Modeling Residual Dependencies in Latent Variable Models with Copulas

Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Psychologie door Johan Braeken
Promotor: Prof. dr. F. Tuerlinckx
Co-promotor: Prof. dr. P. De Boeck

2008
Modeling Residual Dependencies in Latent Variable Models with Copulas

Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Psychologie

doors Johaan Braeken

Promotor: Prof. dr. F. Tuerlinckx
Co-promotor: Prof. dr. P. De Boeck

2008
Latent variable models are popular in many domains in which there are constructs of interests that are measured using a set of manifest indicators. Examples of such models are factor analysis, item response models, and structural equation models. A basic assumption of almost all latent variable models is conditional independence such that the latent variables are considered to completely explain the association between the manifest indicators. However, this assumption is such a strict requirement that it is unlikely to be met completely for most datasets, because it is quite common that a subset of indicators share some additional feature besides the constructs of interest. As a result there remains association between these indicators even after accounting for the latent variables. Ignoring these residual dependencies may lead to biased inferences, due to model misspecification. Solutions for this problem are rather obvious and straightforward to implement in normal linear models, but this is not the case for categorical data models.

This dissertation introduces a novel modeling approach to account for residual dependencies. By including copula functions within the current latent variable model tradition, a model is developed that more properly accounts for the dependence, while still retaining simple structure and clarity. Copula functions originated in mathematics, from the investigation of relations between multidimensional distributions and their lower-dimensional margins, but remained a unexplored tool within the latent variable modeling framework. Therefore the focus is on investigating the possibilities of the use of copulas within latent variable models. A major advantage of the copula approach has to do with its upward compatibility and specifically the reproducibility of the marginal model structure.

In Chapter 1, the foundations of the copula approach are explained and applied to the well-known unidimensional Rasch model for binary manifest variables. A comparison between competing approaches is made and it is shown that the key element here is the double flexibility of the copula approach (i.e., margins for the individual variables and dependence structure for their joint behaviour). In Chapter 2, a latent variable model is developed to accommodate the specific demands inherent to a birth defects screening instrument. A unidimensional two-parameter logistic model is used to model the presence/absence of specific anomalies. Covariates for the sample units, in this case newborns, are included explicitly in the marginal model, and the latent trait distribution is made flexible by adopting a finite mixture approach. Residual dependencies are again accounted for by copulas. Another generalization is proposed in Chapter 3, where the copula approach is extended towards the case of ordinal manifest variables hereby allowing for correlated errors in structural equation models for categorical data. Besides the presence of polytomous response categories, another point of difference with the previous chapters is that the model, applied to a contextualized personality inventory, is now multidimensional. In Chapter 4, the issue of model selection is investigated. There is a wide variety of possible copula functions and it is not always clear which one to choose. Several model selection criteria are proposed and their performance investigated in a simulation study. It appears that simple model selection criteria, although not suited at first sight, often outperform other more complicated ones. Chapter 5 contains the description of a Matlab toolbox for fitting explanatory item response theory models for binary data. There is a lack of this type of software for Matlab, and the toolbox also includes the novel copula approach for residual dependencies.
Latente variabele modellen zijn populair in vele domeinen waarin constructen voorkomen die indirect gemeten worden door middel van een set manifeste indicatoren. Voorbeelden van zulke modellen zijn factor analyse, item-responsmodellen, en structurele vergelijkingsmodellen. Een basis assumptie van bijna alle latente variabele modellen is conditionale onafhankelijkheid, zodat de latente variabelen verondersteld worden de volledige afhankelijkheid tussen de manifeste indicatoren volledig te verklaren. Het is echter onwaarschijnlijk dat de meeste datasets perfect voldoen aan deze stricte vereiste, omdat het gebruikelijk is dat een subset van indicatoren een gemeenschappelijk kenmerk delen buiten de doelconstructs. Het gevolg hiervan is dat er associatie overblijft tussen deze indicatoren zelfs na controle voor de latente variabelen. Het negeren van deze residuele afhankelijkheden kan leiden tot vertekende inferenties door modelmisspecificatie. Oplossingen voor dit probleem zijn vrij voor de hand liggend en rechttoe rechtaan te implementeren in normale lineaire modellen, maar dit is zeker niet het geval voor categorische datamodellen.

Dit proefschrift introduceert een nieuwe modelbenadering om deze residuele afhankelijkheden in rekening te brengen. Door copulafuncties binnen bestaande latente variabele modellen in te passen, wordt een model ontwikkeld dat beter de afhankelijkheidstructuur vat, maar toch een eenvoudige structuur en helderheid behoudt. Copulae ontstonden binnen de wiskunde bij het bestuderen van de relatie tussen multidimensionele distributies en hun lager-dimensionele marges, maar zijn binnen latente variabele modellen onbekend terrein. De focus ligt dan ook op het onderzoeken van de mogelijkheden van het gebruik van deze copulae binnen latente variabele modellen. Een groot voordeel van deze copulabenadering ligt in de bovenwaardse compatibiliteit met de reproduceerbaarheid van de marginale modellstructuur.

In hoofdstuk 1 worden de fundamenten van de copulabenadering gelegd binnen het bekende unidimensionele Rasch model voor binaire manifeste variabelen. In vergelijking met alternatieve benaderingen is het sleutelelement de dubbele flexibiliteit van de copula (marges voor de individuele variabelen en afhankelijkheidstructuur voor hun gezamenlijk gedrag). In hoofdstuk 2, wordt een model ontwikkeld dat covariate informatie over de steekproefeenheden expliciet in het marginale model opneemt of impliciet via een finite mixeduredistributie voor de latente variabele. Binnen dit model van een screeningsinstrument voor geboortedefecten worden residuele afhankelijkheden tussen manifeste variabelen aangepakt via copulae. In hoofdstuk 3, wordt de copulabenadering gegeneraliseerd naar het geval van ordinale manifeste variabelen binnen een structureel vergelijkingsmodel voor een gecontextualiseerde persoonlijkheidsvragenlijst. De combinatie van een meetmodel voor manifeste variabelen en een multidimensioneel structureel model voor de onderliggende set van latente variabelen, wordt uitgebreid met copulae voor de residuele afhankelijkheden. In hoofdstuk 4, wordt het thema modelselectie aangesneden voor de met copulae uitgebreide latente variabele modellen. Verschillende selectecriteria worden voorgesteld en hun perfor- mantie onderzocht in een simulatiestudie. Het blijkt dat ondanks dat eenvoudige model selectie criteria op het eerste gezicht niet geschikt lijken voor deze toepassing, ze toch vaak hun meer ingewikkelde soortgenoten overtreffen. In hoofdstuk 5, wordt een matlab toolbox voorgesteld voor het calibreren van item-responsmodellen voor binaire data, inclusief onze eigen copula-uitbreiding.
Reflectie en dankwoord

Dit proefschrift is het gevolg van een viertal jaar “ontdekken van en spelen met” psychometrische en statistische modellen. Bijna vanzelfsprekend was er tussendoor het nodige geworstel en gevoel als het even niet zo snel wou vlotten als ik op het oog had, maar toch was er telkens die eenvoudige voldoening wanneer ik een probleem weggewerkt had of een drempel verzet. Het neerschrijven van dit werk is misschien wel het minst lonende van je activiteiten als doctoraatstudent. Er is geen concreet probleem meer op te lossen, enkel papieren op te vullen met woorden en formules die hopelijk je boodschap kunnen communiceren aan het beperkte publiek binnen je vakdomein. Op zich is dit best nog wel een uitdaging, alleen valt het wat tegen dat schrijven een soms traag proces is en het wetenschappelijk publicatieproces op snelheid kan geklopt worden door zowel een schildpad als een slak. Met andere woorden feedback is hier veel minder snel en eenduidig dan tijdens je andere werkzaamheden. Niettemin, ben ik toch meer dan tevreden dat ik me hier ook doorgeworsteld heb en durf ik zelfs stellen dat ik trots ben op het eindproduct dat je nu voor je ligt. Een woord van dank tot iedereen die direct of indirect bijgedragen heeft om dit mogelijk te maken, is dan ook op zijn plaats.

Om te beginnen wil ik de onderzoeksgroep en een aantal andere mensen aan de K.U.Leuven bedanken voor de kansen die ze me aangeboden hebben en het feit dat ze mee een rol hebben gespeeld in het pad dat mijn leven uiteindelijk gekozen heeft. Mijn eerste jaar aan de universiteit behaalde ik hilarisch genoeg twee maal een 5 op 20 in eerste zit, en dit juist op de vakken statistiek en psychometrie! Ik heb dit gelukkig achteraf nog rechtgezet, met een 17 op 20 voor psychometrie in tweede zit bijvoorbeeld. Een vreemde ommekeer, deels te wijten aan de trage aanpassing aan het universiteitsleven, maar misschien ook wel door de goede begeleiding van een zekere repetitor die erin slagde om mijn “licht” en logica te doen branden op dit vlak. Ik kwam haar later nog tegen als practicumbegeleidster bij variantie-analyse in de licenties, en leerde haar nog later tijdens mijn doctoraat kennen als een toffe collega, dus Katrijn alvast bedankt. Verder wil ik nog professor Gert Storms bedanken voor het begeleiden van mijn licentie thesis en voor mij in de wereld van de data-analyse te werpen. Dit was ook de aanleiding om voor een stage-onderwerp te kiezen dat met data-analyse te maken had, waardoor ik in contact kwam met mijn huidige
promotor Francis Tuerlinckx. Het stage-onderzoek was mijn eerste kennismaking met psychometrische modellen, en dankzij de vlotte samenwerking en begeleiding van Francis was dit een meevaller. Ik wil hier ook de mensen van de toenmalige stagebureau aan de visbokaal in het PSI bedanken om van heel het stage/onderzoek doen een tof gebeuren te maken in een meer dan gezellige sfeer. Aline, Greet, Lieven, Leen en Leen merci!

Aangezien het onderzoek doen me zo goed meeviel en het triumviraat (Iven Van Mechelen, Paul de Boeck, Francis Tuerlinckx) aan het hoofd van de onderzoeksgroep kwantitatieve psychologie en individuele verschillen wel iets in me zagen, kwam ik uiteindelijk op het eerste verdiep van het PSI in blok C terecht. Iven, bedankt voor de goede ontvangst binnen de groep en de kansen die je doctoraatstudenten aanbiedt om deel te nemen aan allerlei congressen en conventies. De samenwerking voor het vak seminarie data-analyse was een aangename ervaring en bovendien gaf mijn opdracht voor dit vak me een leuke afwisseling tussen mijn dagelijks doctoraatsbezigheden. Paul, bedankt voor de uitstraling die je hebt naar buiten toe. Je contacten binnen de IRT wereld brachten enkele interessante mensen naar Leuven, waar ik anders enkel de naam van zou kennen uit de vakliteratuur. De relevante en vaak scherpzinnige vragen die je stelde tijdens het wekelijkse seminarie, werden zeer geapprecieerd en brachten vaak een grijns of glimlach op mijn gezicht en dit ondanks de doorgaans niet-licht-verteerbare onderwerpen die gepresenteerd werden. Francis, bedankt voor deze kans die je me aanbood. Bij aanvang van het doctoraat stond je bijna continu klaar voor uitleg en verklaring bij mijn waarschijnlijk onophoudelijke stroom vragen en kleine problemen. Dit maakte dat ik onmiddellijk goed uit de startblokken vloog en vlot ingewerkt geraakte in de mixed models en andere gecompliceerde materie. Ik denk dat we samen best wel heel wat werk verzet hebben en, zoals ik eerder al stelde, trots mogen zijn op het resultaat.

Verder wil ik nog mijn bureaugenootjes bedanken voor het goede gezelschap. In het begin vreesde ik voor het soort volk waarmee ik een bureau zou moeten delen, maar tot mijn verrassing kreeg ik met Tineke vertrouwd en aangenaam gezelschap vanuit de stageperiode, een meevaller dus. Wat later kregen we met Zita nog wat extra culturele diversiteit in huis, waardoor we uiteindelijk toch een gevarieerd en interessant gezelschap vormden op onze bureau.

Ook de opeenvolgende secretaresses van de onderzoeksgroep, Emmy en Véronique verdienen een woord van lof voor al hun hulp bij administratieve beslommeringen en praktische zaken, maar bovenal voor de dagelijkse energie die ze beiden uitstraalden. Mijn computerwerkzaamheden en berekeningen werden bovendien ook nog een vlot op de rails gezet door onze ICT-guru Kristof. Vooral het wegijs maken op de cluster was een hele openbaring en versnelde het lopen van analyses en simulatiestudies enorm, merci!
Hoewel ik niet de meest netwerkende collega ben, wil ik toch mijn algemene appreciatie uitspreken tegen over een aantal collega’s in het PSI. Jan, voor een uitstekende en toffe kompaan te blijken gedurende de begeleiding van het seminarie data-analyse, maar ook de andere leden van de onderzoeksgroep voor de optimale sfeer, feestelijk gezelschap op congressen, en voor de occasionele toffe babbels. Joke om me een klein beetje te betrekken bij haar interessant, maar gevoelig en gecompliceerd crosscultureel onderzoek, en gewoon om een aangename Hagelandse lotgezel te zijn in dit academisch wereldje.

Ik mag me gelukkig prijzen dat ik naast mijn onderzoeksbezigheden meer dan voldoende tijd vrij wist te houden voor de belangrijke zaken in het leven, kameraden en vrienden met andere woorden. Dienen hier zeker vermeld te worden: Van het badminton volkje, mijn kameraad van in de licenties Hans, en de huidige vaste kiek met Julie, Natascha en Steven; de ganse “de Curve”-crew natuurlijk, voor de talrijke fijne avonden die voornamelijk bestonden uit kaarten, kickeren, voetbal kijken, bijbabbelen en de nodige drank tegen de dorst; Eveline, Evlin, Joris, en Nele voor een luisterend oor en het op tijd en stond wat samen onnozel doen.

Als laatste wil ik mijn ouders en familie bedanken voor hun onvoorwaardelijke steun, en de opportuniteit en vrijheid die me werd gegeven bij het uitzoeken van wat ik zou doen bij de verschillende stappen in mijn studies. Bovendien vormde thuis op het platteland een schitterend baken van rust waar ik graag vertoefde in het weekend.

Het afronden van dit doctoraat betekent ook het afsluiten van een stap in mijn leven, maar meer nog is het het begin van een nieuwe uitdaging. Ik kijk er alvast naar uit en hoop dat ze zo goed meevalt als deze.

- The Journey is the Destination -

Johan
## CONTENTS

### General introduction

References ........................................ xxv

### Copulas for residual dependency

1.1 Introduction .................................. 2
1.2 Copula theory ................................ 7
1.3 Copula model for residual dependencies ....... 12
1.4 Statistical inference in the copula model ...... 17
1.5 Applications .................................. 18
  1.5.1 Application 1: Small reading test ...... 19
  1.5.2 Application 2: Verbal aggression data ... 23
1.6 Discussion .................................. 28
References ........................................ 32

### A latent variable framework for teratology studies

2.1 Introduction .................................. 38
2.2 Method ....................................... 42
  2.2.1 Latent variable framework ............... 42
  2.2.2 Specification of covariates ............... 45
  2.2.3 Specification of the latent distribution ... 47
  2.2.4 Specification of the joint distribution ... 48
2.3 Model inference .............................. 55
2.4 Application: BAT study .................... 56
  2.4.1 Model specification ....................... 57
  2.4.2 Simulation study ......................... 61
4.4.3 Results: Dimensional increasing LID ......... 139
4.4.4 Results: Dimensional decreasing LID ......... 141
4.4.5 Results: Model complexity ................. 141
4.5 Discussion .................................. 143
References ...................................... 147

5 Fitting explanatory IRT models: a Matlab IRTm toolbox 155
5.1 Introduction .................................. 156
  5.1.1 IRTm toolbox .......................... 158
  5.1.2 General use ............................ 159
5.2 Model building by means of design matrices .... 164
  5.2.1 Accounting for covariate information .... 170
5.3 Conditional independence and copulas ............ 180
5.4 Data simulation .............................. 187
5.5 Discussion .................................. 188
References ...................................... 190
5.6 Appendix: The copula IRT model ................. 192
5.7 Tables ..................................... 195

Epilogue ........................................ 199
References ...................................... 203
General introduction

Latent variable models

For many constructs of interest in the social sciences, educational measurement, psychology, biology, and economics, no direct method exists for measurement. Nonetheless, examples of such constructs are legion, for instance think of political attitudes, abilities, personal traits, or product preferences. To get a hold of such constructs, researchers gather observable variables (hereafter called indicators, manifest variables, or items) which they hope will provide indirect evidence for the constructs of interest. Latent variable models are statistical models built to quantify and help objectify this type of inference by deriving a small set of latent unobserved variables that is underlying to the set of manifest variables and should reflect the constructs of interest. Well-known instances of this type of modeling approach are factor analysis (Spearman, 1904; Thurstone, 1947), latent class models (Lazarsfeld & Henry, 1968), and item response models (Lord & Novick, 1968).

Although the foundations of latent variable models were laid several years ago, it has taken quite some time before they became widely applied. This is mainly due to the statistical nature of these models as well as the sometimes complicated computations and algorithms needed to estimate latent variable models. With the recent advances in computation speed and optimization algorithms and the availability of general purpose software able to fit latent variable models, this has become less of an issue. An additional factor that contributes to ongoing developments and innovations within latent variable models is the insight in
the interrelations between the different types of latent variable models, which allows the construction of a general modeling framework (see e.g., Skrondal & Rabe-Hesketh, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; De Boeck & Wilson, 2004) that is able to make the connection with the mainstream statistical literature.

Despite these developments, it is still the case that latent variable models for categorical manifest variables, compared to normal manifest variables, require an analytically less straightforward and thus computationally more complicated treatment. In this dissertation, the data sources will in fact be of a categorical nature. A first dataset consists of the binary responses (correct or right) of pupils completing a reading-comprehension test in an educational measurement setting (Chapter 1). A second dataset is based upon binary responses from a verbal aggression questionnaire used in a personality study (Chapter 1). A third motivating dataset originates from a birth defect screening instrument, in which the observed binary outcomes are the presence/absence of certain anomalies in newborns (Chapter 2). A fourth dataset contains ordinal (multi-categorical) outcomes reflecting the judged plausibility of specific behaviours within a personality study (Chapter 3). Besides these real-life datasets, simulated datasets that allow to explicitly control data properties of interest will also be considered (Chapter 4 and 5).

**Conditional independence and residual dependence**

A basic principle behind almost all latent variable models is that the set of latent variables is considered to be sufficient to explain the heterogeneity between the data sample units, as well as the homogeneity between the manifest variables. Note that the first is sometimes referred to as individual differences and the latter comprehends the dependency structure of the manifest variables. These aspects of the model illustrate the attractivity of this framework for scientific studies within psychology, educational measurement and many other areas. This basic principle is formalized in the local stochastic or conditional independence assump-
tion of latent variable models, which states that conditional on the set of latent variables, there is no remaining dependency between the manifest variables.

However, this conditional independence is such a strict requirement that it is unlikely to be met completely with a low number of latent variables for most datasets. It is common that subsets of manifest variables share common features which results in these variables partially showing dependence due to aspects not included in the model, and hence, not only because they relate to the latent variables of focus. This is called local item dependence or residual dependency. Examples of such subsets that may show local dependencies are legion: indicators sharing the same sub-theme, stimulus material, or question format. Ignoring these residual dependencies between manifest variables may lead to biased results and biased inferences because the model fails to correctly account for the dependence structure among the manifest variables (see e.g., Chen & Thissen, 1997; Sireci, Thissen, & Wainer, 1991; Yen, 1993).

The key problem with locally dependent indicators can be made more explicit when one recognizes that the information they provide on the construct of interest is partly redundant, such that part of the information is covered more than once. Taking the extreme case example of duplicated indicators, illustrates this point. Being able to spell the same word correctly over and over again, will not allow us to learn much more about this person’s general spelling proficiency, and our mathematical model should take this into account.

A drastic solution to the problem would be to delete all but one indicator within a subset of manifest variables that display residual dependency. However, this is often not an option because the indicators are not perfectly redundant, and as such one would throw away not only overlapping information, but also new information. Furthermore, because of external obligations, these indicators might sometimes even be required from a contents validity perspective. You do not want a narrow set of manifest variables with limited predictive value towards
the construct of interest, and thus, attempts have to be made to develop latent variable models that properly solve this issue.

Most modeling approaches dealing with violations of conditional independence in latent variable models are proposed by psychometricians in the context of item response theory. There is the notable exception of correlated errors in structural equation modeling, which will also be dealt with in this dissertation. Almost all model approaches slightly change their level of analysis from the indicator level to the subset level and change the marginal distribution within the probability model for the individual indicators.

Let us discuss three types of well-known residual dependency approaches. First, the score-based approach (Thissen, Steinberg, & Mooney, 1989) models the sum score of subset outcomes as a polytomous indicator, instead of each subset indicator individually. The charm of this method is that it has a straightforward logic and implementation, for instance, it can make use of existing latent variable models for polytomous data, and their corresponding machinery and software. The counterpart is that using a summary measure implies that one looses the individual information contained in the set of manifest variables. Second, the conditional combination model (Hoskens & De Boeck, 1997), is similar to loglinear modeling approaches for categorical data, and conditions on the other outcomes within the subset of indicators. Although the individual subset outcomes are left intact, the addition of interaction terms to the joint probability model of the subset does change the marginal probability model for an individual indicator, making it dependent on the outcomes on the other indicators within that subset. Third, the testlet model (Bradlow, Wainer, & Wang, 1999) conditions on additional subset-specific latent variables, also leaves the individual subset outcomes intact, but again, the addition of subset-specific latent variables to the joint probability model of the subset outcomes does change the marginal probability model for an individual outcome as it is now dependent on the latent variables specific to the subsets the indicator
belongs to. Thus, in all three modeling approaches these changes to the regular conditional independence model are not merely a change of perspective, but also change the interpretation of traditional model parameters for the marginal probabilities of outcomes on an individual manifest variable in the latent variable model.

The copula approach

This dissertation takes a marginal modeling approach to the problem of residual dependencies. The aim is to produce a latent variable model that accounts for the dependency structure of the manifest variables, while still retaining relative simple structure, parsimony and clarity. The resulting models will have the property that no essential changes to the marginal probability structure of the model need to be made, i.e. they have reproducible margins. Therefore, the interpretation of traditional parameters within the latent variable model do not change, this in contrast to the most accessible alternative modeling approaches. This reproducibility property will be one of the main advantages of the approach.

The key tool is based upon the introduction of copula functions within the stream and tradition of current latent variable models. Copula functions (Nelsen, 1999) originated in the mathematical domain of probabilistic metric spaces, when investigating relations between multidimensional distributions and their lower-dimensional margins. Recently, copulas have found applications for continuous data in risk assessment and actuarial analyses (see e.g. Cherubini, Luciano, & Vecchiato, 2004; McNeil, Frey, & Embrechts, 2006), but they still remain a relatively unexplored tool for multivariate modeling, especially in the case of categorical data (exceptions are Meester & MacKay, 1994; Trégouët et al., 1999). The primary focus of the dissertation is to work on the usefulness of this new and promising multivariate modeling tool that copula functions are, in order to tackle the problem of local dependencies. The main advantage of the copula approach lies in the double
flexibility it offers: There is a large flexibility in defining the marginal probability model for individual indicators, but also a large flexibility in dependence structure for the joint behavior of a set of manifest variables.

Dissertation structure

The dissertation booklet is a collection of five self-contained manuscripts all devoted to the study of copulas in latent variable modeling. Inherently, this formula implies that there is some overlap between the chapters, in our case especially regarding the introduction of the copulas. As previously mentioned, data and applications differ across the chapters: Simulated data as well as data from educational measurement, personality psychology, and biomedical research are analyzed throughout the dissertation.

The first three chapters show an evolution in our development of the copula approach within the broad class of latent variable models. In each chapter the demands and requirements for a suitable latent variable model are increased and therefore also the model complexity.

In Chapter 1, the foundations of the copula approach are build and the method is applied to the Rasch model (Rasch, 1960), the most well-known psychometric model for binary scored items. It is shown how the new copula Rasch model can be interpreted and a comparison is made with other alternative models for local item dependencies. The double flexibility of the copula approach, and the related reproducibility property, will play a central role in this chapter.

In Chapter 2, it is shown how current latent variable models can accommodate the specific demands that are made towards a measurement instrument used in the typical setting of birth defects screening. This involves etiological questions about the influence of exposure, and also the issue of local item dependencies is relevant again (e.g., because some of the items in the screening test relate to the same genetic or physical cause). On a model level this includes the incorporation of covariate information explicitly in the marginal model or implicitly by means of
a finite mixture distribution for the latent variable, as well as the inclusions of copula functions within this more general latent variable model. Note that a two-parameter logistic model of the item response theory tradition is chosen as the starting point for these model developments.

In Chapter 3, the copula approach is extended to the case of ordinal outcomes. The context of a contextualized personality questionnaire allows to illustrate the particularities of dependency structures for ordered category data. Based upon substantive theory we test a mediation model about how anger mediates the relation between hostility and aggression. Furthermore, item response theory is linked up with another latent variable approach often referred to as SEM (short for structural equation modeling), which proposes both a measurement model for the manifest variables (items) and a structural model for the underlying set of latent variables. Thus, besides the transition from binary to ordinal data, also the transition from a unidimensional to a multidimensional latent variable model is made.

The last two chapters deal with methodological issues that arise when applying these newly developed copula IRT models. The focus is here upon unidimensional models for binary outcomes. In Chapter 4, the issue of model selection is investigated. The central question here is whether one can differentiate accurately between different dependency structures as represented by the different copula functions. The applicability and performance of some new model selection heuristics (such as minimum description length) are evaluated. The results of an extensive simulation study indicate that simple model selection criteria, although not suited at first sight, often outperform other more complicated ones. In Chapter 5, a Matlab toolbox for fitting explanatory IRT models for binary data is introduced. This toolbox is a product of some of my programming efforts while trying to develop and implement the copula latent variable models and partially reflects what I learnt about explanatory IRT modeling during the process of completing this dissertation.
Finally the dissertation is concluded with an epilogue. While the copula approach is still in its infancy and we had the privilege to help it make its first steps in the world of latent variables models, this epilogue will point out some challenges and future directions.

As mentioned above, the dissertation booklet is a collection of five self-contained manuscripts. Chapter 1 is published in Psychometrika (Braeken, Tuerlinckx, & De Boeck, 2007). Chapter 2 has been submitted to a biostatistical journal. Chapter 3 has just recently been finished and is about to be submitted, as is Chapter four. Chapter five requires some structural changes before publication. This chapter is best read while at the same time having access to a computer (with MATLAB installed). The software is freely available. All chapters have Johan Braeken, Francis Tuerlinckx, and Paul De Boeck as authors, except for Chapter 3 where Peter Kuppens is a fourth author.
References


Abstract: Most item response theory models are not robust to violations of conditional independence. However, several modeling approaches (e.g., conditioning on other responses, additional random effects) exist that try to incorporate local item dependencies, but they have some drawbacks such as the non-reproducibility of marginal probabilities and resulting interpretation problems. In this paper, a new class of models making use of copulas to deal with local item dependencies is introduced. These models belong to the bigger class of marginal models in which margins and association structure are modeled separately. It is shown how this approach overcomes some of the problems associated with other local item dependency models.
Chapter 1

1.1 Introduction

In most measurement situations test items are developed with the intention to provide unique information about the ability or skill being measured. Items are, however, often related through a common theme, stimulus material or formulation. Examples are legion; reading comprehension items linked to the same reading passage and multi-stage items, for instance, are widespread. Responses on these items will partially show dependency due to their “artificial” common ground and not only because they relate to the skill or ability intended to be measured (Ferrara, Huynh, & Michaels, 1999; Yen, 1993). If one wants to get a “pure” assessment of this skill or ability, measurement models should take this noise, often labeled as local item dependency, into account.

A well-known measurement model used in different areas of research in psychology and educational measurement is the Rasch model (Rasch, 1960). For a person \( p \) \((p = 1, \ldots, P)\) and item \( i \) \((i = 1, \ldots, I)\), a binary random variable \( Y_{pi} \) is defined and the probability of a realization \( y_{pi} \) equals:

\[
\Pr(Y_{pi} = y_{pi}|\theta_p) = \frac{\exp(y_{pi}(\theta_p - \beta_i))}{1 + \exp(\theta_p - \beta_i)},
\]

where \( \theta_p \) is the proficiency of person \( p \) and \( \beta_i \) is the difficulty of the item \( i \). For the moment we will assume that \( Y_{pi} = 1 \) corresponds to a correct response and \( Y_{pi} = 0 \) to an incorrect one (but see the second application in the Application section).

For the remainder of the manuscript, it will be convenient to define the Rasch model as a latent threshold model in which an underlying (latent) continuous variable \( X_{pi} \) is logistically distributed with mean \( \theta_p - \beta_i \), a scale parameter equal to one and a threshold parameter set at 0 (see e.g., De Boeck & Wilson, 2004). More formally, this can be written as follows: \( X_{pi} = \theta_p - \beta_i + \varepsilon_{pi} \), with \( \varepsilon_{pi} \) following a standard logistic distribution (i.e., location zero and scale one) such that \( \varepsilon_{pi} \) can be called a latent residual for person \( p \) and item \( i \).
Furthermore it can be deduced that:

\[
\Pr(Y_{pi} = 1|\theta_p) = \Pr(X_{pi} > 0|\theta_p) = \Pr(\varepsilon_{pi} > -\theta_p + \beta_i)
\]

\[
= 1 - F(-\theta_p + \beta_i)
\]

\[
= 1 - \frac{\exp(-\theta_p + \beta_i)}{1 + \exp(-\theta_p + \beta_i)} = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}.
\]

Before turning to some basic characteristics of the Rasch model, we will first introduce some necessary notation. In general, we will make a notational distinction between a random variable (written in capitals; e.g., \(Y_{pi}\) or \(X_{pi}\)) and its realization (e.g., \(y_{pi}\) or \(x_{pi}\)). An exception to this rule pertains to random variables denoted with Greek letters (e.g., \(\varepsilon_{pi}\)) for which the same notation will be used to denote random variable and realization (a standard convention in the IRT literature). The cumulative distribution function (cdf) of a random variable \(X\) evaluated for an arbitrary value \(x\) will be denoted as \(F_X \equiv F_X(x)\). The joint cdf of a vector of random variables \(X = (X_1, \ldots, X_I)\) will be referred to as \(F_X \equiv F_X(x_1, \ldots, x_I)\). In some situations, it is necessary to express explicitly the conditioning on the latent trait \(\theta\) (e.g., \(F_X|\theta(x|\theta)\)).

