the second step of `rfImpute`, the random forest method (Breiman, 2001) is used on the complete data.

9.1.4 SAS Procedure: PROC MI

The SAS procedure MI (SAS Institute Inc., 2004), or PROC MI, creates, as do the previous R packages, multiply imputed data sets for incomplete multivariate data. It uses methods that incorporate appropriate variability across the m imputations. The method of choice depends on the patterns of missingness and three different possibilities are available.

- The *Markov Chain Monte Carlo* (MCMC) approach for arbitrary missing data patterns, uses an MCMC – which simulates random draws from non-standard distributions via Markov chains (Schafer, 1997) – to create a small number of independent draws of the missing data from a predictive distribution, and these draws are then used for multiple-imputation inference. In PROC MI using MCMC, multiple imputations are drawn from a Bayesian predictive distribution for normal data. The method can also be used to impute enough values to make the missing data pattern monotone, enabling the subsequent application of a more flexible imputation method.

- The *regression method* for monotone missing data fits a regression model for each incomplete variable, with the other variables as covariates. Based on the fitted regression coefficients, a new regression model is simulated from the posterior predictive distribution of the parameters and is used to impute the missing values for each variable (Rubin, 1987). The process is repeated sequentially for variables with missing values.

- The *propensity score method* for monotone missing data, generates, for each incomplete variable, a propensity score – generally defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin, 1983) – for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group.

9.1.5 Stochastic Expectation-Maximization (EM) Algorithm

The stochastic EM (or SEM) algorithm, described in Section 8.1.2, was developed by Celeux and Diebolt (1985) as a computational scheme to facilitate maximum likelihood estimation for incomplete data problems. As such, in a technical sense, the SEM algorithm in itself is not an imputation technique, but can, in fact, be used indirectly to impute missing observations. That is, once the resulting SEM generated Markov Chain attains reasonable convergence, a final pseudo-complete sample can be obtained by substituting, for the missing observations, random draws from the estimated conditional density of the missing values given the observed values at the final SEM parameter estimates.

The rationale behind the use of the SEM algorithm for imputing missing data is in fact equivalent to that of SAS PROC MI using the MCMC method. The latter, however, lends application primarily for multivariate Gaussian outcomes, whereas the SEM approach can be specifically tailored to admit more general forms (Sotto *et al.*, 2009b; Gad and Ahmed, 2006; Jolani and Ganjali, 2007).

9.2 Simulation Study

To evaluate the performance of the different routines in fitting ignorable and non-ignorable models for Gaussian longitudinal data with dropout, simulations for both scenarios were conducted. Data generating models, simulation settings and subsequent results under each case are described in the following subsections. A related simulation setting was already presented in Sotto *et al.* (2009b), with primary emphasis on the SEM algorithm as an estimation, rather than as an imputation, strategy. Moreover, these authors did not consider alternative software tools, neither did they address the MAR situation.

For the simulation study, assessment of the various imputation procedures shall be done in two respects. First, resulting parameter estimates arising from the imputed data using the different routines will be compared with the true parameters used in generating the simulated data. In addition, since imputations at the level of the individual are also of particular interest here, the performance of the different routines will be based on a so-called *standardized squared imputation bias* (SSIB) and the variability of this measure. The SSIB is calculated as follows:

$$SSIB_{ij} = \frac{\left(Y_{ij}^{A} - Y_{ij}^{I}\right)^{2}}{N - \sum\limits_{i=1}^{N} R_{ij}}, \tag{9.1}$$

where Y_{ij}^{A} is the actual generated value, Y_{ij}^{I} denotes the imputed value and the denominator is simply the number of missing values at time point j. This measure is calculated for all subjects in the data set having missing values and the sum over all subjects, for a particular time point, can be viewed as a measure indicating how biased the imputations for that time point are for each data set in the simulation.

9.2.1 Missing Not At Random

For the simulations under MNAR, in addition to direct likelihood, 5 imputation routines will be considered: the SEM algorithm, the MI procedure from SAS, and the Hmisc, random forest and Amelia II packages from R. While PROC MI, Hmisc and

Amelia II are known to be designed to deal with MAR models, non-ignorable missingness can be handled under direct likelihood, SEM and possibly random forest. The impact, in terms of parameter estimates and standard errors, will be further evaluated using two different strategies. In particular, Strategy I consists of imputing the missing observations first, then performing the analysis for each imputed data set, and finally combining the results; Strategy II, on the other hand, entails imputing the missing values and subsequently using the average of the multiple imputations to perform the final analysis. It should be noted that either Strategy I or Strategy II can be used for all the methods, except direct likelihood, for which neither strategy is applicable since no imputation of missing values is done.

9.2.1.1 Data Generation

For the simulation, trivariate Gaussian outcomes $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$, a treatment indicator $x_i (x_i = 0, 1)$, and 3 time periods $t_{ij} (t_{ij} = 1, 2, 3)$, were assumed to be related according to the following:

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_0 \,+\, \beta_t\, t_{ij} \,+\, \beta_{xt}\, x_i\, t_{ij}, \tag{9.2}$$

where $\beta_0 = 2.0, \beta_t = 0.5$, and $\beta_{xt} = 1.0$, giving rise to the following mean structure

$$E(\boldsymbol{Y}_i|x_i) \;\equiv\; \boldsymbol{\mu}_i \;=\; \begin{cases} (2.5, 3.0, 3.5)' & \text{for } x_i = 0 \\ (3.5, 5.0, 6.5)' & \text{for } x_i = 1 \end{cases},$$

with a pre-defined unstructured covariance matrix for $\boldsymbol{Y}_i$ given by:

$$\boldsymbol{\Sigma}_i \;=\; \begin{pmatrix} 0.7 & 0.4 & 0.2 \\ 0.4 & 0.6 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}.$$

A total of $S = 100$ samples, each of size $N = 300$ (equally distributed between the two treatment arms), was then generated from $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

To induce monotone missingness of an MNAR nature, in line with the Diggle and Kenward (1994) model (Section 3.7.4), a logistic dropout model of the form

$$\text{logit } P(D_i = j|D_i \geq j, y_{i,j-1}, y_{ij}, \boldsymbol{\psi}) \;=\; \psi_0 \,+\, \psi_c\, y_{ij} \,+\, \psi_p\, y_{i,j-1}, \tag{9.3}$$

was considered, with $\psi_0 = -4$, $\psi_p = -0.5$ and $\psi_c = 1$, the latter two denoting the coefficients for the previous and the current measurements, respectively. Over the $S = 100$ samples, dropout model (9.3) yields about 55% completers (pattern OOO) and about 22% each for the other two dropout patterns (OOM and OMM).

Table 9.2: *Simulated Data (Missing Not At Random). True parameter values, average parameter estimates* (est) *and standard errors* (s.e.) *for $S = 100$ complete samples, for direct likelihood* (DL)*, stochastic EM* (SEM) *and random forest* (RF)*, for the Diggle-Kenward model.*

Effect	Par	True	Complete	DL	**Strategy I**		**Strategy II**	
					SEM	RF	SEM	RF
				est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)
Measurement Model								
Intercept	β_0	2.0000	2.0005	2.016 (0.071)	2.032 (0.067)	2.062 (0.072)	2.032 (0.071)	2.062 (0.076)
Time	β_t	0.5000	0.4991	0.489 (0.033)	0.479 (0.028)	0.454 (0.034)	0.480 (0.033)	0.454 (0.034)
Treatment	β_{xt}	1.0000	1.0008	0.991 (0.035)	0.976 (0.023)	0.958 (0.039)	0.975 (0.037)	0.959 (0.041)
Variance 1	u_{11}	0.8367	0.8308	0.830 (0.034)	0.830 (0.034)	0.831 (0.036)	0.830 (0.037)	0.831 (0.037)
Covariance 1,2	u_{12}	0.4781	0.4779	0.475 (0.044)	0.472 (0.037)	0.470 (0.039)	0.471 (0.043)	0.470 (0.042)
Covariance 1,3	u_{13}	0.2390	0.2421	0.232 (0.051)	0.228 (0.034)	0.184 (0.034)	0.228 (0.048)	0.184 (0.033)
Variance 2	u_{22}	0.6094	0.6169	0.606 (0.029)	0.604 (0.022)	0.566 (0.025)	0.550 (0.032)	0.566 (0.029)
Covariance 2,3	u_{23}	0.3047	0.3080	0.302 (0.048)	0.298 (0.030)	0.235 (0.030)	0.271 (0.047)	0.235 (0.033)
Variance 3	u_{33}	0.5916	0.5909	0.582 (0.033)	0.576 (0.020)	0.539 (0.022)	0.565 (0.030)	0.543 (0.027)
Dropout Model								
Intercept	ψ_0	-4.0000	-4.1597	-3.718 (0.415)	-3.914 (0.345)	-3.743 (0.395)	-4.049 (0.533)	-3.741 (0.480)
Previous	ψ_p	-0.5000	-0.1008	-0.282 (0.192)	0.015 (0.136)	0.121 (0.146)	-0.093 (0.188)	0.113 (0.164)
Current	ψ_c	1.0000	0.8088	0.792 (0.168)	0.678 (0.110)	0.670 (0.118)	0.789 (0.176)	0.740 (0.127)

Table 9.3: *Simulated Data (Missing Not At Random). True parameter values, average parameter estimates* (est) *and standard errors* (s.e.) *for* $S = 100$ *complete samples, for PROC MI, Amelia II and Hmisc, for the Diggle-Kenward model.*

Effect	Par	True	Complete	PROC MI	Amelia II	Hmisc
				est (s.e.)	est (s.e.)	est (s.e.)
Measurement Model						
Intercept	β_0	2.0000	2.0005	2.072 (0.071)	2.280 (0.080)	2.075 (0.071)
Time	β_t	0.5000	0.4991	0.453 (0.033)	0.325 (0.041)	0.453 (0.034)
Treatment	β_{xt}	1.0000	1.0008	0.945 (0.041)	0.833 (0.045)	0.939 (0.040)
Variance 1	u_{11}	0.8367	0.8308	0.831 (0.037)	0.835 (0.038)	0.831 (0.037)
Covariance 1,2	u_{12}	0.4781	0.4779	0.472 (0.044)	0.480 (0.047)	0.467 (0.045)
Covariance 1,3	u_{13}	0.2390	0.2421	0.231 (0.055)	0.212 (0.056)	0.218 (0.051)
Variance 2	u_{22}	0.6094	0.6169	0.545 (0.030)	0.578 (0.038)	0.555 (0.031)
Covariance 2,3	u_{23}	0.3047	0.3080	0.275 (0.054)	0.215 (0.100)	0.275 (0.053)
Variance 3	u_{33}	0.5916	0.5909	0.463 (0.031)	0.697 (0.055)	0.466 (0.029)
Dropout Model						
Intercept	ψ_0	-4.0000	-4.1597	-3.598 (0.457)	-2.811 (0.523)	-3.611 (0.572)
Previous	ψ_p	-0.5000	-0.1008	0.103 (0.173)	1.061 (0.170)	0.120 (0.158)
Current	ψ_c	1.0000	0.8088	0.554 (0.129)	-0.400 (0.134)	0.544 (0.128)

9.2.1.2 Results

For each of the $S = 100$ generated samples with non-ignorable missingness, models (9.2) and (9.3) were fitted using 6 different approaches: direct likelihood, the SEM algorithm, the random forest approach, the MI procedure from SAS, and the Hmisc and Amelia II packages from R, with $M = 5$ imputations for the latter 5 cases. Focus shall first be restricted on the 3 approaches that are possibly capable of handling non-random missingness, namely, direct likelihood, SEM and random forest, the results of which, under both strategies, are presented in Table 9.2.

It can be observed that these 3 approaches produce generally comparable results, with somewhat larger bias under random forest, implying some sensitivity of this method to non-ignorable missingness. Biases, as well as standard errors, are also generally larger in magnitude for the parameters of the dropout model. With respect to two different strategies, similarity in their respective results suggests that the impact of using the two different strategies does not seem to be too large. Thus, for the other approaches, only the results obtained using Strategy II will be reported.

In Table 9.3, the results obtained for the other three approaches are shown. In this particular scenario, PROC MI and Hmisc produced comparable results in terms of estimates and standard errors for all parameters. Of considerable concern are the large biases observed in the Amelia II estimates for many of the parameters of both the measurement and dropout models, indicating that this package should be used with care, given that conclusions can really be affected, at least as far as dealing with MNAR is concerned.

A general remark about this simulation study is related to the results obtained for the dropout model, in particular, the estimated value obtained for the parameter ψ_p, denoting the coefficient for the previous outcome. First, it is worth noting the discrepancy between the value used for simulation and the value obtained for the complete data (average of $S = 100$ samples), indicating some amount of finite-sampling bias. Moreover, all methods except direct likelihood produced estimated values above zero, though the true parameter was negative, which might imply potential problems in imputing MNAR missing values under these routines. Analysis of the complete data, however, indicated this parameter to be insignificant ($p = 0.6391$), which is also reflected by five of the methods applied, thus possibly explaining why the methods failed to yield a satisfactory estimate for ψ_p.

Focusing now on the SSIB for all 5 methods (excluding direct likelihood), it can be deduced from Figure 9.1 that the sum of SSIB ranges from 0.2 to 1.6 for the second time point (having around 22% missing observations), indicating that the total

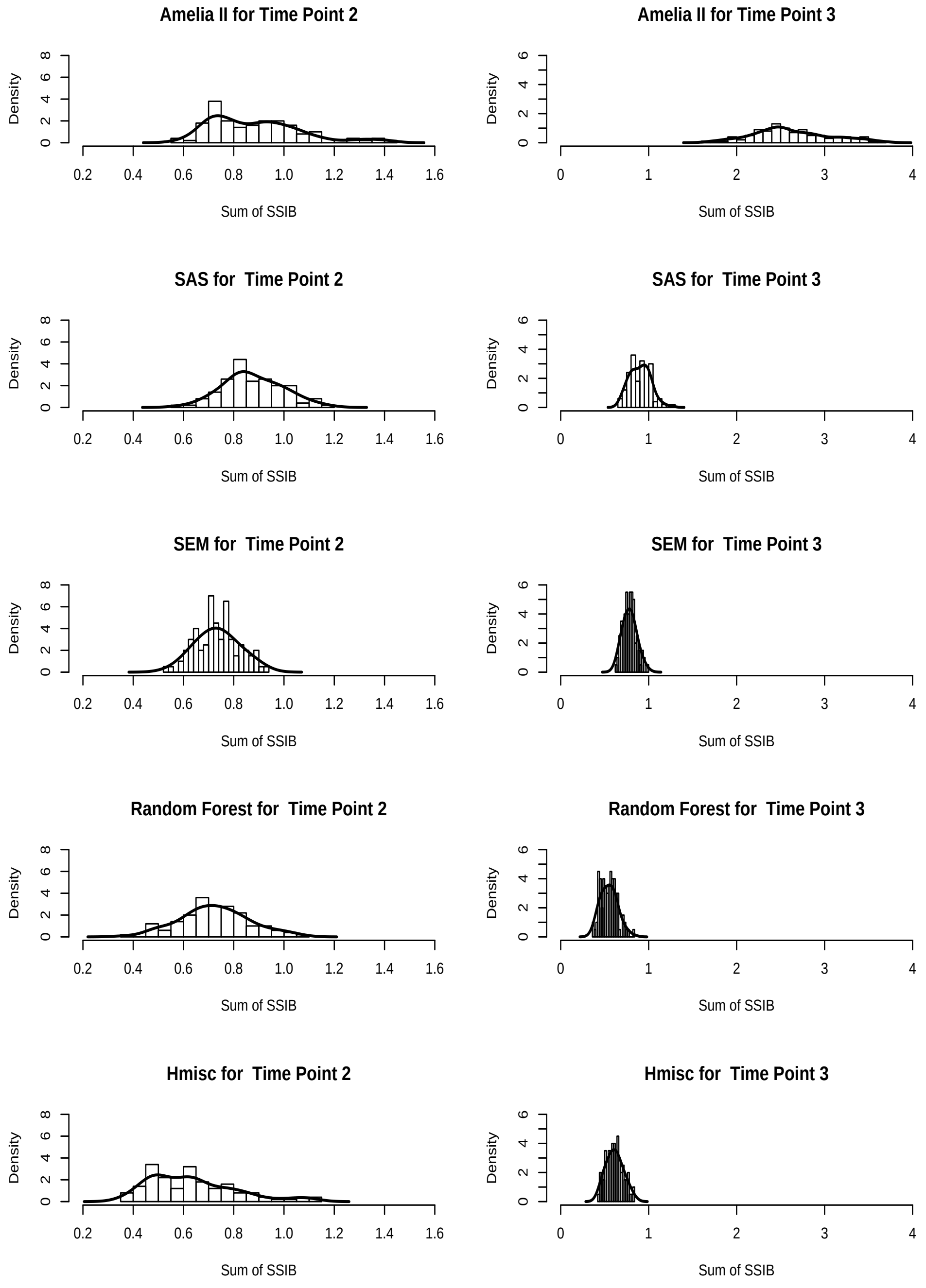


Figure 9.1: *Simulated Data (Missing Not At Random). Histograms and kernel densities for the average SSIB (over 100 samples), for time points 2 and 3, for each of the imputation strategies.*

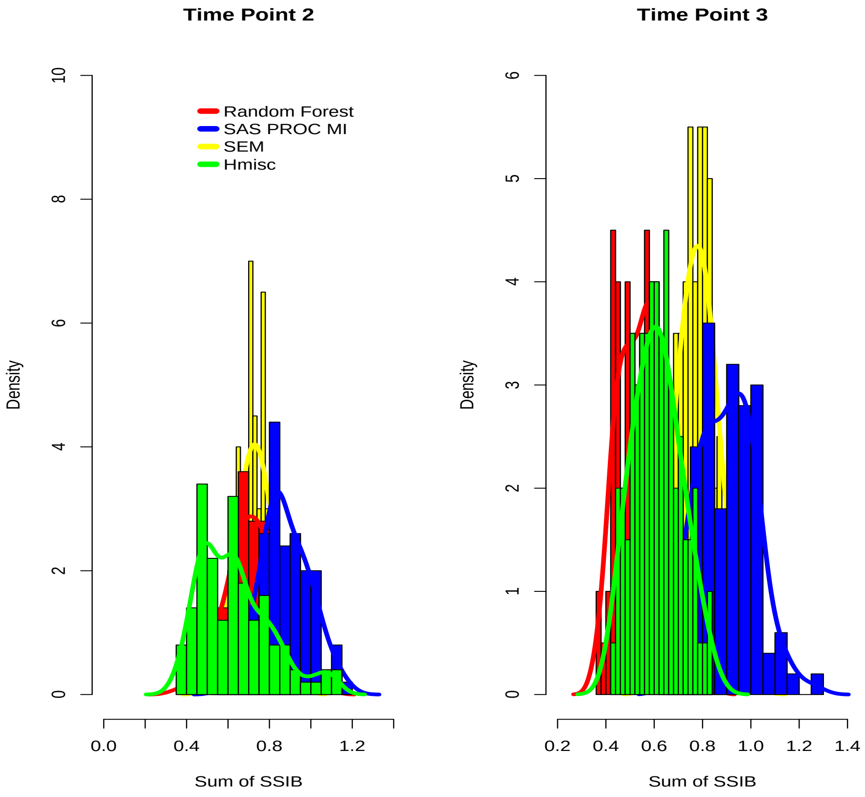


Figure 9.2: *Simulated Data (Missing Not At Random). Comparison of histograms and kernel densities for the average SSIB (over 100 samples), for time points 2 and 3, for the random forest, PROC MI, SEM and Hmisc procedures.*

squared bias should be between 13 and 103 units. Amelia II shows larger variability for the sum of SSIB, while SEM seems to have the smallest variability. In terms of median location, Hmisc shows the smallest value (0.602), while PROC MI and Amelia II produced similar values (0.865 and 0.869, respectively). The kernel densities for time point 3, on the other hand, exhibit similar ranges of values for the PROC MI, SEM, random forest, and Hmisc imputation techniques, while Amelia II is severely affected by the increase in percentage of missingness at this time point, having a lower bound for the sum of SSIB that is even larger than the upper bounds for the other 4 methods. Zooming in on PROC MI, SEM, Hmisc, and random forest (Figure 9.2), practically no differences between the four routines can be seen for time point 2 with respect to the range of SSIB values, but differences are observed for time point 3, with clearly larger values for SSIB under SAS PROC MI.