A basic assumption of the Rasch model is local stochastic or conditional independence:

\[
\Pr(Y_p = y_p|\theta_p) = \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi}|\theta_p),
\]

where \(Y_p\) is the random vector of responses for person \(p\) on the set of \(I\) items (and \(y_p\) is the vector of corresponding realizations). Note that in the latent threshold model formulation above, the conditional assumption is equivalent to uncorrelated logistic error terms \(\varepsilon_{pi}\) across items such that \(F_{\varepsilon_p}(\varepsilon_{p1}, \ldots, \varepsilon_{pI}) = \prod_{i=1}^{I} F_{\varepsilon_{pi}}(\varepsilon_{pi})\), where \(F_{\varepsilon_p}\) is the joint distribution function of the latent residuals for person \(p\).

The Rasch model, and more in general, most item response theory models are not robust to violations of local stochastic independence. Violations of the assumption are called local item dependencies.
or residual dependencies (both terms will be used interchangeably in this manuscript). Local item dependencies can seriously affect the estimation of the model parameters (see e.g., Chen & Thissen, 1997; Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001a; Yen, 1984, 1993). To illustrate this intuitively, consider the extreme case wherein several slightly differently phrased questions are used in one test, this is almost equivalent to asking a single question. This redundancy situation will lead to an inflation in or “double counting of” information (Ip, Wang, De Boeck, & Meulders, 2004). The effects of this redundancy are for instance the underestimation of the standard error of a person’s proficiency parameter $\theta_p$ (Junker, 1991) and a positively biased test information function. For the effect on the item discrimination parameter in a two-parameter logistic model (2PLM), see Masters (1988). In general, when an item response model that assumes local independency is used for a test that suffers from local item dependencies, this can result in biased estimates for both person and item parameters, and, the test information function and other diagnostics that assume conditional independence.

Once local item dependency problems have been detected, there are several possible approaches to model them; for an overview see Tuerlinckx and De Boeck (2004). Two typical problems that are associated with some of the most popular models for local item dependencies are non-reproducibility and the impossibility of interpreting $\beta_i$ as the difficulty of item $i$. To illustrate these problems, we take the constant combination interaction model (CCI) of Hoskens and De Boeck (1997) as a starting point and for simplicity, only the case with two items (1 and 2) showing residual dependency is considered. The joint probability of response $(y_{p1}, y_{p2})$ then equals:

$$
\Pr(Y_{p1} = y_{p1}, Y_{p2} = y_{p2} | \theta_p) = \frac{\exp(y_{p1}(\theta_p - \beta_1) + y_{p2}(\theta_p - \beta_2) + y_{p1}y_{p2}\lambda)}{1 + \exp(\theta_p - \beta_1) + \exp(\theta_p - \beta_2) + \exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}.
$$
Note that conditional on $\theta_p$, the parameter $\lambda$ expresses the conditional log odds ratio for the item pair given the responses on all other items, hence the CCI model only reduces to the familiar Rasch model in case $\lambda$ equals zero (see Tuerlinckx & De Boeck, 2004, pp.303-304, for the interpretation of higher-order associations between 3 or more items).

Calculating the probability of answering the first item correctly gives (the marginal item characteristic curve or ICC for item 1):

$$\Pr(Y_{p1} = 1|\theta_p) = \sum_{y_{p2}=0}^{1} \Pr(Y_{p1} = 1, Y_{p2} = y_{p2}|\theta_p)$$

$$= \frac{\exp(\theta_p - \beta_1) + \exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}{1 + \exp(\theta_p - \beta_1) + \exp(\theta_p - \beta_2) + \exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}.$$  \hspace{1cm} (1.1)

It can be seen that, when $\lambda \neq 0$ (hence residual dependency between items 1 and 2 exists), the marginal probability of responding correctly to item 1 is not a Rasch model anymore (i.e., Equation 1.2 is not a logistic function). Therefore, it is said that the marginals are not reproducible\(^1\). Moreover, the parameter $\beta_1$ looses its natural interpretation of difficulty parameter because it is not simply the location on the latent trait for which the probability of responding correctly is 0.5 (see also Ip, 2002). Both problems are illustrated in the panels of Figure 1.1 where the item characteristic curve of item 1 derived from the CCI model (see Equation 1.2) is plotted for a certain configuration of parameters. The item characteristic Rasch curve is plotted for comparison reasons in each panel by a dotted line. It can be seen that the probability of responding correctly to item 1 also depends on $\lambda$ and $\beta_2$ (cfr., Equation 1.2). The same kinds of problems also occur in random-effect testlet models (Bradlow, Wainer, & Wang, 1999) and other types of conditional models (e.g., the dynamic Rasch model of Verhelst & Glas, 1993).

\(^1\)Note that we employ the concept of reproducibility here in an intuitive fashion. More formal definitions can be found in (Ip, 2002; Fitzmaurice, Laird, & Rotnitzky, 1993).
A class of models that does not suffer from the aforementioned problems are the so-called marginal models (see e.g., Molenberghs & Verbeke, 2005). In these models, the univariate margins and the dependence structure are modeled separately. In consequence, these models will not suffer of the problem of not-reproducible marginals. Examples proposed in the literature are the Bahadur-Ip model (Bahadur, 1961; Ip, 2000, 2001) and the hybrid kernel approach developed by Ip (2002) that can be related to the work of Holland (1990) and generalized log-linear models (Cox, 1972; Zhao & Prentice, 1990). If we want to make use of logistic marginals, as is the case for the Rasch model, a logical approach would be to use a multivariate random effect logit model analogous to the multivariate random effect probit model. Unfortunately, there is no single or generally accepted multivariate logistic distribution (see e.g., Kotz, Balakrishnan, & Johnson, 2000), analogous to the multivariate normal distribution, which can be used as a starting point for introducing dependencies between the latent residuals. However, with copula functions this paper will introduce a convenient tool to model dependency and
construct multivariate distributions. The residual dependencies will be taken into account by these copulas without changing anything to the marginal model part (i.e., item characteristic curve and interpretation of the parameters). A residual dependency model that preserves Rasch logistic margins will be used as an obvious and prototypical working example of the copula approach.

The remainder of the paper is organized as follows. First, some basic theory with respect to copulas will be discussed. We will continue with the formulation of a copula model for residual dependencies that has Rasch logistic margins and is quite straightforward in its approach and interpretation. Then this copula model will be applied to two data sets containing residual dependencies. Finally, the paper is closed with a discussion.

1.2 Copula theory

In order that this paper should be reasonably self-contained, we begin with introducing copula functions as a mathematical concept. A more thorough overview can be found in the reference works by Joe (1997) or by Nelsen (1999). In mathematics, a copula (Latin for link or tie) defines a function that relates, or “couples” a multivariate distribution function to its univariate margins.

It is said that an $R$-dimensional copula is a function $C : [0, 1]^R \rightarrow [0, 1]$ with the following properties:

1. For every vector $u \in [0, 1]^R$, $C(u)$ is increasing in each component $u_r$ with $r \in 1, 2, \ldots, R$.

2. For every vector $u \in [0, 1]^R$, $C(u) = 0$ if at least one coordinate of the vector is 0 and $C(u) = u_r$ if all the coordinates of the vector are equal to one except the $r$-th one.

3. For every $a, b \in [0, 1]^R$ with $\forall r \in \{1, 2, \ldots, R\}, a_r \leq b_r$, given a hypercube $B = [a, b] = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_R, b_R]$ whose
vertices lie in the domain of $C$, and with $V_C(B) \geq 0$. The volume $V_C(B)$ is defined as:

$$V_C(B) = \sum_{k_1=1}^{2} \sum_{k_2=1}^{2} \ldots \sum_{k_R=1}^{2} (-1)^{k_1+k_2+\ldots+k_R} C\left(d_1^{(k_1)}, d_2^{(k_2)}, \ldots, d_R^{(k_R)}\right)$$

(1.2)

where $d_r^{(1)} = a_r$ and $d_r^{(2)} = b_r$ ($r = 1, \ldots, R$).

An important result from the theory of copulas is Sklar’s Theorem (Sklar, 1959). The theorem states that for any $R$-dimensional distribution function $F_X$ with univariate margins $F_{X_1}, F_{X_2}, \ldots, F_{X_R}$ there exists a copula function $C$ such that this multivariate distribution $F_X$ can be represented as a function of its margins through this copula: $F_X = C(F_{X_1}, F_{X_2}, \ldots, F_{X_R})$. The key idea is that any existing multivariate distribution can be reformulated according to this copula presentation, and conversely, based upon any kind of univariate margins, a joint (i.e., multivariate) distribution can be constructed by means of a copula function. Note that we will make use of exactly this part in our modeling approach to construct a multivariate distribution using copulas, such that the univariate margins of this newly constructed multivariate distribution are still the original univariate distributions we started from. Making use of the second property of the copula definition, it can easily be deduced that the univariate marginal distribution for $X_r$ equals $C(1, \ldots, 1, F_{X_r}, 1, \ldots, 1) = F_{X_r}$. Hence, in this way an association between the $R$ random variables is allowed while preserving the univariate margins.

For each joint distribution with margins $F_{X_1}, F_{X_2}, \ldots, F_{X_R}$, and constructed by means of a copula the following applies:

$$W(F_{X_1}, \ldots, F_{X_R}) \leq C(F_{X_1}, \ldots, F_{X_R}) \leq M(F_{X_1}, \ldots, F_{X_R}),$$

where $W(F_{X_1}, \ldots, F_{X_R}) = \max(F_{X_1} + \ldots + F_{X_R} - R + 1, 0)$, and $M(F_{X_1}, \ldots, F_{X_R}) = \min(F_{X_1}, \ldots, F_{X_R})$. 

-8-
Copulas for residual dependency

The functions $W(F_{X_1},\ldots,F_{X_R})$ and $M(F_{X_1},\ldots,F_{X_R})$ correspond to the Fréchet-Hoeffding lower and upper bound (Fréchet, 1951; Hoeffding, 1940) and define the maximum negative and positive dependency of a joint distribution that can be obtained given fixed margins. These bounds can be used to indicate the range of dependency a copula function can capture.

A wide variety of possible copula functions exist which allow for fitting a wide range of dependency types (see e.g., Joe, 1993; Nelsen, 1999). Our modeling approach will focus on the class of Archimedean copulas (Genest & MacKay, 1986; Joe, 1993; Nelsen, 1999). Archimedean copulas have a simple structure and can be written as:

$$C(u_1,\ldots,u_R) = \psi^{-1}\left(\sum_{r=1}^{R} \psi(u_r)\right),$$

where $\psi : [0,1] \to [0,\infty]$ is a continuous strictly decreasing function, called the generator function, such that $\psi(0) = \infty$ and $\psi(1) = 0$, and $\psi^{-1}$ is completely monotonic on $[0,\infty)$, such that $(-1)^k \frac{d^k}{dt^k} \psi^{-1}(t) \geq 0$ $\forall t \in [0,\infty)$ and $k \in \mathbb{N}$. Archimedean copulas have nice symmetry properties, as there are permutation symmetry, $C(u_1,u_2) = C(u_2,u_1)$, and associativity, $C(u_1,u_2,u_3) = C(u_1,C(u_2,u_3)) = C(C(u_1,u_2),u_3)$; which makes them especially attractive for modeling symmetrically dependent data. Furthermore, notice that the independence case can also be rewritten in a convenient way under an Archimedean copula representation:

$$F_X = \prod_{r=1}^{R} F_{X_r} = C(F_{X_1},\ldots,F_{X_R}) = \exp\left(-\sum_{r=1}^{R}[-\log(F_{X_r})]\right),$$

with $\psi(u) = -\log(u)$ and $\psi^{-1}(t) = \exp(-t)$. This copula is also known as the product copula, denoted by $\Pi$ (Frees & Valdez, 1998; Nelsen, 1999).
Two Archimedean copula families are presented in this paper:

1. Frank’s copula (Frank, 1979):

\[
C(F_{X_1}, F_{X_2}, \ldots, F_{X_R}) = \frac{-1}{\delta} \log \left( \frac{1 - \prod_{r=1}^{R} (1 - \exp(-\delta F_{X_r}))}{\prod_{r=1}^{R-1} (1 - \exp(-\delta))} \right),
\]

with for \( R = 2 \): \( \delta \in \mathbb{R}/\{0\} \), if \( \delta \to -\infty \) then \( C \to W \), if \( \delta \to 0 \) then \( C \to \Pi \), if \( \delta \to \infty \) then \( C \to M \); and for \( R > 2 \): \( \delta > 0 \), if \( \delta \to 0 \) then \( C \to \Pi \), if \( \delta \to \infty \) then \( C \to M \);


\[
C(F_{X_1}, F_{X_2}, \ldots, F_{X_R}) = \left( \sum_{r=1}^{R} F_{X_r}^{-\delta} - R + 1 \right)^{-\frac{1}{\delta}}
\]

with \( \delta > 0 \), if \( \delta \to 0 \) then \( C \to \Pi \), if \( \delta \to \infty \) then \( C \to M \).

Both copula families will be used extensively throughout this paper. We discuss below why these two functional forms were choosen. The parameter \( \delta \) that defines each of these Archimedean copulas has an interpretation of an association parameter. For instance, when the value of \( \delta \) increases, the dependence captured by the copula gets closer to the theoretical maximum positive dependence for \( F_X \) (i.e., \( M \), the Fréchet-Hoeffding upper-bound).

When we consider standard logistic marginals Frank’s and Cook-Johnson copula lead to different joint distributions. In Figure 1.2 we illustrate the degree and kind of dependence by depicting for the two-item case the contour lines of several latent bivariate distributions \( F_{X_p|\theta_p} \) with standard logistic margins \( F_{X_{p1}|\theta_p} \) and \( F_{X_{p2}|\theta_p} \) constructed by means of a Frank’s copula (upper 4 panels) and Cook-Johnson copula (lower 4 panels) for varying values of the association parameter \( \delta \). Notice that the copulas can capture a broad range of dependency and differ in the type of dependence they induce; for instance, Cook-Johnson copula has a prominent lower tail (i.e., more formally, \( C(u, \ldots, u)/u \) converges to a constant \( c \) in \([0, 1]\) as \( u \to 0 \), Joe, 1993; Nelsen, 1999), while Frank’s copula leads to a similar kind of dependence in both tails.
Figure 1.2: Bivariate logistic density contour plots for different copula models.

(a) Frank’s copula
(b) Cook-Johnson copula
1.3 Copula model for residual dependencies

With $\theta_p$ being the proficiency of person $p$ ($p = 1, \ldots, P$), $\beta_i$ being the difficulty of the item $i$ ($i = 1, \ldots, I$), latent variable $X_{pi} = \theta_p - \beta_i + \varepsilon_{pi}$, and observed binary variable $Y_{pi} = I(X_{pi} > 0) = I(\varepsilon_{pi} > -\theta_p + \beta_i)$, the vector $\varepsilon_p = (\varepsilon_{p1}, \ldots, \varepsilon_{pI})^T$ is assumed to have an $I$-variate distribution with standard logistic univariate margins $F_{\varepsilon_{pi}}$. The conditional independence assumption of the Rasch model implies that the logistic error terms $\varepsilon_{pi}$ are uncorrelated across items, such that $F_{\varepsilon_p}(\varepsilon_{p1}, \ldots, \varepsilon_{pI}) = \prod_{i=1}^I F_{\varepsilon_{pi}}(\varepsilon_{pi})$. However, when residual dependencies within different subsets of items are detected, a different dependency structure for $\varepsilon_p = (\varepsilon_{p1}, \ldots, \varepsilon_{pI})^T$ is required, and consequently a more appropriate $I$-variate distribution $F_{\varepsilon_p}(\varepsilon_{p1}, \ldots, \varepsilon_{pI})$ is needed. The marginal distributions $F_{\varepsilon_{pi}}$ are univariate logistic distributions because we want the marginal probabilities of a correct response to be a Rasch model. The unknown $I$-variate joint distribution $F_{\varepsilon_p}(\varepsilon_{p1}, \ldots, \varepsilon_{pI})$ will be constructed from these margins by means of copula functions.

Consider $S$ disjoint subsets of $\{1, \ldots, I\}$ denoted as $J_1, \ldots, J_S$, where $J_s$ has cardinality $I_s$. The vector of error terms $\varepsilon_p$ is similarly divided into subsets $\varepsilon_p^{(1)}, \ldots, \varepsilon_p^{(S)}$ where $\varepsilon_p^{(s)} = (\varepsilon_{pi}, i \in J_s)$. The different subsets are independent, and the variables in a subset $\varepsilon_p^{(s)}$ are assumed exchangeable. Subsets of items can be chosen based upon diagnostic tests for residual dependency, context of items, or substantive theory.

The joint probability of the response vector of a person $p$ is:

$$\Pr(Y_p|\theta_p) = \prod_{s=1}^{S} \Pr(Y_{pi} = y_{pi}, i \in J_s|\theta_p),$$

with the joint probability of responses on item subset $J_s$ equal to:

$$\Pr(Y_{pi} = y_{pi}, i \in J_s|\theta_p) = \Pr(d_{pi}^{(1)} < \varepsilon_{pi} \leq d_{pi}^{(2)}, i \in J_s|\theta_p),$$

where for $y_{pi} = 1$, $d_{pi}^{(1)} = -\theta_p + \beta_i$ and $d_{pi}^{(2)} = \infty$, and for $y_{pi} = 0$, $d_{pi}^{(1)} = -\infty$ and $d_{pi}^{(2)} = -\theta_p + \beta_i$. If the cardinality of subset $J_s$ is larger
than one \((I_s > 1)\), \(\Pr(d_{pi}^{(1)} < \varepsilon_{pi} \leq d_{pi}^{(2)}, i \in J_s|\theta_p)\) is evaluated from the copula \(C_S(\cdot, \delta_s)\) for \((\varepsilon_{pi}, i \in J_s)\) as:

\[
\Pr(d_{pi}^{(1)} < \varepsilon_{pi} \leq d_{pi}^{(2)}, i \in J_s|\theta_p) = \sum_{k_1=1}^{2} \ldots \sum_{k_{Is}=1}^{2} (-1)^{k_1+\ldots+k_{Is}} C_s \left( F_{\varepsilon p_1}(d_{p1}^{(k_1)}), \ldots, F_{\varepsilon p_{Is}}(d_{p_{Is}}^{(k_{Is})}) \right).
\]

For clarity, assume \(I = 2\) and \(J_s = \{1, 2\}\), the equations above simplify as follows for a \((0, 0)\)-response:

\[
\begin{align*}
\Pr(Y_{p1} = 0, Y_{p2} = 0|\theta_p) &= \Pr(d_{pi}^{(1)} < \varepsilon_{pi} \leq d_{pi}^{(2)}, i \in J_s|\theta_p) \\
&= \Pr(d_{p1}^{(1)} < \varepsilon_{p1} \leq d_{p1}^{(2)}), d_{p2}^{(1)} < \varepsilon_{p2} \leq d_{p2}^{(2)}|\theta_p) \\
&= \Pr(-\infty < \varepsilon_{p1} \leq -\theta_p + \beta_1, -\infty < \varepsilon_{p2} \leq -\theta_p + \beta_2) \\
&= C_s \left( F_{\varepsilon p_1}(-\infty), F_{\varepsilon p_2}(-\infty) \right) - C_s \left( F_{\varepsilon p_1}(-\theta_p + \beta_1), F_{\varepsilon p_2}(-\infty) \right) - C_s \left( F_{\varepsilon p_1}(-\infty), F_{\varepsilon p_2}(-\theta_p + \beta_2) \right) + C_s \left( F_{\varepsilon p_1}(-\theta_p + \beta_1), F_{\varepsilon p_2}(-\theta_p + \beta_2) \right) \\
&= 0 - 0 - 0 + C_s \left( F_{\varepsilon p_1}(-\theta_p + \beta_1), F_{\varepsilon p_2}(-\theta_p + \beta_2) \right) \\
&= C_s \left( F_{X_{p1}|\theta_p(0|\theta_p)}, F_{X_{p2}|\theta_p(0|\theta_p)} \right).
\end{align*}
\]

The last equality follows from the definition of \(X_{pi}\). The other probabilities are then:

\[
\begin{align*}
\Pr(Y_{p1} = 1, Y_{p2} = 1|\theta_p) &= 1 - F_{X_{p1}|\theta_p(0|\theta_p)} - F_{X_{p2}|\theta_p(0|\theta_p)} + \\
&\quad C_s \left( F_{X_{p1}|\theta_p(0|\theta_p)}, F_{X_{p2}|\theta_p(0|\theta_p)} \right) \\
\Pr(Y_{p1} = 1, Y_{p2} = 0|\theta_p) &= F_{X_{p2}|\theta_p(0|\theta_p)} - C_s \left( F_{X_{p1}|\theta_p(0|\theta_p)}, F_{X_{p2}|\theta_p(0|\theta_p)} \right) \\
\Pr(Y_{p1} = 0, Y_{p2} = 1|\theta_p) &= F_{X_{p1}|\theta_p(0|\theta_p)} - C_s \left( F_{X_{p1}|\theta_p(0|\theta_p)}, F_{X_{p2}|\theta_p(0|\theta_p)} \right)
\end{align*}
\]

Figure 1.2 offers an intuitive insight in these calculations by presenting the contour lines of bivariate logistic densities constructed by means of both Frank’s copula and Cook-Johnson copula. Each contour plot is
divided into quadrants made up by the solid lines drawn at the latent thresholds (the dashed lines indicating the marginal means \(\theta_p - \beta_i\)). In order to calculate the joint probabilities from the joint distribution functions of the latent random variables \(X_{p1|\theta_p}\) and \(X_{p2|\theta_p}\), the volume under the density for the corresponding quadrant needs to be calculated (see e.g., Mood, Graybill, & Boes, 1974, and see also Equation 2).

Once the partitioning of items into subsets is given, and the copula families are given, the following set of parameters has to be estimated: the item parameters \(\beta_i\) \((i = 1, \ldots, I)\), the distributional parameters of the latent trait (the person’s proficiency \(\theta_p\), is generally, but not necessarily, assumed to be normally distributed with mean zero and unknown variance \(\sigma^2\)), and the association parameters \(\delta_1, \ldots, \delta_S\). The regular Rasch model arises as a special case when \(S = I\) and each subset \(J_s\) has size 1; or when the different \(C\) are assumed to be the product copula \(\Pi\) (equivalent to independence). As an example, consider a test with \(I = 7\) items where items 1 and 2 exhibit some symmetric residual dependence, the set of items 3 to 5 also form a dependent subset (independent of the first) and items 6 and 7 do not show any violation of the general conditional independence assumption and are independent of the first two subsets. Thus, \(\{1, \ldots, 7\}\) is partitioned as: \(J_1 = \{1, 2\}, J_2 = \{3, 4, 5\}, J_3 = \{6\}\) and \(J_4 = \{7\}\). The proposed \(I\)-variate distribution for the error component vector \(\varepsilon_p = (\varepsilon_{p1}, \ldots, \varepsilon_{pI})^T\) is then: 

\[
F_{\varepsilon_p}(\varepsilon_{p1}, \ldots, \varepsilon_{pI}) = C_1(F_{\varepsilon_{p1}}(\varepsilon_{p1}), F_{\varepsilon_{p2}}(\varepsilon_{p2})) \times C_2(F_{\varepsilon_{p3}}(\varepsilon_{p3}), F_{\varepsilon_{p4}}(\varepsilon_{p4}), F_{\varepsilon_{p5}}(\varepsilon_{p5})) \times F_{\varepsilon_{p6}}(\varepsilon_{p6}) \times F_{\varepsilon_{p7}}(\varepsilon_{p7}).
\]

Because a copula model is a form of marginal modeling, a broad range of association structures (by means of different copula functions) for the subsets showing residual dependency can be compared without fundamentally changing the base model of the marginal probabilities and thus preserving the individual item characteristic curves of the items. In the previous example, \(C_1\) could be either Frank’s copula or Cook-Johnson copula, and the same holds for \(C_2\).
In order to illustrate that the introduction of the copula can take residual dependencies into account, the odds ratio (conditional on $\theta_p$) for items 1 and 2 involved in the copula $C$ can be computed as follows:

$$\text{OR}(\theta_p) = \frac{\Pr(Y_{p1} = 1, Y_{p2} = 1|\theta_p) \Pr(Y_{p1} = 0, Y_{p2} = 0|\theta_p)}{\Pr(Y_{p1} = 1, Y_{p2} = 0|\theta_p) \Pr(Y_{p1} = 0, Y_{p2} = 1|\theta_p)}$$

$$= \frac{(1 - F_{X_{p1}|\theta_p}(0|\theta_p) - F_{X_{p2}|\theta_p}(0|\theta_p) + C) C}{(F_{X_{p2}|\theta_p}(0|\theta_p) - C) (F_{X_{p1}|\theta_p}(0|\theta_p) - C)},$$

with $C = C(F_{X_{p1}|\theta_p}(0|\theta_p), F_{X_{p2}|\theta_p}(0|\theta_p))$. Using Frank’s copula, the value of the log odds ratio is then computed for several values of $\delta$ and for $\theta_p$ ranging from -4 to 4 and the result is shown in the upper panel of Figure 1.3. For Cook-Johnson copula, the same procedure is followed and this result is shown in the lower panel of the same figure. For ease of demonstration the two margins were set equal to one another, with difficulty parameters $\beta_1$ and $\beta_2$ equal to zero, so that the log odds ratio (conditional on $\theta_p$) was only a function of the copula’s association parameter and the marginal probabilities as determined by $\theta_p$. From both panels it can be seen that when the value of the copula parameter $\delta$ rises the log odds ratio also increases, indicating the copula parameter’s function as an association measure.

In both the upper and lower panel of Figure 1.3, it appears that for a fixed value of the copula parameter $\delta$, there is a dependency between the log odds ratio and the value of the latent trait. Frank’s copula shows a more static residual dependence between the two items (the log odds ratio is more or less constant, unless for large values of $\delta$). On the contrary, Cook-Johnson copula has a dimensional residual dependence structure because the log odds ratio increases as $\theta_p$ increases (for an example of a residual dependency structure that is dimension dependent see e.g., Bell, Pattison, & Withers, 1988, and also the second application in this manuscript). Our attention is restricted in this manuscript to Frank’s and Cook-Johnson copulas because they show such distinct and interesting patterns of association. Different models for static and
Figure 1.3: The conditional log odds ratio given different values of the copula parameter.

(a) Frank’s copula

(b) Cook-Johnson copula
dimensional dependence structures have also been proposed by Hoskens and De Boeck (1997) in their framework of conditional residual dependency models (Hoskens & De Boeck, 1997, use the labels ‘constant combination’ and ‘dimension-dependent’ interaction models). Notice that the dimensional residual dependence induced by Cook-Johnson copula is positively accelerated, whereas in the dimension-dependent interaction model this can be either linearly increasing or decreasing.

1.4 Statistical inference in the copula model

We followed a marginal maximum likelihood (MML; see De Boeck & Wilson, 2004) approach to fit the Rasch copula model. Over all persons, and items, the likelihood to be optimized is:

\[
\prod_{p=1}^{P} \int_{\theta_p} \prod_{s=1}^{S} \left[ \Pr(d_{pi}^{(1)} < x_{pi} \leq d_{pi}^{(2)} , i \in J_s | \theta_p) \right] \phi(\theta_p | \sigma^2) d\theta_p.
\]

Usually the log is taken for numerical reasons. The negative of this loglikelihood is minimized using a quasi-newton algorithm and the approximation of the intractable integral with respect to the distribution of \(\theta_p\) is carried out with a Gauss-Hermite quadrature. In extensive simulations and applications we have found that usually 21 nodes are sufficient for a good recovery and to obtain stable results.

With respect to model checking and selection, the tools usually applied in nonlinear mixed models are available (Wald, score and likelihood ratio tests). The independence model (i.e., a Rasch model) is nested within the copula model both for Frank’s copula as Cook-Johnson copula. Therefore, a test of the null hypothesis that the association parameter equals a particular value (0 for both Frank’s and Cook-Johnson copula) is a comparison between the Rasch model and a certain copula model.

Note that for Frank’s copula, Meester (1991) showed that once \(R\) exceeds 2, the lower bound of the copula parameter \(\delta\) needs to be adapted
in function of $R$. For any $R$ this adapted lower-bound is always strictly less than zero, thus not restricting the positive association range and technically leaving the position of the independence point in the interior of the parameter space\(^2\). However for Cook-Johnson copula the independence case lies on the boundary of the parameter space since $\delta$ cannot be lower than zero. Therefore the appropriate reference distribution for the likelihood ratio test statistic comparing the independence and the Cook-Johnson model is not a chi-square with 1 degree of freedom but a mixture of two chi-square distributions, with 0 and 1 degree of freedom respectively. When more copulas and consequently more association parameters are involved, deducing the appropriate mixture of chi-square distributions to function as reference distribution may get very complicated (see e.g., Self & Liang, 1987). Therefore, we will rely in all cases on the traditional reference distribution, which yields a more conservative test of the null hypothesis.

Attention has to be given to the selection of a particular copula function. Frank’s and Cook-Johnson copula do not have a nested relation and consequently, the selection is best based on methods such as the AIC (Akaike, 1973) or BIC (Schwarz, 1978).

### 1.5 Applications

In this section, we will apply the Rasch copula model to two data sets. One data set originates from ability testing and the other from a research on verbal aggression. In both cases, the presence of residual dependencies is to be expected based on the design of the study.

\(^2\)The range of negative dependency in the case that $R > 2$ is rather limited and attaching a meaningful interpretation to negative dependency between three or more items is not very likely. Hence, we do not consider the use of Frank’s copula for negative residual dependency between more than 2 items for reasons of clarity.
1.5.1 Application 1: Small reading test

A group of high school students interested in studying law in college \( (P = 441) \) answered six multiple choice questions about a text on the president and the separation of powers in the United States of America. The answers were recoded in right and false for ease of demonstration and scored 1 and 0, respectively. Various types of tools have been developed to detect local item or residual dependencies (see e.g., Chen & Thissen, 1997; Holland & Rosenbaum, 1986; Rosenbaum, 1984; Yen, 1984; or for a recent comparison see Tate, 2003). In previous analyses (Tuerlinckx & De Boeck, 2001b) it is shown that two pairs of items showed residual dependencies: items 1 and 6, and items 4 and 5. Here we fit a series of models including the simple Rasch model and possible residual dependence models using Frank’s and Cook-Johnson copula in different combinations. The fits of these models are assessed using the appropriate tools outlined above: the likelihood ratio test and/or the AIC. The list of estimated models and results on the estimated association parameters are given in Table 1.1.

From the results in Table 1.1 it can be seen that, according to the AIC, the most appropriate model is that copula model with a separate Frank’s copula for each item subset (Model 2; Frank:\( J_1\{1, 6\}, \) Frank:\( J_2\{4, 5\}, \) and II : \( J_3\{2, 3\} \)). If we compare Model 2 to the regular Rasch model (Model 1; II : \( J_1\{1, 2, 3, 4, 5, 6\} \)), there is overwhelming evidence against the null hypothesis that both association parameters are equal to zero \( (LRT = 49, df = 2, p < 0.0001) \). Both association parameters differ significantly from zero \( (p < 0.01 \text{ and } 0.05, \text{ for } \delta_1 \text{ and } \delta_2 \text{ respectively}) \).