9.2.2 Missing At Random

For the simulations under MAR, 4 different imputation routines will be considered: the MI procedure from SAS, as well as the Hmisc, Amelia II and random forest packages from R. The SEM approach is not further considered, since in this setting, the SEM algorithm basically uses the same methodology as that of PROC MI.

9.2.2.1 Data Generation

For the simulations under the MAR setting, the same measurement model (Model 9.2) as that used for the MNAR case (Section 9.2.1.1) was considered, and $S = 100$ samples, each of size $N = 300$, equally distributed between the two treatment arms, were again generated from $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

To induce MAR missingness in the outcomes, a logistic dropout model similar to that of the Diggle and Kenward (1994) model, but without dependence on the current possibly unobserved outcome, was used, i.e.,

$$\text{logit}\ P(D_i = j | D_i \geq j, y_{i,j-1}, \boldsymbol{\psi}) \ = \ \psi_0 \ + \ \psi_p \, y_{i,j-1}, \qquad (9.4)$$

where $\psi_0 = -3$ and $\psi_p = 0.5$. Over the $S = 100$ samples, this dropout model produced a percentage of completers (pattern OOO) ranging from 48 and 67%, and ranging from 13 and 29% each for the other two dropout patterns (OOM and OMM).

9.2.2.2 Results

For each of the $S = 100$ generated samples, models (9.2) and (9.4) were fitted to the $M = 5$ completed data sets obtained using 4 different approaches – the MI procedure from SAS, as well as the Hmisc, Amelia II and random forest packages from R, the results of which are presented in Table 9.4. Starting with the parameters of the measurement model, estimates and standard errors for β_0 and u_{11} are almost equivalent across the 4 routines, as might be expected since these parameters involve the outcome at the first time point, which is always observed and never imputed. For $\beta_t, \beta_{xt}, u_{13}, u_{23}$ and u_{33}, the magnitudes of the Amelia II estimates are somewhat different from those for the other 3 methods. In general, these parameters involve the last outcome, having the largest proportion of missing values, which seems to affect the imputations produced under Amelia II. For the other two parameters, u_{12} and u_{22}, the 4 procedures give more or less comparable estimates and standard errors. Finally, it can also be observed that the standard errors under the random forest approach indicate slightly better precision.

Table 9.4: *Simulated Data (Missing At Random). True parameter values, average parameter estimates* (est) *and standard errors* (s.e.) *for $S = 100$ complete samples, for direct likelihood* (DL)*, PROC MI, Amelia II, random forest* (RF) *and Hmisc, for the Diggle-Kenward model.*

Effect	Par	True	Complete	DL	PROC MI	Amelia II	Hmisc	RF
				est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)
Measurement model								
Intercept	β_0	2.000	2.000	2.002 (0.070)	1.999 (0.071)	2.099 (0.077)	2.003 (0.071)	2.009 (0.075)
Time	β_t	0.500	0.499	0.499 (0.033)	0.500 (0.036)	0.423 (0.039)	0.505 (0.035)	0.488 (0.036)
Treatment	β_{xt}	1.000	1.001	0.998 (0.033)	1.000 (0.037)	0.967 (0.039)	0.993 (0.038)	1.004 (0.037)
Variance 1	u_{11}	0.837	0.831	0.830 (0.034)	0.830 (0.037)	0.831 (0.037)	0.830 (0.037)	0.830 (0.037)
Covariance 1,2	u_{12}	0.478	0.478	0.478 (0.045)	0.479 (0.045)	0.494 (0.048)	0.481 (0.048)	0.486 (0.043)
Covariance 1,3	u_{13}	0.239	0.242	0.238 (0.051)	0.240 (0.055)	0.188 (0.047)	0.221 (0.053)	0.222 (0.032)
Variance 2	u_{22}	0.609	0.617	0.615 (0.028)	0.568 (0.028)	0.602 (0.041)	0.592 (0.037)	0.585 (0.028)
Covariance 2,3	u_{23}	0.305	0.308	0.306 (0.047)	0.285 (0.047)	0.204 (0.072)	0.283 (0.052)	0.239 (0.034)
Variance 3	u_{33}	0.592	0.591	0.589 (0.032)	0.483 (0.029)	0.567 (0.048)	0.486 (0.028)	0.462 (0.026)
Dropout model								
Intercept	ψ_0	-3.000	-2.940	-3.011 (0.354)	-2.997 (0.359)	-3.214 (0.415)	-3.222 (0.417)	-2.990 (0.362)
Previous	ψ_p	0.500	0.582	0.506 (0.090)	0.596 (0.091)	0.648 (0.102)	0.650 (0.104)	0.596 (0.092)

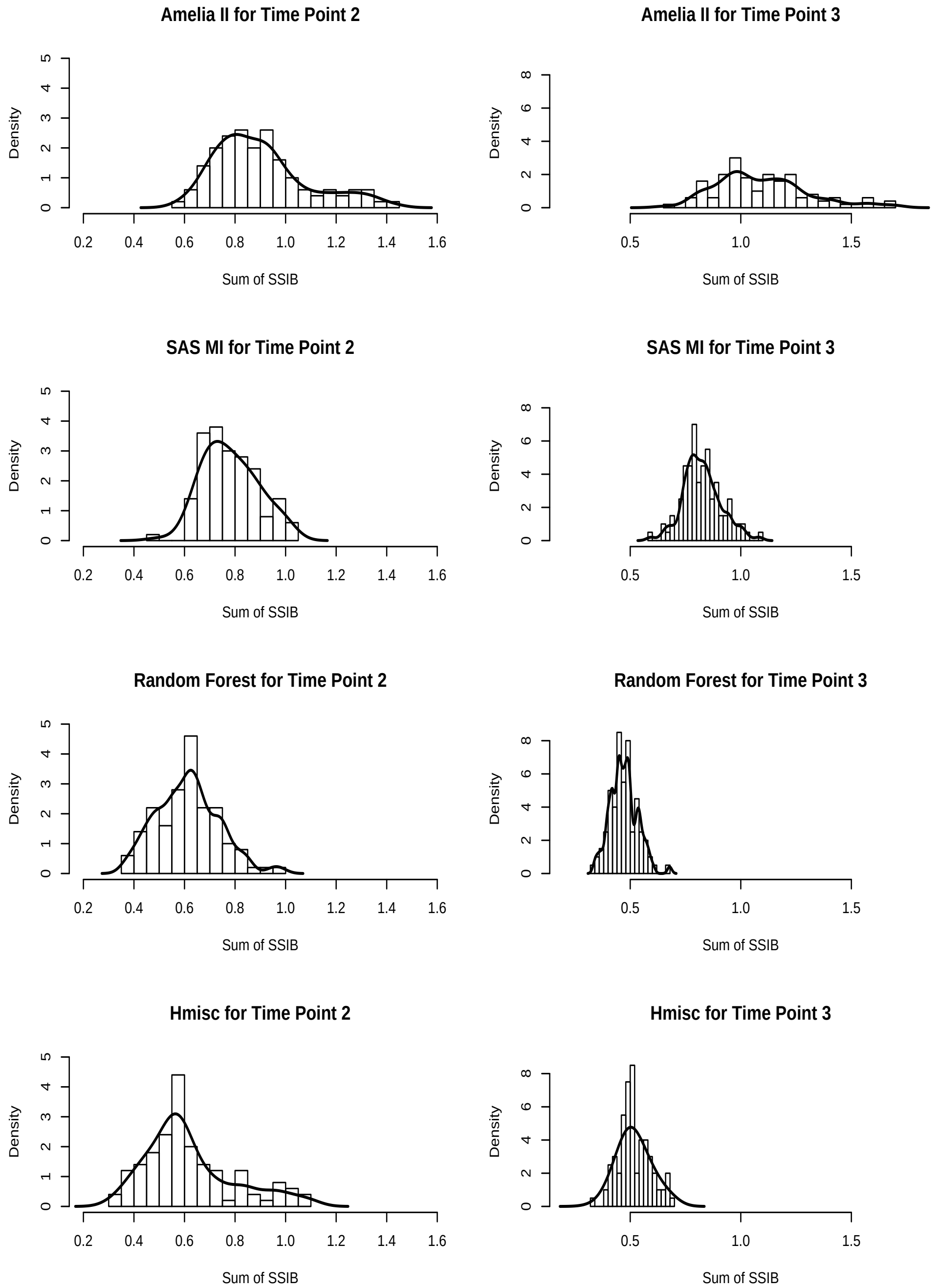


Figure 9.3: *Simulated Data (Missing At Random). Histograms and kernel densities for the average SSIB (over 100 samples), for time points 2 and 3, for each of the imputation strategies.*

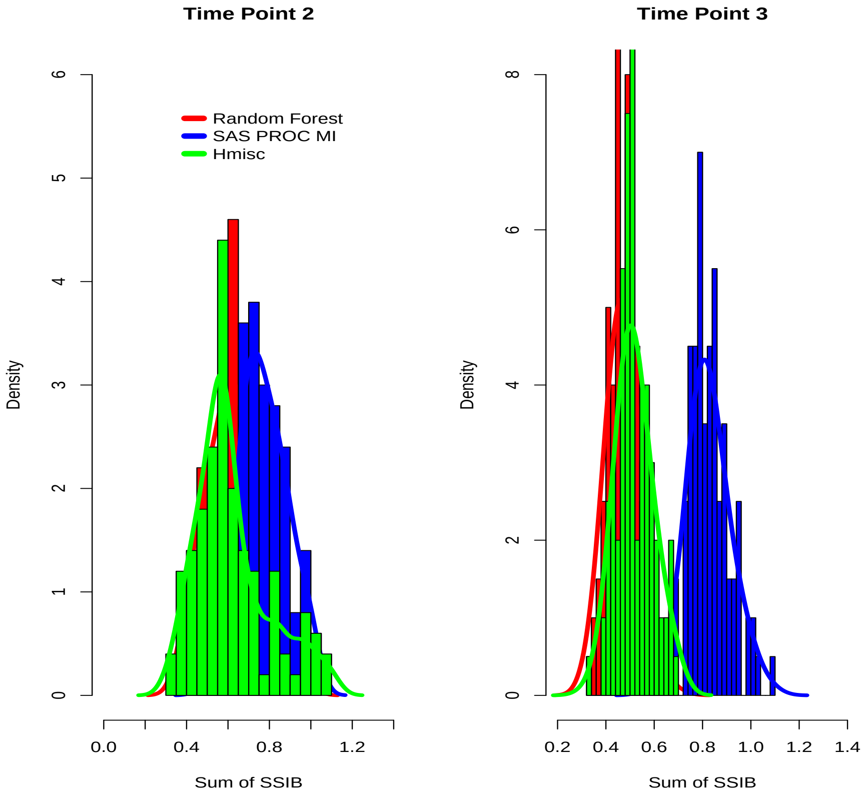


Figure 9.4: *Simulated Data (Missing At Random). Comparison of histograms and kernel densities for the average SSIB (over 100 samples), for time points 2 and 3, for random forest, PROC MI and Hmisc procedures.*

A much more distinct pattern is observed for the 4 imputation methods for the dropout model. For both parameters, ψ_0 and ψ_p, SAS PROC MI and the random forest approach yield very similar estimates as well as standard errors, and these estimates are not only very close to the estimates obtained using the complete samples, but are also only very slightly biased with respect to the true parameter values. The Amelia II and Hmisc packages, on the other hand, exhibit more biased estimates (with respect to both the true and complete values), and in addition, substantially larger standard errors, than the other two techniques.

The histograms and kernel densities for the four routines (Figure 9.3) show wider ranges of SSIB for Amelia II, while the random forest method shows narrower distributions for SSIB. It is also worthwhile to note that for SAS PROC MI, SSIB ranges are practically invariant to the time point considered, whereas for Hmisc and random forest, SSIB values tend to be smaller for time point 3, which could possibly be due to the smaller variability in the outcome at this time point. Figure 9.4 further highlights the similarity of SSIB ranges for SAS PROC MI, Hmisc and random forest at time point 2. For time point 3, however, the smaller variability of the outcome seems to play a role in the SSIB values for Hmisc and random forest.

9.3 Case Study: Age-Related Macular Degeneration Trial

The performance of the various imputation routines of interest was further examined using the ARMD data (Section 2.1). In particular, the subset of 226 ARMD patients having either a complete set of responses or dropout type of missingness (i.e., excluding all non-monotone or fully-missing sequences) was considered for analysis. Restriction to the monotone cases was considered primarily due to the infeasibility of fitting a non-response model for non-monotone missingness, which would require at least a few observed responses in each of the $2^4 = 16$ non-monotone missingness patterns, for every level of the covariate(s). This would not be possible since there were only very few cases (<5%) with intermittent missingness.

For the 4 repeated measurements, a saturated means model was formulated, with the treatment (interferon-α or placebo) received by the patient, denoted x_i, as the primary covariate of interest, and assuming an unstructured covariance for the repeated measures, i.e.,

$$E(Y_{ij}|x_i, t_{ij}, \boldsymbol{\beta}) \;=\; \beta_{0j} + \beta_{1j}\, x_i\, I_{t_{ij}}(j) \qquad \text{and} \qquad \Sigma_i \;=\; [\sigma_{rc}]\,,$$

for $i = 1, 2, \ldots, 226$ and $j, r, c = 1, 2, 3, 4$. For dropout, an MNAR logistic model of the form (9.3) was formulated. These models, measurement and dropout, were fitted to the ARMD data using direct likelihood and using $M = 5$ imputations obtained from the SEM algorithm, SAS PROC MI, as well as the Amelia II, Hmisc and random forest packages of R. The results, under Strategy I, are summarized in Table 9.5.

For the measurement model, all routines produced quite comparable results in terms of parameter estimates. Slight discrepancies, however, were observed for the parameters of the dropout model, with the intercept as estimated by direct likelihood somewhat different compared to the rest of the methods. The estimated values for the previous and current measurements are also different, but all methods point towards non-significant effects. In terms of standard errors, Amelia II produces, in general, slightly larger values than the rest of the methods applied. Overall conclusions were generally consistent across all methods.

Table 9.5: *Age-Related Macular Degeneration Trial (Missing Not At Random). Parameter estimates* (est) *and standard errors* (s.e.) *for direct likelihood* (DL), *stochastic EM* (SEM), *Amelia II, Hmisc, random forest* (RF) *and PROC MI, for the Diggle-Kenward model.*

Effect	Par	DL	SEM	Amelia II	Hmisc	RF	PROC MI
		est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)	est (s.e.)
Measurement model							
Intercept 4	β_{01}	53.974 (1.467)	54.000 (1.470)	54.000 (1.470)	54.000 (1.470)	54.000 (1.470)	54.000 (1.470)
Intercept 12	β_{02}	52.955 (1.594)	52.979 (1.594)	52.976 (1.598)	53.038 (1.597)	52.940 (1.585)	52.991 (1.596)
Intercept 24	β_{03}	49.058 (1.733)	49.075 (1.716)	49.287 (1.757)	48.976 (1.739)	49.251 (1.656)	49.252 (1.727)
Intercept 52	β_{04}	43.583 (1.816)	43.582 (1.730)	44.541 (1.806)	43.732 (1.753)	44.197 (1.586)	44.367 (1.751)
Treatment 4	β_{11}	-3.127 (2.091)	-3.108 (2.097)	-3.108 (2.098)	-3.108 (2.100)	-3.108 (2.098)	-3.108 (2.098)
Treatment 1	β_{12}	-4.676 (2.279)	-4.659 (2.265)	-4.851 (2.280)	-4.611 (2.292)	-4.561 (2.262)	-4.496 (2.278)
Treatment 24	β_{13}	-3.799 (2.487)	-3.791 (2.434)	-3.911 (2.616)	-3.916 (2.468)	-3.872 (2.363)	-3.760 (2.492)
Treatment 52	β_{14}	-5.664 (2.628)	-5.729 (2.432)	-5.376 (2.544)	-5.386 (2.482)	-5.301 (2.263)	-5.213 (2.527)
Variance 1	u_{11}	15.742 (0.739)	15.764 (0.741)	15.764 (0.800)	15.764 (0.786)	15.764 (0.742)	15.764 (0.809)
Covariance 1,2	u_{12}	14.264 (0.918)	14.293 (0.917)	14.279 (0.991)	14.331 (0.975)	14.171 (0.910)	14.288 (1.002)
Covariance 1,3	u_{13}	13.752 (1.067)	13.768 (1.029)	13.853 (1.137)	13.648 (1.108)	13.744 (0.989)	13.823 (1.141)
Covariance 1,4	u_{14}	11.166 (1.202)	11.169 (1.103)	11.058 (1.228)	10.658 (1.181)	10.877 (1.008)	10.999 (1.224)
Variance 2	u_{22}	9.412 (0.450)	9.421 (0.431)	9.425 (0.478)	9.364 (0.464)	9.384 (0.440)	9.404 (0.483)
Covariance 2,3	u_{23}	6.810 (0.773)	6.844 (0.741)	6.852 (0.823)	6.705 (0.810)	6.627 (0.758)	6.853 (0.822)
Covariance 2,4	u_{24}	7.239 (1.045)	7.380 (0.887)	7.127 (1.031)	6.984 (0.998)	7.074 (0.861)	6.835 (1.028)
Variance 3	u_{33}	10.303 (0.503)	10.289 (0.473)	10.403 (0.528)	10.570 (0.535)	10.441 (0.498)	10.228 (0.525)
Covariance 3,4	u_{34}	8.839 (0.885)	8.830 (0.731)	8.829 (0.855)	9.010 (0.846)	8.651 (0.731)	8.809 (0.854)
Variance 4	u_{44}	10.079 (0.528)	10.004 (0.441)	10.153 (0.515)	9.901 (0.482)	9.795 (0.455)	9.977 (0.512)
Dropout model							
Intercept	ψ_0	-1.797 (0.468)	-1.423 (0.395)	-1.489 (0.494)	-1.336 (0.413)	-1.428 (0.398)	-1.596 (0.408)
Previous	ψ_p	-0.035 (0.023)	-0.038 (0.011)	-0.0003 (0.024)	-0.030 (0.016)	-0.014 (0.013)	-0.013 (0.014)
Current	ψ_c	0.011 (0.022)	0.014 (0.012)	-0.019 (0.020)	0.005 (0.014)	-0.007 (0.013)	-0.003 (0.014)

9.4 Discussion

In this chapter, the performance of various imputation routines – SAS PROC MI, Amelia II, Hmisc, SEM and random forest – was investigated in a simulation exercise and in a real life example. Comparisons of resulting estimates and standard errors, for both the measurement and dropout models, were conducted. In addition, for the simulations, two different missingness scenarios were studied: one considering non-ignorable missingness, and another, missing at random. The merits of the different routines were further assessed with respect to a so-called individual imputation bias.

It is important to note that some of these methods are not designed to deal with MNAR, thus requiring caution in their application under such. These methods, however, were nevertheless implemented here under such settings, in order to evaluate the impact of their use in situations under which they are not expected to work well. In general, Amelia II, SAS PROC MI and Hmisc showed some degree of sensitivity to the underlying missingness mechanism in the data, because their resulting estimates, particularly for the dropout model, were inconsistent with those from the other approaches. The Amelia II package seemed the most affected by this fact – producing an estimate for the previous measurement that was positive and significant. The rest of the approaches though provided consistent conclusions, implying insignificant association between the dropout indicator and the previous measurement.

With respect to the individual imputation bias for the 5 imputation approaches, smaller values were generally observed for the random forest and Hmisc packages as compared to the rest of the methods applied. Moreover, results clearly indicated that the performance of Amelia II was severely affected for the case of non-ignorable missingness. It was further observed that the ability of the procedures to yield reasonable individual imputations for monotone non-ignorable missing values seemed to depend on two factors, namely, the amount of missingness, as well as the amount of variability in the response measurements.

The availability of easy-to-use software-based imputation routines has perhaps been a key factor in the widespread popularity of imputation as a means of dealing with missing data. The review conducted in this chapter, however, stipulates a few cautionary remarks regarding their use. Not only should such routines be applied properly, but careful consideration is also warranted regarding the choice of the assumed underlying missingness mechanism, which can ultimately drive the resulting inferences. It is further recommended that imputation approaches be used as tools within a sensitivity analysis, so as to obtain more insight about the data at hand, and consequently, more confidence in the resulting conclusions.

10 Conclusion

Research studies that involve the collection of data are often, though perhaps not intentionally, subject to some form of incompleteness, thus necessitating the application of modified analysis techniques. As a consequence, methodology for handling missing data has advanced significantly over the recent decades, and can be foreseen to continue doing so in the coming years. In the past, limitations – both methodological, as well as computational – have prohibited the use of more sophisticated approaches, and analyses of incomplete data have historically revolved around simple methods such as complete case analysis. Recent developments, however, have broadened the possibilities available to the researcher – a number of which have been examined in this thesis – and the consideration of such should supersede the use of simple techniques that have been shown to yield substandard results. At best, these naïve approaches might be used as preliminary steps or supplements to more appropriate methods.

Despite, however, the availability of a wide array of courses of analysis and corresponding computational routines for their implementation, the modeling of incomplete data remains less than straightforward. This is primarily due to the fact that the inferential validity of applicable techniques is hinged on unverifiable assumptions regarding the underlying missing data mechanism. A large part of this thesis has been devoted to emphasize this, underscoring the importance of *sensitivity analysis* in modeling incomplete data.

The different methodologies presented in this thesis have been considered to highlight either (or both) of two objectives. On the one hand, a number of chapters (Chapters 5, 6, 8 and 9) serve to demonstrate, review and/or propose modifications to existing approaches, geared towards providing useful and important information regarding their application. Chapters 4, 7, and 8, on the other hand, introduce techniques that might serve as tools within the context of a sensitivity analysis, thereby broadening the possibilities under such. A brief overview of the resulting conclusions for the pertinent chapters is now presented.

The concept of an MAR counterpart model to a fitted MNAR model (Molenberghs *et al.*, 2008) was presented and developed in Chapter 4. Such a counterpart is equivalent to the fitted MNAR model in the sense that its fit to the observed data reproduces that of the original MNAR model. Several implications follow from this result. First, the idea of an MAR counterpart to an MNAR model further underscores the infeasibility of using the fit of an MNAR model for or against MAR – a theme similar to arguments previously articulated in Gill, van der Laan and Robins (1997) and Schafer and Graham (2002). In addition, though the original MNAR model and its MAR counterpart agree with respect to their fit to the observed data, one can examine differences in their fits to the unobserved complete data, thereby providing better understanding as to which parts of the data drive the missingness towards non-ignorability. Moreover, in cases where a parametric MAR model fails to fit the observed data sufficiently well, a better-fitting, more versatile MNAR model, along with its MAR version, might be considered. Finally, MAR counterparts can very well serve as additional components of a sensitivity analysis.

In Chapter 5, MAR marginal models for incomplete binary longitudinal data were considered, with particular focus on variations of the semi-parametric approach of Liang and Zeger (1986) using *generalized estimating equations* (GEE). Owing to the frequentist nature of GEE, rendering it essentially valid only under the somewhat unrealistic setting of MCAR, two modifications for the MAR case were examined, namely, weighted GEE (Robins, Rotnitzky and Zhao, 1995) and GEE combined with multiple imputation (Rubin, 1987) or MI-GEE. While asymptotic simulations demonstrated the theoretical properties of the methods, small-sample simulations provided evidence for the practical use of either approach (Beunckens, Sotto and Molenberghs, 2008). The attractive properties of weighted GEE were barely reproduced for small samples, even with an entirely correctly specified analysis, with, under some scenarios, weighted GEE yielding far less precise estimates than those under a misspecified MI-GEE analysis. Moreover, in contrast with MI-GEE, which demonstrated some amount of robustness to misspecifications in either the imputation or measurement model, the weighted GEE approach exhibited sensitivity to misspecifications in either the dropout or measurement model. In line with a number of related references (Scharfstein, Rotnitzky and Robins, 1999; Clayton *et al.*, 1998; Carpenter, Kenward and Vansteelandt, 2006), these results reinforce the strengths of multiple imputation over weighting, specifically in their application to generalized estimating equations.

Pattern-mixture models (PMM) for categorical data with monotone missingness were the subject of Chapter 6. Through the use of simulated data, the approach proposed by Jansen and Molenberghs (2007), which makes use of identifying restrictions to fit PMMs, was examined. Additionally, asymptotic variances for the marginalized effects estimates of the same authors were derived (Sotto *et al.*, 2009a). Results indicated that the precision of the estimates was largely contingent on the amount of missingness within the various dropout patterns, implying, perhaps, a preference for an identification scheme based on the more amply filled patterns. Nevertheless, even within the sparse patterns, precision of the main parameters of interest generally seemed reasonable, with the treatment effect at the last time point being the most precisely estimated parameter, regardless of the dropout setting or the identifying restriction used. With respect to the marginalized effects estimates, those proposed by Jansen and Molenberghs (2007) exhibited slightly less bias and better precision than the direct linear approach (Park and Lee, 1999), which is a viable route for marginalizing pattern-specific estimates arising from a PMM for continuous (Gaussian) outcomes.

Chapter 7 explored the use of influence measures as possible tools for sensitivity. Influence-based sensitivity approaches, such as global and local influence, which revolve around observation-varying ideas, were combined with model-based tools for sensitivity, e.g., non-parametric bounds, a family of identified models, and intervals resulting from over-specified models, yielding a considerably comprehensive sensitivity analysis conducted on the plebiscite-related questions added to the Slovenian Public Opinion Survey (Beunckens *et al.*, 2009). Generalized local influence measures were derived and variations to existing ideas regarding the nature of perturbations were further introduced.

Fitting likelihood-based non-ignorable models for longitudinal data often presents numerical challenges due to the need to integrate the full-data likelihood over the missing values. In Chapter 8, the use of a computational algorithm to address such difficulties was proposed. Sotto *et al.* (2009b) considered the application of the so-called *stochastic expectation-maximization* or SEM algorithm (Celeux and Diebolt, 1985) – a variation of the highly popular expectation-maximization (EM) algorithm – to fit likelihood-based models for longitudinal data with non-ignorable missingness. The substitution of the E-step with an S(imulation)-step to fill in the missing values is particularly advantageous in the latter setting, as completion of the data altogether obviates the need for any integration in the computation of the likelihood. Within a simulation study, stochastic and non-stochastic solutions were compared and results indicated fairly stable results for the former, with slightly better precision than the latter. Moreover, the SEM algorithm not only provides a computationally efficient alternative to direct likelihood, but also admits greater flexibility in the choice for the joint model of the outcomes and the non-response indicators, since one need not be restricted by numerical complexities arising from more complicated forms of likelihoods. Finally, the SEM approach, when not taken as the primary course of analysis, can further serve as an additional tool within a sensitivity analysis.

The focus of Chapter 9 of this thesis is the comparison of various software implementations for the imputation of missing data. Under simulated settings, these imputation routines were assessed not only with respect to resulting estimates for the prescribed model, but also in terms of a measure of individual imputation bias, i.e., the amount of bias in the imputation evaluated at the level of individual. While the procedures generally led to estimates and standard errors that were fairly comparable in magnitude, differences were observed in the amount of individual imputation bias. Moreover, the performance of the various routines seemed to depend on two factors, namely, the amount of missingness, as well as the amount of variability in the response measurements.

The ideas developed in this thesis were presented to feature and provide further insight on currently existing methodology useful for the analysis of incomplete data. The choice regarding the type of analysis to be employed is usually dictated by several pragmatic considerations, which typically include the particular scientific objectives, computational issues, as well as the magnitude of incompleteness expected to be present in the data. Because such a choice can often be daunting to the researcher, the examination of the different procedures considered herein has hopefully provided additional useful information regarding their use. Alongside highlighting the relative merits of various techniques available to the researcher, the thesis also emphasized the practicality and importance of conducting so-called *sensitivity analyses*. Several tools for such were introduced, explored and illustrated. Especially in the face of uncertainty about the underlying missing data mechanism, the application of several approaches can only increase the researcher's confidence in the resulting conclusions, and is therefore strongly recommended.

Bibliography

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (ed), *Studies in Item Analysis and Prediction.* Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

Baker, S.G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics* **51**, 1042–1052.

Baker, S.G., Rosenberger, W.F. and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine* **11**, 643–657.

Barnett, V. (2002). *Sample Survey: Principles and Methods (3rd ed.).* London: Arnold.

Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2007). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics* **64**, 96–105.

Beunckens, C., Sotto, C. and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis* **52**, 1533–1548.

Beunckens, C., Sotto, C., Molenberghs, G. and Verbeke, G. (2009). A multifaceted sensitivity analysis of the Slovenian public opinion survey. *Journal of the Royal Statistical Society: Series C* **58**, 171–196.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. and Cutler, A. (2009). The randomForest Package. *The Comprehensive R Archive Network Website.* Accessed 8 February 2009 from <http://cran.r-project.org/web/packages/randomForest/randomForest.pdf>.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees.* New York: Chapman & Hall.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B* **22**, 302–306.

Burton, A. and Altman, D.G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer* **91**, 4–8.

Carpenter, J.R., Kenward, M.G. and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A* **3**, 571–584.

Celeux, G., Chauveau, D. and Diebolt, J. (1995). On stochastic versions of the EM algorithm. Rhône-Alpes: Institut National de Recherche en Informatique et en Automatique (France); 1995 March. 22 p. Research Report No.: 2514.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* **2**, 73–82.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports* **41**, 119–134.

Chen, T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30**, 629–642.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling (with discussion). *Journal of the Royal Statistical Society: Series B* **60**, 71–102.

Cook, R.D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statististical Society: Series B* **2**, 133–169.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman & Hall.

Crouchley, R. and Ganjali, M. (2002). The common structure of several recent statistical models for dropout in repeated continuous responses. *Statistical Modeling* **2**, 39–62.

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B* **39**, 1–38.

Diebolt, J. and Ip, E.H.S. (1996). Stochastic EM: method and application. In Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds), *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Diggle, P.J. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Journal of the Royal Statistical Society: Series C* **43**, 49–93.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association* **89**, 463–475.

Gad, A.M. and Ahmed, A.S. (2006). Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. *Computational Statistics and Data Analysis* **50**, 2702–2714.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

Gill, R.D., van der Laan, M.J. and Robins, J.M. (1997). Coarsening at random: characterizations, conjectures and counterexamples. In Lin, D.Y. and Fleming, T.R. (eds), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis.* New York: Springer.

Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society: Series B* **81**, 477–482.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986). Selection modelling versus mixture modelling with non-ignorable nonresponse. In Wainer, H. (ed), *Drawing Inferences from Self Selected Samples.* New York: Springer.

Goetz, A. (1970). *Introduction to Differential Geometry.* Reading, MA.: Addison Wesley.

Harrell, F.E. (2009). The Hmisc Package. *The Comprehensive R Archive Network Website.* Accessed 8 February 2009 from <http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>.

Hogan, J.W. and Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16**, 239–258.

Honaker, J., King, G. and Blackwell, M. (2008). Amelia II: A Program for Missing Data. *Amelia Software Website.* Accessed 20 March 2008 from <http://gking.harvard.edu/amelia>.

Horton, N.J. and Kleinman, K.P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* **61**, 79–90.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician* **55**, 244–254.

Horton, N.J. and Switzer, S.S. (2005). Statistical methods in the Journal (research letter). *New England Journal of Medicine* **353**, 1977–1979.

Ip, E.H.S. (1994). A stochastic EM estimator in the presence of missing data - theory and applications. Technichal report, Department of Statistics, Stanford University.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006). Analyzing incomplete binary longitudinal clinical trial data. *Statistical Science* **21**, 52–69.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G. and Kenward, M.G. (2006). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis* **50**, 830–858.

Jansen, I. and Molenberghs, G. (2007). Pattern-mixture models for categorical outcomes with non-monotone missingness. *Submitted for publication.*

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H. and Van Steen, K. (2003). A Local influence approach applied to binary data from a psychiatric study. *Biometrics* **59**, 410–419.

Jolani, S. and Ganjali, M. (2007). Analyis of longitudinal continuous response data with dropout: use of stochastic EM algorithm. Paper presented at the 56th Session of the International Statistical Institute, Lisbon, Portugal.

Kenward, M.G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* **17**, 2723–2732.

Kenward, M.G., Goetghebeur, E. and Molenberghs, G. (2001). Sensitivity analysis of incomplete categorical data. *Statistical Modelling* **1**, 31–48.

Kenward, M.G. and Molenberghs, G. (2009). Last observation carried forward: a crystal ball? *Submitted for publication.*

Kenward, M.G., Molenberghs, G. and Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika* **90**, 53–71.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review* **95**, 49–69.

Laird, N.M. (1994). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C* **43**, 84.

Laird, N.M. and Ware, J.H. (1982). Rnadom effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lang, J.B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* **84**, 447–451.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics* **54**, 570–582.

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

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R.J.A. (1994a). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

Little, R.J.A. (1994b). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C* **43**, 78.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

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

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B* **44**, 226–232.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons, Inc.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **10**, 538–573.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T. and Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine* **21**, 1023–1042.

Michiels, B., Molenberghs, G. and Lipsitz, S.R. (1999a). A pattern-mixture odds ratio model for incomplete categorical data. *Communications in Statistics: Theory and Methods* **28**, 2843–2869.

Michiels, B., Molenberghs, G. and Lipsitz, S.R. (1999b). Selection models and pattern-mixture models for incomplete categorical data with covariates. *Biometrics* **55**, 978–983.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B* **70**, 371–388.

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons, Inc.

Molenberghs, G., Kenward, M.G. and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C* **50**, 15–29.

Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* **84**, 33–44.

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

Molenberghs, G., Michiels, B., Kenward, M.G. and Diggle, P.J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica* **52**, 153–161.

Molenberghs, G., Thijs, H., Jansen, I, Beunckens, C., Kenward, M.G., Mallinkrodt, C. and Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **5**, 445–464.

Molenberghs, G., Thijs, H., Kenward, M.G. and Verbeke, G. (2003). Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistica Neerlandica* **57**, 112–135.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Park, T. and Lee, S.Y. (1999). Simple pattern-mixture models for longitudinal analysis with missing observations: analysis of urinary incontinence data. *Statistics in Medicine* **18**, 2933–2941.

Pharmacological Therapy for Macular Degeneration Study Group (1997). Interferon α-IIA is ineffective for patienst with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology* **115**, 865–872.

Rizopoulos D., Verbeke G. and Molenberghs G. (2008). Shared parameter models under random-effects misspecification. *Biometrika* **95,** 63–74.

Robins, J.M., Rotnitzky, A. and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

Robins, J.M., Rotnitzky, A. and Scharfstein, D.O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran, M.E. and Berry, D.A. (eds), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rotnitzky, A., Scharfstein, D., Su, T.L. and Robins, J.M. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **50**, 1163–1170.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72**, 538–543.

Rubin, D.B. (1978). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce, 1–23.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Rubin, D.B. (1994). Discussion to Diggle, P.J. and Kenward, M.G.: Informative dropout in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C* **43**, 80–82.

Rubin, D.B., Stern, H.S. and Vehovar, V. (1995). Handling "don't know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* **90**, 822–828.

SAS Institute Inc. (2004). The MI Procedure. *SAS OnlineDoc 9.1.2.* Cary, NC: SAS Institute Inc. Accessed 20 November 2008 from <http://support.sas.com/onlinedoc/912/getDoc/statug.hlp/mi_index.htm>.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schafer, J.L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* **57**, 19–35.

Schafer, J.L. and Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods* **2**, 147–177.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with comments). *Journal of the American Statistical Association* **94**, 1096–1146.

Sotto, C., Beunckens, C., Molenberghs, G., Jansen, I. and Verbeke, G. (2009a). Marginalizing pattern-mixture models for categorical data subject to monotone missingness. *Metrika* **69**, 305–336.