We also looked at the Mantel-Haenszel (MH) test as a diagnostic tool (Mantel & Haenszel, 1959) for detecting residual dependencies as proposed by Rosenbaum (1984); Ip (2001); Scott and Ip (2002). The main idea is to test for equal odds ratio between groups of persons and the group division is based upon the estimated latent trait of \( \theta_p \) in the model under investigation. Significant deviations from the hypothesis
<table>
<thead>
<tr>
<th>Model</th>
<th>Association parameter(s)</th>
<th>AIC</th>
<th>( \hat{\delta}_1 )</th>
<th>( \hat{\delta}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch</td>
<td>( {3,2} = {6,1} )</td>
<td>3.148</td>
<td>0.29</td>
<td>0.56</td>
</tr>
<tr>
<td>Frank</td>
<td>( {4,1} = {1,6} )</td>
<td>3.101</td>
<td>0.30</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 1.1: Rasch and Frank Copula models for the small reading test data.
of equal odds ratios will give an indication of the presence of residual dependency between pairs of items while controlling for the dimension $\theta_p$. Applying the Rasch model to the reading data set gives significant MH tests for item pairs $\{1,6\}$ and $\{4,5\}$: $z = 6.38$ ($p < .0001$) and $z = 2.33$ ($p = .03$), respectively. When we repeat this procedure after taking into account the residual dependencies through a copula model (i.e., Model 2), the MH tests for both item pairs become non-significant: $z = 1.37$ ($p = .16$) for item pair $\{1,6\}$ and $z = -0.61$ ($p = .33$) for item pair $\{4,5\}$. Thus it appears that the copula model has taken into account the existing residual dependencies.

The other parameter estimates for Models 1 and 2 are given in detail in Table 1.2. It can be seen that the estimates of the item parameters do not change very much between the Rasch and the Rasch copula model: There is a very slight tendency for the copula estimates to be shrunk more towards zero than the Rasch model estimates. However, the estimate of the standard deviation of the person ability distribution in Rasch copula model is considerably smaller than the equivalent estimate under the Rasch model. Thus, the location of the items on the continuum is hardly affected by introducing the copulas, but the scale of the distribution for the $\theta_p$ is shrunken somewhat. Therefore, after taking into account the residual dependencies, there is less differentiation among the examinees. Because of the large association parameter estimates together with the lower AIC value for the copula model, we may conclude that not taking into account the residual dependencies leads to a wrong inference about the population distribution of the person abilities. Consequently, the predictions of the individual $\theta_p$ may be less accurate under the Rasch model.

For reasons of comparison we also fitted the interaction models of Hoskens and De Boeck (1997) to this dataset and combinations of both this conditional approach and the copula approach. The parameter estimates and AIC for the best fitting model of both these model classes are shown in Table 1.2. A constant combination interaction for both
### Table 1.2: Parameter estimates for the Rasch model and the Rasch copula model, and the Constant combination interaction model and a combination of the previous.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>Estimates</th>
<th>Standard errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rasch</strong></td>
<td>$\beta_1$</td>
<td>-0.36</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>0.99</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>$\beta_4$</td>
<td>-1.38</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$</td>
<td>-1.50</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>$\beta_6$</td>
<td>-0.02</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Copula</strong></td>
<td>$\sigma_{\theta}$</td>
<td>1.28</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>CCI</strong></td>
<td>$\lambda_1$</td>
<td>1.47</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>1.26</td>
<td>0.27</td>
</tr>
<tr>
<td><strong>AIC</strong></td>
<td></td>
<td>3157</td>
<td></td>
</tr>
</tbody>
</table>

Frank: $J_1 = \{1, 6\}$; $\delta_1 = 3.56 (0.62)$

Frank: $J_2 = \{1, 6\}$; $\delta_2 = 2.53 (0.73)$

CCI: $J_1 = \{1, 6\}$; $\lambda_1 = 1.47 (0.24)$

CCI: $J_2 = \{1, 6\}$; $\lambda_2 = 1.26 (0.27)$

---

*Note: Estimates between parentheses*
items subsets was assessed as the best fitting of the interaction models. Compared to its copula counterpart it showed slightly better fit, but also slightly larger standard errors for the item parameters; and as expected the item location parameters were affected by applying the CCI model, and shifted along the latent dimension. Both interaction parameters differ significantly from zero ($p < 0.01$ and $0.05$, for $\lambda_1$ and $\lambda_2$, respectively). The best fitting model however, was a hybrid model with Frank’s copula for one item subset, and a constant combination interaction for the second item subset. Based on these fitting results it is hard to make a conclusive statement about which modeling approach should be preferred for this dataset; this does not come as a surprise for us considering the moderate sample size, and especially given the small subset sizes.

1.5.2 Application 2: Verbal aggression data

As a second example, we analyze the data from a behavioral questionnaire (Vansteelandt, 2000) on verbal aggression. The questionnaire consists of 24 items and all items refer to verbally aggressive reactions in a frustrating situation, and it was administered from 316 persons. Four different situations were used to construct the 24 items: ‘A bus fails to stop for me’, ‘I miss a train because a clerk gave me faulty information’, ‘The grocery store closes just as I am about to enter’, and ‘The operator disconnects me when I had used up my last 10 cents for a call’. The description of the situation is followed by a statement about the behavioral mode of a verbal aggressive reaction (‘I would want to ’ or ‘I would’) and then three verbal aggressive reactions: cursing, scolding, and shouting (the questionnaire was translated from Dutch, and shouting refers to an expressive aggressive reaction in Dutch). The original items had three response categories (‘yes’, ‘perhaps’, and ‘no’) but we dichotomized them (‘yes’ and ‘perhaps’ were scored 1 and ‘no’ was scored 0; see also De Boeck & Wilson, 2004, for an extensive study of this data set).
Because the 24 items are clustered in four groups of six items due to the common situations, we may expect residual dependencies among the items that are related to a similar situation. Making use of graphical exploratory techniques, Tuerlinckx and De Boeck (2004) showed that there are indeed residual dependencies in this dataset. In order to take the residual dependencies originating from the same situation into account, we fitted a model with four copulas, one for each situation (each containing six items), and Rasch marginals. Again we have fitted a series of models with different functional form for the copula. The best fitting Rasch copula model has a Frank’s copula for situation 1 (bus fails to stop, Frank: \( J_1 \{1, 2, 3, 13, 14, 15\} \)) and situation 2 (train miss, Frank: \( J_2 \{4, 5, 6, 16, 17, 18\} \)), and a Cook-Johnson copula for situation 3 (store closes, C-J: \( J_3 \{7, 8, 9, 19, 20, 21\} \)) and 4 (disconnection, C-J: \( J_4 \{10, 11, 12, 22, 23, 24\} \)). The results are presented in Table 1.3. From the results in Table 1.3, one can see that the Rasch copula model outperforms the regular Rasch model. Note that while the interpretation of the individual item parameters remains the same in the Rasch copula model (i.e., all \( \beta \)'s can be seen as the points on the latent continuum where the probability of a yes-response is 0.50), the model clearly has a significant better fit than the regular Rasch model (\( LRT = 197, df = 4, p < 0.0001 \)). All four association parameters, \( \delta_1, \delta_2, \delta_3 \) and \( \delta_4 \), differ significantly from zero (\( p < 0.05 \)).

Contrary to the first example, the four copulas in the model do not all have the same functional form: For the first two situations Frank’s copula was judged to be the appropriate one and for the last two situations we have chosen Cook-Johnson. As a consequence, the verbal aggressive reactions in the former situations show static local item dependence (i.e., the log odds ratio does not depend on the latent verbal aggression \( \theta_p \); see Figure 1.3) and the latter dimensional local item dependence (i.e., the log odds ratio increases as the latent verbal aggression \( \theta_p \) increases; see Figure 1.3). Note that this is why we do not report Mantel-Haenszel test statistics here, as the procedure assumes constant odds ratio over
groups contrary to what we find in this data set. Although the residual dependencies are basically considered as a nuisance in this paper, we find here a case in which an interesting observation from a substantive point of view occurs. The first two situations are situations where someone else is to blame (the bus driver, the clerk), while the last two are self-blame situations. It appears that highly verbally aggressive people want to and/or tend to show the whole range of verbally aggressive reactions in the self-to-blame situations more than can be expected based solely from their general degree of verbal aggressiveness. In contrast, persons with a lower $\theta_p$-value do not exhibit a lot of residual dependency in self-to-blame situations. For other-to-blame situations, the log odds ratio is basically independent of the verbal aggressiveness $\theta_p$.

The data set also contained information about the gender of the participants and notice that both in the regular Rasch model and the Rasch copula model the latent verbal aggression $\theta_p$ is regressed on gender (gender is dummy coded in a variable $G_p$ which takes a a value 1 for males and 0 for females). This resulted in the following model for the margins: latent variable $X_{pi} = \theta_p + g_p \gamma - \beta_i + \varepsilon_{pi}$, with observed binary variable $Y_{pi} = (X_{pi} > 0) = I(\varepsilon_{pi} > -\theta_p - g_p \gamma + \beta_i)$. Incorporating this additional covariate information only changes the formulation of the marginal model part, but not the model of the residual dependence structure as formulated by the copulas. Another advantage of the copula approach is that the effect can still be interpreted using the (log) odds ratio, as is the case for Rasch models with latent regression of $\theta_p$ (see De Boeck & Wilson, 2004, for more examples). Accordingly, this model can be labeled a ‘latent regression Rasch’ copula model. For the verbal aggression data, $\gamma$ the estimated effect of gender under the copula model is 0.30 on the logit scale, with a standard error of 0.19. From this we can conclude that males are not significantly more inclined to verbal aggression than females. The estimated odds ratio is 1.35 with a 95% confidence interval ranging from 0.93 to 1.96. Notice that this hardly differs with the estimate of gender under the regular Rasch model; here
Table 1.3: Rasch and Rasch Copula model, and a constrained Constant combination pairwise interaction model for the verbal aggression data.

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Rasch Estimate (SE)</th>
<th>Rasch copula Estimate (SE)</th>
<th>Restricted CCI Estimate (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-1.13 (0.17)</td>
<td>-1.15 (0.17)</td>
<td>-0.04 (0.19)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.47 (0.16)</td>
<td>-0.50 (0.16)</td>
<td>0.68 (0.19)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.01 (0.16)</td>
<td>-0.01 (0.16)</td>
<td>1.22 (0.20)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-1.66 (0.18)</td>
<td>-1.67 (0.18)</td>
<td>-0.59 (0.19)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.62 (0.18)</td>
<td>-0.64 (0.16)</td>
<td>0.58 (0.19)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.08 (0.16)</td>
<td>0.06 (0.16)</td>
<td>1.37 (0.21)</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.44 (0.16)</td>
<td>-0.42 (0.16)</td>
<td>0.15 (0.15)</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.78 (0.16)</td>
<td>0.66 (0.16)</td>
<td>1.50 (0.18)</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>1.61 (0.17)</td>
<td>1.53 (0.17)</td>
<td>2.44 (0.21)</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-0.99 (0.17)</td>
<td>-0.95 (0.16)</td>
<td>-0.09 (0.17)</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.44 (0.16)</td>
<td>0.37 (0.16)</td>
<td>1.59 (0.19)</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>1.13 (0.17)</td>
<td>1.09 (0.16)</td>
<td>2.41 (0.22)</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-1.13 (0.17)</td>
<td>-1.13 (0.17)</td>
<td>-0.04 (0.19)</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>-0.30 (0.16)</td>
<td>-0.33 (0.16)</td>
<td>0.88 (0.20)</td>
</tr>
<tr>
<td>$\beta_{15}$</td>
<td>0.96 (0.17)</td>
<td>0.93 (0.16)</td>
<td>2.25 (0.23)</td>
</tr>
<tr>
<td>$\beta_{16}$</td>
<td>-0.78 (0.16)</td>
<td>-0.79 (0.15)</td>
<td>0.40 (0.19)</td>
</tr>
<tr>
<td>$\beta_{17}$</td>
<td>0.15 (0.16)</td>
<td>0.10 (0.16)</td>
<td>1.45 (0.21)</td>
</tr>
<tr>
<td>$\beta_{18}$</td>
<td>1.57 (0.18)</td>
<td>1.52 (0.17)</td>
<td>3.02 (0.26)</td>
</tr>
<tr>
<td>$\beta_{19}$</td>
<td>0.30 (0.16)</td>
<td>0.27 (0.15)</td>
<td>0.97 (0.16)</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>1.59 (0.18)</td>
<td>1.46 (0.17)</td>
<td>2.42 (0.21)</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>3.06 (0.24)</td>
<td>2.91 (0.23)</td>
<td>4.04 (0.29)</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-0.62 (0.16)</td>
<td>-0.64 (0.16)</td>
<td>0.34 (0.17)</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>0.47 (0.16)</td>
<td>0.41 (0.15)</td>
<td>1.63 (0.19)</td>
</tr>
<tr>
<td>$\beta_{24}$</td>
<td>2.09 (0.19)</td>
<td>1.99 (0.18)</td>
<td>3.51 (0.26)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.35 (0.19)</td>
<td>0.30 (0.19)</td>
<td>0.19 (0.12)</td>
</tr>
<tr>
<td>$\sigma_\theta$</td>
<td>1.37 (0.07)</td>
<td>1.31 (0.07)</td>
<td>0.80 (0.06)</td>
</tr>
</tbody>
</table>

Frank: $J_1; \delta_1$, Frank: $J_2; \delta_2$, C-J: $J_3; \delta_3$, C-J: $J_4; \delta_4$, CCpwI: $J_1; \lambda_1$, CCpwI: $J_2; \lambda_2$, CCpwI: $J_3; \lambda_3$, CCpwI: $J_4; \lambda_4$

AIC: 8123, 7928, 7970

(standard error between parentheses)
the estimated odds ratio is 1.42 with a 95% confidence interval ranging from 0.98 to 2.06.

Again, for reasons of comparison we fitted the interaction models by Hoskens and De Boeck (1997) on this dataset. Given 6 items in a subset, this would lead to 15 \(= 6!/(4!2!)\) second order interaction terms, 20 \(= 6!/(3!3!)\) third order interaction terms, 15 \(= 6!/(4!2!)\) forth order interaction terms, and 6 \(= 6!/(5!1!)\) fifth order interaction terms. In other words, conditioning on a large amount of items can lead to a quite extensive model that might need to be restricted in some way to remain practical. To be able to compare the copula approach with an equally parsimonious model, two types of constrained models were chosen. In the first type of constrained model the interaction terms for an item subset were restricted to be all equal; in the second type of constrained model the second-order interactions for an item subset were restricted to be equal and all the other higher-order interactions were restricted to zero. The latter type of models will be referred to as pairwise interaction models. Both models result in one interaction parameter for each item subset, in correspondence with one copula parameter for each item subset in the copula approach. Both constant combination as dimension dependent interactions were considered. The pairwise interaction model with a constant combination interaction (CcpwI) for each item subset was the best fitting of the possible interaction models; its parameter estimates are displayed in Table 1.3. Notice, how the item parameter estimates are not comparable anymore to the regular Rasch, or Rasch copula model, and that they have slightly larger standard errors. The estimates of the interaction parameters are low, but all significant. The estimate of the standard deviation of the person ability distribution in Rasch copula model is considerably smaller than the equivalent estimate under the Rasch model, or the Rasch copula model. The estimated effect of gender \(\gamma\) under the model is 0.19 on the logit scale, with a standard error of 0.12. The estimated odds ratio is 1.21 with a 95% confidence interval ranging from 0.95 to 1.53. This is slightly smaller than the cor-
responding estimates for the Rasch and Rasch copula, but still within both their 95% confidence interval. All four interaction parameters, $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$, differ significantly from zero ($p < 0.01$). The overall goodness-of-fit as indicated by the AIC for this interaction model is 7970, an improvement compared to the regular Rasch model (AIC = 8123), but underperforming compared to the Rasch copula model (AIC = 7928). Overall it appears that large item subsets favour the copula approach with regard to both practical application and goodness-of-fit, whereas the conditional approach bumps into its boundaries.

1.6 Discussion

In this paper we have introduced the use of copulas for modeling residual dependencies. The basic idea is that we start out with a specific continuous latent response process (conditional on the random effects). Instead of assuming that given the random effects, these latent response processes are independent, they are coupled through a copula function. As a consequence, the discretized observed random variables (i.e., responses to the items) derived from coupled latent processes will exhibit residual dependency. In this way, the residual dependencies can be taken into account without altering the marginal item characteristic curves (which were typically Rasch models in our illustration of the approach).

It has been shown that the proposed copula model overcomes some of the problems in existing models for residual dependencies. The model has the property of reproducibility because the univariate marginals are Rasch models and the item parameters can still be interpreted as difficulty parameters. This reproducibility property (see also Liang, Zeger, & Qaqish, 1992), or what McCullagh calls ‘upward compatibility’ (McCullagh, 1989), means that we can remain within a well-known and conceptually attractive framework as offered by Rasch ICCs when applying a Rasch copula model.
The different dependence structures of the varying copula functions enable the modeling of static as well as dimensional residual dependencies. The general class of Archimedean copulas (with Frank and Cook-Johnson as special cases) is limited to the case in which the dependency is symmetric. Therefore, Archimedean copulas are less suited to model serial dependence as in learning phenomena.

Building on the basic model proposed in this paper, extensions to other more complicated models can be easily implemented. For instance, the univariate margins do not need to be restricted to the Rasch model: One can easily insert the 2PLM or 3PLM, and both within-subset constant and varying covariates can be easily added. Moreover, the method can be applied to response variables other than binary. Instead of dichotomizing the latent continuous response variable, as has been done throughout this paper, we can also use two or more cutoffs and consequently end up with a polytomous observed random variable (if the underlying distribution is logistic, the model for the categorical data is a version of the graded response model; Samejima, 1969). The copula approach can be easily extended beyond binary variables to polytomous variables, in which case the number of possible interval points \( d_{pk}^{(k)} \) in the distribution function of Equation 2 increases with increasing categories. An advantage of the copula approach is that even for these polytomous items, taking residual dependencies into account means that only a few parameters have to be added. However, note that the treatment of residual dependencies in polytomous models is much more complex than in the binary case. In the latter case one has only a single association or odds ratio (conditional on \( \theta_p \)), but when considering the residual dependencies between two items with \( M_1 \) and \( M_2 \) categories, respectively, there are \((M_1 - 1)(M_2 - 1)\) (global or local) odds ratios. For example, one might hypothesize that in some cases the residual association becomes stronger with higher categories. The potential of the copula approach to accommodate for these kinds of phenomena will be part of the focus of future work.
The flexibility of the copula IRT models brings up a model selection issue. In practice one has to choose the most adequate model for the marginal probabilities and find out which copula function shows the best representation of the residual dependence structure. Because the copula IRT model is fully likelihood based, this practical issue is resolved by using the traditional likelihood based model selection tools (e.g., likelihood ratio tests, AIC, BIC). The construction of specific diagnostic tools within the IRT framework for assessing the fit of both the margins and the copula may be desirable and is an open field where much work can be done.

This rather straightforward and parsimonious copula approach can make use of existing and known IRT models for the marginals, without changing anything about the interpretation of the marginal model part. Adding or removing a residual dependent item or item set also doesn't influence the marginal model part and the interpretation of the model parameters, as for instance would be the case when applying a conditional type of residual dependency model (in which the marginal probability model of an individual residual dependent item is dependent on other than the individual item parameter; see Equation 1.2). Obviously these characteristics make it a lot easier to explain these types of copula models to practitioners that have a basic and conceptual understanding of item response theory, but are less familiar with mathematical statistics and more complicated models that are available to approach the problem of local item dependency.

An alternative approach would be to fix the margins to be normally distributed instead of following a logistic distribution, leading to multivariate random effect probit models (Ashford & Sowden, 1970). This would allow the introduction of residual dependencies by estimating the variance-covariance matrix of the multivariate normal distribution. Note that it can be shown that this approach fits in the copula modeling framework: the multivariate normal distribution can be reformulated as a joint distribution constructed by means of a so-called Gaussian copula.
with univariate normally distributed margins. However, the computational burden of a multivariate probit model is quite high due to the multiple integrals that need to be approximated. In contrast, the copula approach presented in this manuscript avoids these extra integrals, and is not restricted to a linear dependency structure as implied by the use of the multivariate normal distribution. Thus the copula approach allows for a greater variety of models for the margins and models, i.e., copulas, for the dependency structure.

Although, based upon modeling possibilities, technical statistical properties and interpretability, the copula IRT model is clearly distinguished from other local item dependency models, general differences and similarities between the application of copula IRT models and their conditional (or random effects) counterparts are worthwhile to receive further investigation.
References


Rasch, G. (1960). Probabilistic models for some intelligence and achieve-
ment tests. Copenhagen: Danish Institute for Educational Research.


Abstract: A latent variable framework is presented to model the characteristic multivariate binary anomaly data as provided in some teratology studies. Building upon earlier studies in this domain and item response theory, we propose a model with a continuous underlying dimension. The conditional independence assumption, common to most latent variable models, is relaxed by means of copula functions. In addition, the distribution of the latent severity is modeled through a finite mixture of normals and can therefore be quite flexible. Covariates effects can be general or anomaly specific. It is shown that these elaborations of the traditional latent variable model provide extra insights for a data set from the Boston Anticonvulsant Teratogenesis study. The proposed framework allows to answer substantive questions concerning general and anomaly-specific exposure effects of covariates, interrelations between anomalies, and objective diagnostic measurement.
2.1 Introduction

Several external agents such as chemicals, hyperthermia, or viruses, can cause abnormalities during the development of a fetus. These external agents play an important role in the domain of teratology, where researchers are diagnosing birth defects and investigating the causal processes or etiology behind it (‘tera’ being Greek for monster). About 7 to 10% of all children will require extensive medical care to diagnose or treat a birth defect. Although significant progress has been made in identifying etiologic causes of some birth defects, approximately 65% still have no known or identifiable cause (O’Rahilly & Muller, 1996). This last fact can be ascribed to the inherent complexity of the domain: Few human teratogens have one single well-defined effect. Human teratogens rather generate a set of (possibly partly overlapping) birth defects out of a variety of deficiencies, malformations, and anomalies. Therefore, multiple outcomes have to be assessed in teratology studies.

The data set motivating this manuscript is the known Boston Anticonvulsant Teratogenesis study (BAT; Holmes et al., 2001) in which the effect of anticonvulsant drug use during pregnancy on a variety of birth defects is evaluated. Each of the 687 infants in the study was assessed on the presence or absence of several anomalies (resulting in a multivariate binary data string for each infant). Out of the 687 infants, there were 168 (24%) whose mothers took anticonvulsants during pregnancy, 73 (11%) whose mothers suffered from epileptic seizures but did not continue to take the anticonvulsant medication during pregnancy, and 446 (65%) infants functioning as control group (i.e., whose mothers did not suffer from epileptic seizures, nor took anticonvulsant drugs). The data set under consideration consists of 10 anomalies, going from facial anomalies (e.g., a depressed nasal bridge) to growth indicators (e.g., a small head) and other physical features such as hypoplastic toenails. Summary statistics and abbreviations are presented in Table 2.1. Although most of these anomalies are not of clinical importance them-
selves, they are of clinical interest due to their possible predictive power as markers of more serious, but not yet emerged anomalies and further developmental problems (Holmes et al., 1987). Note that the dataset is based on the same clinical study as analyzed by Legler and Ryan (1997) and that we consider the same 10 anomalies.

Table 2.1: Summary statistics for 10 anomalies from the BAT.

<table>
<thead>
<tr>
<th>Anomaly</th>
<th>Percentage ($n = 687$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hypoplastic toenails</td>
<td>0.179</td>
</tr>
<tr>
<td>2 Hypoplastic fingernails</td>
<td>0.108</td>
</tr>
<tr>
<td>3 Depressed nasal bridge</td>
<td>0.138</td>
</tr>
<tr>
<td>4 Tapered fingernails</td>
<td>0.108</td>
</tr>
<tr>
<td>5 Anteverted nostrils</td>
<td>0.102</td>
</tr>
<tr>
<td>6 Small head</td>
<td>0.112</td>
</tr>
<tr>
<td>7 Short birth length</td>
<td>0.058</td>
</tr>
<tr>
<td>8 Low birth weight</td>
<td>0.131</td>
</tr>
<tr>
<td>9 Clinodactyly of the hand</td>
<td>0.128</td>
</tr>
<tr>
<td>10 Broad nasal bridge</td>
<td>0.133</td>
</tr>
</tbody>
</table>

In our modeling approach, we work with latent variable models from item response theory (IRT) as they have been developed for problems in educational measurement. This is no coincidence, because also in teratology, one is confronted with a measurement problem. The anomalies may be considered as the items and the infants as the test takers. Some anomalies occur more frequently (i.e., the easy "items") while others are more rare (i.e., the difficult "items"). When modeling the person-specific probabilities of being affected by an anomaly, we will make use of an underlying latent trait, as in IRT, on which each of the infants (and anomalies) can be placed. The latent trait can be considered the severity of affect dimension. In what follows, we will discuss the three main research questions in a teratology study and link them with several aspects of our modeling approach.

First of all, there is a diagnostic goal involving a measurement problem. A quantification has to be made of how severely an infant is affected. This severity measure will serve as an objective assessment that can be of use for public health and welfare support, and policy makers.
Prior research (Legler & Ryan, 1997) introduced a latent variable model, in which the latent trait was assumed to follow a Poisson distribution in the population and reflects the unobserved severity at which an infant is affected. We will follow along these lines and propose a more general nonlinear mixed model alternative for this latent variable model. Borrowing from recent advances in linear mixed models in biostatistics (Verbeke & Lesaffre, 1996; Muthén et al., 2002), we will adopt a finite mixture distribution as population distribution for the unobserved severity to avoid possibly stringent parametric restrictions for the distribution of the random effect, i.e., the latent trait. Furthermore, such a mixture distribution may reveal hidden groups (latent classes) of infants in the data, yet still maintains quantitative inter-infant differences.

The second, etiological, goal is motivated by the fact that newborns of epileptic mothers have more chance of showing adverse birth outcomes. Interest goes out to differentiating two possible causes: Is this due to the mere presence of maternal epilepsy or an artefact of the related anti-convulsant drug therapy? In this regard, global tests and methods for multiple comparison between covariate groups have already been explored (Legler, Lefkopoulu, & Ryan, 1995; Lefkopoulu & Ryan, 1993). We propose to add these infant covariates to the nonlinear mixed model in a regression-like fashion to comply to the interest in exposure group effects. Furthermore, borrowing from the model tradition in educational measurement, we will make clear how to find out whether the covariate effect is anomaly specific or has a more general nature. In educational measurement, interest often goes out to determining whether the test items function in a similar way across known groups of pupils, for instance to verify whether a test is not biased towards a specific ethnic group. This is known as differential item functioning (DIF) (Holland & Wainer, 1993; De Boeck & Wilson, 2004). A similar reasoning can be followed in a teratology context where the exposure groups can be seen as the ethnic groups. In this way, one wants to find out whether the anomalies are differently affected in the different exposure groups.
As a third goal, there is interest in how the adverse birth outcomes relate to one another. Patterns of anomalies typical for severely affected infants can be helpful in identifying or defining a syndrome, as well as providing new insights in potential underlying biological processes. A statistical clustering technique (Everitt, 1993) might offer a means of analysis to answer this question. However, the same question can also be dealt with within a mixed model context. One source of relatedness between the anomalies is of course the unobserved severity as is reflected in the latent variable model. Furthermore, because anomalies may cluster together for various reasons, known and unknown, it can be expected that the commonly made assumption of conditional independence will be violated in this case study. Thus even after accounting for the dependency due to the latent severity, there still remain residual dependencies among specific anomalies. This type of situation has a long tradition in psychometrics and item response theory (IRT), and is commonly referred to as the problem of local item dependencies (Lord & Novick, 1968). These local dependencies lead to misspecified models when assuming conditional independence, and hence, alternative models (Hoskens & De Boeck, 1997; Wainer, Bradlow, & Wang, 2007; Tuerlinckx & De Boeck, 2004) have been proposed to counter the negative bias effects of the unaccounted residual dependencies. Recently, Braeken, Tuerlinckx, and De Boeck (2007) introduced a Rasch model extended with copula functions to accommodate for this type of local dependencies within an educational measurement context. The copula approach has a distinctive advantage over earlier methods that the main effects terms in the models are interpreted marginally, instead of conditionally on the added terms to account for the residual dependencies. We will focus on this fresh approach to tackle the issue of residual dependencies between specific anomalies.

In this paper, we will present an IRT model that combines all three modeling aspects outlined previously (random effect mixtures, anomaly specific covariate effects and copulas), and therefore has the advantage
that the three substantive goals put forward, can be reached simultaneously. The main contribution of this paper will be twofold. First, we will discuss a new copula-based method to take into account residual dependencies in IRT models. The copula approach was previously introduced by (Braeken et al., 2007) in the more restrictive context of a Rasch model with a normal random effects density. In this paper, the approach is extended to more general nonlinear mixed models. Second, we present the integration of various model developments arising from diverse fields such as biostatistics, educational measurement and IRT, and their application to the field of teratology which provides quite unique and characteristic datasets. As such, we will arrive at a general latent variable model to be used in teratology studies. With respect to this integration of modeling aspects, it should be noted that they are not independent of each other. For instance, failing to accommodate a violation of the conditional independence assumption may introduce bias into the model estimates for the covariates, leading to potentially erroneous conclusions.

The outline of the paper is as follows. In the following section, the specific latent variable model will be discussed in detail. First, the inclusion of covariates is described and then the finite mixture model for the latent trait will be outlined. In a third subsection, the recently developed copula approach to accommodate deviations from conditional independence will be explained. In the Application section, the BAT data will be analyzed within the proposed latent variable model framework. Simulation and data analytical results will be presented, before turning to a more general conclusion.

2.2 Method

2.2.1 Latent variable framework

Let \( Y_p = (Y_{p1}, Y_{p2}, \ldots, Y_{pl})^T \) represent the \((I \times 1)\) binary anomaly outcome vector (i.e., with \(I\) anomalies) for the \(p\)th infant. Assume \(Y_{pi}\) is a
manifest indicator of the event that some unobserved latent continuous variable $X_{pi}$ (following a logistic distribution with location 0 and scale 1) exceeds a threshold, which can be taken to be zero without loss of generality. The event $Y_{pi} = 1$ indicates the presence of the $i$th anomaly in infant $p$ and $Y_{pi} = 0$ indicates the absence. Specifically let

$$Y_{pi} = I(X_{pi} > 0), \text{ with } X_{pi} = \alpha_i(\theta_p - \beta_i) + \varepsilon_{pi},$$

such that

$$\Pr(Y_{pi} = y_{pi} | \theta_p) = \Pr(X_{pi} > 0 | \theta_p) = \Pr(\varepsilon_{pi} > -\alpha_i(\theta_p - \beta_i))$$

$$= 1 - F_{\varepsilon_{pi}}(-\alpha_i(\theta_p - \beta_i))$$

$$= \frac{\exp(y_{pi}\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))},$$

with $\theta_p$ being an infant-specific intercept representing the unobserved severity at which an infant $p$ has been affected (a larger value means that the infant is more severely affected), the parameters $\alpha_i$ and $\beta_i$ representing the discrimination and severity degree of the $i$th anomaly, respectively. The random variables $\varepsilon_{pi}$ are also called the latent residuals. This model is called the two-parameter logistic model (2PL) (Birnbaum, 1968) and is a standard model from item response theory.

Figure 2.1: Illustration of the model parameter interpretation.
Figure 2.1 illustrates the function of the parameters $\alpha_i$ and $\beta_i$ with respect to the logistic item characteristic curve $\text{Pr}(Y_{pi} = 1|\theta_p)$ over the range of the latent variable $\theta_p$. The larger the severity degree $\beta_i$ is, the lower the chance of having the $i$th anomaly for an infant with severity $\theta_p$. Notice that the value of $\beta_i$ is exactly the location on the latent scale $\theta_p$ where the probability of having anomaly $i$ is equal to a half (i.e., $\beta_i$ is the location of the logistic curve). The discrimination parameter $\alpha_i$ indicates the diagnostic value of anomaly $i$ for the underlying severity dimension. The larger $\alpha_i$, the better the discrimination between severely affected infants and less severely affected infants using anomaly $i$.

If it is assumed that the density of the latent variable $\theta_p$ is $h(\theta_p; \zeta)$, the resulting model for $\mathbf{Y}_p = (Y_{p1}, Y_{p2}, \ldots, Y_{pI})^T$, the joint outcome vector is a nonlinear mixed model (NLMM; see e.g. Molenberghs & Verbeke, 2005) in which $\theta_p$ is an infant-specific intercept. The dependency within the multivariate binary outcome vector $\mathbf{Y}_p$ of an infant $p$ and the heterogeneity among infants have been taken into account by assuming that $\theta_p$ is a random effect (Larsen, Petersen, Budtz-Jørgensen, & Endahl, 2000). The marginal probability of $\mathbf{Y}_p = \mathbf{y}_p$ is:

$$\text{Pr}(\mathbf{Y}_p = \mathbf{y}_p) = \int_{\theta_p} \text{Pr}(\mathbf{Y}_p = \mathbf{y}_p|\theta_p)h(\theta_p; \zeta)d\theta_p$$

$$= \int_{\theta_p} \prod_{i=1}^I \text{Pr}(Y_{pi} = y_{pi}|\theta_p)h(\theta_p; \zeta)d\theta_p$$

where $h(\theta_p; \zeta)$ is the density of the latent distribution of the severity of affect $\theta_p$, parameterized by the vector $\zeta$. We will show that this latent distribution can take a parametric specification (e.g., a normal distribution), or can be specified in a semi-parametric way (e.g., a finite mixture distribution). In the next few subsections we will go into more detail about the assumptions behind this basic model, highlight possible problems and outline model extensions to capture specific features of the BAT dataset.
2.2.2 Specification of covariates

With the availability of additional covariate information, represented by the infant-specific vector $Z_p = (Z_{p1}, Z_{p2}, \ldots, Z_{pJ})^T$, it is of interest to estimate the extent to which the severity of being affected $\theta_p$, is related to these covariates. From a clinical perspective, the inclusion of such covariate information in the model offers a way of exploring possible risk and protective factors for teratogenesis (i.e., the development of anomalies). For instance, it may be that infants exposed in utero to anticonvulsant drugs will have on average a higher severity of affect $\theta_p$ than infants not belonging to this group. Under the assumption that the distribution of $\theta_p$ is a standard normal (to fix the origin and scale of the latent variable), $h(\theta_p; \zeta) = \phi(\theta_p; \mu, \sigma^2)$, covariates can be taken into account as follows (Mislevy, 1987; De Boeck & Wilson, 2004):

$$
Pr(Y_p = y_p \mid Z_p) = \int_{\theta_p} Pr(Y_p = y_p \mid \theta_p) \phi \left( \theta_p; \sum_{j=1}^{J} Z_{pj} \lambda_j, 1 \right) d\theta_p. \quad (2.1)
$$

Model 2.1 can be seen as a decomposition of the severity of affect $\theta_p$ in a fixed part $\sum_{j=1}^{J} Z_{pj} \lambda_j$, representing the effect of the covariates, and a random part $\theta_p^* \sim N(0, 1)$, representing the residual latent infant-specific severity not accounted for by the known covariates:

$$
X_{pi} = \alpha_i(\theta_p^* - \beta_i + \sum_{j=1}^{J} Z_{pj} \lambda_j) + \varepsilon_{pi}. \quad (2.2)
$$

A regression weight $\lambda_j$ for categorical covariates like the exposure groups, can be interpreted as a main effect of the covariate $Z_{pj}$ on the latent severity $\theta_p$, corresponding to the average change on the logit scale one would expect when a newborn has the $j$th covariate or infant characteristic, relative to when that characteristic is absent, conditional upon $\theta_p^*$ and keeping all other covariates fixed. Hence, when the covariate $Z_{pj}$ indicates membership to an exposure group, $\lambda_j$ will indicate whether or not the exposure group differs in mean severity of affect from the control.
group. However to quantify the effect on a specific anomaly \(i\), the main effect is scaled following the importance of the anomaly for the severity \(\theta_p\), leading to a multiplicative change in the odds of having birth defect \(i\) by \(\exp(\alpha_i \lambda_j)\).

Possibly there is no general covariate effect, but is there still an effect for specific anomalies, or perhaps the general effect is inconsistent in direction for a specific anomaly. If one is interested in anomaly-specific effects of the covariates, one can opt for the following formulation:

\[
X_{pi} = \alpha_i (\theta_p^* - \beta_i + \sum_{j=1}^{J} Z_{pj} \lambda_{ji}) + \varepsilon_{pi}. \tag{2.3}
\]

A useful measure for anomaly-specific group differences is \(\exp(\alpha_i \lambda_{ji})\). The interpretation is similar as above, but now the covariate effect is really anomaly specific. This can be easily deduced by setting all \(\alpha_i\) at 1 and observing that the change in the odds are \(\exp(\lambda_{ji})\).

To identify anomaly-specific effects for a covariate \(Z_{pj}\), a step-by-step likelihood ratio test procedure can be followed where model 2.2 is compared to a combination of model 2.2 and model 2.3, in which for one anomaly a \(\lambda_{ji}\) is added to the general effect \(\lambda_j\), while for the other anomalies the anomaly-specific covariate effects are fixed to 0. This approach is successfully used in the study of differential item functioning in item response theory (see e.g., Holland & Wainer, 1993). While the detection and estimation of a general anomaly effect is based on the combined information present in the multivariate anomaly data, for a single anomaly-specific effect only an isolated and more limited part of the data can be used, hence resulting in less precision for the parameter \(\lambda_{ji}\). The technique described by Benjamini and Hochberg (1995) will be used to control the false discovery rate of an anomaly-specific effect in this multiple comparison testing situation.
2.2.3 Specification of the latent distribution

It is generally assumed that the random intercept follows a normal distribution with mean zero and variance 1. Alternatively, one could set the variance free and restrict one discrimination parameter $\alpha_i$ to 1 to keep the model identified. The restricted $\alpha_i$ would then function as reference to the other $\alpha_i$’s. Note however, that the first formulation will be used throughout the paper. The standard practice of choosing this specific parametric distribution form is mainly motivated by mathematical convenience and the fact that it is provided by most commercial statistical software packages. However, this assumption puts an important constraint on the shape of the distribution of the random effect and a misspecification of the distribution can lead to biased parameter estimates in the model (Agresti, Ohman, & Caffo, 2004; Hartford & Davidian, 2000; Neuhaus, Hauck, & Kalbfleisch, 1992).

In a teratogenesis study, the majority of the infants are at most mildly affected and thus scoring positively on only few outcomes. Given that the overall incidence of adverse birth defects is likely to be low, one can expect that the common normality assumption of the latent trait will be questionable. A positively skewed distribution for the severity dimension (i.e., the random intercept) might be more suitable for these data.

To accommodate possible problems due to misspecifying the latent distribution, we choose to rely on finite mixture distributions, which allow for a more flexible and data-driven latent distribution (Verbeke & Lesaffre, 1996; Aitkin, 1999). Let the density $h(\theta_p; \zeta)$ be a mixture density consisting of $G$ normal distributions, with mean vector $\mu = (\mu_1, \ldots, \mu_G)^T$, a common variance $\sigma^2 = 1$, and component weights $\pi_1, \ldots, \pi_g, \ldots, \pi_G$ such that $\sum_{g=1}^{G} \pi_g = 1$. Each component distribution of the mixture will cover a local area of the true distribution, hereby enabling the modeling of a quite complex distribution. Extending our model with a finite mixture distribution of normals for the random effect
\( \theta_p \) results in the following marginal probability of the outcome vector:

\[
\Pr(Y_p = y_p) = \sum_{g=1}^{G} \pi_g \int_{\theta_{pg}} \Pr(Y_p = y_p | \theta_p) \phi(\theta_{pg}; \mu_g, 1) d\theta_{pg},
\]

(2.4)

with \( \phi(\theta_{pg}; \mu_g, 1) \) denoting the normal density with mean \( \mu_g \) and variance 1 for the \( g^{th} \) component of the latent trait \( \theta_p \). For identification reasons, the first component mean \( \mu_1 \) is restricted to zero so that the origin of the latent variable remains fixed. In the presence of covariates as in model 2.1, it will be assumed that it is the random part, the residual infant severity \( \theta^*_p \) that follows a mixture distribution. Hence, the covariate effects \( \lambda_j \) are not allowed to vary over the \( G \) components.

Besides an increased flexibility, a finite mixture distribution also allows to investigate and reveal possible hidden grouping in the data. In such a case, infants can be ascribed to the component for which they have the highest posterior probability to belong to, resulting in a classification of infants over components. By characterizing the typical infants belonging to a specific component, researchers might attempt to interpret the underlying reason of the component separation and hereby possibly gaining new unexpected insights in the data and the teratogenic processes involved.

### 2.2.4 Specification of the joint distribution

Traditionally, the joint distribution of the outcome vector \( Y_p \) in latent variable models is formulated under conditional independence:

\[
\Pr(Y_p = y_p | \theta_p) = \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi} | \theta_p),
\]

(2.5)

The anomaly outcomes are assumed to be independent realizations conditional upon the latent severity \( \theta_p \) (in addition to the covariate effects). Hence, the dependency in the data is ascribed to the fact that repeated measurements were taken from the same infant \( p \). Taking the perspec-
tive of the continuous latent variable $X_{pi}$, the conditional independence assumption can be formulated as having uncorrelated error terms $\varepsilon_{pi}$ (over anomalies).

This assumption allows for a mathematical convenient way of modeling the joint probability of the anomaly outcome vector, but might not always be plausible. In a teratology context for instance, it is likely that even after accounting for the common dependence due to an infant’s severity of affect, specific anomalies will show some extra association or residual dependency. This can be the case when they have a similar origin in common, like for instance the same body part, a genetic link or some kind of environmental factor. For instance, in the BAT the growth indicators can be expected to show this type of extra association.

Failing to correctly account for the associations in the data, implies a misspecified model. The most prominent effects of not acknowledging the redundant information in the data are that the standard error of the estimated infants severity $\hat{\theta}_p$ is underestimated (resulting in an overly optimistic measurement instrument) (Junker, 1991), and that discrimination parameters are overestimated or even diverge to an unrealistic large value (hence anomalies are falsely considered to be better indicators of the underlying severity dimension) (Masters, 1988; Tuerlinckx & De Boeck, 2001a). More in general, when ignoring violations of the conditional independence assumption, both fixed and random effects, their estimated standard errors, and associated inferences are potentially biased (Chen & Thissen, 1997), a phenomenon that is well-known in psychometrics. In this manuscript we propose the use of copula functions to account for residual dependencies in the data by formulating a more proper joint distribution than in Equation 2.5, hereby correcting for the abovementioned bias problems.

An $I$-dimensional copula is a function $C : [0, 1]^I \rightarrow [0, 1]$ with the following properties: (1) For every vector $u \in [0, 1]^I$, $C(u)$ is increasing in each component $u_i$ with $i \in \{1, \ldots, I\}$, (2) and $C(u) = 0$ if at least
one coordinate of the vector is 0 and \( C(u) = u_i \) if all the coordinates of
the vector are equal to one except the \( i \)-th one. (3) For every \( \mathbf{a}, \mathbf{b} \in [0, 1]^I \)
with \( \forall i, a_i \leq b_i \), given a hypercube \( \mathbf{B} = [\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \ldots \times [a_I, b_I] \)
whose vertices lie in the domain of \( C \), and with \( V_C(\mathbf{B}) \geq 0 \). The volume
\( V_C(\mathbf{B}) \) is defined as:

\[
V_C(\mathbf{B}) = \sum_{k_1=1}^{2} \ldots \sum_{k_I=1}^{2} (-1)^{k_1+\ldots+k_I} C(d_{1}^{(k_1)}, \ldots, d_{I}^{(k_I)})
\]

where \( d_{i}^{(1)} = a_i \) and \( d_{i}^{(2)} = b_i \) \( (i = 1, \ldots, I) \).

An important theorem, that makes clear the application possibilities in multivariate modeling, is due to Sklar (Sklar, 1959) (for a thorough overview of copula theory, see Nelsen, 1998, and Joe, 1997). For
any \( I \)-dimensional distribution function \( F_X \) with univariate margins
\( F_{X_1}, \ldots, F_{X_I} \) there exists a copula function \( C \) such that this multivariate distribution \( F_X \) can be represented as a function of its margins through this copula: \( F_X = C (F_{X_1}, \ldots, F_{X_I}) \). In the continuous case, this copula reformulation of an existing multivariate distribution is unique, for discrete random variables the copula \( C \) is uniquely deter-
mined on \( \text{Ran}(F_{X_1}) \times \ldots \times \text{Ran}(F_{X_I}) \), the Cartesian product of ranges
of the margins.

Furthermore, the converse of the theorem also holds. Thus given a
set of univariate margins \( F_{X_1}, \ldots, F_{X_I} \), a multivariate (i.e., joint) dis-
tribution \( F_X \) can be constructed by adopting a specific copula \( C \), such
that \( F_X = C (F_{X_1}, \ldots, F_{X_I}) \). This part of the theorem often finds appli-
cations in econometrics and actuarial sciences. Given the second prop-
erty in the copula definition, it can easily be deduced that the univari-
ate marginal distribution for \( X_i \) equals \( C (1, \ldots, 1, F_{X_i}, 1, \ldots, 1) = F_{X_i} \).
Hence, in this way an association between the \( I \) random variables is
allowed through the copula \( C \), while still preserving the original univari-
ate margins we started from. In the discrete case, the joint probability
\( \Pr(\mathbf{X}) \) of the multivariate distribution \( F_X = C (F_{X_1}, \ldots, F_{X_I}) \) can be
A latent variable framework for teratology studies derived using a recursive formula based upon quadrant probabilities (see e.g., property 3 above or Mood, Graybill, & Boes, 1974). The copula approach introduced by Braeken et al. (2007), for a Rasch model in an educational measurement context, makes use of exactly this theorem to formulate a more proper joint distribution that accounts for the association structure in the data. We extend and generalize this approach here to the setting of nonlinear mixed models.

To make the ideas explicit, consider \( S \) disjoint subsets of the set of anomalies \( \{1, \ldots, I\} \) denoted as \( J_1, \ldots, J_S \), where \( J_s \) has cardinality \( I_s \). While all anomalies are indicators of the same latent construct, anomalies within a subset \( J_s \) show some residual dependency. The different subsets are assumed independent given the latent variable \( \theta_p \), and the anomaly outcomes in a subset \( Y_p^{(s)} \) are assumed exchangeable. The model we propose is then

\[
\Pr(Y_p = y_p|\theta_p) = \prod_{s=1}^{S} \Pr_s(Y_p^{(s)}|\theta_p),
\]

where \( \Pr_s(Y_p^{(s)}|\theta_p) \) is the joint probability of the anomaly outcomes \( Y_p^{(s)} \) in subset \( J_s \).

If the cardinality of subset \( J_s \) is larger than one \( (I_s > 1) \), \( \Pr_s(Y_p^{(s)}|\theta_p) \) is the joint probability of the outcome vector of subset \( s \) evaluated from the copula function \( C_s \) for all \( Y_{pi} \in J_s \) under the following formula:

\[
\Pr_s(Y_{p1} = y_{p1}, \ldots, Y_{pI_s} = y_{pI_s}|\theta_p) = \\
\Pr_s(d^{(1)}_{p1} < Y_{p1} \leq d^{(2)}_{p1}, \ldots, d^{(1)}_{pI_s} < Y_{pI_s} \leq d^{(2)}_{pI_s}|\theta_p) \\
= \sum_{k_1=1}^{2} \cdots \sum_{k_{I_s}=1}^{2} (-1)^{k_1+\ldots+k_{I_s}} C_s\left(F_{Y_{p1}|\theta_p(d^{(k_1)}_{p1})}, \ldots, F_{Y_{pI_s}|\theta_p(d^{(k_{I_s})}_{pI_s})}\right),
\]

where \( d^{(1)}_{pi} \) and \( d^{(2)}_{pi} \) stem from the definition of the distribution functions.
\[ F_{Y_p|\theta_p}(y_{pi}) \]

\[
F_{Y_p|\theta_p}(y_{pi}) = \begin{cases} 
0 & \text{for } y_{pi} < 0 \\
\Pr(Y_{pi} = 0|\theta_p) & \text{for } y_{pi} = 0 \\
1 & \text{for } y_{pi} = 1 
\end{cases}
\]

such that for \( y_{pi} = 1 \), \( d_{pi}^{(1)} = 0 \) and \( d_{pi}^{(2)} = \infty \), and for \( y_{pi} = 0 \), \( d_{pi}^{(1)} = -\infty \) and \( d_{pi}^{(2)} = 0 \).

The regular conditional independence model arises as a special case when \( S = 1 \) and \( C_s \) is the independence copula \( \Pi \left( F_{Y_{p1}|\theta_p}, \ldots, F_{Y_{pI}|\theta_p} \right) = \prod_{i=1}^{I} F_{Y_{pi}|\theta_p} \). In general, a copula function \( C \) is typically a member of a more general family of copulas that are capable of inducing a particular type of association structure. In this study, three copula functions, presented below, will be used to model residual dependencies within subsets of anomalies. These three copulas were chosen because they have a simple functional form, with only a single parameter \( (\delta_s) \) capturing the whole range of positive association (from independency to absolute positive dependency), known multivariate extensions, and together comprise a broad range of possible association structures for the residual dependencies.

1. **Frank copula** (Frank, 1979)

\[
C_s \left( F_{Y_{p1}|\theta_p}, \ldots, F_{Y_{pI}|\theta_p} \right) = \frac{-1}{\delta} \log \left( 1 - \frac{\prod_{i=1}^{I_s} (1 - \exp(-\delta F_{Y_{pi}|\theta_p}))}{\prod_{i=1}^{I_s-1} (1 - \exp(-\delta))} \right)
\]

For \( I_s = 2 \), \( \delta \in \mathbb{R}/\{0\} \), negative values indicate negative association, positive values positive association; for \( I_s > 2 \) the parameter range is constrained to the range of positive association. If \( \delta \to 0 \) then \( C \to \Pi \) (i.e., independency), and if \( \delta \to \infty \), \( C \to M \) (i.e., absolute positive dependency).

The copula characteristic association structure can be illustrated by computing the log odds ratio (conditional on \( \theta_p \)) for two anomalies,
A latent variable framework for teratology studies

1 and 2, in the copula $C_s$

$$\log(\text{OR}(\theta_p)) = \log \left( \frac{\Pr_s(Y_{p1} = 1, Y_{p2} = 1 | \theta_p) \Pr_s(Y_{p1} = 0, Y_{p2} = 0 | \theta_p)}{\Pr_s(Y_{p1} = 1, Y_{p2} = 0 | \theta_p) \Pr_s(Y_{p1} = 0, Y_{p2} = 1 | \theta_p)} \right)$$

$$= \log \left( \frac{\left( 1 - F_{Y_{p1} | \theta_p}(0 | \theta_p) - F_{Y_{p2} | \theta_p}(0 | \theta_p) + C \right) C}{\left( F_{Y_{p2} | \theta_p}(0 | \theta_p) - C \right) \left( F_{Y_{p1} | \theta_p}(0 | \theta_p) - C \right)} \right),$$

with $C = C_s \left( F_{Y_{p1} | \theta_p}(0 | \theta_p), F_{Y_{p2} | \theta_p}(0 | \theta_p) \right)$. This conditional log odds ratio gives an indication of what the odds are that both anomalies cooccur or are jointly absent. A value near 0 means that the two anomalies are almost independent (given $\theta_p$); A large negative value means that if one of both anomalies is present, the other is almost certainly not; a large positive value means that if one of both is present, the other is almost sure to be present as well. This odds ratio was calculated for several values of its association parameter $\delta$, and for $\theta_p$ ranging from $-2$ to 2.

For clarity, $\beta_1$ and $\beta_2$ were fixed to 0, and $\alpha_1$ and $\alpha_2$ were fixed to 1, such that the log odds ratio was only a function of the copula’s association parameter $\delta_s$ and the latent trait $\theta_p$. The result is shown in the left panel of Figure 2.2. Notice that when the value of the copula parameter $\delta$ rises, the log odds ratio also increases, indicating $\delta$’s function as an association measure; and that the odds ratio remains relatively constant over $\theta_p$ with a small peak around $\beta$ the location of the margins.


$$C_s \left( F_{Y_{p1} | \theta_p}, \ldots, F_{Y_{pIs} | \theta_p} \right) = \left( \sum_{i=1}^{I_s} F_{Y_{pi} | \theta_p}^{\gamma} - I_s + 1 \right)^{-\frac{1}{\gamma}}$$

With $\delta > 0$, higher values of the copula parameter indicate stronger positive association. If $\delta \to 0$ then $C \to \Pi$, and if $\delta \to \infty$ then $C \to M$. A similar characterization of the association structure this copula implies, is shown in the middle panel of Figure 2.2. Notice that the odds ratio increases with increasing $\theta_p$.
3. **Gumbel-Hougaard copula** (Gumbel, 1960; Hougaard, 1986)

\[
C \left( F_{Y_{p1}|\theta_p}, \ldots, F_{Y_{pl|\theta_p}} \right) = \exp \left( - \left[ \sum_{i=1}^{l} \left( - \log(F_{Y_{pi}|\theta_p}) \right)^{\delta} \right]^{\frac{1}{\delta}} \right)
\]

\(\delta > 1\), higher values of the copula parameter indicate stronger positive association. If \(\delta \rightarrow 1\) then \(C \rightarrow \Pi\), and if \(\delta \rightarrow \infty\) then \(C \rightarrow M\). The odds ratio characterization of the LID structure this copula implies, is shown in the right panel of Figure 2.2. Notice that the odds ratio decreases with increasing \(\theta_p\).

Figure 2.2: *Log odds ratio plots for 3 Archimedean copulas in the latent variable model.*

Compared to most alternative modeling approaches for residual dependencies (see e.g., Hoskens & De Boeck, 1997; Wainer et al., 2007; Tuerlinckx & De Boeck, 2004), the advantage of the copula formulation is that the parametrisation of each univariate margin (i.e., anomaly) preserves its natural interpretation. This is called the reproducibility property (Ip, 2002; Liang, Zeger, & Qaqish, 1992), which allows for changes to the joint model without having to leave the attractive modeling framework as described by the latent variable model for the univariate margins proposed in Equation 2.1. A wide variety of association structures can be accommodated, and this by means of applying different copula func-
tions for the anomaly subsets. One can easily see that another approach is taken by alternative models which does not retain reproducibility. For instance, the testlet models of Wainer et al. (2007) add an extra subset specific latent trait $\zeta_{ps}$ to the formulation of the margins to capture the residual dependency, such that $X_{pi} = \alpha_i(\theta_p + \zeta_{ps} - \beta_i) + \epsilon_{pi}$ and that the severity degree of an anomaly $\beta_i$ is not anymore the location on the latent severity dimension $\theta_p$ where the probability of having the anomaly $i$ is equal to a half. In other words, this approach changes the individual model characteristics of an anomaly to model the joint behavior of an anomaly set, whereas the copulas take a more natural approach by directly adapting the joint model and leaving the marginal model part intact.

In sum, the copulas are introduced to capture dependence between anomalies over and above the association that can be explained by the common severity of affect of an infant. Besides its use as a vehicle to take into account violations of the conditional independency assumption, copulas can also be used as an indicator of association between specific anomalies and hereby provide additional information about potential common underlying pathways.

2.3 Model inference

The parameters to be estimated can be divided into three groups: The anomaly specific parameters and covariate regression weights ($\alpha_i, \beta_i$ and $\lambda_j$ or $\lambda_{ji}$ where $i = 1, \ldots, I$ and $j = 1, \ldots, J$), the copula dependency parameters ($\delta_1, \ldots, \delta_S$, where $S$ is the number of disjoint anomaly subsets) and the parameters of the latent distribution ($\sigma^2, \mu_2, \ldots, \mu_G$ and $\pi_1, \ldots, \pi_{G-1}$). Note that possibly some of the $\delta$ parameters drop out because not all subsets will be modeled by means of a copula (i.e., the independence subsets). Parameters are collected into vectors $\mathbf{\beta}, \mathbf{\alpha}, \mathbf{\lambda}, \mathbf{\delta}, \mathbf{\mu}$ and $\mathbf{\pi}$. The marginal likelihood under the full model is then:

$$L(\mathbf{\beta}, \mathbf{\alpha}, \mathbf{\lambda}, \mathbf{\delta}, \mathbf{\mu}, \mathbf{\pi}, \sigma^2) = \prod_{p=1}^{P} \sum_{g=1}^{G} \pi_g \int \prod_{s=1}^{S} \Pr_{s}(Y_p^{(s)}|\theta_{pg}) \phi(\theta_{pg}; \mu_g, \sigma^2) d\theta_{pg}.$$
All analyses reported in this paper are carried out following full information marginal maximum likelihood (FIML). The intractable integral with respect to the distribution of the latent variable $\theta_p$ is approximated using a non-adaptive Gauss-Hermite quadrature (with 20 nodes). If $G = 1$, the applied optimization algorithm is a quasi-Newton method. In case $G > 1$, the optimization is performed using a generalized Expectation-Maximization (EM) algorithm with a quasi-newton iteration to solve the M-step (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997). Multiple starting points were used for the EM algorithm to safeguard against the danger of local minima. Making use of Bayes theorem, empirical Bayes estimates of the latent severity $\theta_p$ can be computed (Bock & Lieberman, 1970; Baker & Kim, 2004) as

$$\hat{\theta}_p = \sum_{g=1}^{G} \frac{\pi_g}{\int_{\theta_{pg}} \prod_{s=1}^{S} \Pr_s(Y^{(s)}_p | \theta_{pg}) \phi(\theta_{pg}; \mu_g, \sigma^2) d\theta_{pg}} \int_{\theta_{pg}} \prod_{s=1}^{S} \Pr_s(Y^{(s)}_p | \theta_{pg}) \phi(\theta_{pg}; \mu_g, \sigma^2) d\theta_{pg}.$$ 

in which all other model parameters are set at their estimates obtained in the previous optimization. With respect to model checking and selection, the tools usually applied in the context of mixed (e.g., Wald, score or likelihood ratio tests) and mixture models (e.g., information criteria like the BIC) are also applicable here. The estimation procedures were implemented in Matlab (R2007, The Mathworks).

### 2.4 Application: BAT study

As a start of the analysis, the standard latent variable model in Equation 2.1 was fitted to the BAT data. Although the model reached convergence, the estimates and standard errors of the discrimination parameter $\alpha_i$ of anomalies 6, 7, and 8 take relatively large values, and the severity degree $\beta_i$ of anomalies 3 and 10 are also affected (see Model 1 in Table 2.2). These observations indicate, as expected, that the model is hampered by model specification issues (e.g., important missing covariates and/or the presence of residual dependency).
2.4.1 Model specification

A first explorative attempt to deal with the specification issues, was to extend the standard model with a finite mixture distribution for the severity of affect $\theta_p$. This resulted in an improved fit and the previously mentioned estimation problems occur to a lesser degree (see Model 2 in Table 2.2). The finite mixture consists of 2 components with means 0 and $-10.96$, common variance equal to 1, and component weights 0.54 and 0.46, respectively. This configuration leads to a clear bimodality in the distribution of an infant’s severity of affect $\theta_p$. After classifying the infants into these two components based upon their maximum posterior component probability, the second component can be characterized as gathering unaffected infants (i.e., all infants without any anomaly and a few infants with a single anomaly present, being either a depressed or a broad nasal bridge), while the first component gathers the affected infants, which contains the majority (68%) of infants belonging to the anticonvulsant-exposure group. There is no difference between the two components with respect to seizure history and gender.

When extra relevant covariate information is available from the start, we can directly include it into the model, as is done in Equation 2.1. Adding the three covariates (i.e., exposed in utero to anticonvulsants, gender, and maternal seizure-history) to both the standard and the finite-mixture model further improves the model fit. Of all four fitted models, the model without a mixture but with covariates is preferred based upon the BIC. Looking in detail to the mixture model with covariates (Model 4), it can be seen why this is the case: The second component’s mean increased to $-3.76$ and the difference in presence of treatment groups leveled out. This is an example of a finite mixture model that picks up an unmodeled covariate effect (and where the need for more than one component disappears when accounting for the relevant covariate). The second component remained the class of the hardly affected infants.