Sotto, C., Beunckens, C., Molenberghs, G. and Kenward, M.G. (2009b). MCMC-based estimation methods for continuous longitudinal data with non-random (non-) monotone missingness. *Submitted for publication.*

Stoker, J.J. (1969). *Differential Geometry.* New York: John Wiley & Sons, Inc.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G. and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3**, 245–265.

Thijs, H., Molenberghs, G. and Verbeke, G. (2000). The milk protein trial: influence analysis of the dropout process. *Biometrical Journal* **42**, 617–646.

Troxel, A.B., Harrington, D.P. and Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society: Series C* **47**, 425–438.

Tsonaka R., Verbeke G. and Lesaffre E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* **65**, 81–87.

Van Damme, P., Van Herck, K., Van der Wielen, M., Broodhaers, S., Bauer, T. and Jilg, W. (2006). 20 year follow-up of hepatitis B vaccination in institutionalized mentally retarded subjects. Centre for the Evaluation of Vaccination, University of Antwerp, Belgium; Institut für Medizinische Microbiologie und Hygiene, Universität Regensburg, Germany.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16**, 953–979.

Van Steen, K., Molenberghs, G., Verbeke, G. and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal* **1**, 125–142.

Verbeke, G., Lesaffre, E. and Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal* **35**, 419–434.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M.G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics* **57**, 7–14.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* **85**, 699–704.

Welsh, A.H. (1996). *Aspects of Statistical Inference.* New York: John Wiley & Sons, Inc.

Wu, M.C. and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* **45**, 939–955.

Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* **44**, 175–188.

Appendix A

Derivation of Local Influence Measures on A Function of the Parameters

In line with the notations used in Section 7.2, let $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$ denote the log-likelihood corresponding to a particular model. Under the null model, $\boldsymbol{\omega} = \boldsymbol{\omega}_o = (0, 0, \ldots, 0)'$. Further denote by $\widehat{\boldsymbol{\phi}}$ be the maximum likelihood estimator for $\boldsymbol{\phi}$, obtained by maximizing $\ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)$. For any general function of the parameters, $Z(\boldsymbol{\phi})$, say a particular predicted cell count, the corresponding maximum likelihood estimator under the null model is $Z(\widehat{\boldsymbol{\phi}})$.

Consider now the perturbed model, where $\boldsymbol{\omega} \neq (0, 0, \ldots, 0)'$. Let $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ denote the maximum likelihood estimator for $\boldsymbol{\phi}$, obtained by maximizing $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$, and $Z(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})$ is the corresponding maximum likelihood estimator for the predicted cell count of interest (or any other function of the parameters).

The difference now between the null model and the perturbed model in terms of the predicted cell count is:

$$D_k(\boldsymbol{\omega}) = Z_k(\widehat{\boldsymbol{\phi}}) - Z_k(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}), \qquad \text{for } k = 1, 2, \ldots, 16,$$

where k represents the index of the particular predicted cell count of interest.

One can look at the surface $(\boldsymbol{\omega}, D_k(\boldsymbol{\omega}))$, which is dim$(\boldsymbol{\omega})$+1 = 10-dimensional, since there are 9 observed cells on which a perturbation can be introduced. Now, suppressing k from notation, say

$$\alpha(\boldsymbol{\omega}) = (\alpha_1(\boldsymbol{\omega}), \alpha_2(\boldsymbol{\omega})) = (\boldsymbol{\omega}, D_k(\boldsymbol{\omega})).$$

$\alpha(\boldsymbol{\omega})$ is a surface in $\mathbb{R}^{q+1}$, where q=dim$(\boldsymbol{\omega})$=9.

How does $\alpha(\boldsymbol{\omega})$ deviate from its tangent plane at $\boldsymbol{\omega}_o$?

- *Normal curvatures* are curvatures of normal sections which are formed by the intersection of the surface with planes containing the vector that is normal (orthogonal) to the tangent plane at $\boldsymbol{\omega}_o$.
- These curvatures are used to characterize the behavior of the influence around $\boldsymbol{\omega}_o$.

How does one represent a normal section?

- First consider a straight line passing through $\boldsymbol{\omega}_o$:

  $$\boldsymbol{\omega}(a) = \boldsymbol{\omega}_o + a\boldsymbol{h},$$

  where $a \in \mathbb{R}$ and $\boldsymbol{h} \in \mathbb{R}^9$ with $\|\boldsymbol{h}\| = 1$.
- This is a <u>lifted line</u> on the surface/graph $\alpha(\boldsymbol{\omega})$ through $\alpha(\boldsymbol{\omega}_o)$. Each <u>direction $\boldsymbol{h}$</u> specifies a lifted line, and each lifted line corresponds to a <u>normal section</u>.

The *tangent plane* at $\boldsymbol{\omega}_o$ is spanned by columns of $(q+1) \times q$ matrix V with elements

$$\left.\frac{\partial \alpha_i(\boldsymbol{\omega})}{\partial \omega_j}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}, \quad \text{for} \quad \left\{ \begin{array}{l} i = 1, 2, \ldots, q+1 \\ j = 1, 2, \ldots, q \end{array} \right. .$$

Thus, $V^T = (I_q, \dot{D})$ with

$$\dot{D}_j = \left.\frac{\partial D(\boldsymbol{\omega})}{\partial \omega_j}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}.$$

For $\boldsymbol{\omega} = \boldsymbol{\omega}_o$, $\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$.

This means $\dot{D}(\boldsymbol{\omega}_o) = 0$ or D has local minimum at $\boldsymbol{\omega}_o$, i.e.,

$$\left.\frac{\partial D(\boldsymbol{\omega})}{\partial \omega_j}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} = 0.$$

Consequently,

- $V^T = (I_q, \mathbf{0})$.
- The subspace orthogonal to the tangent plane is spanned by the basis vector for $\mathbb{R}^{q+1}$, i.e., $(0, 0, \ldots, 0, 1)$, with a 1 in the last position, and zeros elsewhere.
- A normal section can now be viewed as that portion of the influence graphs, $\alpha(\boldsymbol{\omega})$, cut out by the plane spanned by the vectors $(0, 0, \ldots, 0, 1)$ and $(\boldsymbol{h}',0)$.

It follows that each lifted line $\alpha(\boldsymbol{\omega}(\alpha))$ is a normal section.

Note that a *plane curve*, i.e., a curve in 2 dimensions, can be expressed as:

$$\underline{\text{Plane Curve}}\text{: } f(\boldsymbol{\omega}) = (f_1(\boldsymbol{\omega}), f_2(\boldsymbol{\omega})).$$

From Stoker (1969) and Goetz (1970), the curvature of a plane curve at $\boldsymbol{\omega}_o$ is:

$$C = \frac{\left|\dot{f}_1\ddot{f}_2 - \dot{f}_2\ddot{f}_1\right|}{(\dot{f}_1^2 + \dot{f}_2^2)^{3/2}},$$

with derivatives evaluated at $\boldsymbol{\omega}_o$.

This can be applied to the plane curve $\rho'(a) = (a, D(\boldsymbol{\omega}(a)))$ at $a = 0$, which is then the lifted line $\alpha(\boldsymbol{\omega}(a))$ in rotated coordinates!

The individual curvatures in the family of plane curves ρ, obtained by letting $\boldsymbol{h}$ range over all unit vectors in $\mathbb{R}^9$, form the basis for our characterization of α.

$$\begin{aligned}
\dot{\rho}_1 &= \frac{\partial \rho_1(a)}{\partial a} = \frac{\partial a}{\partial a} = 1 \qquad \text{and} \qquad \ddot{\rho}_1 = 0 \\
\dot{\rho}_2 &= \left.\frac{\partial \rho_2(a)}{\partial a}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} = \left.\frac{\partial D(\boldsymbol{\omega}(a))}{\partial a}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \\
&= \left(\left.\frac{\partial D(\boldsymbol{\omega}(a))}{\partial \boldsymbol{\omega}(a)}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}\right)\left(\left.\frac{\partial \boldsymbol{\omega}(a)}{\partial a}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}\right) \\
&= \underbrace{\left.\frac{\partial D(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}} \times \boldsymbol{h} \\
&= 0 \quad , \quad \text{since } \boldsymbol{\omega}_o \text{ is a local minimum}
\end{aligned}$$

So, $\dot{\rho}_2 = 0$.

Deriving now $\ddot{\rho}_2$, by the Chain Rule,

$$\ddot{\rho}_2 \quad \equiv \quad \ddot{D}(\boldsymbol{\omega}(a)) \quad = \quad \frac{\partial^2 D(\boldsymbol{\omega}(a))}{\partial a^2} \quad = \quad \frac{\partial \boldsymbol{\omega}(a)}{\partial a} \cdot \frac{\partial^2 D(\boldsymbol{\omega}(a))}{\partial (\boldsymbol{\omega}(a))^2} \cdot \frac{\partial \boldsymbol{\omega}(a)}{\partial a}.$$

Thus,

$$\begin{aligned} \ddot{D}(\boldsymbol{\omega}(a))\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} &= \boldsymbol{h}' \quad \frac{\partial^2 D(\boldsymbol{\omega}(a))}{\partial (\boldsymbol{\omega}(a))^2}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \quad \boldsymbol{h} \\ &= \boldsymbol{h}' \quad \frac{\partial^2 D(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}^2}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \quad \boldsymbol{h}. \end{aligned}$$

This yields

$$c_{\boldsymbol{h}} \quad = \quad |\ddot{\rho}_2| \quad = \quad \left|\ddot{D}(\boldsymbol{\omega}(a))\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}\right| \quad = \quad \left|\, \boldsymbol{h}' \, \ddot{F} \, \boldsymbol{h} \,\right|, \tag{A.1}$$

where $\ddot{F}$ is a $q \times q$ matrix with elements

$$[\ddot{F}]_{jk} \quad = \quad \frac{\partial^2 D(\boldsymbol{\omega})}{\partial \omega_j \, \partial \omega_k}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}.$$

To evaluate $\ddot{F}$,

$$[\ddot{F}]_{jk} \;=\; \frac{\partial^2 D(\boldsymbol{\omega})}{\partial \omega_j \, \partial \omega_k}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \;=\; \frac{\partial^2 \left(Z(\widehat{\boldsymbol{\phi}}) - Z(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})\right)}{\partial \omega_j \, \partial \omega_k}\Bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \;=\; -\frac{\partial^2 Z(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})}{\partial \omega_j \, \partial \omega_k}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o}.$$

Applying the Chain Rule twice gives:

$$\ddot{F} \quad = \quad -J' \, \ddot{Z} \, J, \tag{A.2}$$

where J is a $p \times q$ matrix with elements

$$[J]_{ij} \quad = \quad \frac{\partial \hat{\phi}_{i,\boldsymbol{\omega}}}{\partial \omega_j}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o},$$

for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, q$, and $\ddot{Z}$ is a $p \times p$ matrix with elements

$$[\ddot{Z}]_{il} \quad = \quad \frac{\partial^2 Z(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}})}{\partial \hat{\phi}_{i,\boldsymbol{\omega}} \, \partial \hat{\phi}_{l,\boldsymbol{\omega}}}\bigg|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \qquad \text{or} \qquad [\ddot{Z}]_{il} \quad = \quad \frac{\partial^2 Z(\boldsymbol{\phi})}{\partial \phi_i \, \partial \phi_l}\bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}.$$

Substituting expression (A.2) into (A.1) gives

$$\begin{aligned} c_{\boldsymbol{h}} &= \left|\, \boldsymbol{h}' \, (-J' \, \ddot{Z} \, J) \, \boldsymbol{h} \,\right| \\ c_{\boldsymbol{h}} &= \left|\, \boldsymbol{h}' \, J' \, \ddot{Z} \, J \, \boldsymbol{h} \,\right|. \end{aligned}$$

Since $\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}$ is the maximum likelihood estimator for $\boldsymbol{\phi}_{\boldsymbol{\omega}}$ under the perturbed model,

$$\left.\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} = 0.$$

Differentiating this with respect to $\boldsymbol{\omega}$ and evaluating at $\boldsymbol{\omega} = \boldsymbol{\omega}_o$,

$$\frac{\partial}{\partial \boldsymbol{\omega}} \left(\left.\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} \right) = 0.$$

Recall the following result:

$$\text{If} \quad f(\omega) = f\left(g(\omega), h(\omega)\right), \quad \text{then} \quad \frac{\partial f}{\partial \omega} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial \omega} + \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial \omega}.$$

Letting

$$f\left(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}, \boldsymbol{\omega}\right) = \left.\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}},$$

the previous result implies:

$$\frac{\partial}{\partial \boldsymbol{\omega}} \left(f(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}, \boldsymbol{\omega}) \right) = \frac{\partial f(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}, \boldsymbol{\omega})}{\partial \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} \cdot \frac{\partial \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}} + \frac{\partial f(\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}, \boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \cdot \frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\omega}} = 0,$$

or, in terms of the log-likelihood,

$$\frac{\partial}{\partial \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} \left(\left.\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} \right) \cdot \frac{\partial \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}} + \frac{\partial}{\partial \boldsymbol{\omega}} \left(\left.\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}} \right) = 0.$$

Evaluating at $\boldsymbol{\omega} = \boldsymbol{\omega}_o$ gives:

$$\begin{aligned} 0 &= \left(\left.\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)}{\partial \phi_k \partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}} \right) \cdot \left(\left.\frac{\partial \widehat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=\boldsymbol{\omega}_o} \right) + \left(\left.\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}, \boldsymbol{\omega}=\boldsymbol{\omega}_o} \right) \\ 0 &= \ddot{L}\, J \quad + \quad \boldsymbol{\Delta}, \end{aligned}$$

from which

$$J = -(\ddot{L})^{-1}\, \boldsymbol{\Delta}.$$

Hence, a final expression for $c_{\boldsymbol{h}}$ can be obtained as:

$$\begin{aligned} c_{\boldsymbol{h}} &= \left| \boldsymbol{h}'\, J'\, \ddot{Z}\, J\, \boldsymbol{h} \right| \\ c_{\boldsymbol{h}} &= \left| \boldsymbol{h}'\, \boldsymbol{\Delta}'\, (\ddot{L})^{-1}\, \ddot{Z}\, (\ddot{L})^{-1}\, \boldsymbol{\Delta}\, \boldsymbol{h} \right|, \end{aligned}$$

with

- ▷ $\|\boldsymbol{h}\| \;=\; 1,$

- ▷ $\boldsymbol{\Delta}$ is a $p \times q$ matrix with elements:

$$[\Delta]_{ij} \quad = \quad \left.\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_i \, \partial \omega_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}},\boldsymbol{\omega}=\boldsymbol{\omega}_o},$$

- ▷ $\ddot{L}$ is a $p \times p$ matrix with elements:

$$[\ddot{L}]_{il} \quad = \quad \left.\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega}_o)}{\partial \phi_i \, \partial \phi_l}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}, \qquad \text{and,}$$

- ▷ $\ddot{Z}$ is a $p \times p$ matrix with elements:

$$[\ddot{Z}]_{il} \quad = \quad \left.\frac{\partial^2 Z(\boldsymbol{\phi})}{\partial \phi_i \, \partial \phi_l}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}.$$

Appendix B

Derivatives of the Log-Likelihood Function for the BRD Models

B.1 Local Influence with Perturbations in the Model Parameters

Log-Likelihood Expression

The observed-data log-likelihood for the BRD family of models is given by:

$$\begin{aligned}\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) \;=\; & \sum_{j,k} Z_{11,jk}\, \ln \pi_{11,jk} \;+\; \sum_{j} Z_{10,j+}\, \ln \pi_{10,j+} \;+\\ & \sum_{k} Z_{01,+k}\, \ln \pi_{01,+k} \;+\; Z_{00,++}\, \ln \pi_{00,++},\end{aligned} \tag{B.1}$$

where the cell probabilities, $\pi_{r_1 r_2,jk}$, via a selection model factorization, can be expressed in terms of the marginal response probabilities multiplied by the conditional probabilities of the response indicators given the outcomes, i.e.,

$$\pi_{r_1 r_2,jk} \;=\; p_{jk}\, q_{r_1 r_2|jk}, \tag{B.2}$$

with,

$$\begin{aligned} p_{jk} &= P(Y_1 = j, Y_2 = k) \qquad\qquad \text{and}\\ q_{r_1 r_2|jk} &= \frac{\exp\{\alpha_{jk}(1-r_1) + \beta_{jk}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{jk}) + \exp(\beta_{jk}) + \exp(\alpha_{jk} + \beta_{jk} + \gamma)}.\end{aligned} \tag{B.3}$$

Alternatively, using the cell indexing system in Table 7.2, terms in (B.1) can be indexed by the labels for the observed cells, $c = 1, 2, \ldots, 9$, thereby yielding the following re-expression:

$$\ell(\boldsymbol{\phi}|\boldsymbol{\omega}) \;=\; \sum_{c=1}^{9} Z_c \ln \pi_c \;=\; \sum_{c=1}^{9} L_c. \tag{B.4}$$

The 2$^{\text{nd}}$ partial derivatives of $\ell(\boldsymbol{\phi}|\boldsymbol{\omega})$ with respect to the elements of $\boldsymbol{\omega}$ and the elements of $\boldsymbol{\phi}$ are required, i.e.,

$$\frac{\partial^2 \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\phi_s \partial\omega_i} \;=\; \frac{\partial^2}{\partial\phi_s\partial\omega_i}\sum_{c=1}^{9} Z_c \,\ln \pi_c \;=\; \sum_{c=1}^{9} \frac{\partial^2 Z_c \ln \pi_c}{\partial\phi_s\partial\omega_i} \;=\; \sum_{c=1}^{9}\frac{\partial^2 L_c}{\partial\phi_s\partial\omega_i}. \tag{B.5}$$

Taking derivatives first with respect to $\boldsymbol{\omega}$:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \omega_i} &= \sum_{c=1}^{9} \frac{\partial L_c}{\partial \omega_i} \\
&= \sum_{c=1}^{9} \frac{\partial Z_c \ln \pi_c}{\partial \omega_i} \\
&= \sum_{c=1}^{9} Z_c \, \frac{\partial \ln \pi_c}{\partial \omega_i} \\
&= \sum_{c=1}^{9} Z_c \, \frac{\partial p_c q_{r_1 r_2|c}}{\partial \omega_i} \\
&= \sum_{c=1}^{9} Z_c \, p_c \, \frac{\partial q_{r_1 r_2|c}}{\partial \omega_i} \\
\frac{\partial \ell(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \omega_i} &= \sum_{j=1}^{2}\sum_{k=1}^{2} Z_{r_1 r_2, jk} \, p_{jk} \, \underbrace{\frac{\partial q_{r_1 r_2|jk}}{\partial \omega_i}}_{\text{see Section B.1.1}}
\end{aligned}$$

Differentiating further with respect to $\boldsymbol{\phi}$:

$$\begin{aligned}
\frac{\partial^2 \ell_{(}\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_s \partial \omega_i} &= \sum_{c=1}^{9} Z_c \, \frac{\partial}{\partial \phi_s} \left(p_c \, \frac{\partial q_{r_1 r_2|c}}{\partial \omega_i} \right) \\
&= \sum_{c=1}^{9} Z_c \, \left(\frac{\partial p_c}{\partial \phi_s} \, \frac{\partial q_{r_1 r_2|c}}{\partial \omega_i} \; + \; p_c \, \frac{\partial^2 q_{r_1 r_2|c}}{\partial \phi_s \partial \omega_i} \right) \\
&= \sum_{j=1}^{2}\sum_{k=1}^{2} Z_{r_1 r_2, jk} \, \left(\frac{\partial p_{jk}}{\partial \phi_s} \, \frac{\partial q_{r_1 r_2|jk}}{\partial \omega_i} \; + \; p_{jk} \, \frac{\partial^2 q_{r_1 r_2|jk}}{\partial \phi_s \partial \omega_i} \right) \\
\frac{\partial^2 \ell_{(}\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial \phi_s \partial \omega_i} &= \sum_{j=1}^{2}\sum_{k=1}^{2} Z_{r_1 r_2, jk} \, \left[\underbrace{\frac{\partial p_{jk}}{\partial \phi_s}}_{\S\text{ B.1.3}} \, \underbrace{\frac{\partial q_{r_1 r_2|jk}}{\partial \omega_i}}_{\S\text{ B.1.1}} + p_{jk} \, \underbrace{\frac{\partial}{\phi_s} \left(\frac{\partial q_{r_1 r_2|jk}}{\partial \omega_i} \right)}_{\S\text{ B.1.2}} \right]
\end{aligned}$$

The BRD Models

Each of the 9 BRD models consists of a different set of parameters. For ease of notation, let

$$p_{11} \equiv p_1, \qquad p_{12} \equiv p_2, \qquad p_{21} \equiv p_3, \qquad \text{and} \qquad p_{22} = 1 - p_1 - p_2 - p_3.$$

The respective parameter vectors for each of the 9 BRD models are enumerated in the following table.

BRD Model		Parameter Vector
		$\boldsymbol{\phi} = (p_1, p_2, p_3, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)'$
BRD1:	$(\alpha_{..}, \beta_{..})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{..}, \beta_{..}, \gamma)'$
BRD2:	$(\alpha_{..}, \beta_{j.})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{..}, \beta_{1.}, \beta_{2.}, \gamma)'$
BRD3:	$(\alpha_{.k}, \beta_{..})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{..}, \gamma)'$
BRD4:	$(\alpha_{..}, \beta_{.k})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{..}, \beta_{.1}, \beta_{.2}, \gamma)'$
BRD5:	$(\alpha_{j.}, \beta_{..})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{..}, \gamma)'$
BRD6:	$(\alpha_{j.}, \beta_{j.})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{1.}, \beta_{2.}, \gamma)'$
BRD7:	$(\alpha_{.k}, \beta_{.k})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{.1}, \beta_{.2}, \gamma)'$
BRD8:	$(\alpha_{j.}, \beta_{.k})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{.1}, \beta_{.2}, \gamma)'$
BRD9:	$(\alpha_{.k}, \beta_{j.})'$	$\boldsymbol{\phi} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{1.}, \beta_{2.}, \gamma)'$

Within Figure 3.1, a total of 12 model nestings are possible, and for each case, the log-likelihood expression (B.1) and the cell probabilities defined in (B.2) retain the same expressions. However, the general expression (B.3) for $q_{r_1 r_2 | jk}$ will change depending on the particular BRD model pair of interest. These expressions shall now be defined for each of the possible nested model pairs.

BRD1 *vs.* BRD2

$$\begin{aligned}
\text{BRD1:}\quad (\alpha_{..},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD1}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\gamma)' \\
\text{BRD2:}\quad (\alpha_{..},\beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD2}} = (p_1,p_2,p_3,\alpha_{..},\beta_{1.},\beta_{2.},\gamma)' \\
&\qquad\quad \text{with:}\quad \beta_{1.}=\beta_{..} \quad\text{and}\quad \beta_{2.}=\beta_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD2}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\beta_{..}+\omega_i,\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|1k} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..})+\exp(\alpha_{..}+\beta_{..}+\gamma)} \\
q_{r_1r_2|2k} &= \frac{\exp\{\alpha_{..}(1-r_1)+(\beta_{..}+\omega_i)(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..}+\omega_i)+\exp(\alpha_{..}+\beta_{..}+\omega_i+\gamma)}
\end{aligned}$$

BRD1 *vs.* BRD3

$$\begin{aligned}
\text{BRD1:}\quad (\alpha_{..},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD1}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\gamma)' \\
\text{BRD3:}\quad (\alpha_{.k},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD3}} = (p_1,p_2,p_3,\alpha_{.1},\alpha_{.2},\beta_{..},\gamma)' \\
&\qquad\quad \text{with:}\quad \alpha_{.1}=\alpha_{..} \quad\text{and}\quad \alpha_{.2}=\alpha_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD3}} = (p_1,p_2,p_3,\alpha_{..},\alpha_{..}+\omega_i,\beta_{..},\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|j1} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..})+\exp(\alpha_{..}+\beta_{..}+\gamma)} \\
q_{r_1r_2|j2} &= \frac{\exp\{(\alpha_{..}+\omega_i)(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..}+\omega_i)+\exp(\beta_{..})+\exp(\alpha_{..}+\omega_i+\beta_{..}+\gamma)}
\end{aligned}$$

BRD1 *vs.* BRD4

$$\begin{aligned}
\text{BRD1:}\quad (\alpha_{..},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD1}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\gamma)' \\
\text{BRD4:}\quad (\alpha_{..},\beta_{.k})' &\implies \boldsymbol{\phi}_{\text{BRD4}} = (p_1,p_2,p_3,\alpha_{..},\beta_{.1},\beta_{.2},\gamma)' \\
&\qquad\quad \text{with:}\quad \beta_{.1}=\beta_{..} \quad\text{and}\quad \beta_{.2}=\beta_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD4}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\beta_{..}+\omega_i,\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|j1} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..})+\exp(\alpha_{..}+\beta_{..}+\gamma)} \\
q_{r_1r_2|j2} &= \frac{\exp\{\alpha_{..}(1-r_1)+(\beta_{..}+\omega_i)(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..}+\omega_i)+\exp(\alpha_{..}+\beta_{..}+\omega_i+\gamma)}
\end{aligned}$$

BRD1 *vs.* BRD5

$$\begin{aligned}
\text{BRD1:}\quad (\alpha_{..},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD1}} = (p_1,p_2,p_3,\alpha_{..},\beta_{..},\gamma)' \\
\text{BRD5:}\quad (\alpha_{j.},\beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD5}} = (p_1,p_2,p_3,\alpha_{1.},\alpha_{2.},\beta_{..},\gamma)' \\
&\qquad \text{with:}\quad \alpha_{1.}=\alpha_{..} \quad\text{and}\quad \alpha_{2.}=\alpha_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD5}} = (p_1,p_2,p_3,\alpha_{..},\alpha_{..}+\omega_i,\beta_{..},\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|1k} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{..})+\exp(\alpha_{..}+\beta_{..}+\gamma)} \\
q_{r_1r_2|2k} &= \frac{\exp\{(\alpha_{..}+\omega_i)(1-r_1)+\beta_{..}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..}+\omega_i)+\exp(\beta_{..})+\exp(\alpha_{..}+\omega_i+\beta_{..}+\gamma)}
\end{aligned}$$

BRD2 *vs.* BRD6

$$\begin{aligned}
\text{BRD2:}\quad (\alpha_{..},\beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD2}} = (p_1,p_2,p_3,\alpha_{..},\beta_{1.},\beta_{2.},\gamma)' \\
\text{BRD6:}\quad (\alpha_{j.},\beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD6}} = (p_1,p_2,p_3,\alpha_{1.},\alpha_{2.},\beta_{1.},\beta_{2.},\gamma)' \\
&\qquad \text{with:}\quad \alpha_{1.}=\alpha_{..} \quad\text{and}\quad \alpha_{2.}=\alpha_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD6}} = (p_1,p_2,p_3,\alpha_{..},\alpha_{..}+\omega_i,\beta_{1.},\beta_{2.},\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|1k} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{1.}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{1.})+\exp(\alpha_{..}+\beta_{1.}+\gamma)} \\
q_{r_1r_2|2k} &= \frac{\exp\{(\alpha_{..}+\omega_i)(1-r_1)+\beta_{2.}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..}+\omega_i)+\exp(\beta_{2.})+\exp(\alpha_{..}+\omega_i+\beta_{2.}+\gamma)}
\end{aligned}$$

BRD2 *vs.* BRD9

$$\begin{aligned}
\text{BRD2:}\quad (\alpha_{..},\beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD2}} = (p_1,p_2,p_3,\alpha_{..},\beta_{1.},\beta_{2.},\gamma)' \\
\text{BRD9:}\quad (\alpha_{.k},\beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD9}} = (p_1,p_2,p_3,\alpha_{.1},\alpha_{.2},\beta_{1.},\beta_{2.},\gamma)' \\
&\qquad \text{with:}\quad \alpha_{.1}=\alpha_{..} \quad\text{and}\quad \alpha_{.2}=\alpha_{..}+\omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD9}} = (p_1,p_2,p_3,\alpha_{..},\alpha_{..}+\omega_i,\beta_{1.},\beta_{2.},\gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1r_2|j1} &= \frac{\exp\{\alpha_{..}(1-r_1)+\beta_{j.}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..})+\exp(\beta_{j.})+\exp(\alpha_{..}+\beta_{j.}+\gamma)} \\
q_{r_1r_2|j2} &= \frac{\exp\{(\alpha_{..}+\omega_i)(1-r_1)+\beta_{j.}(1-r_2)+\gamma(1-r_1)(1-r_2)\}}{1+\exp(\alpha_{..}+\omega_i)+\exp(\beta_{j.})+\exp(\alpha_{..}+\omega_i+\beta_{j.}+\gamma)}
\end{aligned}$$

BRD3 *vs.* BRD7

$$
\begin{aligned}
\text{BRD3:}\quad (\alpha_{.k}, \beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD3}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{..}, \gamma)' \\
\text{BRD7:}\quad (\alpha_{.k}, \beta_{.k})' &\implies \boldsymbol{\phi}_{\text{BRD7}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{.1}, \beta_{.2}, \gamma)' \\
&\qquad \text{with:}\quad \beta_{.1} = \beta_{..} \quad \text{and} \quad \beta_{.2} = \beta_{..} + \omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD7}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{..}, \beta_{2.} + \omega_i, \gamma)'
\end{aligned}
$$

$$
\begin{aligned}
q_{r_1 r_2|j1} &= \frac{\exp\{\alpha_{.1}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{.1}) + \exp(\beta_{..}) + \exp(\alpha_{.1} + \beta_{..} + \gamma)} \\
q_{r_1 r_2|j2} &= \frac{\exp\{\alpha_{.2}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{.2}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{.2} + \beta_{..} + \omega_i + \gamma)}
\end{aligned}
$$

BRD3 *vs.* BRD9

$$
\begin{aligned}
\text{BRD3:}\quad (\alpha_{.k}, \beta_{..})' &\implies \boldsymbol{\phi}_{\text{BRD3}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{..}, \gamma)' \\
\text{BRD9:}\quad (\alpha_{.k}, \beta_{j.})' &\implies \boldsymbol{\phi}_{\text{BRD9}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{1.}, \beta_{2.}, \gamma)' \\
&\qquad \text{with:}\quad \beta_{1.} = \beta_{..} \quad \text{and} \quad \beta_{2.} = \beta_{..} + \omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD9}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{..}, \beta_{2.} + \omega_i, \gamma)'
\end{aligned}
$$

$$
\begin{aligned}
q_{r_1 r_2|1k} &= \frac{\exp\{\alpha_{.k}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{.k}) + \exp(\beta_{..}) + \exp(\alpha_{.k} + \beta_{..} + \gamma)} \\
q_{r_1 r_2|2k} &= \frac{\exp\{\alpha_{.k}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{.k}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{.k} + \beta_{..} + \omega_i + \gamma)}
\end{aligned}
$$

BRD4 *vs.* BRD7

$$
\begin{aligned}
\text{BRD4:}\quad (\alpha_{..}, \beta_{.k})' &\implies \boldsymbol{\phi}_{\text{BRD4}} = (p_1, p_2, p_3, \alpha_{..}, \beta_{.1}, \beta_{.2}, \gamma)' \\
\text{BRD7:}\quad (\alpha_{.k}, \beta_{.k})' &\implies \boldsymbol{\phi}_{\text{BRD7}} = (p_1, p_2, p_3, \alpha_{.1}, \alpha_{.2}, \beta_{.1}, \beta_{.2}, \gamma)' \\
&\qquad \text{with:}\quad \alpha_{.1} = \alpha_{..} \quad \text{and} \quad \alpha_{.2} = \alpha_{..} + \omega_i \\
&\implies \boldsymbol{\phi}_{\text{BRD7}} = (p_1, p_2, p_3, \alpha_{..}, \alpha_{..} + \omega_i, \beta_{.1}, \beta_{.2}, \gamma)'
\end{aligned}
$$

$$
\begin{aligned}
q_{r_1 r_2|j1} &= \frac{\exp\{\alpha_{..}(1-r_1) + \beta_{.1}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{.1}) + \exp(\alpha_{..} + \beta_{.1} + \gamma)} \\
q_{r_1 r_2|j2} &= \frac{\exp\{(\alpha_{..} + \omega_i)(1-r_1) + \beta_{.2}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..} + \omega_i) + \exp(\beta_{.2}) + \exp(\alpha_{..} + \omega_i + \beta_{.2} + \gamma)}
\end{aligned}
$$

BRD4 *vs.* BRD8

$$\begin{aligned}
\text{BRD4:} \quad (\alpha_{..}, \beta_{.k})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD4}} = (p_1, p_2, p_3, \alpha_{..}, \beta_{.1}, \beta_{.2}, \gamma)' \\
\text{BRD8:} \quad (\alpha_{j.}, \beta_{.k})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD8}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{.1}, \beta_{.2}, \gamma)' \\
&\qquad \text{with:} \quad \alpha_{1.} = \alpha_{..} \quad \text{and} \quad \alpha_{2.} = \alpha_{..} + \omega_i \\
&\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD8}} = (p_1, p_2, p_3, \alpha_{..}, \alpha_{..} + \omega_i, \beta_{.1}, \beta_{.2}, \gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1 r_2|1k} &= \frac{\exp\{\alpha_{..}(1-r_1) + \beta_{.k}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..}) + \exp(\beta_{.k}) + \exp(\alpha_{..} + \beta_{.k} + \gamma)} \\
q_{r_1 r_2|2k} &= \frac{\exp\{(\alpha_{..} + \omega_i)(1-r_1) + \beta_{.k}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{..} + \omega_i) + \exp(\beta_{.k}) + \exp(\alpha_{..} + \omega_i + \beta_{.k} + \gamma)}
\end{aligned}$$

BRD5 *vs.* BRD6

$$\begin{aligned}
\text{BRD5:} \quad (\alpha_{j.}, \beta_{..})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD5}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{..}, \gamma)' \\
\text{BRD6:} \quad (\alpha_{j.}, \beta_{j.})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD6}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{1.}, \beta_{2.}, \gamma)' \\
&\qquad \text{with:} \quad \beta_{1.} = \beta_{..} \quad \text{and} \quad \beta_{2.} = \beta_{..} + \omega_i \\
&\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD6}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{..}, \beta_{..} + \omega_i, \gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1 r_2|1k} &= \frac{\exp\{\alpha_{1.}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{1.}) + \exp(\beta_{..}) + \exp(\alpha_{1.} + \beta_{..} + \gamma)} \\
q_{r_1 r_2|2k} &= \frac{\exp\{\alpha_{2.}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{2.}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{2.} + \beta_{..} + \omega_i + \gamma)}
\end{aligned}$$

BRD5 *vs.* BRD8

$$\begin{aligned}
\text{BRD5:} \quad (\alpha_{j.}, \beta_{..})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD5}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{..}, \gamma)' \\
\text{BRD8:} \quad (\alpha_{j.}, \beta_{.k})' \quad &\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD8}} = (p_1, p_2, p_3, \alpha_{1.}, \alpha_{2.}, \beta_{.1}, \beta_{.2}, \gamma)' \\
&\qquad \text{with:} \quad \beta_{.1} = \beta_{..} \quad \text{and} \quad \beta_{.2} = \beta_{..} + \omega_i \\
&\Longrightarrow \quad \boldsymbol{\phi}_{\text{BRD8}} = (p_1, p_2, p_3, \alpha_{..}, \alpha_{..} + \omega_i, \beta_{.1}, \beta_{.2}, \gamma)'
\end{aligned}$$

$$\begin{aligned}
q_{r_1 r_2|j1} &= \frac{\exp\{\alpha_{j.}(1-r_1) + \beta_{..}(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{j.}) + \exp(\beta_{..}) + \exp(\alpha_{j.} + \beta_{..} + \gamma)} \\
q_{r_1 r_2|j2} &= \frac{\exp\{\alpha_{j.}(1-r_1) + (\beta_{..} + \omega_i)(1-r_2) + \gamma(1-r_1)(1-r_2)\}}{1 + \exp(\alpha_{j.}) + \exp(\beta_{..} + \omega_i) + \exp(\alpha_{j.} + \beta_{..} + \omega_i + \gamma)}
\end{aligned}$$

B.1.1 Derivatives of $q_{r_1 r_2|jk}$ with respect to $\boldsymbol{\omega}$

BRD1 *vs.* BRD2

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{+0|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|2k}\,(1 - q_{+0|2k}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|2k}\, q_{+0|2k} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\,(1 - q_{+0|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD1 *vs.* BRD3

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|j2}\,(1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\,(1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD1 *vs.* BRD4

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{+0|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|j2}\,(1 - q_{+0|j2}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|j2}\, q_{+0|j2} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\,(1 - q_{+0|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD1 *vs.* BRD5

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|2k}\,(1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\,(1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD2 *vs.* BRD6

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|2k}\, (1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\, (1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD2 *vs.* BRD9

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|j2}\, (1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\, (1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD3 *vs.* BRD7

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{+0|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|j2}\, (1 - q_{+0|j2}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|j2}\, q_{+0|j2} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\, (1 - q_{+0|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD3 *vs.* BRD9

$$\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{+0|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|2k}\, (1 - q_{+0|2k}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|2k}\, q_{+0|2k} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\, (1 - q_{+0|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}$$

BRD4 *vs.* BRD7

$$
\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|j2}\, q_{0+|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|j2}\,(1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\,(1 - q_{0+|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}
$$

BRD4 *vs.* BRD8

$$
\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ -q_{10|2k}\, q_{0+|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ q_{01|2k}\,(1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\,(1 - q_{0+|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}
$$