The extreme values of some discrimination parameters indicated the
Table 2.2: Models for the BAT data.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No mixture</td>
<td>Mixture</td>
<td>No mixture</td>
<td>Mixture</td>
<td>No mixture</td>
</tr>
<tr>
<td>No covariates</td>
<td>No covariates</td>
<td>Covariates</td>
<td>Covariates</td>
<td>Covariates</td>
</tr>
<tr>
<td>No copulas</td>
<td>No copulas</td>
<td>No copulas</td>
<td>No copulas</td>
<td>Copulas</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>$2.46 (0.52)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$2.43 (0.40)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$4.97 (2.02)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$2.54 (0.42)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>$3.37 (0.81)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>$1.32 (0.08)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>$1.65 (0.09)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>$1.30 (0.09)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>$3.14 (0.72)$</td>
<td></td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>$18.25 (18.71)$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>$0.67 (0.16)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>$1.02 (0.21)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>$0.38 (0.16)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>$0.96 (0.20)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>$0.70 (0.19)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>$3.95 (0.93)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>$4.85 (1.08)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td>$3.01 (0.56)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_9$</td>
<td>$0.66 (0.17)$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>$0.10 (0.11)$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_2$</td>
<td>$-10.96 (2.51)$</td>
<td>$-10.96 (2.51)$</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>$0.46 (0.05)$</td>
<td>$0.46 (0.05)$</td>
</tr>
<tr>
<td>$\lambda_{anticonvulsants}$</td>
<td>$1.19 (0.14)$</td>
<td>$1.19 (0.14)$</td>
</tr>
<tr>
<td>$\lambda_{gender}$</td>
<td>$0.00 (0.11)$</td>
<td>$0.00 (0.22)$</td>
</tr>
<tr>
<td>$\lambda_{seizurehistory}$</td>
<td>$0.46 (0.19)$</td>
<td>$0.94 (0.40)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>$1.42 (0.15)$</td>
<td></td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>$13.51 (1.99)$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>BIC</th>
<th>-2LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>4765</td>
<td>2317</td>
</tr>
<tr>
<td>22</td>
<td>4726</td>
<td>2291</td>
</tr>
<tr>
<td>23</td>
<td>4722</td>
<td>2286</td>
</tr>
<tr>
<td>25</td>
<td>4725</td>
<td>2281</td>
</tr>
<tr>
<td>25</td>
<td>4469</td>
<td>2153</td>
</tr>
</tbody>
</table>
potential presence of residual dependencies. To further explore and detect violations of conditional independence, various diagnostic tools have been developed (Holland & Rosenbaum, 1986; Chen & Thissen, 1997; Yen, 1984; Tate, 2003). In this manuscript the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) will be used to identify pairs of anomaly outcomes that exhibit residual dependency beyond what is accounted for by a single latent trait. The main idea is to test for an equal odds ratio deviating from one (i.e., independence) across different groups of infants. Because under the conditional independence assumption, the items are independent given the latent trait $\theta_p$, the group division in the MH procedure is based upon a proxy for the severity of affect (e.g., the sum score over anomalies or a preliminary model-based estimate $\hat{\theta}_p$). The MH test statistic will give an indication of the presence of residual dependency between the anomalies while controlling for the latent severity $\theta_p$. Because we regard the MH test procedure in this context mainly as an exploratory and crude screening tool to detect residual dependencies and not as a formal statistical test, we will not report $p$-values. Two subsets of anomalies were demarcated (assisted by a hierarchical clustering of the MH statistics), the anomaly pair $J_1 = \{1, 2\}$ (MH = 4.6), and the triplet $J_2 = \{6, 7, 8\}$ (all 3 MH > 6.9). Both subsets appear to make sense from a substantive perspective as well, subset $J_1$ refers to hypoplastic nails and subset $J_2$ gathers the growth indicators.

The latent variable model with covariates was extended with copulas to account for the presence of the anomaly subsets. Selecting the best-fitting combination of possible copula functions, resulted in Gumbel-Hougaard copula for subset $J_1$, and Frank copula for subset $J_2$. Results are shown in model 5 of Table 2.2. The chances of both finger- and toenail hypoplasia to cooccur or be jointly absent are higher than can be expected based upon an infant’s severity of affect $\theta_p$ ($\hat{\theta}_1 = 1.42, p < 0.0001$), and infants who are less severely affected (low $\theta_p$) show a higher tendency to have either both anomalies present/absent, than only one of both (see Figure 2.3). The growth indicators also show a very large
tendency to cooccur or jointly be absent ($\hat{\delta}_2 = 13.51, p < 0.0001$), with a slightly higher tendency for more severely affected infants (high $\theta_p$) (see Figure 2.3). Note that in Figure 2.3 only the part of the severity of affect dimension $\theta_p$ is shown, on which infants are positioned in the current dataset. This explains the differences with the previously shown example curves in Figure 2.2. A two-component mixture model with covariates and copulas resulted in approximately the same minus log-likelihood value as for the one-component alternative. Because it does not provide new information, detailed results on this mixture solution are omitted.

Figure 2.3: *Log odds ratio plots for the two copula subsets in the final model.*

![Log odds ratio plots](image)

To further verify possible model misspecification, an alternative model with a bidimensional latent trait (Bock, Gibbons, & Muraki, 1988) $\theta_p = \{\theta_{p1}, \theta_{p2}\}$ was fitted in Mplus (Muthén & Muthén, 1998-2007). This model includes a double set of discrimination parameters $\alpha_i = \{\alpha_{i1}, \alpha_{i2}\}$, one for each latent trait, that can be considered as weighting the importance of anomaly $i$ for each dimension of the latent severity of affect (similar to factor loadings in factor analysis). Although the bidimensional latent trait model has a better fit than the unidimensional model (LL = 2286 vs 2149), it performs similarly to our proposed copula extended unidimensional model (LL = 2149 vs 2153). Further-
more, the copula model is more parsimonious (25 vs 37 parameters) as it is reflected in the lower BIC (4469 vs 4539), and computationally more efficient as it only requires one integral to be approximated. Multi-dimensional latent traits can lead to prohibitive computation and given the relatively small amount of anomalies, the number of dimensions that can be efficiently estimated, is limited). Given the above statistical reasons, and the clinical interest and requirement to have one summary composite measure for severity, our use of an unidimensional trait to capture an infant’s latent severity is warranted.

Yet another model specification issue that has to be assessed is the impact of the covariates on the various anomalies. After running the sequential likelihood ratio procedure combined with the Benjamini-Hochberg technique for each of the $j$ covariates, none of the anomaly-specific covariate effects $\lambda_{ji}$ were retained additionally on the general covariate effects $\lambda_j$. Therefore it can be concluded that model 2.2 with a general main effect $\lambda_j$ for each covariate $j$ suffices for the BAT study.

These various model assessment results suggest that the variation and association structure in the data are adequately accounted for by capturing the residual dependencies and the covariate information in the model. Consequently, we will go into a bit more detail with regard to the chosen copula model and its results in the next subsections.

### 2.4.2 Simulation study

We observed that the all-zero anomaly patterns are prominently present in the BAT study and that the proportions of the 10 birth defects (see Table 2.1) are asymmetric, with the anomalies being much more commonly absent than present. Consequently, severity degree of most anomalies, that is their location on the latent dimension, will be rather high. Hence, the scale constructed based upon this set of anomalies will differentiate well between severely affected infants (i.e., with a higher sum score over anomalies), but will show less differentiation between infants that are hardly showing any sign of being affected. This is because not
much information is present about the latter area of the latent severity dimension. Conversely, more precise estimates can be made for anomaly parameters at a location where lot of infants are positioned. In our case, the focus of the anomaly set is at the upper part of the severity dimension, which serves the desired purpose of the diagnostic assessment, and the sample of infants is mainly located at the lower part, which resembles the population distribution. Thus, for the mixture distribution the same principle will hold, when a component mean is located at the upper part of the severity scale, it will be more easily detected than when it is at the lower part of the scale, due to the asymmetry in the amount of information present at each side.

A small recovery simulation study for the most complex model under consideration was set up to illustrate that the proposed latent variable framework works even in this teratology-specific context which explicitly poses a challenge to model optimization and estimation. Data were generated following a 2PL copula model with one general covariate and an underlying mixture distribution for the latent trait. The discriminations $\alpha_i$ of the anomalies were independently drawn from an uniform distribution on the interval $[0,4,1.6]$, the severity degree $\beta_i$ of an anomaly was drawn from a normal distribution with mean 2, variance 1, and truncated at zero to span the positive domain. The anomaly subsets were taken to be $J_1 = \{1,2\}$, in which the residual dependency originates from a Gumbel-Hougaard copula with parameter $\delta_1 = 1.3$, and $J_2 = \{6,7,8\}$, in which the residual dependency originates from a Frank copula with parameter $\delta_2 = 10$. This will result in diagnostic tests based upon anomalies with similar behavior to the set used in the BAT study. The known covariate was drawn from a Bernoulli distribution with mean 0.5, with $\lambda_j$ set at 1.3. Two conditions were created to induce a hidden grouping of infants. For 30% of the infants their residual severity of affect $\theta_p$ was drawn from a standard normal distribution ($\mu_1 = 0, \sigma^2 = 1$), whereas for the other 70% this was from a normal distribution with mean $\mu_2$ and variance $\sigma^2$ equal to one. In the first condition, the mean of this
second component is equal to 2, resembling to a situation with a hidden group that is relatively severely affected. In the second condition, the mean of the second component is equal to \(-2\), comparable to situation with a large hidden group of infants that is almost not affected.

Table 2.3: Recovery results for the latent variable framework in a teratology context.

<table>
<thead>
<tr>
<th>Condition</th>
<th>(\mu_2 = 2)</th>
<th>(\mu_2 = -2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Signed Bias</td>
<td>Absolute Bias</td>
</tr>
<tr>
<td>(\alpha_i) non-subset</td>
<td>0.029</td>
<td>0.156</td>
</tr>
<tr>
<td>(\hat{\alpha}_i) (J_1{1, 2})</td>
<td>0.020</td>
<td>0.159</td>
</tr>
<tr>
<td>(\hat{\alpha}_i) (J_2{6, 7, 8})</td>
<td>0.001</td>
<td>0.166</td>
</tr>
<tr>
<td>(\beta_i) non-subset</td>
<td>-0.151</td>
<td>0.385</td>
</tr>
<tr>
<td>(\hat{\beta}_i) (J_1{1, 2})</td>
<td>-0.156</td>
<td>0.378</td>
</tr>
<tr>
<td>(\hat{\beta}_i) (J_2{6, 7, 8})</td>
<td>-0.131</td>
<td>0.366</td>
</tr>
<tr>
<td>(\delta_1) (J_1{1, 2})</td>
<td>-0.011</td>
<td>0.076</td>
</tr>
<tr>
<td>(\delta_2) (J_2{6, 7, 8})</td>
<td>0.017</td>
<td>0.884</td>
</tr>
<tr>
<td>(\lambda_j)</td>
<td>-0.044</td>
<td>0.148</td>
</tr>
<tr>
<td>(\hat{\mu}_2)</td>
<td>-0.195</td>
<td>0.376</td>
</tr>
<tr>
<td>(\hat{\pi}_2)</td>
<td>-0.013</td>
<td>0.078</td>
</tr>
<tr>
<td>(LL_{G=2} - LL_{G=1})</td>
<td>-3.721</td>
<td>-1.112</td>
</tr>
</tbody>
</table>

The averaged signed and absolute bias, based upon 100 replicated datasets simulated under the above settings, are given in Table 2.3 for the various parameter sets. Overall, the data generating model seems to be quite nicely recovered, including the mixture component parameters. However, as expected, recovery is worse in the low-information condition \((\mu_2 = -2)\). Whereas in the high-information condition \((\mu_2 = 2)\) the mixture model is strongly separated from the one-component model alternative as can been seen in the difference of their log likelihood \((LL_{G=2} - LL_{G=1})\), this separation is less readily present in the low-information condition.

2.4.3 Model assessment

To assess the selected model for the BAT dataset, two series of datasets were simulated using the model parameters corresponding to the esti-
mated model 3 and model 5 in Table 2.2, giving rise to a condition in which conditional independence holds (indep) and a condition in which an additional association structure is present (copula). Datasets of both conditions were then fitted following both model 3 and model 5. These simulation results will prove to be useful for supporting the chosen model for the BAT study and clarifying the data-analytical results. Table 2.4 displays the average bias and root mean squared error for different parameter sets in this second simulation. In case the underlying model is the regular independence model, both the independence and copula model perform at equal foot. However, in case the underlying model is the copula model, the conditional independence model clearly underperforms.

To illustrate more clearly what can happen when we ignore the extra association structure present in the BAT study, we take a look at the covariate effects $\lambda_j$ in both model 3 and model 5 for the BAT dataset. In the latter model, only the covariate information on in utero exposure to anticonvulsant drugs was indicated as being significantly related to an infant’s severity of affect. On the other hand, when we would stick to the erroneous conditional independence model, one would conclude that, besides in utero exposure to anticonvulsants, having a mother with seizure history also has a significant effect on an infant’s severity of affect. This is a striking difference in inference exactly on one of the main research questions behind the BAT study. The mixture results of model 2 support our final copula model as the components only differed in the presence of infants belonging to the anticonvulsants drugs exposure group, and not with regard to the seizure-history group. Furthermore, we conjecture that, because a cluster-level covariate such as treatment and the joint conditional distribution both pertain to between-individual variation, ignoring violations of conditional independence compromises quality of estimation of the corresponding regression coefficient (supported by the simulation results which show on average larger bias and root mean squared error for the covariate parameters $\lambda$ when wrongfully
### Table 2.4: Simulation results on the comparison between independence and copula model.

<table>
<thead>
<tr>
<th>Estimated</th>
<th>Simulated</th>
<th>Bias</th>
<th>indep</th>
<th>RMSE</th>
<th>copula</th>
<th>Bias</th>
<th>indep</th>
<th>RMSE</th>
<th>copula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>indep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{\alpha} ) non-subset</td>
<td>0.008</td>
<td>0.011</td>
<td>0.164</td>
<td>0.165</td>
<td>-0.123</td>
<td>0.003</td>
<td>0.273</td>
<td>0.201</td>
<td></td>
</tr>
<tr>
<td>( \hat{\alpha} ) ( J_1 {1, 2} )</td>
<td>0.052</td>
<td>0.037</td>
<td>0.425</td>
<td>0.426</td>
<td>0.167</td>
<td>0.004</td>
<td>0.741</td>
<td>0.336</td>
<td></td>
</tr>
<tr>
<td>( \hat{\alpha} ) ( J_2 {6, 7, 8} )</td>
<td>0.014</td>
<td>0.017</td>
<td>0.192</td>
<td>0.193</td>
<td>1.353</td>
<td>0.014</td>
<td>2.195</td>
<td>0.128</td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta} ) non-subset</td>
<td>0.146</td>
<td>0.140</td>
<td>0.646</td>
<td>0.642</td>
<td>0.442</td>
<td>0.217</td>
<td>1.024</td>
<td>0.724</td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta} ) ( J_1 {1, 2} )</td>
<td>0.012</td>
<td>0.021</td>
<td>0.179</td>
<td>0.180</td>
<td>-0.023</td>
<td>0.064</td>
<td>0.443</td>
<td>0.235</td>
<td></td>
</tr>
<tr>
<td>( \hat{\beta} ) ( J_2 {6, 7, 8} )</td>
<td>0.125</td>
<td>0.116</td>
<td>0.684</td>
<td>0.670</td>
<td>-3.591</td>
<td>0.562</td>
<td>3.788</td>
<td>2.793</td>
<td></td>
</tr>
<tr>
<td>( \hat{\lambda} ) (averaged over ( j ))</td>
<td>-0.002</td>
<td>0.001</td>
<td>0.146</td>
<td>0.147</td>
<td>-0.118</td>
<td>0.013</td>
<td>0.227</td>
<td>0.166</td>
<td></td>
</tr>
<tr>
<td>( \hat{\delta}_1 ) ( J_1 {1, 2} )</td>
<td>.</td>
<td>0.148</td>
<td>.</td>
<td>0.030</td>
<td>.</td>
<td>0.055</td>
<td>.</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>( \hat{\delta}_2 ) ( J_2 {6, 7, 8} )</td>
<td>.</td>
<td>0.381</td>
<td>.</td>
<td>0.268</td>
<td>.</td>
<td>0.252</td>
<td>.</td>
<td>3.418</td>
<td></td>
</tr>
<tr>
<td>( \theta_p )</td>
<td>0.003</td>
<td>0.002</td>
<td>0.859</td>
<td>0.859</td>
<td>-0.002</td>
<td>-0.002</td>
<td>0.852</td>
<td>0.884</td>
<td></td>
</tr>
</tbody>
</table>

*100 replications per condition*
assuming conditional independence).

Table 2.5: *Parametric bootstrap assessments: Subset item patterns.*

<table>
<thead>
<tr>
<th></th>
<th>observed freq</th>
<th>indep freq</th>
<th>copula freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_1{1,2}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_{ps} = (1,1)$</td>
<td>48</td>
<td>39</td>
<td>48</td>
</tr>
<tr>
<td>$Y_{ps} = (0,0)$</td>
<td>538</td>
<td>529</td>
<td>538</td>
</tr>
<tr>
<td>$J_2{6,7,8}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_{ps} = (1,1,1)$</td>
<td>28</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>$Y_{ps} = (0,0,0)$</td>
<td>567</td>
<td>511</td>
<td>568</td>
</tr>
</tbody>
</table>

The generated datasets also allow for parametric bootstrap assessments to make more obvious the importance of accounting for this extra association structure. For instance, Table 2.5 shows the joint behavior of the anomalies in the detected subsets. The frequencies of the all-one and all-zero anomaly-subset patterns observed in the data of the BAT study most closely resemble the frequencies of these patterns in data generated under the copula model. The more accurate representation of detected anomaly subsets is again support for the selected copula model. Table 2.6 shows the observed sum score (over anomalies) distribution in the BAT data, compared to the resulting sum score distribution under both the independence as well as the copula datagenerating model. The $\chi^2$ statistic indicates that again the copula model resembles the most to the data of the BAT study. The copula model especially shows more accurate predictions (cfr., smaller $\chi^2$ values) for high sum score patterns, i.e., infants that are highly affected, and which are of clinical importance.

Without taking into account the residual dependencies a measurement instrument appears to be artificially more precise than it actually is. This can also be seen in the simulation results with regard to the infants latent severity $\theta_p$, in which the average bias and the average root mean squared error of $\theta_p$ over the sample of infants is given (RMSE = 0.852 under the erroneous conditional independence model while RMSE = 0.884 under the correctly specified copula model). For the BAT study, the difference between the standard error under the conditional indepen-
Table 2.6: *Parametric bootstrap assessments: Sum score.*

<table>
<thead>
<tr>
<th>sum score</th>
<th>observed freq</th>
<th>indep freq</th>
<th>$\chi^2$</th>
<th>copula freq</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>296</td>
<td>293</td>
<td>0.038</td>
<td>300</td>
<td>0.045</td>
</tr>
<tr>
<td>1</td>
<td>194</td>
<td>190</td>
<td>0.083</td>
<td>177</td>
<td>1.482</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>97</td>
<td>5.030</td>
<td>94</td>
<td>3.973</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>50</td>
<td>0.023</td>
<td>55</td>
<td>0.354</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>27</td>
<td>4.409</td>
<td>30</td>
<td>2.352</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>15</td>
<td>0.370</td>
<td>16</td>
<td>0.289</td>
</tr>
<tr>
<td>$\geq$6</td>
<td>11</td>
<td>16</td>
<td>1.949</td>
<td>14</td>
<td>0.914</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>1.70</td>
<td></td>
<td>1.33</td>
</tr>
</tbody>
</table>

dence model (model 3) and the copula model (model 5) is plotted against the standard error under the copula model in Figure 2.4. Here one can see that precision varies quite a lot over the latent dimension, but is always (artificially) more precise under the independence model. Because decisions about individual infants can be based on the estimate of the severity, a correct assessment of its uncertainty is necessary for an adequate measurement instrument, and as such a copula model is more appropriate for use in the BAT study. As mentioned earlier, measurement is more precise at the upper part of the scale, which corresponds with the intended diagnostic purposes.

Figure 2.4: *Standard errors of the latent severity in the presence of residual dependencies under a conditional independence model and a copula model.*

Figure 2.5 plots the infant severity index against its corresponding standard error over the scale. It is more important to have a precise es-
timate of the most severely affected infants, where as it is less of an issue to where the hardly affected infants are located on the lower scale end. This also relates to the observation that the location of the anomalies on the scale, i.e., their severity degree, is less precisely estimated based upon the copula simulated data than under the conditional independence generated datasets (see the RMSE of the $\beta_i$’s in Table 2.3). In the former condition less information is present in the data than in the latter, yet each condition consists of basically the same raw data structure; it is the association structure in the data that makes the difference.

Figure 2.5: *Precision of measurement over the infant severity index scale.*

2.5 Discussion

A FIML latent variable framework for studying teratology problems, that goes beyond standard assumptions, and is motivated by scientific questions, has been presented. The core of the model is a nonlinear mixed model with anomaly-specific discrimination and severity degree parameters. Instead of rashly accepting the standard approach of a normal distributed latent variable, a more flexible data-driven way of specifying the latent distribution is proposed by adopting a finite mixture approach and by incorporating covariates into the model. If needed, the covariate effects can be allowed to be anomaly-specific. Finally, the model can also handle residual dependencies stemming from similarities
between anomalies that are not accounted for by traditional conditional independence models. These dependencies are modeled using copulas, because this allows the univariate margins, and their interpretation, to be left unchanged.

The results of our analysis of the teratology data clearly suggest that in utero exposure to anticonvulsants is a high risk factor for having an infant with birth defects. There was no evidence for an effect of having a mother with a seizure history, nor an effect of gender on a higher risk for birth defects, unless a misspecified model was used. This last observation highlights the strong relation between statistically biased models and substantively unrealistic models in a latent variable context. Making adequately use of and exploring more of the information in the data, should always lead to improved analysis and inference. The same conclusion on the covariate effects can be found in the round-up article on the more extensive BAT project by Holmes et al. (2001).

However, the present study does not allow us to make the bold conclusion that epilepsy has no influence on birth defect risks, and that the higher risk is solely to be ascribed to the related anticonvulsant medication therapy followed by most epileptic women. To fully exclude the influence of epilepsy, one should control for the strength of the epileptic disease in the respective participants of the study. Clinical studies have suggested that our set of 10 anomalies might even be predictive for a weaker cognitive development (Holmes, Coull, Dorfman, & Rosenberger, 2005). It would be interesting to further explore the relation between the measurement scale we obtained, the infant severity index $\theta_p$, and other external criteria to further validate the model.
Chapter 2

References


Frank, M. J. (1979). On the simultaneous associativity of \( F(x, y) \) and \( x + y - F(x, y) \). *Aequationes Mathematica, 19*, 194–226.


Chapter 2

291–297.


models: a generalized linear and nonlinear approach (pp. 289–316). New York: Springer.


Correlated error terms in structural equation models for categorical data

Abstract: For structural equation models with categorical data, there is one practice regularly applied for Gaussian data that is not easily implemented: the technique of correlated errors. The problem lies mainly in the absence of a categorical analogue to the multivariate normal distribution and the presence of not-closed form formulas in structural equation models for categorical data. We will present a novel method to handle correlated errors that has a parsimonious structure and leads to straightforward interpretations without a too excessive computational burden. This method is based upon the concept of copula functions and will be introduced with a dataset of ordinal responses originating from a contextualized personality study.
3.1 Introduction

Traits are an important conceptual device in personality psychology and the study of individual differences. A trait is defined as a relatively permanent individual characteristic that is generalizable across situations and contexts, and forms the foundation of a person’s typical behavior. Yet, since Mischel (1968) (also see Mischel & Shoda, 1995) demonstrated the lack of behavioral situational consistency, a long-standing debate ensued about the importance of such traits, because behavior occurs in a context and is not only determined by some general intrinsic personal characteristics. An important lesson drawn from the past debate is that it is important to acknowledge the context in which behavior takes place. This change in the understanding and position of traits has also influenced the measurement practice. The assessment of traits relies most often on questionnaires. Taking into account the context of behavior, it makes sense then to assess and study personality not in a vacuum (i.e., with contextless general indicators and questions), but to use a measurement instrument that looks at different situations that provide relevant context for a person’s behavior.

Research that seeks to validate a measurement instrument or test theories about relations between traits, often makes use of confirmatory factor analysis (CFA) or the more general structural equation models (SEM). Such structural equation models are helpful in testing specific structural relations (i.e., regressions) between the traits. Although structural equation models were originally only useful for normally distributed observations, the data spectrum is nowadays broadened and it can handle also categorical data types within the same framework.

However, in the context of SEM with categorical data, there is one practice regularly applied with normal data that is not easily implemented: correlated errors. Such correlated errors are often used in an opportunistic way to achieve model fit, but in other cases they are necessary to consider. For instance, in the contextualized questionnaires,
several indicators may refer to the same situation and this creates additional association beyond what is accounted for by the latent traits. Such additional association is readily modeled in a normal SEM by allowing the distributions of the error terms, which are otherwise assumed to be uncorrelated, to correlate with one another. Such correlations between the errors are straightforward to implement by adding additional covariances to the covariance matrix. Such a simple approach is not available in SEM with categorical variables because of two reasons.

A first reason for this lack of simplicity is situated on the level of interpretation. The normal distribution has the attractive property that its univariate margins are also normal, irrespective of the correlations. Most other multivariate distributions do not share this property with the normal distribution. Stated otherwise, changing the dependence structure has no influence on the univariate marginal distributions. This is rarely the case with categorical data models. Take as an example a logistic regression model for a bivariate binary outcome. If the two outcomes cannot be assumed to be independent, then one can add an interaction parameter in order to obtain a bivariate distribution. However, the univariate margins (i.e., the probability of a success on one of the outcomes) do not have a logistic regression form any more (see e.g., Braeken, Tuerlinckx, & De Boeck, 2007, see also the concept of upward compatibility as discussed in McCullagh, 1989). Thus, by altering the association structure of the model, the interpretation at the level of the marginals changes.

A second reason why correlated errors are not easily accommodated in SEM for categorical data is situated on the computational level. Structural equation models were developed as extensions of classical linear regression and path analysis, and hence, they are more or less tailor-made for the case of normally distributed variables. To accommodate for categorical variables a similar logic can be followed as in extending the linear regression model to for instance logistic regression or more in general the generalized linear model (see e.g., Skrondal & Rabe-Hesketh,
2004). However, whereas in the case of normal distributed variables the variance-covariance matrix and mean structure are sufficient statistics to describe their distribution (Mood, Graybill, & Boes, 1974; Knol & Berger, 1991), there is no such direct analogue in the case of categorical variables. This makes the computational algorithms in SEM for categorical data much more complex (for a recent overview see e.g., Wirth & Edwards, 2007). The challenge of dimensionality, i.e., approximating multidimensional integrals without a closed form solution, is the major obstacle in applications of SEM to categorical data.

This is also the reason why taking the alternative perspective that what underlies the additional associations are other individual difference factors, does not avoid the issue. Such an approach requires the inclusion of more latent variables leading to a high computational cost because of this multidimensionality issue. Furthermore, such an approach can easily end up in an overfactorization of a given dataset, resulting in exact and/or empirical model identification problems (see e.g., Rindskopf, 1984; Kenny & Kashy, 1992; Wothke, 1996; Eid, 2000).

Our goal is to present an approach that can handle correlated errors, but at the same time leads to straightforward interpretations without a too excessive computational burden. Given the often small sample sizes and already complex models in behavioral sciences, a parsimonious structure will be pursued. Our approach will be based on copula functions as they have been introduced recently in the psychometric literature by Braeken et al. (2007). However, the approach in Braeken et al. (2007) is only suited for binary data and item response models. In this paper, we will extend their method to handle ordered polytomous data and to include structural relations between the latent variables. To facilitate the introduction of the topic, we start from a specific research problem, to be explained in the next section.
3.1.1 Data and substantive research questions

Aggression is at the basis of much societal and personal suffering (Berkowitz, 1993). As a consequence, the personality trait of aggression has been studied extensively in personality psychology. Over the years, researchers have come to conceptualize aggressiveness as consisting of several, interrelated components that reflect the affective, behavioral, and cognitive aspects that are involved in aggression. This is the so-called ABC model (see e.g., Martin, Watson, & Wan, 2000). Recent definitional attempts and studies of aggressiveness consistently include three constructs: Behavioral aggression, anger, and hostility. The behavioral aggression construct involves “hurting or harming others” and represents the instrumental component of aggressiveness. Anger includes “physiological arousal and preparation for aggression” and represents the affective component. Finally, hostility consists of “feelings of ill will and injustice” and represents the cognitive component of aggression. Yet, most questionnaires used in personality studies consider aggression at the general trait-level only, leaving out potential contextual effects in the assessments of aggression, contextual effects that evidently can be considered to exist (for an overview, see e.g., Anderson & Bushman, 2002).

Our motivating dataset arises from a study of aggression that assessed the different subcomponents of aggression in different situational contexts. As such, it allows to study aggression while taking into account the possible context-dependency of aggressive behavior. This in contrast to prior studies by for instance Buss and Perry (1992) and Martin et al. (2000) that were based upon typical general trait questionnaires. In particular, our example study assessed each aggression component in three different situational contexts: In the first situation, you find out that a friend spreads gossip and speaks ill about you behind your back. In the second situation, you are waiting for already more than 30 minutes on a terrace for the local waiter to bring the drink you ordered. In the third situation, you are out on a hiking trip with friends and the next day
when setting up camp the construction poles of the tent appear to have gone lost. This contextualized personality instrument was administered in a directed imagery study, in which participants were asked to read each situational description, imagine as vividly as possible that they found themselves in this situation, and subsequently indicate to what extent they would display each of the three aggressive subcomponents.

The design of our aggression inventory is fully crossed (three trait components by three situations), resulting in a total of nine indicators. The first set of three indicators refers to the first situation (gossip), the second set of three refers to second situation (terrace), and the third set of three refers to the third situation (tent). Within such a set of three, the first question is an indicator of anger, the second an indicator of behavioral aggression, and the third an indicator of hostility. Each question was answered on a Likert scale ranging from 0 (I would not react like that at all) to 6 (I would totally react like that). The inventory was completed by 340 university students who participated in the study in fulfillment of class requirements. To reduce the interpretational and computational problems associated with seven category items but at the same time retain some of the ordinal information, we decided to collapse categories into three. A 0 or 1 response was recoded as 0, a 2 or 3 response was recoded as 1, and a 4 to 6 response was recoded to 2.

With respect to these data, the primary question of substantive interest is due to Buss and Perry (1992) and can be investigated within a SEM framework. It is hypothesized that anger is a psychological bridge between the instrumental component of aggression, and the cognitive component, hostility. Anger often precedes aggressive behavior and makes the manifestation of such behavior more likely. Hostility is seen as some residual ill-will after anger has cooled down. As such, the relation between hostility and behavioral aggression might be fully ascribed to the mediating role of anger. Structural equation models can be used to investigate the relations described in this process theory.
When taking into account the additional dependencies coming from the contextualized nature of the measurement instrument, it might be asked how this influences the inferences with respect to this substantive research question. For instance, does accounting for the associations from the same situations change, in a qualitative or quantitative way, the directed relations between the different trait components? To investigate this, we will develop a method to handle the residual dependencies and illustrate its use with respect to the data and research questions stated above.