BRD5 *vs.* BRD6

$$
\begin{aligned}
\frac{\partial q_{r_1 r_2|1k}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|2k}}{\partial \omega_i} &= \begin{cases} -q_{11|2k}\, q_{+0|2k} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|2k}\,(1 - q_{+0|2k}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|2k}\, q_{+0|2k} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|2k}\,(1 - q_{+0|2k}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}
$$

BRD5 *vs.* BRD8

$$
\begin{aligned}
\frac{\partial q_{r_1 r_2|j1}}{\partial \omega_i} &= 0 \\
\frac{\partial q_{r_1 r_2|j2}}{\partial \omega_i} &= \begin{cases} -q_{11|j2}\, q_{+0|j2} & \text{for } r_1 = 1 \text{ and } r_2 = 1 \\ q_{10|j2}\,(1 - q_{+0|j2}) & \text{for } r_1 = 1 \text{ and } r_2 = 0 \\ -q_{01|j2}\, q_{+0|j2} & \text{for } r_1 = 0 \text{ and } r_2 = 1 \\ q_{00|j2}\,(1 - q_{+0|j2}) & \text{for } r_1 = 0 \text{ and } r_2 = 0 \end{cases}
\end{aligned}
$$

B.1.2 Derivatives of $q_{r_1 r_2|jk}$ with respect to ϕ

Derivatives of $q_{11|jk}$

$$\frac{\partial q_{11|jk}}{\partial p_{j'k'}} = 0, \quad \forall \quad (j',k') \quad \text{and} \quad \forall \quad (j,k)$$

$$\frac{\partial q_{11|jk}}{\partial \alpha_{j'k'}} = \begin{cases} -q_{11|jk}\, q_{0+|jk} & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{11|jk}}{\partial \beta_{j'k'}} = \begin{cases} -q_{11|jk}\, q_{+0|jk} & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{11|jk}}{\partial \gamma} = -q_{11|jk}\, q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{10|jk}$

$$\frac{\partial q_{10|jk}}{\partial p_{j'k'}} = 0, \quad \forall \quad (j',k') \quad \text{and} \quad \forall \quad (j,k)$$

$$\frac{\partial q_{10|jk}}{\partial \alpha_{j'k'}} = \begin{cases} -q_{10|jk}\, q_{0+|jk} & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{10|jk}}{\partial \beta_{j'k'}} = \begin{cases} q_{10|jk}\,(1-q_{+0|jk}) & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{10|jk}}{\partial \gamma} = -q_{10|jk}\, q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{01|jk}$

$$\frac{\partial q_{01|jk}}{\partial p_{j'k'}} = 0, \quad \forall \quad (j',k') \quad \text{and} \quad \forall \quad (j,k)$$

$$\frac{\partial q_{01|jk}}{\partial \alpha_{j'k'}} = \begin{cases} q_{01|jk}\,(1-q_{0+|jk}) & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{01|jk}}{\partial \beta_{j'k'}} = \begin{cases} -q_{01|jk}\,q_{+0|jk} & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{01|jk}}{\partial \gamma} = -q_{01|jk}\,q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{00|jk}$

$$\frac{\partial q_{00|jk}}{\partial p_{j'k'}} = 0, \quad \forall \quad (j',k') \quad \text{and} \quad \forall \quad (j,k)$$

$$\frac{\partial q_{00|jk}}{\partial \alpha_{j'k'}} = \begin{cases} q_{00|jk}\,(1-q_{0+|jk}) & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{00|jk}}{\partial \beta_{j'k'}} = \begin{cases} q_{00|jk}\,(1-q_{+0|jk}) & \text{if } (j',k')=(j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{00|jk}}{\partial \gamma} = q_{00|jk}\,(1-q_{00|jk}), \quad \forall \quad (j,k)$$

B.1.3 Derivatives of p_{jk} with respect to ϕ

BRD1 *vs.* BRD2-5

$$\frac{\partial p_{jk}}{\partial p_1} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \alpha_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_2} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=2 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \beta_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_3} = \begin{cases} 1 & \text{for } j=2 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \gamma_{..}} = 0$$

BRD2 *vs.* BRD6&9

$$\frac{\partial p_{jk}}{\partial p_1} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \alpha_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_2} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=2 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \beta_{j.}} = 0, \text{ for } j=1,2$$

$$\frac{\partial p_{jk}}{\partial p_3} = \begin{cases} 1 & \text{for } j=2 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \gamma_{..}} = 0$$

BRD3 *vs.* BRD7&9

$$\frac{\partial p_{jk}}{\partial p_1} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \alpha_{.k}} = 0, \text{ for } k=1,2$$

$$\frac{\partial p_{jk}}{\partial p_2} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=2 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \beta_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_3} = \begin{cases} 1 & \text{for } j=2 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \gamma_{..}} = 0$$

BRD4 *vs.* BRD7&8

$$\frac{\partial p_{jk}}{\partial p_1} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \alpha_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_2} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=2 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \beta_{.k}} = 0, \text{ for } k=1,2$$

$$\frac{\partial p_{jk}}{\partial p_3} = \begin{cases} 1 & \text{for } j=2 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \gamma_{..}} = 0$$

BRD5 *vs.* BRD6&8

$$\frac{\partial p_{jk}}{\partial p_1} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \alpha_{j.}} = 0, \text{ for } j=1,2$$

$$\frac{\partial p_{jk}}{\partial p_2} = \begin{cases} 1 & \text{for } j=1 \text{ and } k=2 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \beta_{..}} = 0$$

$$\frac{\partial p_{jk}}{\partial p_3} = \begin{cases} 1 & \text{for } j=2 \text{ and } k=1 \\ -1 & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases} \qquad \frac{\partial p_{jk}}{\partial \gamma_{..}} = 0$$

B.2 Local Influence with Perturbations in Cell Probabilities

Log-Likelihood Expression

The perturbed observed-data log-likelihood for the BRD family of models is given by:

$$\begin{aligned}\ell_\omega \;=\; & \sum_{j,k}\left(Z_{11,jk}+N\omega_{11,jk}\right)\ln\pi_{11,jk} \;+\; \sum_{j}\left(Z_{10,j+}+N\omega_{10,j+}\right)\ln\pi_{10,j+} \;+\\ & \sum_{k}\left(Z_{01,+k}+N\omega_{01,+k}\right)\ln\pi_{01,+k} \;+\; \left(Z_{00,++}+N\omega_{00,++}\right)\ln\pi_{00,++},\end{aligned}$$

where the cell probabilities, $\pi_{r_1r_2,jk}$, via a selection model factorization, can be expressed in terms of the marginal response probabilities multiplied by the conditional probabilities of the response indicators given the outcomes, i.e.,

$$\pi_{r_1r_2,jk} \;=\; p_{jk}\, q_{r_1r_2|jk},$$

with,

$$\begin{aligned}p_{jk} \;&=\; P(Y_1=j,Y_2=k) \qquad\qquad \text{and}\\ q_{r_1r_2|jk} \;&=\; \frac{\exp\left\{\alpha_{jk}(1-r_1)+\beta_{jk}(1-r_2)+\gamma(1-r_1)(1-r_2)\right\}}{1+\exp\left(\alpha_{jk}\right)+\exp\left(\beta_{jk}\right)+\exp\left(\alpha_{jk}+\beta_{jk}+\gamma\right)}.\end{aligned}$$

As before, the 2nd partial derivatives of ℓ_ω with respect to the elements of $\boldsymbol{\omega}$ and the elements of $\boldsymbol{\phi}$ are required, i.e.,

$$\frac{\partial^2\ell_\omega}{\partial\phi_i\,\partial\omega_{r_1r_2,jk}}.$$

Consider taking first derivatives with respect to $\boldsymbol{\omega}$,

$$\frac{\partial\ell_\omega}{\partial\omega_{r_1r_2,jk}} \;=\; \ln\pi_{r_1r_2,jk},$$

and then further differentiating with respect to $\boldsymbol{\phi}$ yields:

$$\frac{\partial^2\ell_\omega}{\partial\phi_i\,\partial\omega_{r_1r_2,jk}} \;=\; \frac{1}{\pi_{r_1r_2,jk}}\underbrace{\frac{\partial\pi_{r_1r_2,jk}}{\partial\phi_i}}_{\text{see Section B.2.2}}.$$

B.2.1 Derivatives of the Log-Likelihood ℓ_ω with respect to $\boldsymbol{\omega}$ and with respect to $\boldsymbol{\phi}$

For $r_1 = 1$ and $r_2 = 1$:

$$\frac{\partial \ell_\omega}{\partial \omega_{11,jk}} = \ln \pi_{11,jk}$$

$$\frac{\partial^2 \ell_\omega}{\partial \phi_i \, \partial \omega_{11,jk}} = \frac{1}{\pi_{11,jk}} \frac{\partial \pi_{11,jk}}{\partial \phi_i}$$

For $r_1 = 1$ and $r_2 = 0$:

$$\frac{\partial \ell_\omega}{\partial \omega_{10,j+}} = \ln \pi_{10,j+} = \ln\left(\pi_{10,j1} + \pi_{10,j2}\right)$$

$$\frac{\partial^2 \ell_\omega}{\partial \phi_i \, \partial \omega_{10,j+}} = \frac{1}{\pi_{10,j+}} \frac{\partial \pi_{10,j+}}{\partial \phi_i} = \frac{1}{\pi_{10,j+}} \left(\frac{\partial \pi_{10,j1}}{\partial \phi_i} + \frac{\partial \pi_{10,j2}}{\partial \phi_i} \right)$$

For $r_1 = 0$ and $r_2 = 1$:

$$\frac{\partial \ell_\omega}{\partial \omega_{01,+k}} = \ln \pi_{01,+k} = \ln\left(\pi_{01,1k} + \pi_{01,2k}\right)$$

$$\frac{\partial^2 \ell_\omega}{\partial \phi_i \, \partial \omega_{01,+k}} = \frac{1}{\pi_{01,+k}} \frac{\partial \pi_{01,+k}}{\partial \phi_i} = \frac{1}{\pi_{01,+k}} \left(\frac{\partial \pi_{01,1k}}{\partial \phi_i} + \frac{\partial \pi_{01,2k}}{\partial \phi_i} \right)$$

For $r_1 = 0$ and $r_2 = 0$:

$$\frac{\partial \ell_\omega}{\partial \omega_{00,++}} = \ln \pi_{00,++} = \ln\left(\pi_{00,11} + \pi_{00,12} + \pi_{00,21} + \pi_{00,22}\right)$$

$$\begin{aligned}\frac{\partial^2 \ell_\omega}{\partial \phi_i \, \partial \omega_{00,++}} &= \frac{1}{\pi_{00,++}} \frac{\partial \pi_{00,++}}{\partial \phi_i} \\ &= \frac{1}{\pi_{00,++}} \left(\frac{\partial \pi_{00,11}}{\partial \phi_i} + \frac{\partial \pi_{00,12}}{\partial \phi_i} + \frac{\partial \pi_{00,21}}{\partial \phi_i} + \frac{\partial \pi_{00,22}}{\partial \phi_i} \right)\end{aligned}$$

B.2.2 Derivatives of $\pi_{r_1r_2,jk}$ with respect to ϕ

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial p_{11}} = \begin{cases} q_{r_1r_2|jk} & \text{for } j=1 \text{ and } k=1 \\ -q_{r_1r_2|jk} & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial p_{12}} = \begin{cases} q_{r_1r_2|jk} & \text{for } j=1 \text{ and } k=2 \\ -q_{r_1r_2|jk} & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial p_{21}} = \begin{cases} q_{r_1r_2|jk} & \text{for } j=2 \text{ and } k=1 \\ -q_{r_1r_2|jk} & \text{for } j=2 \text{ and } k=2 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial\alpha_{j'k'}} = \underbrace{\frac{\partial\pi_{r_1r_2,jk}}{\partial q_{r_1r_2|jk}}}_{=p_{jk}} \quad \underbrace{\frac{\partial q_{r_1r_2|jk}}{\partial\alpha_{j'k'}}}_{\text{see § B.2.3}}$$

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial\beta_{j'k'}} = \underbrace{\frac{\partial\pi_{r_1r_2,jk}}{\partial q_{r_1r_2|jk}}}_{=p_{jk}} \quad \underbrace{\frac{\partial q_{r_1r_2|jk}}{\partial\beta_{j'k'}}}_{\text{see § B.2.3}}$$

$$\frac{\partial\pi_{r_1r_2,jk}}{\partial\gamma} = \underbrace{\frac{\partial\pi_{r_1r_2,jk}}{\partial q_{r_1r_2|jk}}}_{=p_{jk}} \quad \underbrace{\frac{\partial q_{r_1r_2|jk}}{\partial\gamma}}_{\text{see § B.2.3}}$$

B.2.3 Derivatives of $q_{r_1 r_2|jk}$ with respect to ϕ

Derivatives of $q_{11|jk}$

$$\frac{\partial q_{11|jk}}{\partial \alpha_{j'k'}} = \begin{cases} -q_{11|jk}\, q_{0+|jk} & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{11|jk}}{\partial \beta_{j'k'}} = \begin{cases} -q_{11|jk}\, q_{+0|jk} & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{11|jk}}{\partial \gamma} = -q_{11|jk}\, q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{10|jk}$

$$\frac{\partial q_{10|jk}}{\partial \alpha_{j'k'}} = \begin{cases} -q_{10|jk}\, q_{0+|jk} & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{10|jk}}{\partial \beta_{j'k'}} = \begin{cases} q_{10|jk}\, (1 - q_{+0|jk}) & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{10|jk}}{\partial \gamma} = -q_{10|jk}\, q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{01|jk}$

$$\frac{\partial q_{01|jk}}{\partial \alpha_{j'k'}} = \begin{cases} q_{01|jk}\, (1 - q_{0+|jk}) & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{01|jk}}{\partial \beta_{j'k'}} = \begin{cases} -q_{01|jk}\, q_{+0|jk} & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{01|jk}}{\partial \gamma} = -q_{01|jk}\, q_{00|jk}, \quad \forall \quad (j,k)$$

Derivatives of $q_{00|jk}$

$$\frac{\partial q_{00|jk}}{\partial \alpha_{j'k'}} = \begin{cases} q_{00|jk}\, (1 - q_{0+|jk}) & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{00|jk}}{\partial \beta_{j'k'}} = \begin{cases} q_{00|jk}\, (1 - q_{+0|jk}) & \text{if } (j',k') = (j,k) \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial q_{00|jk}}{\partial \gamma} = q_{00|jk}\, (1 - q_{00|jk}), \quad \forall \quad (j,k)$$

Appendix C

Derivation of Asymptotic Variances of Marginalized Estimates of Pattern-Mixture Models

Marginalized Parameters

For the case of 3 time points, and hence, 3 dropout patterns, the marginalized parameters of a PMM as defined by Jansen and Molenberghs (2007) are

$$A_{JM} = \operatorname{logit}\left(\sum_{k=1}^{3} \pi_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}}\right) \text{ and } B_{JM} = \frac{\sum_{k=1}^{3} \pi_k \beta_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}} \frac{1}{1+e^{\alpha_k}}}{\left(\sum_{k=1}^{3} \pi_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}}\right)\left(\sum_{k=1}^{3} \pi_k \frac{1}{1+e^{\alpha_k}}\right)}, \quad \text{(C.1)}$$

with corresponding estimators given by

$$\widehat{A}_{JM} = \operatorname{logit}\left(\sum_{k=1}^{3} \widehat{\pi}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}}\right) \quad \text{and} \quad \widehat{B}_{JM} = \frac{\sum_{k=1}^{3} \widehat{\pi}_k \widehat{\beta}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}} \frac{1}{1+e^{\widehat{\alpha}_k}}}{\left(\sum_{k=1}^{3} \widehat{\pi}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}}\right)\left(\sum_{k=1}^{3} \widehat{\pi}_k \frac{1}{1+e^{\widehat{\alpha}_k}}\right)}.$$

The marginalized parameters of a PMM as defined by Park and Lee (1999) are

$$A_{PL} = \sum_{k=1}^{3} \pi_k \alpha_k \qquad \text{and} \qquad B_{PL} = \sum_{k=1}^{3} \pi_k \beta_k, \tag{C.2}$$

with corresponding estimators

$$\widehat{A}_{PL} = \sum_{k=1}^{3} \widehat{\pi}_k \widehat{\alpha}_k \qquad \text{and} \qquad \widehat{B}_{PL} = \sum_{k=1}^{3} \widehat{\pi}_k \widehat{\beta}_k. \tag{C.3}$$

Asymptotic Variances of Marginalized Parameter Estimates

For the case considered here, the parameter vector of interest is:

$$\boldsymbol{\theta} = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \pi_1, \pi_2, \pi_3)',$$

where α_k and $\beta_k, k = 1, 2, 3$, denote the intercept and treatment effect, respectively, for the k^{th} pattern, and π_k denotes the probability that a subject belongs to the k^{th} pattern.

For A_{JM} and B_{JM} (Jansen and Molenberghs, 2007)

Define the following expressions:

$$\begin{aligned} U \equiv f_1(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}}, \\ V \equiv f_2(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \frac{1}{1+e^{\alpha_k}}, \quad \text{and} \\ W \equiv f_3(\boldsymbol{\theta}) &= \sum_{k=1}^{3} \pi_k \beta_k \frac{e^{\alpha_k}}{1+e^{\alpha_k}} \frac{1}{1+e^{\alpha_k}}, \end{aligned}$$

with corresponding estimators given by:

$$\begin{aligned} \widehat{U} \equiv f_1(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}}, \\ \widehat{V} \equiv f_2(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \frac{1}{1+e^{\widehat{\alpha}_k}}, \qquad \text{and} \\ \widehat{W} \equiv f_3(\widehat{\boldsymbol{\theta}}) &= \sum_{k=1}^{3} \widehat{\pi}_k \widehat{\beta}_k \frac{e^{\widehat{\alpha}_k}}{1+e^{\widehat{\alpha}_k}} \frac{1}{1+e^{\widehat{\alpha}_k}}. \end{aligned}$$

Note that A_{JM} and B_{JM} can be expressed in terms of the above expressions as

$$A_{JM} = \text{logit}(U) = \ln\left(\frac{U}{1-U}\right) \qquad \text{and} \qquad B_{JM} = \frac{W}{UV}.$$

Further, define $\boldsymbol{Q}$ as the vector

$$\boldsymbol{Q} = \begin{bmatrix} U \\ V \\ W \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \\ f_3(\boldsymbol{\theta}) \end{bmatrix} \quad \text{with estimator} \quad \widehat{\boldsymbol{Q}} = \begin{bmatrix} \widehat{U} \\ \widehat{V} \\ \widehat{W} \end{bmatrix} = \begin{bmatrix} f_1(\widehat{\boldsymbol{\theta}}) \\ f_2(\widehat{\boldsymbol{\theta}}) \\ f_3(\widehat{\boldsymbol{\theta}}) \end{bmatrix}.$$

Using the delta method, the asymptotic covariance of $\widehat{\boldsymbol{Q}}$ can be obtained as:

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{Q}}) &= \left(\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\theta}}\right)' \text{Cov}(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\theta}}\right), \\ (3 \times 3) &= (3 \times 9)\ \ (9 \times 9)\ \ (9 \times 3) \end{aligned} \tag{C.4}$$

where:

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{\theta}}) &= \left[\text{Cov}(\widehat{\theta}_i, \widehat{\theta}_j)\right]_{i,j=1,2,\ldots,9} \\ &= \begin{bmatrix} \text{Var}(\widehat{\theta}_1) & \text{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) & \cdots & \text{Cov}(\widehat{\theta}_1, \widehat{\theta}_9) \\ \text{Cov}(\widehat{\theta}_2, \widehat{\theta}_1) & \text{Var}(\widehat{\theta}_2) & \cdots & \text{Cov}(\widehat{\theta}_2, \widehat{\theta}_9) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\widehat{\theta}_9, \widehat{\theta}_1) & \text{Cov}(\widehat{\theta}_9, \widehat{\theta}_2) & \cdots & \text{Var}(\widehat{\theta}_9) \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}(\widehat{\alpha}_1) & \text{Cov}(\widehat{\alpha}_1, \widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\alpha}_1, \widehat{\pi}_3) \\ \text{Cov}(\widehat{\beta}_1, \widehat{\alpha}_1) & \text{Var}(\widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\beta}_1, \widehat{\pi}_3) \\ \text{Cov}(\widehat{\alpha}_2, \widehat{\alpha}_1) & \text{Cov}(\widehat{\alpha}_2, \widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\alpha}_2, \widehat{\pi}_3) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\widehat{\pi}_3, \widehat{\alpha}_1) & \text{Cov}(\widehat{\pi}_3, \widehat{\beta}_1) & \cdots & \text{Var}(\widehat{\pi}_3) \end{bmatrix}, \end{aligned}$$

and

$$\left(\frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\theta}}\right)' = \left[\frac{\partial f_i}{\partial \theta_j}\right]_{\substack{i=1,2,3, \\ j=1,2,\ldots,9}} = \begin{bmatrix} \frac{\partial U}{\partial \alpha_1} & \frac{\partial U}{\partial \beta_1} & \cdots & \frac{\partial U}{\partial \pi_3} \\ \frac{\partial V}{\partial \alpha_1} & \frac{\partial V}{\partial \beta_1} & \cdots & \frac{\partial V}{\partial \pi_3} \\ \frac{\partial W}{\partial \alpha_1} & \frac{\partial W}{\partial \beta_1} & \cdots & \frac{\partial W}{\partial \pi_3} \end{bmatrix}.$$

Multiplying out the matrices yields the following expression for the $(i,j)^{\text{th}}$ element of $\text{Cov}(\widehat{\boldsymbol{Q}})$:

$$\begin{aligned}
\left[\text{Cov}(\widehat{\boldsymbol{Q}})\right]_{i,j} &= \sum_{s=1}^{9}\sum_{t=1}^{9}\frac{\partial f_i}{\partial\theta_s}\frac{\partial f_j}{\partial\theta_t}\text{Cov}(\widehat{\theta}_s,\widehat{\theta}_t) \\
&= \sum_{s=1}^{9}\frac{\partial f_i}{\partial\theta_s}\frac{\partial f_i}{\partial\theta_s}\text{Var}(\widehat{\theta}_s)+\sum_{s\neq t}\frac{\partial f_i}{\partial\theta_s}\frac{\partial f_j}{\partial\theta_t}\text{Cov}(\widehat{\theta}_s,\widehat{\theta}_t)
\end{aligned}$$

In terms of the pattern proportions and the pattern-specific parameters, and their corresponding estimates,

$$\begin{aligned}
\left[\text{Cov}(\widehat{\boldsymbol{Q}})\right]_{i,j} &= \sum_{k=1}^{3}\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\alpha_k}\text{Var}(\widehat{\alpha}_k) \\
&+ \sum_{k=1}^{3}\frac{\partial f_i}{\partial\beta_k}\frac{\partial f_j}{\partial\beta_k}\text{Var}(\widehat{\beta}_k) \\
&+ \sum_{k=1}^{3}\frac{\partial f_i}{\partial\pi_k}\frac{\partial f_j}{\partial\pi_k}\text{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3}\left(\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\beta_k}+\frac{\partial f_i}{\partial\beta_k}\frac{\partial f_j}{\partial\alpha_k}\right)\text{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&+ \sum_{k=1}^{3}\left(\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\pi_k}+\frac{\partial f_i}{\partial\pi_k}\frac{\partial f_j}{\partial\alpha_k}\right)\text{Cov}(\widehat{\alpha}_k,\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3}\left(\frac{\partial f_i}{\partial\beta_k}\frac{\partial f_j}{\partial\pi_k}+\frac{\partial f_i}{\partial\pi_k}\frac{\partial f_j}{\partial\beta_k}\right)\text{Cov}(\widehat{\beta}_k,\widehat{\pi}_k) \qquad (\text{C.5}) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\alpha_\ell}+\frac{\partial f_i}{\partial\alpha_\ell}\frac{\partial f_j}{\partial\alpha_k}\right)\text{Cov}(\widehat{\alpha}_k,\widehat{\alpha}_\ell) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\beta_k}\frac{\partial f_j}{\partial\beta_\ell}+\frac{\partial f_i}{\partial\beta_\ell}\frac{\partial f_j}{\partial\beta_k}\right)\text{Cov}(\widehat{\beta}_k,\widehat{\beta}_\ell) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\pi_k}\frac{\partial f_j}{\partial\pi_\ell}+\frac{\partial f_i}{\partial\pi_\ell}\frac{\partial f_j}{\partial\pi_k}\right)\text{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\beta_\ell}+\frac{\partial f_i}{\partial\beta_\ell}\frac{\partial f_j}{\partial\alpha_k}\right)\text{Cov}(\widehat{\alpha}_k,\widehat{\beta}_\ell) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\alpha_k}\frac{\partial f_j}{\partial\pi_\ell}+\frac{\partial f_i}{\partial\pi_\ell}\frac{\partial f_j}{\partial\alpha_k}\right)\text{Cov}(\widehat{\alpha}_k,\widehat{\pi}_\ell) \\
&+ \sum_{k<\ell}\left(\frac{\partial f_i}{\partial\beta_k}\frac{\partial f_j}{\partial\pi_\ell}+\frac{\partial f_i}{\partial\pi_\ell}\frac{\partial f_j}{\partial\beta_k}\right)\text{Cov}(\widehat{\beta}_k,\widehat{\pi}_\ell)
\end{aligned}$$

- Independence of pattern-specific estimates across patterns implies, $\forall k \neq \ell$,

$$\mathrm{Cov}(\widehat{\alpha}_k, \widehat{\alpha}_\ell) \;=\; 0, \quad \mathrm{Cov}(\widehat{\beta}_k, \widehat{\beta}_\ell) = 0 \quad \text{and} \quad \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_\ell) = 0.$$

- Independence of estimates for pattern probabilities and (measurement model) pattern-specific estimates further implies:

$$\begin{aligned}\mathrm{Cov}(\widehat{\alpha}_k, \widehat{\pi}_k) &= 0 \quad \text{and} \quad \mathrm{Cov}(\widehat{\beta}_k, \widehat{\pi}_k) = 0, \quad \forall k = 1,2,3,\\ \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\pi}_\ell) &= 0 \quad \text{and} \quad \mathrm{Cov}(\widehat{\beta}_k, \widehat{\pi}_\ell) = 0, \quad \forall k \neq \ell.\end{aligned}$$

Thus, $\left[\mathrm{Cov}(\widehat{\boldsymbol{Q}})\right]_{i,j}$ reduces to

$$\begin{aligned}\left[\mathrm{Cov}(\widehat{\boldsymbol{Q}})\right]_{i,j} &= \sum_{k=1}^{3} \frac{\partial f_i}{\partial \alpha_k}\frac{\partial f_j}{\partial \alpha_k}\mathrm{Var}(\widehat{\alpha}_k)\\ &\quad + \sum_{k=1}^{3} \frac{\partial f_i}{\partial \beta_k}\frac{\partial f_j}{\partial \beta_k}\mathrm{Var}(\widehat{\beta}_k)\\ &\quad + \sum_{k=1}^{3} \frac{\partial f_i}{\partial \pi_k}\frac{\partial f_j}{\partial \pi_k}\mathrm{Var}(\widehat{\pi}_k)\\ &\quad + \sum_{k=1}^{3} \left(\frac{\partial f_i}{\partial \alpha_k}\frac{\partial f_j}{\partial \beta_k} + \frac{\partial f_i}{\partial \beta_k}\frac{\partial f_j}{\partial \alpha_k}\right)\mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k)\\ &\quad + \sum_{k<\ell} \left(\frac{\partial f_i}{\partial \pi_k}\frac{\partial f_j}{\partial \pi_\ell}\frac{\partial f_i}{\partial \pi_\ell}\frac{\partial f_j}{\partial \pi_k}\right)\mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell).\end{aligned} \tag{C.6}$$

Note that for $k = 1,2,3$,

$$\begin{array}{lll}\dfrac{\partial U}{\partial \alpha_k} = \pi_k \dfrac{e^{\alpha_k}}{(1+e^{\alpha_k})^2} & \dfrac{\partial U}{\partial \beta_k} = 0 & \dfrac{\partial U}{\partial \pi_k} = \dfrac{e^{\alpha_k}}{1+e^{\alpha_k}}\\[2ex] \dfrac{\partial V}{\partial \alpha_k} = \pi_k \dfrac{-e^{\alpha_k}}{(1+e^{\alpha_k})^2} & \dfrac{\partial V}{\partial \beta_k} = 0 & \dfrac{\partial V}{\partial \pi_k} = \dfrac{1}{1+e^{\alpha_k}}\\[2ex] \dfrac{\partial W}{\partial \alpha_k} = \pi_k\beta_k \dfrac{e^{\alpha_k}(1-e^{\alpha_k})}{(1+e^{\alpha_k})^3} & \dfrac{\partial W}{\partial \beta_k} = \pi_k \dfrac{e^{\alpha_k}}{(1+e^{\alpha_k})^2} & \dfrac{\partial W}{\partial \pi_k} = \beta_k \dfrac{e^{\alpha_k}}{(1+e^{\alpha_k})^2}\end{array} \tag{C.7}$$

Asymptotic variances can then be obtained by using (C.7) in (C.6), for specific values of i and j.

Variance of $\widehat{U}$

$$
\begin{aligned}
\mathrm{Var}(\widehat{U}) &= \Big[\mathrm{Cov}(\widehat{\boldsymbol{Q}})\Big]_{1,1} \\
&= \sum_{k=1}^{3} \frac{\partial U}{\partial \alpha_k}\frac{\partial U}{\partial \alpha_k}\mathrm{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \beta_k}\frac{\partial U}{\partial \beta_k}\mathrm{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \pi_k}\frac{\partial U}{\partial \pi_k}\mathrm{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} \left(\frac{\partial U}{\partial \alpha_k}\frac{\partial U}{\partial \beta_k} + \frac{\partial U}{\partial \beta_k}\frac{\partial U}{\partial \alpha_k} \right) \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \sum_{k<\ell} \left(\frac{\partial U}{\partial \pi_k}\frac{\partial U}{\partial \pi_\ell} + \frac{\partial U}{\partial \pi_\ell}\frac{\partial U}{\partial \pi_k} \right) \mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell) \\
&= \sum_{k=1}^{3} \left(\frac{\partial U}{\partial \alpha_k}\right)^2 \mathrm{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \left(\frac{\partial U}{\partial \beta_k}\right)^2 \mathrm{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \left(\frac{\partial U}{\partial \pi_k}\right)^2 \mathrm{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} 2\left(\frac{\partial U}{\partial \alpha_k}\frac{\partial U}{\partial \beta_k} \right) \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \ + \sum_{k<\ell} 2\left(\frac{\partial U}{\partial \pi_k}\frac{\partial U}{\partial \pi_\ell} \right) \mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\widehat{U}) &= \sum_{k=1}^{3} {\pi_k}^2 \frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\mathrm{Var}(\widehat{\alpha}_k) \ + \ \sum_{k=1}^{3} \frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^2}\mathrm{Var}(\widehat{\pi}_k) \\
&+ \ 2\sum_{k<\ell} \frac{e^{\alpha_k}e^{\alpha_\ell}}{(1+e^{\alpha_k})(1+e^{\alpha_\ell})}\mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}
$$

Variance of $\widehat{V}$

$$
\begin{aligned}
\mathrm{Var}(\widehat{V}) &= \Big[\mathrm{Cov}(\widehat{\boldsymbol{Q}})\Big]_{2,2} \\
&= \sum_{k=1}^{3} \frac{\partial V}{\partial \alpha_k}\frac{\partial V}{\partial \alpha_k}\mathrm{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial V}{\partial \beta_k}\frac{\partial V}{\partial \beta_k}\mathrm{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial V}{\partial \pi_k}\frac{\partial V}{\partial \pi_k}\mathrm{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} \left(\frac{\partial V}{\partial \alpha_k}\frac{\partial V}{\partial \beta_k} \frac{\partial V}{\partial \beta_k}\frac{\partial V}{\partial \alpha_k} \right) \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \sum_{k<\ell} \left(\frac{\partial V}{\partial \pi_k}\frac{\partial V}{\partial \pi_\ell} + \frac{\partial V}{\partial \pi_\ell}\frac{\partial V}{\partial \pi_k} \right) \mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell) \\
&= \sum_{k=1}^{3} \left(\frac{\partial V}{\partial \alpha_k}\right)^2 \mathrm{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \left(\frac{\partial V}{\partial \beta_k}\right)^2 \mathrm{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \left(\frac{\partial V}{\partial \pi_k}\right)^2 \mathrm{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} 2\left(\frac{\partial V}{\partial \alpha_k}\frac{\partial V}{\partial \beta_k} \right) \mathrm{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \ + \sum_{k<\ell} 2\left(\frac{\partial V}{\partial \pi_k}\frac{\partial V}{\partial \pi_\ell} \right) \mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\widehat{V}) &= \sum_{k=1}^{3} {\pi_k}^2 \frac{(-e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\mathrm{Var}(\widehat{\alpha}_k) \ + \ \sum_{k=1}^{3} \frac{1}{(1+e^{\alpha_k})^2}\mathrm{Var}(\widehat{\pi}_k) \\
&+ \ 2\sum_{k<\ell} \frac{1}{(1+e^{\alpha_k})(1+e^{\alpha_\ell})}\mathrm{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}
$$

Variance of $\widehat{W}$

$$\begin{aligned}
\operatorname{Var}(\widehat{W}) &= \Big[\operatorname{Cov}(\widehat{\boldsymbol{Q}})\Big]_{3,3} \\
&= \sum_{k=1}^{3} \frac{\partial W}{\partial \alpha_k}\frac{\partial W}{\partial \alpha_k}\operatorname{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial W}{\partial \beta_k}\frac{\partial W}{\partial \beta_k}\operatorname{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial W}{\partial \pi_k}\frac{\partial W}{\partial \pi_k}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3}\left(\frac{\partial W}{\partial \alpha_k}\frac{\partial W}{\partial \beta_k} + \frac{\partial W}{\partial \beta_k}\frac{\partial W}{\partial \alpha_k}\right)\operatorname{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \sum_{k<\ell}\left(\frac{\partial W}{\partial \pi_k}\frac{\partial W}{\partial \pi_\ell} + \frac{\partial W}{\partial \pi_\ell}\frac{\partial W}{\partial \pi_k}\right)\operatorname{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell) \\
&= \sum_{k=1}^{3}\left(\frac{\partial W}{\partial \alpha_k}\right)^2\operatorname{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3}\left(\frac{\partial W}{\partial \beta_k}\right)^2\operatorname{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3}\left(\frac{\partial W}{\partial \pi_k}\right)^2\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} 2\left(\frac{\partial W}{\partial \alpha_k}\frac{\partial W}{\partial \beta_k}\right)\operatorname{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \ + \sum_{k<\ell} 2\left(\frac{\partial W}{\partial \pi_k}\frac{\partial W}{\partial \pi_\ell}\right)\operatorname{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

$$\begin{aligned}
\operatorname{Var}(\widehat{W}) &= \sum_{k=1}^{3} {\pi_k}^2{\beta_k}^2\frac{(e^{\alpha_k})^2(1-e^{\alpha_k})^2}{(1+e^{\alpha_k})^6}\operatorname{Var}(\widehat{\alpha}_k) \ + \sum_{k=1}^{3} {\pi_k}^2\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\operatorname{Var}(\widehat{\beta}_k) \\
&+ \sum_{k=1}^{3} {\beta_k}^2\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\operatorname{Var}(\widehat{\pi}_k) \ + \ 2\sum_{k=1}^{3} {\pi_k}^2\beta_k\frac{(e^{\alpha_k})^2(1-e^{\alpha_k})}{(1+e^{\alpha_k})^5}\operatorname{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \ 2\sum_{k<\ell}\beta_k\beta_\ell\frac{e^{\alpha_k}e^{\alpha_\ell}}{(1+e^{\alpha_k})^2(1+e^{\alpha_\ell})^2}\operatorname{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

Covariance of $\widehat{U}$ and $\widehat{V}$

$$\begin{aligned}
\operatorname{Cov}(\widehat{U}, \widehat{V}) &= \Big[\operatorname{Cov}(\widehat{\boldsymbol{Q}})\Big]_{1,2} \\
&= \sum_{k=1}^{3} \frac{\partial U}{\partial \alpha_k}\frac{\partial V}{\partial \alpha_k}\operatorname{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \beta_k}\frac{\partial V}{\partial \beta_k}\operatorname{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \pi_k}\frac{\partial V}{\partial \pi_k}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3}\left(\frac{\partial U}{\partial \alpha_k}\frac{\partial V}{\partial \beta_k} + \frac{\partial U}{\partial \beta_k}\frac{\partial V}{\partial \alpha_k}\right)\operatorname{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \sum_{k<\ell}\left(\frac{\partial U}{\partial \pi_k}\frac{\partial V}{\partial \pi_\ell} + \frac{\partial U}{\partial \pi_\ell}\frac{\partial V}{\partial \pi_k}\right)\operatorname{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

$$\begin{aligned}
\operatorname{Cov}(\widehat{U}, \widehat{V}) &= \sum_{k=1}^{3} {\pi_k}^2(-1)\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\operatorname{Var}(\widehat{\alpha}_k) \ + \sum_{k=1}^{3} \frac{e^{\alpha_k}}{(1+e^{\alpha_k})^2}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k<\ell}\frac{e^{\alpha_k}+e^{\alpha_\ell}}{(1+e^{\alpha_k})(1+e^{\alpha_\ell})}\operatorname{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