3.1.2 Organization of the paper

First, we will outline the structural equation modeling approach for ordinal data, and consequently apply this approach to the aggression data. Second, we will deal with the dependency issue and the possibility of correlated errors. Copula functions will be presented as a novel way to account for the possibility of correlated errors, or in other words remaining unaccounted dependency between the indicators. We will round up with applying this copula approach to accommodate the specific contextualized structure underlying our personality study and explore what the introduction of copulas can learn us for the data at hand. Finally, to conclude the paper some room is made for a more general discussion.
3.2 SEM for ordinal variables

If participant \( p \) \((p = 1, \ldots, P \text{ with } P = 340)\) judges the plausibility of a specific reaction on an indicator \( i \) \((i = 1, \ldots, I \text{ with } I \text{ and } I = 9)\) in category \( q \) \((\text{with values } q = 0, \ldots, Q - 1 \text{ and } Q = 3)\), then we write \( Y_{pi} = q \). Because of the ordinal response scale, the plausibility of the reaction increases with increasing category level \( q \). The modeling proceeds with assuming that the categorical ordinal outcome \( Y_{pi} \) is a discrete manifestation of an underlying latent continuous variable \( \tilde{Y}_{pi} \). Next, a measurement model will be defined for the vector of latent outcomes \( \tilde{Y}_p \) and consequently, the latent outcome \( \tilde{Y}_{pi} \) is linked to the observed categorical outcome \( Y_{pi} \) by means of a threshold model. Furthermore, a structural model for the latent factors defined through the measurement model can be formulated.

3.2.1 Measurement model for the latent outcomes \( \tilde{Y}_p \)

The model for a latent outcome \( \tilde{Y}_{pi} \) is as follows:

\[
\tilde{Y}_{pi} = \mu_i + \sum_{d=1}^{D} \lambda_{id}\eta_{pd} + \varepsilon_{pi} \tag{3.1}
\]

where \( \mu_i \) is the general effect of indicator \( i \), \( \eta_{pd} \) is the score of person \( p \) on common factor \( d \) \((d = 1, \ldots, D)\), \( \lambda_{id} \) the factor loading of indicator \( i \) on common factor \( d \), and \( \varepsilon_{pi} \) the unique factor score or residual error term for indicator \( i \) of person \( p \). Thus, it is assumed that the set of indicators are manifestations of a smaller set of latent factors, explaining the common information shared by these indicators, and a measurement error part, giving rise to (factor) scores that are a more reliable measure compared to the single indicators.
In matrix form, the model for all responses of person \( p \) reads as:

\[
\tilde{Y}_p = \mu + \Lambda \eta_p + \varepsilon_p
\]

where \( \eta_p \) is normally distributed: \( \eta_p \sim N(\vartheta, \Sigma) \). For the distribution of the residual \( \varepsilon_{pi} \), we will opt for a logistic distribution with mean 0 and variance \( \psi_i \). The choice for a logistic distribution instead of a normal distribution is motivated by the fact that the logistic distribution has nicer computational and interpretational properties (see below). The \( I \) residuals are independent and this is the conditional independence or local stochastic independence assumption of the model. The \( I \) variances can be put in a diagonal matrix, denoted as \( \Psi \).

These settings imply that (making use of the laws of total expectation and total variance):

\[
E(\tilde{Y}_p) = E\left(E(\tilde{Y}_p|\eta_p)\right) = \mu + \Lambda \vartheta \tag{3.2}
\]

and

\[
\text{Var}(\tilde{Y}_p) = E\left(\text{Var}(\tilde{Y}_p|\eta_p)\right) + \text{Var}\left(E(\tilde{Y}_p|\eta_p)\right) \tag{3.3}
\]

\[
= \Lambda \Sigma \Lambda' + \Psi.
\]

So far this is analogue to traditional factor analysis for continuous observable variables.

### 3.2.2 Threshold model for the latent outcomes \( \tilde{Y}_p \)

Consider the observed ordinal outcome \( Y_{pi} \) with \( Q \) categories with values \( q \) going from 0 to \( Q - 1 \). The latent outcome \( \tilde{Y}_{pi} \) is linked to this observed categorical outcome \( Y_{pi} \) by means of a threshold model, following the relation

\[
Y_{pi} = q \text{ if } \tau_{iq} < \tilde{Y}_{pi} \leq \tau_{i,q+1}
\]
where the thresholds satisfy the ordinal restriction:

\[-\infty = \tau_{i0} \leq \tau_{i1} \leq \cdots \leq \tau_{iq} \leq \cdots \leq \tau_{i,Q-1} \leq \tau_{iQ} = +\infty.\]

This idea is depicted graphically in Figure 3.1 for $Q = 3$.

Figure 3.1: Threshold model for the latent outcome $\tilde{Y}_{pi}$.

The above relationship gives rise to the following conditional probability formulation for an individual outcome given a latent factor score vector $\eta_p$:

\[
\Pr(Y_{pi} \geq q|\eta_p) = \Pr(\tilde{Y}_{pi} \geq \tau_{iq}|\eta_p) \tag{3.4}
= 1 - \Pr(\tilde{Y}_{pi} < \tau_{iq}|\eta_p)
= 1 - \Pr(\epsilon_{pi} < \tau_{iq} - \mu_i - \sum_{d=1}^{D} \lambda_{id}\eta_{pd}|\eta_p)
= 1 - \Pr\left(\frac{\epsilon_{pi}^*}{\sqrt{\psi_i}} < \frac{1}{\sqrt{\psi_i}}(\tau_{iq} - \mu_i - \sum_{d=1}^{D} \lambda_{id}\eta_{pd})|\eta_p\right)
= G\left[\frac{1}{\sqrt{\psi_i}}(\tau_{iq} - \mu_i - \sum_{d=1}^{D} \lambda_{id}\eta_{pd})\right],
\]

where $\epsilon_{pi}^*$ follows a logistic distribution with variance 1 and its CDF is denoted by $G$ (such that $G(x) = \frac{e^x}{1+e^x}$, where $k = \sqrt{3}\pi$).
Due to the fact that $\tilde{Y}_p$ is unobservable, there is no information present about its metric, and this needs to be set a priori (Bartholomew & Knott, 1999) to identify the model. For instance, it can be seen that there is a trade-off relation between $\mu$ the mean vector of the indicators and $\Lambda \vartheta$, the weighted mean vector of the latent variables $\eta_p$ (a constant can be added to one of them and subtracted from the other without affecting the model). Therefore, it is assumed that $\vartheta = 0$ so that $E(\tilde{Y}_p) = \mu$. From Equation 3.4, it can be seen that an additional constraint is needed to identify the model since there is a trade-off relation between the $\tau_i$s and $\mu_i$. Therefore, we choose to set $\mu_i$ equal to zero. Another identification problem we face is that all thresholds $\tau$ and latent variables $\eta$ can be multiplied by a factor, and at the same time $\psi$ as well, without changing the probability from Equation 3.4. The usual convention here is to further standardize and set $\text{Var}(\tilde{Y}_{pi}) = 1$, such that the unique variance is actually not a free parameter, but a remainder term defined as $\psi_i = (1 - \lambda_i' \Sigma \lambda_i)$, where $(\lambda_i' \Sigma \lambda_i)$ is recognized as the communality or variance explained by the common factors $\eta$ in indicator $i$. In matrix form, this restriction looks as $\Psi = (I_{D \times D} - \Lambda' \Sigma \Lambda)$. The whole set of restrictions result in setting the metric of the latent variables in a standardized format where $E(\tilde{Y}_{pi}) = 0$ and $\text{Var}(\tilde{Y}_{pi}) = 1$.

Now we can find the probability that person $p$ responds to item $i$ exactly in category $q$ as follows:

$$\Pr(Y_{pi} = q | \eta_p) = \Pr(Y_{pi} \geq q | \eta_p) - \Pr(Y_{pi} \geq q + 1 | \eta_p)$$

$$= G[-(\tau_{iq} - \sum_{d=1}^{D} \lambda_{id} \eta_{pd})/\sqrt{\psi_i}]$$

$$- G[-(\tau_{i,q+1} - \sum_{d=1}^{D} \lambda_{id} \eta_{pd})/\sqrt{\psi_i}].$$
The model can be reparametrized such that one obtains a formulation akin to item response models (see e.g., Takane & Leeuw, 1987). This reparametrization is carried out by absorbing the variance of the latent residual $\varepsilon_{pi}$ (the unique variance $\psi_i$) into the thresholds and loadings:

$$
\Pr(Y_{pi} \geq q | \eta_p) = G \left[ -\frac{1}{\sqrt{\psi_i}} \left( \tau_{iq} - \sum_{d=1}^{D} \lambda_{id} \eta_{pd} \right) \right] = G \left[ -(\beta_{iq} - \sum_{d=1}^{D} \alpha_{id} \eta_{pd}) \right] = \frac{\exp \left( \sum_{d=1}^{D} \alpha_{id} \eta_{pd} - \beta_{iq} \right)}{1 + \exp \left( \sum_{d=1}^{D} \alpha_{id} \eta_{pd} - \beta_{iq} \right)},
$$

where the new coefficients $\alpha_{id}$ and $\beta_{iq}$ are defined as follows:

$$
\alpha_{id} = \frac{\lambda_{id}}{\sqrt{\psi_i}}, \quad (3.5)
$$

and

$$
\beta_{iq} = \frac{\tau_{iq}}{\sqrt{\psi_i}}, \quad (3.6)
$$

Conversely, one may express the factor analysis parameters as functions of the item response model parameters$^1$:

$$
\lambda_{id} = \frac{\alpha_{id}}{\sqrt{1 + \alpha_i' \Sigma \alpha_i}} \quad \text{and} \quad \tau_{i,q} = \frac{\beta_{i,q}}{\sqrt{1 + \alpha_i' \Sigma \alpha_i}}
$$

$^1$This can be derived for $\lambda_{id}$ by premultiplying both sides of Equation 3.5 by $\lambda_i' \Sigma$ and then solving for $\lambda_i' \Sigma \lambda_i$ and inserting it back into Equation 3.5 and solving for $\lambda_{id}$.
Both parametrizations are equivalent and it is easy to go from one to the other. For estimation purposes, we will work with the unstandardized coefficients $\alpha$ and $\beta$ in the conditional probability model but for inferential purposes, the traditional factor analysis coefficients $\tau$ and $\lambda$ are perhaps more informative due to their standardized nature, such that $\lambda$ is for instance to be interpreted as a common factor loading taking values between 0 and 1.

### 3.2.3 Structural model for the latent factors $\eta_{pd}$

If a theory is available proposing directional relations between (some of) the latent factors, then structural equation models allow to put a structure on the latent factors similar to what is done in path analysis with observed variables. These regressions among the latent factors are specified in the structural part of the model:

$$\eta = \Gamma \eta + \zeta$$

(3.7)

where $\Gamma$ is a $D \times D$ matrix of regression coefficients where the diagonal elements are always zero, meaning that a factor can not be explained by itself (hence $I_{D \times D} - \Gamma$ is non-singular). In Equation 3.7, $\zeta$ is the $D \times 1$ residual vector of the regressions and is typically assumed to follow a normal distribution with mean zero and variance-covariance matrix $\Xi$:

$$\zeta \sim N(0, \Xi).$$

### 3.2.4 Estimation

All models in this paper will be estimated using a marginal maximum likelihood (MML) approach, in which the latent factors $\eta_p$ are integrated out. Because there are no closed form solutions for the latter integrals, they are approximated numerically using a Gauss-Hermite quadrature with 20 nodes for unidimensional models ($D = 1$) and 5 nodes per dimension for multidimensional models ($D > 1$) (i.e., 20 and $5^D$ summations are carried out per evaluation of the log likelihood, respectively). This
MML method is suitably specified to the particularities of the categorical case, leading to consistent and efficient estimates, and the availability of a full likelihood expression allows for likelihood based inference (Bock, Gibbons, & Muraki, 1988). Note that, where possible, results were verified by means of comparison to model results given by Mplus (Muthén & Muthén, 1998).

Although there are no well established guidelines for which minimal conditions constitute an adequate result of the model fitting steps, a general approach is to establish that the model is identified, that the iterative estimation procedure converges, that all parameter estimates are within the range of permissible values, and that the standard errors of the parameter estimates have reasonable size (Marsh & Grayson, 1995). We will follow these recommendations.

### 3.3 Application: Part I

In this section we will apply a structural equation model to the aggression data set in accordance with theorizing within the ABC personality framework. We follow the suggestion of Buss and Perry (1992) that anger mediates the relation between hostility on the one hand and behavioral aggression on the other hand, and test this process theory in a mediation model. To visualize the models, we make use of path diagrams (see e.g., Bollen, 1989). Single-headed arrows represent directional influences (i.e., regressions) between the latent traits and double-headed arrows represent correlations, the observed variables and error terms are omitted for clarity.

Two specific models will be pitched against each other. The first model (see first panel of Figure 3.2) is a strict mediation model in which there is no direct effect of hostility on behavioral aggression; all influence of hostility runs through anger. This model will be called the strict mediation model. The second model also includes a direct relation between hostility and aggression and will be called the partial mediation
model. The latter model is illustrated in the second panel of Figure 3.2 and is formally equivalent model to a unstructured confirmatory factor analytical model with interrelated factors.

Figure 3.2: Path diagrams of the ABC models.

Because the strict mediation model is nested in the partial one, a likelihood ratio test can be carried out to compare both models. The test indicates that the two models do not differ significantly ($\chi^2 = 0.8$, $df = 1$, $p = 0.371$) and thus that we prefer the simpler strict mediation model. This indicates that our process theory can give a reasonable and more parsimonious account of the relations between the aggression components for the data at hand. In the Figure 3.2 the resulting standardized regression coefficients $\Gamma$ for both models are given, and in Table 3.1, all parameter estimates for the strict mediation model are listed. To verify that our ABC formulation of aggression is not a needlessly complex model, the strict mediation model was compared to the unidimensional model (which is also a nested model). The likelihood ratio test for nested models resulted in a (marginal) significant outcome ($\chi^2 = 6$, $df = 2$, $p = 0.050$). Thus in principle, the ABC models should be rejected in favour of the more parsimonious unidimensional model, which would mean that the distinctions among anger, behavioral aggression, and hostility are not critical in understanding aggressiveness.
Table 3.1: An ABC model of aggression: strict mediation model

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \beta_i^+ = \beta_i^2 - \beta_i^1 \]

Note. The parameter \( \beta_i^+ \) is the unstandardized coefficient of the regression of \( Y \) on \( X \) (\( \beta \) is the standardized version).

Standard errors are given between parentheses and \( p \)-values result from the Wald test.

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.035 (0.331) 0.003 (0.001)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.853 (0.153) 0.000 (0.000)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.987 (0.235) 0.002 (0.001)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( \Gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.564 (0.283)</td>
</tr>
<tr>
<td>B</td>
<td>0.869 (0.153)</td>
</tr>
<tr>
<td>C</td>
<td>1.391 (0.235)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.082</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.082</td>
</tr>
<tr>
<td>B</td>
<td>0.033</td>
</tr>
<tr>
<td>C</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Note. The parameter \( \Gamma \) is the unstandardized coefficient of the regression of \( Y \) on \( X \) (\( \gamma \) is the standardized version). Standard errors are given between parentheses and \( p \)-values result from the Wald test.
Note that the models presented above assume that the latent factors $\eta_p$ are sufficient to explain the association between the individual outcomes. This is often referred to as the local stochastic or conditional independence assumption: Conditional on the latent factors the individual outcomes are assumed to be independent. However, in case there are additional sources of association, this assumption does not hold, and there will be some residual dependence not explained by the model. Conditional independence in the model should be reflected by measurement error terms $\varepsilon_{pi}$ that are uncorrelated\(^2\).

As explained in the introduction, a possible approach to account for these residual dependencies is to add more latent factors. In the context of the example, this means that instead of having only personality traits, we can also include latent situation factors. However, a downside to this solution is that it may lead to overfactorization of a given dataset, leading to identification problems (see e.g., Rindskopf, 1984; Kenny & Kashy, 1992; Wothke, 1996; Eid, 2000). In any case, adding factors in a latent variable model for categorical data increases the computational cost of fitting the model (cfr. the challenge of dimensionality; more intractable integrals need to be approximated numerically). In fact, we applied an ABC model with three added situation factors to the aggression inventory data, but the estimation algorithm failed to converge.

### 3.4 Correlated errors through copulas

Instead of adding latent variables to the model, an alternative approach is to allow for correlated errors. This is straightforward when the residuals $\varepsilon$ are normally distributed because then one may allow for non-zero covariances in the covariance matrix of the residuals. Unfortunately, this approach only seems to work well for normally distributed error

\(^2\)One might carry out exploratory analyses to investigate whether these residual dependencies exist by examining bivariate or higher-order measures of association. However, when there are theoretical reasons for the possibility of residual dependencies and they are of interest, as in our case, a confirmatory approach might be a more direct way to handle the issue.
distributions and requires additional computational effort due to the extra multivariate (not closed form) integrals that are introduced in this way to the model. The same option for logistically distributed errors, as considered in this paper, is not readily available due to the absence of a general multivariate logistic distribution (see e.g., Kotz, Balakrishnan, & Johnson, 2000). In addition, there is another constraint because we want a solution that does not change the marginal probability model for an individual indicators to allow for straightforward interpretations such as in Equation 3.4. In the remainder of this section, we will discuss such an approach, based on copulas. In its basic form, a copula is a function that couples univariate marginal distributions to construct a multivariate distribution, without affecting the univariate margins.

3.4.1 Copula introduction

To introduce what copulas are, we step back to the simpler normal linear model. The reason we do so is because the normal distribution is very familiar to most people, which makes it easier to introduce some key concepts. Consider for reasons of simplicity two continuous random variables \( X_1 \) and \( X_2 \), which are both normally distributed with mean zero and unit variance (the general case is easy to work out by standardizing variables). The fact that \( X_1 \) and \( X_2 \) are both normally distributed can be denoted in many ways, but we will indicate this by saying that their respective cumulative distribution functions are \( F_{X_1}(x_1) = \Phi(x_1) \) and \( F_{X_2}(x_2) = \Phi(x_2) \), where \( \Phi(x_i) \) is the univariate CDF of the random variable \( X_i \) evaluated at \( x_i \). If \( X_1 \) and \( X_2 \) are independent, their joint cumulative distribution, denoted by \( F_X(x_1, x_2) \), is a product of the two univariate distribution functions:

\[
F_X(x_1, x_2) = \Phi(x_1) \times \Phi(x_2)
= \prod_{i=1}^{2} F_{X_i}(x_i)
= C_{\Pi}(F_{X_1}(x_1), F_{X_2}(x_2)).
\]
The function introduced in the last line is the product function, which is nothing more than a way to write a product of two factors: \( C_{\Pi}(u_1, u_2) = u_1 \times u_2 \). The function \( C_{\Pi} \) is actually a copula, but a very simple one (i.e., the independence copula), because it does not imply any dependence between \( X_1 \) and \( X_2 \) but instead preserves the independence between the two variables.

However, let us now look at some properties of the copula \( C_{\Pi} \). The first property is that the copula is a function and its arguments lie between 0 and 1. This is easy to see because the arguments are the CDFs and their range lies by definition between 0 and 1. Second, if at least one of the arguments is exactly zero, the output value of the copula function is also zero (e.g., \( C(0, u_2) = 0 \)). Third, if all arguments except one are equal to 1, then the output value of the copula is equal to that single non-unit value (e.g., \( C(1, u_2) = u_2 \)).

The mentioned properties of the copula function are actually shared by all bivariate and multivariate cumulative distribution functions. In fact, a copula is a multivariate cumulative distribution function but for uniformly distributed random variables. To see this, note that for a normal random variable \( X \), the distribution of \( U = \Phi(X) \) is uniform between 0 and 1 (see e.g., Mood et al., 1974). The specific product copula introduced above is thus the bivariate CDF for two independent uniformly distributed random variables \( U_1 = \Phi(X_1) \) and \( U_2 = \Phi(X_2) \):

\[
C_{\Pi}(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2) = \Pr(U_1 \leq u_1) \times \Pr(U_2 \leq u_2) = \Pr(X_1 \leq \Phi^{-1}(u_1)) \times \Pr(X_2 \leq \Phi^{-1}(u_2)) = \Pr(X_1 \leq x_1) \times \Pr(X_2 \leq x_2),
\]

with \( \Phi^{-1}(u_i) \) being the inverse CDF of \( X_i \) evaluated at \( u_i \) \( (u_i \in [0, 1]) \). The fact that the copula itself is a CDF combined with the third property above leads to the fact that a copula preserves the univariate CDFs. This
can be seen as follows:

\[ C_{\Pi}(u_1, 1) = \Pr(U_1 \leq u_1, U_2 \leq 1) = \Pr(U_1 \leq u_1) \times \Pr(U_2 \leq 1) = \Pr(X_1 \leq \Phi^{-1}(u_1)) \times 1 = \Phi(x_1). \]

After having introduced the most simple copula, we will make matters a little bit more complex. Assume we want to model the dependency between the two normal variables \( X_1 \) and \( X_2 \). The natural choice is to consider a bivariate normal distribution with correlation parameter \( \rho \):

\[ F_{X}(x_1, x_2) = \Phi_2(x_1, x_2; \rho), \quad (3.8) \]

where \( \Phi_2(x_1, x_2; \rho) \) is the standard bivariate normal cumulative distribution function with correlation \( \rho \):

\[
\Phi_2(x_1, x_2; \rho) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( -\frac{t_1^2 - 2\rho t_1 t_2 + t_2^2}{2(1 - \rho^2)} \right) \, dt_1 \, dt_2.
\]

Equation 3.8 can be rewritten as follows:

\[
F_{X}(x_1, x_2) = \Phi_2(x_1, x_2; \rho) \quad = \quad \Phi_2(\Phi^{-1}(\Phi(x_1)), \Phi^{-1}(\Phi(x_2)); \rho) \quad = \quad C_{\Phi}(\Phi(x_1), \Phi(x_2); \rho),
\]

such that the bivariate CDF \( F_{X}(x_1, x_2) \) is written as a function of the univariate CDFs \( \Phi(x_1) \) and \( \Phi(x_2) \). In this case, the copula function is \( C_{\Phi}(u_1, u_2; \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho) \), where \( \Phi^{-1} \) is the inverse of the standard normal cumulative distribution function.

Because it is not too difficult to show that the same properties as for the independence copula hold, using the properties of the univariate and bivariate normal distributions, we will not show these derivations here, except for one. It will be illustrated that also this second copula preserves the univariate Gaussian marginals. In order to show, consider
the following derivation:

\[
C_{\Phi}(\Phi(x_1), 1; \rho) = \Phi_2(\Phi^{-1}(\Phi(x_1)), \Phi^{-1}(1); \rho)
\]

\[
= \Phi_2(x_1, \infty; \rho)
\]

\[
= \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( -\frac{t_1^2 - 2\rho t_1 t_2 + t_2^2}{2(1 - \rho^2)} \right) dt_1 dt_2
\]

\[
= \frac{1}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{x_1} \exp \left( -\frac{1}{2(1 - \rho^2)} \frac{t_1^2}{t_2^2 - 2\rho t_1 t_2} \right) dt_2 dt_1
\]

\[
= \frac{\sqrt{2\pi(1 - \rho^2)}}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{x_1} \exp \left( -\frac{1}{2(1 - \rho^2)} \left( t_1^2 - t_1^2 \rho^2 \right) \right) dt_1
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} \exp \left( -\frac{1}{2} t_1^2 \right) dt_1
\]

\[
= \Phi(x_1),
\]

where we have worked out the integral over \(t_2\) by completing the squares and recognizing a normal integral over the real line. Thus, also the current copula function \(C_{\Phi}\) preserves the marginals, something with which we are already familiar from the bivariate normal case.

To summarize, what we have shown here is that in two cases a bivariate distribution\(^3\) could be reformulated as a function \(C\) that accounts for the dependency structure of \(X_1\) and \(X_2\), and takes as arguments the univariate CDFs of \(X_1\) and \(X_2\). Both functions \(C_{\Pi}\) and \(C_{\Phi}\) are actually two types of copula functions, the independence copula and the Gaussian copula. The main characteristic of such a copula function is that it is itself a multivariate cumulative distribution function of univariate uniformly distributed variables. Hence, copula functions will have similar properties as a regular cumulative distribution function.

\(^3\)Note that the approach described here, can easily be generalized to more dimensions.
3.4.2 Copulas to couple error terms in SEM

The two examples in the previous section illustrate an important theorem in copula theory due to Sklar (Sklar, 1959). The theorem will make clear what the application possibilities are of copulas in multivariate modeling (for a thorough overview of copula theory, see Nelsen, 1999, and Joe, 1997). Sklar’s theorem states that for any $I$-dimensional distribution function $F_{\mathbf{X}}$ with univariate margins $F_{X_1}, \ldots, F_{X_I}$ there exists a copula function $C$ such that this multivariate distribution $F_{\mathbf{X}}$ can be represented as a function of its margins through this copula: $F_{\mathbf{X}} = C(F_{X_1}, \ldots, F_{X_I})$. In the continuous case, this copula reformulation of an existing multivariate distribution is unique, for discrete random variables the copula $C$ is uniquely determined on $\text{Ran}(F_{X_1}) \times \cdots \times \text{Ran}(F_{X_I})$, the Cartesian product of ranges of the margins.

More importantly for modeling purposes, the converse of the theorem also holds. Thus given a set of univariate margins $F_{X_1}, \ldots, F_{X_I}$, a multivariate (i.e., joint) distribution $F_{\mathbf{X}}$ can be constructed by adopting a specific copula $C$, such that $F_{\mathbf{X}} = C(F_{X_1}, \ldots, F_{X_I})$. Different copula functions besides $C_{\Pi}$ and $C_{\Phi}$ exist, and they allow for various types of dependency between the univariate margins. This offers a lot of flexibility to choose a copula function that adequately captures the dependence structure of the data, and has the primary advantage of not sacrificing attractive properties of the margins $F_{X_i}$. For instance, we could have exponentially distributed margins showing a Gaussian-type of dependency structure when these margins are coupled by means of the Gaussian copula $C_{\Phi}$.

To allow for correlated errors in a structural equation model, we will apply copulas to the latent continuous variables $\tilde{Y}_{pi}$’s. Consider two of the underlying latent outcomes in the model, $\tilde{Y}_{p1}$ and $\tilde{Y}_{p2}$. Given their latent factor scores $\eta_p$ and the assumption of logistically distributed error terms $\varepsilon_{p1}$ and $\varepsilon_{p2}$, $\tilde{Y}_{p1}$ and $\tilde{Y}_{p2}$ each follow a univariate logistic distribution for the person $p$, $F_{\tilde{Y}_{p1}|\eta_p}(\tilde{y}_{p1})$ and $F_{\tilde{Y}_{p2}|\eta_p}(\tilde{y}_{p2})$. 

-96-
In the traditional structural equation model as presented earlier, conditional independence is assumed, which gives rise to the multivariate joint distribution:

\[ F_{\tilde{Y}_p|\eta_p}(\tilde{y}_{p1}, \tilde{y}_{p2}) = \prod_{i=1}^{2} F_{\tilde{Y}_{pi}|\eta_p}(\tilde{y}_{pi}), \]

which is now recognized as the earlier introduced product copula \( C_{\Pi} \)

\[ F_{\tilde{Y}_p|\eta_p}(\tilde{y}_{p1}, \tilde{y}_{p2}) = C_{\Pi} \left( F_{\tilde{Y}_{p1}|\eta_p}(\tilde{y}_{p1}), F_{\tilde{Y}_{p2}|\eta_p}(\tilde{y}_{p2}) \right). \]

Accounting for correlated errors can then be done by applying to these univariate margins a copula function \( C \) that allows for a suitable residual dependency structure:

\[ F_{\tilde{Y}_p|\eta_p}(\tilde{y}_{p1}, \tilde{y}_{p2}) = C \left( F_{\tilde{Y}_{p1}|\eta_p}(\tilde{y}_{p1}), F_{\tilde{Y}_{p2}|\eta_p}(\tilde{y}_{p2}); \delta \right), \]

where \( \delta \) refers to the parameter controlling the dependence (e.g., in the Gaussian copula, \( \delta \) equals \( \rho \), the correlation). Examples of other copula functions \( C \) will be discussed later.

This approach is illustrated graphically in Figure 3.3. In the upper-left panel one sees the regular conditional independence case with univariate logistic margins \( F_{\tilde{Y}_{p1}|\eta_p} \) and \( F_{\tilde{Y}_{p2}|\eta_p} \), but the other three panels show three different copulas and the density contour lines clearly show a pronounced and distinct dependence structure. Note that in Figure 3.3, it is not the bivariate distribution function that is shown, but the corresponding bivariate density function.
Figure 3.3: Examples of copula constructed bivariate densities with univariate logistic margins.

(a) Independence copula

(b) Frank copula $\delta = 4.16$

(c) Cook-Johnson copula $\delta = 1.33$

(d) Gumbel-Hougaard copula $\delta = 1.67$

Note. The solid and horizontal lines refer to the category boundaries.
From Figure 3.3, it can also be illustrated how on the basis of the latent variables $\tilde{Y}_{p1}$ and $\tilde{Y}_{p2}$, one may calculate the joint probability distribution for the ordinal outcomes $Y_{p1}$ and $Y_{p2}$. To find these joint probabilities, the volumes under the joint density delineated by the intersections of the thresholds can be computed as follows:

$$\Pr (Y_{p1} = q_1, Y_{p2} = q_2 | \eta_p) =$$

$$\sum_{k_1=0}^{1} \sum_{k_2=0}^{1} (-1)^{k_1+k_2} C \left( F_{\tilde{Y}_{p1}}|\eta_p (\tau_{1,q1+k_1}), F_{\tilde{Y}_{p2}}|\eta_p (\tau_{2,q2+k_2}); \delta \right).$$

where $F_{\tilde{Y}_{pi}}|\eta_p (\tau_{i,q}) = 1 - G \left[ -\frac{1}{\sqrt{\psi_{i1}}} (\tau_{i,q+1} - \sum_{d=1}^{D} \lambda_{id}\eta_{pd}) \right]$. In the previous paragraphs, we have illustrated how the copula approach can be used to model residual dependencies for two indicators but this approach can be generalized to any number of items. In practice, the indicators $Y_{pi} (i = 1, \ldots, I)$ will be divided in $S$ disjoint subsets $J_s$ of size $I_s$ according to their reference to a specific context or situation $s$ in the personality questionnaire. The joint probability of the outcomes in a context-specific subset $J_s$ will then be modeled by means of a copula-constructed joint distribution based upon the margins of the SEM and a specific copula function $C_s$, while conditional independence is assumed to hold between subsets, such that the joint probability model for the whole outcome vector $Y_p$ becomes:

$$\Pr(Y_p = y_p | \eta_p) = \prod_{s=1}^{S} \Pr (Y_{pi} = y_{pi}, i \in J_s | \eta_p)$$

where $\Pr (Y_{pi} = y_{pi}, i \in J_s | \eta_p)$ is the joint probability derived from the joint cumulative distribution distribution function $C_s(F_{\tilde{Y}_{pi}}|\eta_p (\tilde{y}_{pi}), i \in J_s; \delta)$ for the context-specific indicator subset $J_s$.