Covariance of $\widehat{U}$ and $\widehat{W}$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{U},\widehat{W}) &= \Big[\operatorname{Cov}(\widehat{\boldsymbol{Q}})\Big]_{1,3} \\
&= \sum_{k=1}^{3} \frac{\partial U}{\partial \alpha_k}\frac{\partial W}{\partial \alpha_k}\operatorname{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \beta_k}\frac{\partial W}{\partial \beta_k}\operatorname{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial U}{\partial \pi_k}\frac{\partial W}{\partial \pi_k}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} \left(\frac{\partial U}{\partial \alpha_k}\frac{\partial W}{\partial \beta_k} + \frac{\partial U}{\partial \beta_k}\frac{\partial W}{\partial \alpha_k}\right)\operatorname{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&+ \sum_{k<\ell} \left(\frac{\partial U}{\partial \pi_k}\frac{\partial W}{\partial \pi_\ell} + \frac{\partial U}{\partial \pi_\ell}\frac{\partial W}{\partial \pi_k}\right)\operatorname{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{U},\widehat{W}) &= \sum_{k=1}^{3} {\pi_k}^2\beta_k\frac{(e^{\alpha_k})^2(1-e^{\alpha_k})}{(1+e^{\alpha_k})^5}\operatorname{Var}(\widehat{\alpha}_k) \ + \sum_{k=1}^{3} \beta_k\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^3}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} {\pi_k}^2\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\operatorname{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&+ \sum_{k<\ell} e^{\alpha_k}e^{\alpha_\ell}\frac{\beta_k(1+e^{\alpha_\ell})+\beta_\ell(1+e^{\alpha_k})}{(1+e^{\alpha_k})^2(1+e^{\alpha_\ell})^2}\operatorname{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}
$$

Covariance of $\widehat{V}$ and $\widehat{W}$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{V},\widehat{W}) &= \Big[\operatorname{Cov}(\widehat{\boldsymbol{Q}})\Big]_{2,3} \\
&= \sum_{k=1}^{3} \frac{\partial V}{\partial \alpha_k}\frac{\partial W}{\partial \alpha_k}\operatorname{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} \frac{\partial V}{\partial \beta_k}\frac{\partial W}{\partial \beta_k}\operatorname{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} \frac{\partial V}{\partial \pi_k}\frac{\partial W}{\partial \pi_k}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} \left(\frac{\partial V}{\partial \alpha_k}\frac{\partial W}{\partial \beta_k} + \frac{\partial V}{\partial \beta_k}\frac{\partial W}{\partial \alpha_k}\right)\operatorname{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&+ \sum_{k<\ell} \left(\frac{\partial V}{\partial \pi_k}\frac{\partial W}{\partial \pi_\ell} + \frac{\partial V}{\partial \pi_\ell}\frac{\partial W}{\partial \pi_k}\right)\operatorname{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Cov}(\widehat{V},\widehat{W}) &= \sum_{k=1}^{3} {\pi_k}^2(-1)\beta_k\frac{(e^{\alpha_k})^2(1-e^{\alpha_k})}{(1+e^{\alpha_k})^5}\operatorname{Var}(\widehat{\alpha}_k) \ + \sum_{k=1}^{3} \beta_k\frac{e^{\alpha_k}}{(1+e^{\alpha_k})^3}\operatorname{Var}(\widehat{\pi}_k) \\
&+ \sum_{k=1}^{3} {\pi_k}^2(-1)\frac{(e^{\alpha_k})^2}{(1+e^{\alpha_k})^4}\operatorname{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&+ \sum_{k<\ell} \frac{(1+e^{\alpha_k})\beta_\ell e^{\alpha_\ell} + (1+e^{\alpha_\ell})\beta_k e^{\alpha_k}}{(1+e^{\alpha_k})^2(1+e^{\alpha_\ell})^2}\operatorname{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}
$$

Asymptotic Variance of $\widehat{A}_{JM}$

Applying the delta method to obtain the asymptotic variance of $\widehat{A}_{JM}$,

$$\begin{aligned}
\mathrm{Var}(\widehat{A}_{JM}) &= \mathrm{Var}\left[\ln\left(\frac{\widehat{U}}{1-\widehat{U}}\right)\right] \\
&\cong \left[\frac{d}{dU}\ln\left(\frac{U}{1-U}\right)\right]^2 \mathrm{Var}(\widehat{U}) \\
&\cong \left[\left(\frac{1-U}{U}\right)\frac{1}{(1-U)^2}\right]^2 \mathrm{Var}(\widehat{U}) \\
\mathrm{Var}(\widehat{A}_{JM}) &\cong \frac{1}{U^2(1-U)^2}\mathrm{Var}(\widehat{U})
\end{aligned}$$

Asymptotic Variance of $\widehat{B}_{JM}$

Since B_{JM} is a function of U, V and W, the delta method can also be applied to obtain its asymptotic variance as follows:

$$\begin{aligned}
\mathrm{Var}(\widehat{B}_{JM}) &= \mathrm{Var}\left(\frac{\widehat{W}}{\widehat{U}\widehat{V}}\right) \\
&\cong \begin{bmatrix} \frac{\partial B}{\partial U} & \frac{\partial B}{\partial V} & \frac{\partial B}{\partial W} \end{bmatrix}
\begin{bmatrix} \mathrm{Var}(\widehat{U}) & \mathrm{Cov}(\widehat{U},\widehat{V}) & \mathrm{Cov}(\widehat{U},\widehat{W}) \\ \mathrm{Cov}(\widehat{V},\widehat{U}) & \mathrm{Var}(\widehat{V}) & \mathrm{Cov}(\widehat{V},\widehat{W}) \\ \mathrm{Cov}(\widehat{W},\widehat{U}) & \mathrm{Cov}(\widehat{V},\widehat{W}) & \mathrm{Var}(\widehat{W}) \end{bmatrix}
\begin{bmatrix} \frac{\partial B}{\partial U} \\ \frac{\partial B}{\partial V} \\ \frac{\partial B}{\partial W} \end{bmatrix} \\
&\cong \left(\frac{\partial B}{\partial U}\right)^2 \mathrm{Var}(\widehat{U}) + \left(\frac{\partial B}{\partial V}\right)^2 \mathrm{Var}(\widehat{V}) + \left(\frac{\partial B}{\partial W}\right)^2 \mathrm{Var}(\widehat{W}) \\
&\quad + 2\left(\frac{\partial B}{\partial U}\right)\left(\frac{\partial B}{\partial V}\right)\mathrm{Cov}(\widehat{U},\widehat{V}) + 2\left(\frac{\partial B}{\partial U}\right)\left(\frac{\partial B}{\partial W}\right)\mathrm{Cov}(\widehat{U},\widehat{W}) \\
&\quad + 2\left(\frac{\partial B}{\partial V}\right)\left(\frac{\partial B}{\partial W}\right)\mathrm{Cov}(\widehat{V},\widehat{W})
\end{aligned}$$

where:

$$\frac{\partial B}{\partial U} = \frac{-W}{U^2V}, \qquad \frac{\partial B}{\partial V} = \frac{-W}{UV^2} \qquad \text{and} \qquad \frac{\partial B}{\partial W} = \frac{1}{UV}$$

$$\begin{aligned}
\mathrm{Var}(\widehat{B}_{JM}) &\cong \frac{W^2}{U^4V^2}\mathrm{Var}(\widehat{U}) + \frac{W^2}{U^2V^4}\mathrm{Var}(\widehat{V}) + \frac{1}{U^2V^2}\mathrm{Var}(\widehat{W}) \\
&\quad + \frac{2W^2}{U^3V^3}\mathrm{Cov}(\widehat{U},\widehat{V}) - \frac{2W}{U^3V^2}\mathrm{Cov}(\widehat{U},\widehat{W}) - \frac{2W}{U^2V^3}\mathrm{Cov}(\widehat{V},\widehat{W})
\end{aligned}$$

For A_{PL} and B_{PL} (Park and Lee, 1999)

Define the following expressions and their corresponding estimators:

$$Y \equiv f_1(\boldsymbol{\theta}) = \sum_{k=1}^{3} \pi_k \alpha_k \qquad \widehat{Y} \equiv f_1(\widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{3} \widehat{\pi}_k \widehat{\alpha}_k$$

$$Z \equiv f_2(\boldsymbol{\theta}) = \sum_{k=1}^{3} \pi_k \beta_k \qquad \widehat{Z} \equiv f_2(\widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{3} \widehat{\pi}_k \widehat{\beta}_k$$

Note that $A_{PL} = Y$ and $B_{PL} = Z$, with, for $k = 1, 2, 3,$

$$\begin{array}{lll} \dfrac{\partial Y}{\partial \alpha_k} = \pi_k & \dfrac{\partial Y}{\partial \beta_k} = 0 & \dfrac{\partial Y}{\partial \pi_k} = \alpha_k \\[2ex] \dfrac{\partial Z}{\partial \alpha_k} = 0 & \dfrac{\partial Z}{\partial \beta_k} = \pi_k & \dfrac{\partial Z}{\partial \pi_k} = \beta_k \end{array} \tag{C.8}$$

Let $\boldsymbol{G}$ be the vector

$$\boldsymbol{G} = \begin{bmatrix} Y \\ Z \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ f_2(\boldsymbol{\theta}) \end{bmatrix} \quad \text{with estimator} \quad \widehat{\boldsymbol{G}} = \begin{bmatrix} \widehat{Y} \\ \widehat{Z} \end{bmatrix} = \begin{bmatrix} f_1(\widehat{\boldsymbol{\theta}}) \\ f_2(\widehat{\boldsymbol{\theta}}) \end{bmatrix}.$$

Using the delta method, the asymptotic covariance of $\widehat{\boldsymbol{G}}$ can be obtained as:

$$\begin{array}{rcl} \text{Cov}(\widehat{\boldsymbol{G}}) & = & \left(\dfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}\right)' \text{Cov}(\widehat{\boldsymbol{\theta}}) \left(\dfrac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}\right) \\[2ex] (2 \times 2) & = & (2 \times 9) \;\; (9 \times 9) \;\; (9 \times 2) \end{array} \tag{C.9}$$

where:

$$\begin{array}{rcl} \text{Cov}(\widehat{\boldsymbol{\theta}}) & = & \left[\text{Cov}(\widehat{\theta}_i, \widehat{\theta}_j)\right]_{i,j=1,2,\ldots,9} \\[3ex] & = & \begin{bmatrix} \text{Var}(\widehat{\alpha}_1) & \text{Cov}(\widehat{\alpha}_1, \widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\alpha}_1, \widehat{\pi}_3) \\ \text{Cov}(\widehat{\beta}_1, \widehat{\alpha}_1) & \text{Var}(\widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\beta}_1, \widehat{\pi}_3) \\ \text{Cov}(\widehat{\alpha}_2, \widehat{\alpha}_1) & \text{Cov}(\widehat{\alpha}_2, \widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\alpha}_2, \widehat{\pi}_3) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\widehat{\pi}_3, \widehat{\alpha}_1) & \text{Cov}(\widehat{\pi}_3, \widehat{\beta}_1) & \cdots & \text{Var}(\widehat{\pi}_3) \end{bmatrix} \end{array}$$

and

$$\left(\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}\right)' = \left[\frac{\partial f_i}{\partial \theta_j}\right]_{\substack{i=1,2, \\ j=1,2,\ldots,9}} = \begin{bmatrix} \dfrac{\partial Y}{\partial \alpha_1} & \dfrac{\partial Y}{\partial \beta_1} & \cdots & \dfrac{\partial Y}{\partial \pi_3} \\[2ex] \dfrac{\partial Z}{\partial \alpha_1} & \dfrac{\partial Z}{\partial \beta_1} & \cdots & \dfrac{\partial Z}{\partial \pi_3} \end{bmatrix}.$$

Multiplication of these matrices and simplifications arising from the independence assumptions lead to the following expression for the $(i,j)^{\text{th}}$ element of $\text{Cov}(\widehat{\boldsymbol{G}})$, which is similar to that in (C.6).

$$\begin{aligned}
\left[\text{Cov}(\widehat{\boldsymbol{G}})\right]_{i,j} &= \sum_{k=1}^{3} \frac{\partial f_i}{\partial \alpha_k}\frac{\partial f_j}{\partial \alpha_k}\text{Var}(\widehat{\alpha}_k) \;+\; \sum_{k=1}^{3} \frac{\partial f_i}{\partial \beta_k}\frac{\partial f_j}{\partial \beta_k}\text{Var}(\widehat{\beta}_k) \\
&+ \sum_{k=1}^{3} \frac{\partial f_i}{\partial \pi_k}\frac{\partial f_j}{\partial \pi_k}\text{Var}(\widehat{\pi}_k) \qquad\qquad (\text{C.10}) \\
&+ \sum_{k=1}^{3} \left(\frac{\partial f_i}{\partial \alpha_k}\frac{\partial f_j}{\partial \beta_k} + \frac{\partial f_i}{\partial \beta_k}\frac{\partial f_j}{\partial \alpha_k}\right)\text{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \\
&+ \sum_{k<\ell} \left(\frac{\partial f_i}{\partial \pi_k}\frac{\partial f_j}{\partial \pi_\ell} + \frac{\partial f_i}{\partial \pi_\ell}\frac{\partial f_j}{\partial \pi_k}\right)\text{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

Asymptotic variances can then be obtained by using (C.8) in (C.10), for specific values of i and j.

Asymptotic Variance of $\widehat{A}_{PL}$

$$\begin{aligned}
\text{Var}(\widehat{A}_{PL}) &\equiv \text{Var}(\widehat{Y}) = \left[\text{Cov}(\widehat{\boldsymbol{G}})\right]_{1,1} \\
&= \sum_{k=1}^{3}\left(\frac{\partial Y}{\partial \alpha_k}\right)^2\text{Var}(\widehat{\alpha}_k) \;+\; \sum_{k=1}^{3}\left(\frac{\partial Y}{\partial \beta_k}\right)^2\text{Var}(\widehat{\beta}_k) \;+\; \sum_{k=1}^{3}\left(\frac{\partial Y}{\partial \pi_k}\right)^2\text{Var}(\widehat{\pi}_k) \\
&\quad + \sum_{k=1}^{3} 2\left(\frac{\partial Y}{\partial \alpha_k}\frac{\partial Y}{\partial \beta_k}\right)\text{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \;+\; \sum_{k<\ell} 2\left(\frac{\partial Y}{\partial \pi_k}\frac{\partial Y}{\partial \pi_\ell}\right)\text{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

$$\text{Var}(\widehat{A}_{PL}) \equiv \text{Var}(\widehat{Y}) = \sum_{k=1}^{3} {\pi_k}^2\text{Var}(\widehat{\alpha}_k) + \sum_{k=1}^{3} {\alpha_k}^2\text{Var}(\widehat{\pi}_k) + 2\sum_{k<\ell}\alpha_k\alpha_\ell\text{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)$$

Asymptotic Variance of $\widehat{B}_{PL}$

$$\begin{aligned}
\text{Var}(\widehat{B}_{PL}) &\equiv \text{Var}(\widehat{Z}) = \left[\text{Cov}(\widehat{\boldsymbol{G}})\right]_{2,2} \\
&= \sum_{k=1}^{3}\left(\frac{\partial Z}{\partial \alpha_k}\right)^2\text{Var}(\widehat{\alpha}_k) \;+\; \sum_{k=1}^{3}\left(\frac{\partial Z}{\partial \beta_k}\right)^2\text{Var}(\widehat{\beta}_k) \;+\; \sum_{k=1}^{3}\left(\frac{\partial Z}{\partial \pi_k}\right)^2\text{Var}(\widehat{\pi}_k) \\
&\quad + \sum_{k=1}^{3} 2\left(\frac{\partial Z}{\partial \alpha_k}\frac{\partial Z}{\partial \beta_k}\right)\text{Cov}(\widehat{\alpha}_k, \widehat{\beta}_k) \;+\; \sum_{k<\ell} 2\left(\frac{\partial Z}{\partial \pi_k}\frac{\partial Z}{\partial \pi_\ell}\right)\text{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)
\end{aligned}$$

$$\text{Var}(\widehat{B}_{PL}) \equiv \text{Var}(\widehat{Z}) = \sum_{k=1}^{3} {\pi_k}^2\text{Var}(\widehat{\beta}_k) + \sum_{k=1}^{3} {\beta_k}^2\text{Var}(\widehat{\pi}_k) + 2\sum_{k<\ell}\beta_k\beta_\ell\text{Cov}(\widehat{\pi}_k, \widehat{\pi}_\ell)$$

Covariance of $\widehat{A}_{PL}$ and $\widehat{B}_{PL}$

$$\begin{aligned}
\mathrm{Cov}(\widehat{A}_{PL},\widehat{B}_{PL}) &\equiv \mathrm{Cov}(\widehat{Y},\widehat{Z}) = \Big[\mathrm{Cov}(\widehat{\boldsymbol{G}})\Big]_{1,2} \\
&= \sum_{k=1}^{3} \frac{\partial Y}{\partial \alpha_k}\frac{\partial Z}{\partial \alpha_k}\mathrm{Var}(\widehat{\alpha}_k) \\
&\quad + \sum_{k=1}^{3} \frac{\partial Y}{\partial \beta_k}\frac{\partial Z}{\partial \beta_k}\mathrm{Var}(\widehat{\beta}_k) \\
&\quad + \sum_{k=1}^{3} \frac{\partial Y}{\partial \pi_k}\frac{\partial Z}{\partial \pi_k}\mathrm{Var}(\widehat{\pi}_k) \\
&\quad + \sum_{k=1}^{3} \left(\frac{\partial Y}{\partial \alpha_k}\frac{\partial Z}{\partial \beta_k} + \frac{\partial Y}{\partial \beta_k}\frac{\partial Z}{\partial \alpha_k}\right)\mathrm{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&\quad + \sum_{k<\ell} \left(\frac{\partial Y}{\partial \pi_k}\frac{\partial Z}{\partial \pi_\ell} + \frac{\partial Y}{\partial \pi_\ell}\frac{\partial Z}{\partial \pi_k}\right)\mathrm{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(\widehat{A}_{PL},\widehat{B}_{PL}) &\equiv \mathrm{Cov}(\widehat{Y},\widehat{Z}) \\
&= \sum_{k=1}^{3} \alpha_k\beta_k \mathrm{Var}(\widehat{\pi}_k) \; + \sum_{k=1}^{3} {\pi_k}^2\mathrm{Cov}(\widehat{\alpha}_k,\widehat{\beta}_k) \\
&\quad + \sum_{k<\ell} (\alpha_k\beta_\ell + \alpha_\ell\beta_k)\mathrm{Cov}(\widehat{\pi}_k,\widehat{\pi}_\ell)
\end{aligned}$$