With respect to estimation, we still use marginal maximum likelihood. The only difference is that in the calculation of the loglikelihood, we now need to compute the joint outcome probabilities for some subsets of indicators.
3.4.3 Three copula functions

Besides the product copula, three other copula functions will be used in this paper, to be defined below\(^4\).

1. Frank copula (Frank, 1979) Frank copula is defined as:

\[
C (U_1, \ldots, U_{I_s}; \delta) = \frac{-1}{\delta} \log \left( 1 - \frac{\prod_{i=1}^{I_s}(1 - \exp(-\delta U_i))}{\prod_{i=1}^{I_s-1}(1 - \exp(-\delta))} \right)
\]

where \(\delta\) is restricted to the interval \([0, \infty[\). As \(\delta\) approaches 0, the margins become independent (and hence \(C \to C_{\Pi}\)). As \(\delta\) approaches \(\infty\), the copula approximates the upper bound of association between the set of random variables.


\[
C (U_1, \ldots, U_{I_s}; \delta) = \left( \sum_{i=1}^{I_s} U_i^{-\delta} - I_s + 1 \right)^{-\frac{1}{\delta}}
\]

with \(\delta\) restricted to the interval \([0, \infty[\). As \(\delta\) approaches 0, the margins become independent \((C_{\Pi})\). As \(\delta\) approaches \(\infty\), the copula approximates the upper bound of association.

3. Gumbel-Hougaard copula (Gumbel, 1960; Hougaard, 1986) The Gumbel-Hougaard copula is as follows:

\[
C (U_1, \ldots, U_{I_s}; \delta) = \exp \left( - \left[ \sum_{i=1}^{I_s} (- \log(U_i))^{\frac{1}{\delta}} \right] \right)
\]

with \(\delta\) restricted to the interval \([1, \infty[\). The boundary values of this interval, 1 and \(\infty\), correspond to \(C_{\Pi}\) and the maximum association, respectively.

\(^4\)Note that we will not consider the Gaussian copula anymore because it still involves intractable integrals that have to be approximated leading to a slowdown in the computation (although fast and accurate routines exist for the bivariate case, this is less so for the multivariate case).
In contrast to the Gaussian copula, all three copulas described here have a relatively simple and closed-form expression. In addition, the functions are parsimonious with only one dependency parameter $\delta$ that increases with increasing dependency, the simple structure of exchangeability applies, and they are bounded below by the regular independence case $C_H$ and above by absolute positive dependence given fixed margins, i.e., the Fréchet-Hoeffding upper bound $C_M$ (Fréchet, 1951; Hoeffding, 1940)

$$C_M(U_1, \ldots, U_{I_s}) = \min(U_1, \ldots, U_{I_s}).$$

Together, they capture a quite wide spectrum of residual dependency structures, while still allowing us to keep working in the familiar structural equation framework.

On the level of the underlying continuous outcomes $\tilde{Y}_{pi} | \eta_p, i \in J_s$, the imposed dependency structure by each copula function is clearly visible. Frank copula exhibits centralized and symmetric dependency, whereas the two other copulas show a form of tail dependence. Cook-Johnson copula exhibits strong lower tail dependence. Hence, this copula corresponds to a situation where the outcomes $\tilde{Y}_{pi} | \eta_p, i \in J_s$, are known to be highly concordant at low values, but somewhat less concordant at high values. In contrast, Gumbel-Hougaard copula exhibits strong upper tail dependence. Hence, Gumbel-Hougaard copula corresponds to a situation where the outcomes $\tilde{Y}_{pi}$ are known to be highly concordant at high values, but somewhat less concordant at low values (for a more formal treatment of tail dependence, we direct the reader to Joe, 1997). These patterns can also be observed graphically in Figure 3.3. It can be deduced that these tail dependence properties imply that compared to the rather symmetrical Frank copula, Cook-Johnson copula will show a bit more residual dependence for low concordant values, while the opposite is the case for Gumbel-Hougaard copula (more residual dependence for high concordant values).
Dependencies might not only differ across categories, but also across levels of the latent factors $\eta_{pd}$. For example, the residual association between a couple of indicators within a situation might be much stronger for generally aggressive persons than for non-aggressive persons. Notice that our copula approach works on the conditional SEM margins for ordinal data which implies that the exact influence of the adopted copula function is differential according to $\eta_p$, the person’s latent factor scores. As such, different copula functions can accommodate different relationships with the latent traits. The tail dependence property again plays an important role here. For instance, the residual dependence between $Y_{p1}$ and $Y_{p2}$ increases with increasing values of the latent factors $\eta_{pd}$ for Cook-Johnson copula, while the opposit occurs for Gumbel-Hougaard copula (residual dependence increases with decreasing latent factor scores). Due to its symmetrical shape, Frank copula has stronger association around the center of the latent continuum $\eta_p$ and location of the thresholds $\beta_{ic}$. A visual thought experiment in which one shifts the distributions in Figure 3.3 holding the thresholds fixed, and looks at the changes in the covered surface for the different delineated rectangles (which reflect the probability of a specific joint outcome pattern) can shed light to these characteristics.

On the continuous latent level of $\tilde{Y}_p$, the dependency parameter $\delta$ of each copula can be translated to a more familiar association measure as Kendall’s $\tau$ rank correlation, which is formalized as the probability of concordance minus the probability of discordance (Nelsen, 1999). Table 3.2 shows the resulting expressions for the copula functions considered here. This derived measure can be used as a rough indicator of the general strength of association, but it should be kept in mind that it does not fully apply to the discrete case due to the larger presence of ties.
Table 3.2: Copula dependency parameter $\delta$ translated to Kendall’s $\tau$.

<table>
<thead>
<tr>
<th>Copula</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>$1 + 4(D_1(\delta) - 1)/\delta$</td>
</tr>
<tr>
<td>Cook-Johnson</td>
<td>$\delta/(2 + \delta)$</td>
</tr>
<tr>
<td>Gumbel-Hougaard</td>
<td>$1 - 1/\delta$</td>
</tr>
</tbody>
</table>

Note. $D_1$ is the first Debye function, where $D_1(x) = \int_0^x \frac{t^{x-1}}{\exp(t)} dt$.

3.5 Application: Part II

So far the specific contextualized structure of the aggression inventory was of no concern in our model analyses. The question now is, what does happen to our model results and inferences when we do account for context by adopting the above-introduced copula functions in the ordinal SEM models. For instance, does accounting for the associations from the same situations change, in a qualitative or quantitative way, the directed relations between the different trait components?

Including correlated measurement error in the model between indicators relating to the same situation allows to test whether the association between the indicators is sufficiently explained by the common latent factors, in our case the personality traits, or not. Indicators belonging to the same situation will be coupled by means of a copula function, resulting in three subsets $J_1 = \{1, 2, 3\}$, $J_2 = \{4, 5, 6\}$, and $J_3 = \{7, 8, 9\}$. In our model for the contextualized aggression inventory, the degree of dependency imposed by a specific copula $C$ given its parameter $\delta$, can be seen as reflecting the consistency of response within a situation. It is an indication of additional dependence between responses given a person’s relevant traits.

Table 3.3 shows the model fit results of all previously discussed models as well as their competing copula alternatives. A first observation that can be made is that accounting for the contextualized structure leads to a strong increase in the log likelihood, such that the copula alternative is always retained in favour of the regular SEM model under the conditional independence assumption. This indicates that the contextualized structure of the aggression inventory is indeed reflected
in the dependence structure of the data. Furthermore, the best fitting set of copulas according to the model’s log likelihood (i.e., respectively Frank, Cook-Johnson, Frank) remained constant across the three SEM models, indicating the stability of the found additional dependence due to the situational contexts.

Table 3.3: Conditional independence and copula versions of the models of interest.

<table>
<thead>
<tr>
<th>Model</th>
<th>Conditional Independence</th>
<th>Copula</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LogL</td>
<td>♯ parameters</td>
</tr>
<tr>
<td>unidimensional</td>
<td>$-2380.9$</td>
<td>27</td>
</tr>
<tr>
<td>strict mediation</td>
<td>$-2377.9$</td>
<td>29</td>
</tr>
<tr>
<td>partial mediation</td>
<td>$-2377.5$</td>
<td>30</td>
</tr>
</tbody>
</table>

Note. The best fitting copula combination always remained Frank, Cook-Johnson, and Frank.

A second important observation that can be made, is that accounting for the contextualized structure leads to a qualitative change in the chosen model. The difference between the unidimensional model and the ABC models is now a lot more outspoken, in favour of the latter ($\chi^2 = 51.4, df = 2, p < 0.0001$). Among the two ABC models the model of choice according to a likelihood ratio test is the partial mediation model ($\chi^2 = 5, df = 1, p = 0.0253$). This implies that the results contradict the mediation process theory as suggested by Buss and Perry (1992), and are in line with the more general ABC model theory. The related, but separable, factors of angry affect, behavioral aggression, and hostile cognition then reflect the fact that all forms of aggression do not share similar manifestations and consequences. In Figure 3.4 the resulting standardized regression coefficients $\Gamma$ for both ABC models are indicated in path diagrams, and in Table 3.4, all parameter estimates for the preferred partial mediation model are listed.
Table 3.4: An ABC model of aggression: partial mediation copula model

<table>
<thead>
<tr>
<th>d</th>
<th>s</th>
<th>i</th>
<th>α_i</th>
<th>β_{i1}</th>
<th>β_{i2}</th>
<th>λ_{id}</th>
<th>τ_{i1}</th>
<th>τ_{i2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1.359</td>
<td>(0.312)</td>
<td>-4.707</td>
<td>(0.535)</td>
<td>2.574</td>
<td>(0.410)</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>1.251</td>
<td>(0.223)</td>
<td>-1.179</td>
<td>(0.167)</td>
<td>2.286</td>
<td>(0.212)</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>3</td>
<td>2.565</td>
<td>(1.049)</td>
<td>0.378</td>
<td>(0.214)</td>
<td>3.289</td>
<td>(1.042)</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>4</td>
<td>1.727</td>
<td>(0.332)</td>
<td>-1.344</td>
<td>(0.211)</td>
<td>2.862</td>
<td>(0.332)</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>5</td>
<td>1.903</td>
<td>(0.411)</td>
<td>1.150</td>
<td>(0.231)</td>
<td>1.971</td>
<td>(0.305)</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>6</td>
<td>0.834</td>
<td>(0.267)</td>
<td>2.271</td>
<td>(0.234)</td>
<td>1.646</td>
<td>(0.313)</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>7</td>
<td>1.257</td>
<td>(0.246)</td>
<td>-1.783</td>
<td>(0.210)</td>
<td>2.371</td>
<td>(0.233)</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>8</td>
<td>1.419</td>
<td>(0.273)</td>
<td>-0.157</td>
<td>(0.148)</td>
<td>2.086</td>
<td>(0.230)</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>9</td>
<td>0.832</td>
<td>(0.238)</td>
<td>1.959</td>
<td>(0.207)</td>
<td>1.212</td>
<td>(0.211)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p</th>
<th>Γ</th>
<th>J_2</th>
<th>copula</th>
<th>δ</th>
<th>p</th>
<th>τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>p_{BA}</td>
<td>0.706</td>
<td>(0.149)</td>
<td>0.000</td>
<td>0.559</td>
<td>{1, 2, 3}</td>
<td>Frank</td>
</tr>
<tr>
<td>p_{BC}</td>
<td>0.290</td>
<td>(0.126)</td>
<td>0.021</td>
<td>0.218</td>
<td>{4, 5, 6}</td>
<td>Cook-Johnson</td>
</tr>
<tr>
<td>p_{AC}</td>
<td>0.334</td>
<td>(0.133)</td>
<td>0.012</td>
<td>0.317</td>
<td>{7, 8, 9}</td>
<td>Frank</td>
</tr>
</tbody>
</table>

Note. The parameter $\beta_{i+} = \beta_{i2} - \beta_{i1}$.

The parameter $g_{YX}$ is the unstandardized coefficient of the regression of $Y$ on $X$ ($\Gamma$ is the standardized version).

Standard errors are given between parentheses and $p$-values result from the Wald test.
The threshold part of the conditional independence model in Table 3.1 and the copula model in Table 3.4 are rather similar. This observation could be anticipated, because copula functions are a marginal modeling approach (see e.g., Liang, Zeger, & Qaqish, 1992). Looking at the pattern and values of the thresholds $\tau_1$ and $\tau_2$, it is noticeable that the most plausible reaction (i.e., with the lowest threshold values) is anger (A), followed by behavioral aggression (B), and hostility (C), and this for all situations and category intersections. Situation 1 (gossip) is the most salient context which is the most likely to elicit aggressive reactions, followed by situation 3 (tent), and situation 2 (terrace).

The factor loading part of the conditional independence model in Table 3.1 and the copula model in Table 3.4 do show some small differences. Again this observation could be anticipated, because the factor loadings are a reflection of the model-implied underlying dependence structure, changes to this dependence structure by the addition of copulas to capture context effects will also show themselves here. The most noticeable difference is the change in factor loading values for the indicators of hostility (C), in which $\lambda_3$ is corrected upwards and $\lambda_6$ and $\lambda_9$ are both corrected downwards. The loadings on the other factors (A and B) remain relatively unchanged.
The structural part of the conditional independence model in Table 3.1 and the copula model in Table 3.4 do not only change in a qualitative way, from a strict mediation model to a partial mediation model, but also show some quantitative changes. The effect of anger (A) on behavioral aggression (B) is less strong, as is visible in the decreased value of the standardized regression coefficient (i.e., from $\Gamma_{BA} = 0.848$ to $\Gamma_{BA} = 0.559$). The effect of hostility (C) on behavioral aggression (B) is, although still small, now considered to be substantial. The effect of hostility (C) on anger (A) is less strong, as visible in the decreased value of the standardized regression coefficient (i.e., from $\Gamma_{BA} = 0.719$ to $\Gamma_{BA} = 0.317$). After the introduction of the copula functions the component’s mutual interrelations are now corrected for with regard to the association accounted for by the contextual influences. This resembles the situation in multiple regression when another predictor is added into a regression equation.

When we look at the dependence parameter $\delta$ of all 3 copulas they are significantly different from zero, and hence indicating that the dependence structure deviates from the regular conditional independence structure. A rough inspection of the Kendall’s $\tau$ values, points at situation 2 for being the situation where the most residual dependence is present, and situation 1 for being the situation where the least residual dependence is presence. Situation 2 is in fact the situation in which the most public control or constraint is present. Most people tend to react rather calm when waiting a long time on a tavern’s terrace, which can also be seen in the rather large estimated thresholds $\tau_1$ and $\tau_2$ for this situation’s indicators ($i = 4, 5, 6$) compared to the other situations. It is not unlikely then that a large part of people’s reactions is determined by the context or environment here, and not only by their personal traits. In contrast, situation 1 involves a more up-close and personal situation (people gossiping about you). In this type of situation contextual influence is likely to be less pronounced. Frank copula fitted two out of three situations, but not for situation 2. The Cook-Johnson copula for sit-
uation 2 indicates that the residual dependence between the responses is larger here for more aggressive people. Given the public nature of situation 2 (waiting for your drink), it is indeed not unlikely that this particular context will elicit either no reaction or in contrast the whole set of aggressive responses, and that this context-effect is even stronger for people who are in general already more aggressive.

All in all, we must conclude that in our case the difference between either accounting for the residual dependence due to the context, or not, gives rise to both qualitative as well as quantitative changes in related model inferences. In other words, context does matter in personality psychology. Furthermore, the copula approach also managed to provide extra information that allowed us to complete the overall picture of the personality study on aggression.

3.6 Discussion

Copula functions are a new and promising tool for multivariate modeling, also in conjunction with SEM. The primary advantage is that copula functions offer parsimonious simple structure yet flexible dependency relations, without leaving the attractive framework of the existing structural equation methodology. This extra freedom or flexibility copula functions can offer, is a positive thing, but at the same time also a challenge as it broadens the spectrum of modeling possibilities in the structural equation domain. Furthermore, it implies thinking out of the box when considering association between variables, because one is not necessarily limited to work within the default framework of multivariate normal distributions and linear correlation.

It should be noted that so far the copula approach presented here is limited by the requirement of having disjoint subsets of indicators. It might be of interest to investigate whether it is possible to extend this approach to overlapping subsets. Another point of interest is the development of specific modification indices and goodness-of-fit procedures
for the case of copula extended structural equation models. The SEM boundaries to model complexity are not yet recognized. Whether further extensions are merely a matter of technical computation problems or real data-model issues remains an open question, but with the introduction of this general multivariate modeling tool that copulas are, we hope to contribute to ongoing developments in the structural equation domain.
References


Correlated error in SEM


-111-


Chapter 4

Model selection in copula IRT models

Abstract: Modern measurement models of item response theory (IRT) are instantiations of the nonlinear mixed model, in which conditional independence is a key assumption. Recently, a copula IRT model was introduced to account for violations of this assumption within specific item subsets, in which the additional association unmodeled by the latent trait is captured by a specific copula function. A priori information with regard to the choice of this copula function is mostly lacking. In this paper the feasibility of identifying different association structures is investigated within these copula IRT models. Besides evaluating the performance of traditional model selection criteria, relatively unknown approaches such as Minimum Description Length (MDL) and model weights through continuous model expansion are explored.
4.1 Introduction

In educational and psychological measurement, interest often goes to describing and explaining the dependencies between responses on items of a test. To study individual differences in test performance and to construct a measurement scale where both persons as well as items have a position, latent trait models under the label of item response theory (IRT) are applied in these fields.

A well-known item response model for binary responses is the two-parameter logistic (2PL) model (Birnbaum, 1968). Suppose a person $p$ responds to item $j$ and define a random variable $Y_{pj}$ such that $Y_{pj} = 1$ if person $p$ responds correctly to item $i$ and $Y_{pj} = 0$ otherwise. The 2PL item response model states that the probability of a correct response on an item, given the person’s position on the latent trait $\theta_p$, is

$$\Pr(Y_{pj} = 1|\theta_p) = \frac{\exp(\alpha_j(\theta_p - \beta_j))}{1 + \exp(\alpha_j(\theta_p - \beta_j))},$$

where $\alpha_j$ and $\beta_j$ are called the item discrimination and item difficulty, respectively. When viewed as a function of $\theta_p$, this conditional probability function is known as the item characteristic curve in IRT, and clearly is a monotone increasing function, meaning that people with a larger proficiency of the latent trait (i.e., higher values of $\theta_p$) will have more chance of correctly solving the item. Figure 4.1 illustrates the function of the item parameters. The parameter $\beta_j$ indicates the position on the latent scale where the probability of correctly answering the item is a 0.5. For a person $p$ with a given ability $\theta_p$, the probability of answering correctly on items with a smaller $\beta_j$ will be higher than on items with a larger item difficulty $\beta_j$ (given equal $\alpha_j$). The parameter $\alpha_j$ controls the steepness of the logistic function, where a steeper curve leads to a better discrimination between persons with nearby positions on the latent scale. If $\alpha_j = \alpha$, for all items $j$ ($j = 1, \ldots, J$), the 2PL model reduces to the familiar Rasch model (Rasch, 1960).
Under the assumption of conditional or local stochastic independence (LSI) the joint probability of a person’s item response pattern on the full test $Y_p$ can be formulated as a product of the conditional probabilities of the individual items:

$$\Pr(Y_p = y_p | \theta_p) = \prod_{j=1}^{J} \Pr(Y_{pj} = y_{pj} | \theta_p).$$

Agresti, Booth, Hobert, and Caffo (2000) and Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) showed that when assuming a distribution for the latent trait $\theta_p$ this 2PL model, and other item response models, are formally equivalent to nonlinear mixed models (NLMM). This relates these IRT measurement models common to the social sciences, to the broad statistical literature of mixed models for repeated measurements (see e.g., (Diggle, Heagerty, Liang, & Zeger, 2002; Molenberghs & Verbeke, 2005; Fahrmeir & Tutz, 2001)). For instance, when assuming a standard normal distribution for the random intercept $\theta_p$, ...
the marginal probability of a response vector can then be expressed as:

$$Pr(Y_p = y_p) = \int \prod_{j=1}^J Pr(Y_{pj} = y_{pj}|\theta_p) \phi(\theta_p|\mu_\theta, \sigma^2_\theta) d\theta_p,$$

where $\phi(\theta_p|\mu_\theta, \sigma^2_\theta)$ is a normal density with mean $\mu_\theta$ set equal to 0 and variance $\sigma^2_\theta$ equal to 1 to fix the origin and scale of the latent variable, respectively (under the latter restriction, all $\alpha_j$’s can be estimated).

The assumption of LSI is a strong one because it implies that all associations between the responses of a person to the items can be explained by the latent proficiency. This assumption is violated whenever extra dependencies arise within specific subsets of items. Such extra dependencies are due to, for instance, the use of a common sub-theme or the same stimulus material. Failing to correctly account for the additional association in the data (over the proficiency) means that one’s model is misspecified and leads to biases both for fixed and random effects and their estimated standard errors, a phenomenon that is well-known in psychometrics (see e.g., Chen & Thissen, 1997; Junker, 1991; Yen, 1993. In item response theory, such violations are commonly referred to as local item dependencies (LID; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968).

LID is mostly not intended by the test developers. As such, often no a priori information is present about the nature of possible LID problems in a test. However, as being said, ignoring these problems is also no option because it may lead to distorted inferences. Determining the exact nature of the LID problem might help in locating its origin and understanding underlying processes that play a role, hereby making way for test improvement and understanding of the construct one is attempting to measure.

In past research, several approaches to accommodating LID have been proposed (see e.g., Hoskens & De Boeck, 1997; Verhelst & Glas, 1993; Wainer, Bradlow, & Wang, 2007). Recently, we developed a class of item response models that is able to take into account such LID prob-
lems by means of copula functions (Braeken, Tuerlinckx, & De Boeck, 2007). The specific nature of the modeled LID association structure is determined by the specific copula function that is adopted for the item subset showing LID. The major advantage of the copula IRT models is that the models for the probability correct on each separate item are still 2PL models. This property is called reproducibility and it is a major improvement over other type of LID models where the marginal probabilities change together with the association structure and become hard to interpret. However, a priori guidelines for choosing a specific copula function are lacking and therefore it becomes an issue. As such, a model (i.e., copula) selection issue surfaces in this approach to model LID, and one desires an effective objective criterion to aid in deciding which of the competing copula models to apply.

Handling the model selection issue in an appropriate way is important because in measurement settings an item response model is explicitly required to adequately fit a whole range of samples, as a test is obviously used more than once. Furthermore, it is also common place to have item response models that need to generalize towards a larger population than the sample, think of tests for clinical diagnosis or test results used to support public policy makers. However, often multiple types of item response models are available, competing theories need to be compared, or the absence of specific a priori information leaves open a broad range of model options and choices. In these cases many models can be considered, and a model selection issue arises. The issue is to select a meaningful statistical model that most closely resembles the true but unknown data generating process.

In this manuscript we would like to investigate tools for model selection in copula IRT models and to assess the feasibility of choosing the ‘true’, or at least most optimal, underlying copula model. Firstly, the copula IRT model is introduced, and it is shown that different copulas

\footnote{One exception is a hybrid-kernel approach by Ip (2002) which does have this reproducibility property, but has not yet found its way to practitioners in the field.}
imply different LID association structures, either unrelated to the main dependency source $\theta_p$ or indeed related to this latent dimension. Secondly, the model selection issue is discussed more in detail. Relatively new and unexplored model selection approaches as Minimum Description Length (MDL) and model weights through continuous model expansion, are introduced besides the more traditional criteria. Finally, a simulation study is set up to compare the different model selection criteria and to investigate their sensitivity to specific aspects of the data and choice of the copula. The paper concludes with a discussion.

4.2 Copula IRT model

To account for LID, the class of IRT models proposed in Braeken et al. (2007) makes use of copula functions. Copula functions are a new and so far unfamiliar tool in psychometrics, but basically comprehend the intuitive idea of defining a function that relates, or ‘couples’, $J$ univariate margins $F_{X_1}, \ldots, F_{X_J}$ such that a multivariate distribution function $F_X$ emerges with a specific dependency structure. This section intends to introduce the copula IRT model, starting with the very basics of copula theory with regard to their use in statistical modeling. A broader view on copulas as a general dependence concept and modeling tool is available in reference works by Nelsen (1999) and Joe (1997).

A $J$-dimensional copula is a function $C : [0,1]^J \rightarrow [0,1]$ with following properties:

1. For every vector $\mathbf{u} \in [0,1]^J$, $C(\mathbf{u})$ is increasing in each component $u_j$ with $j \in \{1, \ldots, J\}$.

2. For every vector $\mathbf{u} \in [0,1]^J$, $C(\mathbf{u}) = 0$ if at least one coordinate of the vector is 0 and $C(\mathbf{u}) = u_j$ if all the coordinates of the vector are equal to one except the $j$-th one.

3. For every $\mathbf{a}, \mathbf{b} \in [0,1]^J$ with $\forall j, a_j \leq b_j$, given a hypercube $B = [\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \ldots \times [a_J, b_J]$ whose vertices lie in the domain
of $C$, and with $V_C(B) \geq 0$. The volume $V_C(B)$ is defined as:

$$V_C(B) = \sum_{k_1=1}^{2} \ldots \sum_{k_J=1}^{2} (-1)^{k_1+\ldots+k_J} C \left( d_1^{(k_1)}, \ldots, d_J^{(k_J)} \right)$$

where $d_j^{(1)} = a_j$ and $d_j^{(2)} = b_j$ ($i = 1, \ldots, J$).

An important result from the theory of copulas is Sklar’s Theorem (Sklar, 1959). The theorem states that for any $J$-dimensional distribution function $F_X$ with univariate margins $F_{X_1}, \ldots, F_{X_J}$ there exists a copula function $C$ such that this multivariate distribution $F_X$ can be represented as a function of its margins through this copula: $F_X = C(F_{X_1}, \ldots, F_{X_J})$. In the continuous case, this copula reformulation of an existing multivariate distribution is unique. For discrete random variables, the result is a bit more complex. The cumulative distribution of a discrete variable is discontinuous and as such it does not map the variable to the entire $[0,1]$ interval. Thus, the copula $C$ does not need to be uniform over rectangles, meaning that the copula representation is not unique for a multivariate distribution of discrete random variables (for a detailed discussion see Joe, 1997). However, from a modeling perspective, this non-uniqueness problem does not create a problem, because the non-uniqueness does not imply non-identifiableness, as the copula $C$ is still uniquely determined on $\text{Ran}(F_{X_1}) \times \ldots \times \text{Ran}(F_{X_J})$, the range of the margins.

Furthermore, the converse of the theorem also holds. Thus based upon any kind of univariate margins $F_{X_1}, \ldots, F_{X_J}$, a multivariate (i.e., joint) distribution $F_X$ can be constructed by adopting a specific copula $C$, such that $F_X = C(F_{X_1}, \ldots, F_{X_J})$. Given the second property in the copula definition, it can easily be deduced that the univariate marginal distribution for $X_j$ equals $C(1, \ldots, 1, F_{X_j}, 1, \ldots, 1) = F_{X_j}$. Hence, in this way an association between the $J$ random variables is allowed through the copula $C$, while still preserving the original uni-
variate margins we started from. In the continuous case, the joint density of the multivariate distribution $F_X = C(F_{X_1}, \ldots, F_{X_J})$ can be derived as $f_C(X) = c(F_{X_1}, \ldots, F_{X_J}) f(X_1) \times \ldots \times f(X_J)$, where $c(F_{X_1}, \ldots, F_{X_J}) = \frac{\partial^J C(F_{X_1}, \ldots, F_{X_J})}{\partial F_{X_1} \cdots \partial F_{X_J}}$ and $f(X_j)$’s are the univariate densities. In the discrete case, the joint probability $\Pr_C(X)$ of the multivariate distribution $F_X = C(F_{X_1}, \ldots, F_{X_J})$ can be derived using a recursive formula based upon quadrant probabilities (see e.g., Property 3 and Mood, Graybill, & Boes, 1974). The index $C$ in $\Pr_C$ makes clear that the joint probability depends upon the assumed copula function $C$.

Making use of the above properties and Sklar’s theorem, we can account for LID in an item response model by formulating a more appropriate joint probability than under the regular LSI model. To make this more concrete, consider a small mathematics test ($J = 6$), in which a subset (denoted by $s$) consisting of the first two items refer to the same problematical case, and the four other items are separate questions. Moreover, assume that the responses on the subset $s$ will show extra association above the one accounted for by a person’s mathematical proficiency $\theta_p$. For example, if a person is able to understand the key issues in the presented case, it is more likely that correct responses will result from both items (and if not, an incorrect response on both). A model assuming LSI ignores this extra source of dependency and models the joint probability of the set of item responses given a person’s proficiency as the product of the above introduced univariate item margins

$$\Pr(Y_p = y_p | \theta_p) = \Pr(Y_{p1} = y_{p1} | \theta_p) \Pr(Y_{p2} = y_{p2} | \theta_p) \prod_{j=3}^J \Pr(Y_{pj} = y_{pj} | \theta_p).$$

In contrast, the idea behind the copula IRT models is to accommodate for this extra source of dependency by redefining the joint probability of the response vector as
Model selection in copula IRT models

\[
\Pr(Y_p = y_p | \theta_p) = \Pr_C(Y_{p1} = y_{p1}, Y_{p2} = y_{p2} | \theta_p) \prod_{j=3}^{J} \Pr(Y_{pj} = y_{pj} | \theta_p),
\]

in which \( \Pr_C(Y_{p1} = y_{p1}, Y_{p2} = y_{p2} | \theta_p) \) is the joint probability under the copula function (with univariate margins defined as in Equation 4.1). The copula function \( C \) accounts for the extra association (i.e., the LID) between items 1 and 2, beyond the association already induced by the common latent trait. Yet, LSI still holds between items not in the subset, and between the subset and these other items. Specific families of copula functions (that are able to capture different dependency patterns) will be introduced later on in this paper.

The distribution function \( F_{Y_{pj}|\theta_p}(y_{pj}) \) (with \( j = 1, 2 \)) is formulated as:

\[
F_{Y_{pj}|\theta_p}(y_{pj}) = \begin{cases} 
0 & \text{for } y_{pj} < 0 \\
\Pr(Y_{pj} = 0 | \theta_p) & \text{for } y_{pj} = 0 \\
1 & \text{for } y_{pj} = 1.
\end{cases}
\]

Hence, the joint probability can be computed by the following recursive formula:

\[
\Pr_C(Y_{p1} = y_{p1}, Y_{p2} = y_{p2} | \theta_p) = \\
\Pr_C(d_{p1}^{(1)} < Y_{p1} \leq d_{p1}^{(2)}, d_{p2}^{(1)} < Y_{p2} \leq d_{p2}^{(2)} | \theta_p) \quad (4.2)
\]

= \( \sum_{k_1=1}^{2} \sum_{k_2=1}^{2} (-1)^{k_1+k_2} C \left( F_{Y_{p1}|\theta_p}(d_{p1}^{(k_1)}), F_{Y_{p2}|\theta_p}(d_{p2}^{(k_2)}) \right) \).

where for \( y_{pj} = 1 \), \( d_{pj}^{(1)} = 0 \) and \( d_{pj}^{(2)} = \infty \), and for \( y_{pj} = 0 \), \( d_{pj}^{(1)} = -\infty \) and \( d_{pj}^{(2)} = 0. \)
The equations above simplify for a \( \{0, 0\} \) response pattern to:

\[
\Pr_C(Y_{p1} = 0, Y_{p2} = 0 | \theta_p) =
\]

\[
= C \left( F_{Y_{p1}|\theta_p}(-\infty), F_{Y_{p2}|\theta_p}(-\infty) \right) - C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(-\infty) \right)
\]

\[
- C \left( F_{Y_{p1}|\theta_p}(-\infty), F_{Y_{p2}|\theta_p}(0) \right) + C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(0) \right)
\]

\[
= 0 - 0 - 0 + C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(0) \right)
\]

The other probabilities are then:

\[
\Pr_C(Y_{p1} = 1, Y_{p2} = 1 | \theta_p) = 1 - \sum_{j=1}^{2} F_{Y_{pj}|\theta_p}(0) + C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(0) \right)
\]

\[
\Pr_C(Y_{p1} = 1, Y_{p2} = 0 | \theta_p) = F_{Y_{p2}|\theta_p}(0) - C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(0) \right)
\]

\[
\Pr_C(Y_{p1} = 0, Y_{p2} = 1 | \theta_p) = F_{Y_{p1}|\theta_p}(0) - C \left( F_{Y_{p1}|\theta_p}(0), F_{Y_{p2}|\theta_p}(0) \right)
\]

Note that the joint probability under a LSI model can be reformulated under a copula representation as \( \Pr_{\Pi}(Y_{p1} = y_{p1}, Y_{p2} = y_{p2} | \theta_p) \), where \( \Pi \left( F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2}|\theta_p}(y_{p2}) \right) = \prod_{j=1,2} F_{Y_{pj}|\theta_p}(y_{pj}) \) is the independence copula. Furthermore, the upper bound of positive dependency for any joint distribution of given margins, the Fréchet-Hoeffding upper bound (Fréchet, 1951; Hoeffding, 1940) is also a copula function and defined as \( M \left( F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2}|\theta_p}(y_{p1}) \right) = \min(F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2|\theta_p}(y_{p1})}). \)

### 4.2.1 LID structures and copulas

Compared to most alternative modeling approaches for LID in IRT (see e.g., (Hoskens & De Boeck, 1997; Verhelst & Glas, 1993; Wainer et al., 2007; Tuerlinckx & De Boeck, 2004)), this copula IRT model class has the advantage of following a marginal modeling approach in defining the joint response probabilities, hereby preserving the chosen item response model for the individual item responses (i.e., the margins). In practice this property, sometimes called ‘reproducibility’ (Fitzmaurice, Laird, & Rotnitzky, 1993; Ip, 2002), means that it allows the application of most well known item response models without changing anything about their
model formulation, nor interpretation of the parameters. The nature of the modeled LID is determined by the specific copula function adopted to take into account the LID issue. In this study, three specific copula functions will be used to model LID with only a single parameter \( \delta \) capturing the whole range of positive LID (from independence to maximum positive dependence).

Below we present the three types of copula functions. To clarify notation, the index \( j_s \) functions as the subset-item index, running from 1 to \( J_s \) the total number of items in the LID subset \( s \). We will define the copula functions for the case of local item dependencies between two items (1 and 2, \( J_s = 2 \)), but the generalization to more items is straightforward.

1. **Frank copula** (Frank, 1979) can be written as follows:

\[
C \left( F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2}|\theta_p}(y_{p2}) \right) = \frac{-1}{\delta} \log \left( 1 - \prod_{j_s=1}^{J_s} (1 - \exp(-\delta F_{Y_{pjs}|\theta_p}(y_{pj})) \right) \prod_{j_s=1}^{J_s-1} (1 - \exp(\delta)) \right).
\]

For \( J_s = 2, \delta \in \mathbb{R}/\{0\} \), negative values indicate negative association, positive values positive association; for \( J_s > 2 \) the parameter range is constrained to the range of positive association. If \( \delta \to 0 \) then \( C \to \Pi \) (i.e., independence), and if \( \delta \to \infty \), \( C \to M \) (i.e., the Fréchet-Hoeffding upper bound meaning absolute positive dependency).

To illustrate the characteristic LID structure this specific copula implies, the log odds ratio (conditional on \( \theta_p \)) for items 1 and 2 is computed,

\[
\log(\text{OR}(\theta_p)) = \log \left( \frac{\Pr_C(Y_{p1} = 1, Y_{p2} = 1|\theta_p) \Pr_C(Y_{p1} = 0, Y_{p2} = 0|\theta_p)}{\Pr_C(Y_{p1} = 1, Y_{p2} = 0|\theta_p) \Pr_C(Y_{p1} = 0, Y_{p2} = 1|\theta_p)} \right),
\]

and this for several values of its association parameter \( \delta \), and for \( \theta_p \) ranging from \(-2\) to \(2\). For ease of demonstration, a Rasch model formulation was chosen for the item characteristic curves in which the difficulty parameters \( \beta_1 \) and \( \beta_2 \) were fixed to 0, such that the log odds ratio was only a function of the copula’s association parameter \( \delta \) and the latent
trait $\theta_p$. The result is shown in the upper-panel of Figure 4.2. Notice that when the value of the copula parameter $\delta$ rises, the log odds ratio also increases, indicating $\delta$’s function as an association measure; and that the odds ratio remains relatively constant over $\theta_p$. In other words, Frank copula gives rise to dimensional stable LID.

2. Cook-Johnson copula (Clayton, 1978; Cook & Johnson, 1981) looks like:

$$C \left( F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2}|\theta_p}(y_{p2}) \right) = $$

$$\left( \frac{1}{\delta} \sum_{J_s=1}^{J_s} F_{Y_{pjs}|\theta_p}(y_{pj_s}) - J_s + 1 \right)^{-\frac{1}{\delta}}$$

With $\delta > 0$, higher values of the copula parameter indicate stronger positive association. If $\delta \to 0$ then $C \to \Pi$, and if $\delta \to \infty$ then $C \to M$. A similar characterization of the LID structure this copula implies, is shown in the middle-panel of Figure 4.2. Notice that when the value of the copula parameter $\delta$ rises, the log odds ratio also increases, indicating $\delta$’s function as an association measure; and that the odds ratio increases with increasing $\theta_p$. In other words, Cook-Johnson copula gives rise to dimensional increasing LID.

3. Gumbel-Hougaard copula (Gumbel, 1960; Hougaard, 1986) is defined as:

$$C \left( F_{Y_{p1}|\theta_p}(y_{p1}), F_{Y_{p2}|\theta_p}(y_{p2}) \right) = $$

$$\exp \left( - \left[ \sum_{J_s=1}^{J_s} \log(F_{Y_{pjs}|\theta_p}(y_{pj_s})) \right]^{\frac{1}{\delta}} \right)$$

With $\delta > 1$, higher values of the copula parameter indicate stronger positive association. If $\delta \to 1$ then $C \to \Pi$, and if $\delta \to \infty$ then $C \to M$. The odds ratio characterization of the LID structure this copula implies, is shown in the bottom-panel of Figure 4.2. Notice that when the value of the copula parameter $\delta$ rises, the log odds ratio also increases, indicating
δ’s function as an association measure; and that the odds ratio decreases with increasing $\theta_p$. In other words, Gumbel-Hougaard copula gives rise to dimensional decreasing LID.

Figure 4.2: Log odds ratio plots for 3 copula functions.

4.3 Model selection for copula IRT models

Thus, competing models for the LID association structure, as defined by different copula functions, exist and one desires an effective objective criterion to aid in deciding which copula function to apply. This section will discuss the model (i.e., copula) selection issue for copula IRT models in more detail.

In our model selection issue, two cases can be distinguished. First, one may wish to select between an independence structure and a LID structure implied by a particular copula function. A straightforward solution exists because the independence copula is a special case of all copulas considered in this paper. Compared to an independence copula, the only additional parameter in the copula functions we use is the
association parameter $\delta$ that defines the degree of LID within an item subset. Hence, we may work with likelihood ratio tests to compare nested models and the usual asymptotic results apply. Second, one may wish to compare different copulas directly and then things get more complicated. The copula functions presented above (and thus the models using them), only differ in the implied association structure but they do not have a nested relation. Hence, the key question here is how to balance goodness-of-fit and model complexity for these non-nested models, and make the most appropriate choice in selecting a copula function for a problem at hand.

### 4.3.1 Traditional model selection criteria

In cases where one can not rely on a likelihood ratio test between two models, because they do not share a nested relationship some traditional model selection criteria for non-nested models have been proposed to solve this issue. The most popular and wide-applied of these are Akaike’s Information Criterion (AIC; Akaike, 1973) and Schwarz’s Bayes Information Criterion (BIC; Schwarz, 1978).

AIC is defined as minus two times the sum of the log likelihood evaluated at the maximum likelihood estimators $\hat{\xi}$ (and given the data $Y$) and the number of parameters $k$ (length of parameter vector $\xi$) in the model: $AIC = -2LL(\xi|Y) + 2k$. It was developed as an approximation to the maximum cross-entropy, or equivalently the minimal Kullback-Leibler divergence between the true probability distribution of the data generating process and the probability distribution proposed by the candidate model. This distance can be estimated by the maximal average log likelihood of the data given the candidate model. However, it is well known that such a method produces estimation bias due to the fact the same set of observations is used to estimate the model parameters, which in turn estimate the cross-entropy. Under certain regularity conditions (see e.g., Akaike, 1973; Bozdogan, 2000) it can be shown that this bias is well approximated by the model dimension, i.e., the number
of parameters. AIC can also be regarded as the theoretical asymptoti-
cal equivalent of leave-one-out cross-validation (Stone, 1977; De Leeuw,
1989).

BIC is a relative fit measure, defined as minus 2 times the log likeli-
hood plus the natural logarithm of the sample size \( P \) multiplied by the
number of parameters \( k \) in the model: \( BIC = -2LL(\hat{\xi}|Y) + \log(P)k \).
Although BIC looks quite similar to AIC, it was developed from other
background premises. The BIC is an asymptotic approximation to the
Bayes factor (Raftery, 1995). In Bayesian model selection, a model \( M \)
is evaluated based upon their integrated likelihood defined as \( \Pr(Y|M) = \int p(Y|\xi,M)h(\xi|M)d\xi \), where \( p(Y|\xi,M) \) is the regular model likelihood
(also denoted as \( L(\xi|M,Y) \)) and the second part refers to the prior dis-
tribution on the model parameter vector \( \xi \). The ratio of the integrated
likelihoods of two models is the Bayes factor. Under the assumption
of equal prior odds of the two models, the posterior odds of one model
compared to the other given the data reduces to this Bayes factor. This
implies that the model with the largest integrated likelihood is the most
likely to be true given the data. The BIC is the asymptotic approxi-
mation of this integrated likelihood of a model. The BIC attempts to
correct the over-fitting tendency of the AIC criterion.

A problem with these traditional model selection criteria applied to
our model selection issue is that our different copulas have the same
number of parameters. This would imply that all three copula models
under consideration are equally complex , at least according to the AIC
and BIC (both of them give the same result). However, one might won-
der whether this is really the case, model dimension being merely one
aspect of a model’s complexity, and whether one should not only take
into account the number of parameters, but also the functional complex-
ity that might differ from copula to copula. Furthermore, Myung (2000)
showed that when model complexity is not properly taken into account,
this leads to the choice of unnecessarily complex models that overfit
the data. Considering the clearly obvious differences in formulas for the
copulas, and in their corresponding LID association structure, functional
differences might as well be an influential factor, however, a natural way
of quantifying functional differences is lacking. In what follows, model
complexity is further explored and two fresh alternative model selection
approaches are introduced that might be more appropriately designed
for model selection in a copula IRT context than traditional model se-
lection criteria as AIC and BIC.

4.3.2 A MDL copula selection criterion

We propose a relative fit measure based upon the Minimum Description
Length (MDL; Rissanen, 2007; Hansen & Yu, 2001; Grünwald, 2007;
Grünwald, Myung, & Pitt, 2005), an approach rooted in algorithmic
coding theory (Kolmogorov, 1968), to aid in selecting the right copula
function. The basic idea behind the approach is to find the model that
permits the greatest compression of the data in its description. This
means that one seeks a minimal description for both data and model.
The shortest code or description length obtainable by a certain model
is called its ‘stochastic complexity’, which can be decomposed in the
description length of the data given the model and the description length
of the model itself sometimes called ‘parametric complexity’. The first
term is simply operationalized as the loglikelihood of the model given
the data, the second term is less obvious and can be regarded as a
complexity term (cfr. complexity related to the model’s flexibility to fit
various data patterns).

The optimal implementation of the MDL principle that results in an
universal shortest description length for a specific model is the Normal-
ized Maximum Likelihood (Rissanen, 2001):

\[
NML = - \log \left( \frac{L(\hat{\xi}(Y)|Y)}{\int_{X \in \Omega} L(\hat{\xi}(X)|X)dX} \right).
\]

It represents the likelihood of the current data sample \( Y \) given the model
with its optimal estimated parameters \( \hat{\xi}(Y) \) for this data (i.e., the max-
Model selection in copula IRT models

The maximum likelihood, divided by the integral of maximum likelihoods of the model over all possible data samples (i.e., \( \mathbf{X} \in \Omega \)) (if the sample space is discrete, the integral should be replaced by a sum). The denominator in the formula is a ‘normalizing’ factor and relates to the intuitive operationalization of model complexity as the flexibility to fit several possible data patterns. A model that divides the set of observable data \( \Omega \) into a small set with relatively high likelihood, and a large set with relatively small likelihood will lead to a shorter description length than a model that is able to accommodate a large range of data and thus tends to spread the likelihood more evenly among all possible outcomes. In laymen’s terms one could say that the first model is simple (i.e., specific), while the second model is complex (i.e., highly flexible). Unfortunately, in practice this NML can not be computed for many (even standard) statistical models (De Rooij & Grünwald, 2006). Consequently, alternative procedures or approximations are needed, and currently being developed. Solving these computational issues is the greatest challenge in this field to be able to promote this promising model selection approach and allow its application to more complex models.

The most prominent (asymptotic) approximation of the NML within the MDL approach has the following form (Rissanen, 1996):

\[
\text{ANML} = -\log \left( \frac{L(\hat{\xi}(\mathbf{y})|\mathbf{Y})}{\exp(k/2)(P/2\pi) + \int_{\xi \in \Xi} \sqrt{\det(I(\xi))} d\xi} \right)
\]

\[
= -LL(\hat{\xi}|\mathbf{Y}) + (k/2) \log(P/2\pi) + \log \int_{\xi \in \Xi} \sqrt{\det(I(\xi))} d\xi
\]

The first term is again minus the loglikelihood of the the model given the model, hence a badness-of-fit measure, and the last two terms make up the description length of the model itself. In the second term, a penalty for the number of parameters as a function of the sample size is recognized, resembling the more familiar BIC. The third term represents functional complexity as indicated by the integral of the Fisher information over the parameter space \( \Xi \) (notice that the impact of the third
term decreases with increasing sample size). Thus, we arrive at a model selection criterion that, in contrast to AIC and BIC, takes into account the functional differences between the copulas\(^2\). However, the approach is again computationally demanding and we need to resort to simplifying approximations in order to make it feasible. Below, we sketch the basic idea underlying our implementation of the ANML-MDL approach.

1. Restrict attention to the copula parameter \(\delta\) instead of the full parameter vector \(\xi\): In practice this means that we will isolate the copula part from the rest of the model and only consider the integral over the Fisher information regarding the copula’s association parameter \(\delta\); this evidently will simplify the calculation of the ANML criterion for a copula IRT model (for a similar approach see, Pitt, Myung, & Zhang, 2002; Grünwald et al., 2005). This simplification makes sense because the different copula IRT models have the same given margins but different association structures. Hence, as model selection is focused on the copula part of the model, only the model description length related to the specific copula function needs to be computed, and the description length due to the IRT model, which is constant over copula IRT models, cancels out. Consequently, the integral in the ANML can be replaced by \(\log \int_\delta \sqrt{I(\delta)} d\delta\).

2. Use the observed instead of expected Fisher information: Instead of considering the expected Fisher information \(I(\delta)\), we will approximate it by the observed Fisher information, which is under fairly general regularity conditions a good approximation of its expected counterpart, and easier to calculate.

\[
I(\delta) \approx I(\delta) = -\frac{d^2}{d\delta^2} LL(\delta|Y)|\hat{\gamma}
\]

\(^2\)For links between this MDL approach and geometric complexity, and Bayesian inference using Jeffreys prior see Balasubramanian (1997) and Myung, Balasubramanian, and Pitt (2000).
where \( \hat{\gamma} \) are the model’s estimated IRT parameters.

3. An analytical solution for the second derivative of the model’s log likelihood for \( \delta \) is not readily available. The full likelihood of a copula IRT model (Braeken et al., 2007) under a marginal maximum likelihood formulation involves an integral for the latent trait \( \theta_p \) which has no closed-form solution and needs to be approximated by for instance Gauss-Hermite quadrature, and furthermore, the second derivatives of most copula functions themselves are not self-evident, especially when a subset involves more than two items (Note that even automatic differentiation by means of software packages might fail due to memory problems.). To circumvent these computational issues, we resorted to numerical approximation of the second derivative by means of finite differences, which are universally applicable to all copula IRT models.

Despite the simplifications outlined above, we experienced in the case of Frank copula computational problems with the integral of \( \sqrt{I(\delta)} \). An illustration can be found in the upper panel of Figure 4.3 where for Frank copula the function \( f(\delta) = \sqrt{I(\delta)} \) is displayed that needs to be integrated over the parameter space of \( \delta \) (from 0 to \( \infty \)) such that it can serve as a model complexity penalty. This function \( f(\delta) \) becomes instable around \( \delta = 25 \), and even breaks down computationally at even larger values of the copula association parameter while it has not yet attained zero. This breakdown might even vary for different model setups or datasets. A solution to these problems is to put ad hoc restrictions on the range of the parameter space. In practice, these restrictions only make sense if prior knowledge about the parameter range is available (De Rooij & Grünwald, 2006), as it is advisable to put the bounds within reasonable distance of the maximum likelihood estimate \( \hat{\delta} \). This is a not so uncommon problem when implementing this ANML approach, even for more regular and simple statistical models. To avoid an undefined integral and possible instability due to the limiting bounds of the pa-
Chapter 4

Parameter space corresponding to Π and M, the parameter space of δ was consequently restricted based upon Kendall’s tau values of 0.1 and 0.8 for the lower and upper bound ([lb, ub]), respectively. These values comprehend the broad range of equivalent copula association parameter values that can be expected for LID in practice (i.e., [0.91, 18.19], [0.22, 8.00], and [1.11, 5.00] for Frank, Cook-Johnson, and Gumbel-Hougaard copula, respectively). This bounded integral is approximated by means of a trapezoid rule with 1000 nodes.

Figure 4.3: Computational instability of f(δ).

Summarizing, this isolation-based approach will be further referred to as MDL1 and the MDL1 criterion is defined as follows:

\[ MDL1 = -LL(\hat{\xi}|Y) + (k/2)\log(P/2\pi) + \log \int_{lb}^{ub} \sqrt{\det(I(\delta))} d\delta, \]

where lb and ub refer to the lower and upper bound of the parameter space for δ as derived from Kendall’s tau.

Alternatively, we can opt for a rough but slightly more elegant approximation of the third term of the MDL criterion, by considering the
Model selection in copula IRT models

Hessian of the log likelihood at the maximum likelihood estimate \(\hat{\xi}\), conveniently available as by-product of most optimization procedures. This method is denoted as \(MDL2\) and its formal definition is as follows:

\[
MDL2 = -LL(\xi|Y) + k/2 \log(P/2\pi) + \log \sqrt{\det(I(\hat{\xi}))}.
\]

In this way, the complicated integral over the parameter space is approximated by evaluating the integrand at the maximum likelihood estimate which is believed to be the dominant contribution to the integral (Grünwald et al., 2005). A possible advantage of this approach is that we take into account all parts of the model, not only the association part represented by the copula parameter \(\delta\). This Hessian-based approach will be further referred to as \(MDL2\).

4.3.3 A copula-weights selection criterion

A totally different approach to the model selection issue can be taken by considering a continuous model expansion as advocated by Gelman and Rubin (1995). The idea is to construct a ‘super model’ that comprehends all association structures as implied by the various copula functions under consideration. Define a function \(C^\upsilon\) as \(v_1C_1 + v_2C_2 + (1 - v_1 - v_2)C_3\), where \(C_1, C_2\) and \(C_3\) are Frank, Cook-Johnson, and Gumbel-Hougaard copula, respectively, and with \(0 \leq v_k \leq 1\) and \(\sum_{k=1}^{3} v_k = 1\). This convex linear combination of copula functions, is again a copula function (Nelsen, 1999). The application of this new copula function \(C^\upsilon\) can be viewed as a model expansion, and it inherits properties of each of its component copulas, and more or less avoids the need to make a model selection choice.

While this approach is promising, a rule has to be implemented which allows one to make a discrete decision after having fit the copula weights model. One may notice that each of the component copulas is now nested within this ‘super model’, which in principle would open the way for likelihood ratio tests for nested models. However, there exists the
possibility that either all likelihood ratio tests of the component copulas against the super copula would be non-significant, all significant, or a combination with two or more non-significant results. Consequently, this set of likelihood ratio tests does not necessarily lead to a conclusive decision. We prefer to choose an alternative path, and propose to interpret this ‘super model’ as a construction that weighs the applicability of each copula for the data at hand. This last consideration would lead to a procedure where the copula with the largest weight $v_k$ is selected as the preferred model choice and this is also the rule we follow in this paper.

As is often the case in mixture-type of models, the practical implication of $C^v$ may run into an identification problem. In this case, the identification problem is due to a subtle trade off that exists between the degree of association as indicated by the association parameter $\delta_k$ and the component weight $v_k$ of a specific component copula: A larger $\delta_k$ or a larger $v_k$ for a specific copula $C_k$ both increase the relative importance of the copula $C_k$ in the convex linear combination, and may give rise to similar association patterns. Therefore, we opt for a two-step procedure in which the first step consists of estimating the individual association parameter of the different component copula models; in the second step, the resulting estimates are used as fixed association parameters for the component copulas in the convex linear combination: $C^v = v_1C_1(\hat{\delta}_1) + v_2C_2(\hat{\delta}_2) + (1 - v_1 - v_2)C_3(\hat{\delta}_3)$. Note that when association parameters $\delta_1$, $\delta_2$, and $\delta_3$ all point at independence it is logically to stop the procedure at the first step, and conclude that independence is the case.

4.4 Simulation study

To investigate model selection in these copula IRT models in greater detail, and to evaluate the performance of the above presented selection tools, a simulation study will be set up. Datasets will be simulated
following four copula IRT models that only differ in the specific copula being used (i.e., independence, Frank, Cook-Johnson, and Gumbel-Hougaard copula), but not in the characteristics of the individual items in the test. These four data generating processes correspond to the situation of LSI, stable LID, dimensional increasing LID, and dimensional decreasing LID, respectively. Each dataset will be again fitted using each of these alternative copula IRT models. The three factors that were manipulated are sample size (denoted by \( P \)), item difficulty of LID items (denoted by \( B \)), and degree of LID (denoted by \( D \)).

For each of the four situations, data sets consisting of binary item responses for \( P \) persons (sample size, \( P = \{250, 500, 1000\} \)) and \( J = 10 \) items were simulated following a two-parameter-logistic model (2PL). The abilities of the persons (\( \theta_p \)) were generated from a standard normal distribution. Item difficulty parameters \( \beta_j \) were set to be spaced on a scale between \(-1.28 \) and \( 1.28 \) following deciles of a standard normal distribution (with absolute values \( 0, 0.25, 0.52, 0.84, \) and \( 1.28 \)). For simplicity, difficulty of the items is aligned with item number, meaning that for instance \( \beta_1 \) and \( \beta_{10} \) are equal to \(-1.28 \) and \( 1.28 \), respectively, that \( \beta_5 \) and \( \beta_6 \) are equal to \( 0 \), and \( \beta_2 \) and \( \beta_9 \) are equal to \(-0.84 \) and \( 0.84 \), respectively, etc. Item discrimination parameters \( \alpha_j \) were generated from an uniform distribution on the interval \([0.7, 1.3] \).

For each of the three LID situations, one subset of four items was fixed to show LID. The items within the subset were either all of equal average difficulty (\( B = eq \), subset \( s_{eq} = \{4, 5, 6, 7\} \)) or coming from the two poles of the scale (\( B = uneq \), subset \( s_{uneq} = \{1, 2, 9, 10\} \)). The LID itself was generated under copula association parameters that are derived from Kendall’s \( \tau \) values of \( 0.35, 0.50 \) and \( 0.65 \), and correspond to various degrees of dependency (i.e., \( D = low \), \( D = mid \), and \( D = high \), respectively). Note that in case of the LSI situation, copula models were fitted based upon the two different subsets \( s_{eq} \) and \( s_{uneq} \), to allow comparison with a regular LSI model in the absence of true LID.

All this results in a one factor design (sample size, with three levels)
for the LSI case, and factorial designs of three factors (sample size, degree of LID, and subset item difficulty) with the latter factor having two levels, and the first two factors having each three levels) for the LID situations. For each cell of the design, 100 replications were simulated.

Note that data simulation using these particular copula functions is pretty straightforward. Making use of their mixture of power distributions formulation (Oakes, 1982; Marshall & Olkin, 1988), dependent random variables, as arising from these three copulas, can be simulated by means of well-known univariate distributions. Appendix I contains a description of the data simulation algorithms used in the study. The copula IRT models will be fitted by means of marginal maximum likelihood (MML) using a quasi-Newton optimization technique and non-adaptive Gauss-Hermite quadrature with 20 quadrature points (see also, Tuerlinckx, Rijmen, Verbeke, & De Boeck, 2006).

The dependent variable of interest is the number of correctly identified true models in each cell of the simulation design. The factorial design of the simulation study allows the calculation of the effect size measure $\hat{\omega}^2$ (i.e., an unbiased estimate of the proportion of explained variance by an effect in the population, see e.g., Keppel, 1991; Olejnik & Algina, 2003, which will be used to filter out the most important effect(s) that influence model selection. We will follow the guidelines of Cohen (1988) stating that an $\hat{\omega}^2$ of 0.15 can be regarded as a large effect (and 0.06 as a medium effect). Such an approach helps to keep the focus on the important patterns in the data and facilitates a clear presentation of the results.

4.4.1 Results: LSI

For the situation where local stochastic independence is not violated, the likelihood ratio test comparing a copula LID model with the LSI model, did not in any case accept a copula model as the better model. Hence, the traditional likelihood ratio test for nested models obtains a 100% correct identification of the true LSI data generating model.
For reasons of comparison, Table 4.1 shows correct identification results when applying the model selection criteria the deviance (i.e., -2LL), AIC, BIC, MDL1 (i.e., the isolation-based approximation), MDL2 (i.e., the Hessian-based approximation), and $C^\upsilon$ (i.e., the copula weights). Whereas the criteria that take into account model complexity perform very well (i.e., 90% correct identification and more; although AIC is slightly less performant), the deviance and a $\chi^2$ deviation measure, both not correcting for model complexity, are not able to correctly identify LSI as they almost always choose the more flexible copula model that shows a slightly better goodness-of-fit to the data (i.e., in this case mostly Frank copula). Also note that in this LSI situation, the association parameter $\delta$ of either of the three copula functions were generally concluded to be not significantly deviating from their independence limit, and thus this parameter also was indicative for the absence of LID.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Model selection criterion</th>
<th>$-2LL$</th>
<th>AIC</th>
<th>BIC</th>
<th>MDL1</th>
<th>MDL2</th>
<th>$C^\upsilon$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>1</td>
<td>93</td>
<td>112</td>
<td>117</td>
<td>114</td>
<td>107</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>99</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>118</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>93</td>
<td>119</td>
<td>120</td>
<td>120</td>
<td>114</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

*120 replications per condition

4.4.2 Results: Dimensional stable LID

Table 4.2 shows correct identification results when applying the deviance, MDL1, MDL2, and $C^\upsilon$ as model selection criterion on 20 replicated datasets for each condition of the simulation design. Their overall correct identification is on average 14, 14, 18, and 11 respectively. The $\chi^2$ deviation criterion was dropped for further analysis as it continues to severely underperform with an average of 4 correct identifications out of 20. One notices that when selecting on goodness-of-fit alone, one already obtains a correct identification rate of 70%, whereas only the complexity penalty as formulated by means of the Hessian-based MDL2
Table 4.2: Correct model identification of the true dimensional stable IID generating process.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Model selection criterion</th>
<th>-2LL</th>
<th>MDL1</th>
<th>MDL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>υ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>eq</td>
<td>11</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>puneq</td>
<td></td>
<td>14</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>500</td>
<td>eq</td>
<td>14</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>puneq</td>
<td></td>
<td>13</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>1000</td>
<td>eq</td>
<td>14</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>puneq</td>
<td></td>
<td>13</td>
<td>12</td>
<td>19</td>
</tr>
</tbody>
</table>

20 replications per condition.
formulation can improve on these results. The factor with the largest contribution to the explained variance in correct identification of the true underlying model based upon the deviance criterion was the sample size ($\hat{\omega}_P^2 = 0.36$). Correct identification increased from 11 over 13 to 17, for $P = 250$, $P = 500$, and $P = 1000$, respectively. This in contrast to the MDL2 criterion which performed at a high correct identification rate regardless of sample size (from 18 over 18 to 19, for $P = 250$, $P = 500$, and $P = 1000$, respectively). When misidentification occurs, it is mostly Gumbel-Hougaard copula that is preferred above Frank copula.

4.4.3 Results: Dimensional increasing LID

Table 4.3 shows correct identification results when applying the deviance, MDL1, MDL2, and $C^v$ as model selection criterion on 20 replicated datasets for each condition of the simulation design. Their overall correct identification is even more impressive than in the dimensional stable LID situation and is on average 18, 18, 17, and 19, respectively. One notices that when selecting on goodness-of-fit alone, one already obtains a correct identification rate of 90%, whereas only the model weights of $C^v$ can improve on these results. The factor with the largest contribution to the explained variance in correct identification of the true underlying model was for all model selection criteria, the sample size ($\hat{\omega}_P^2 = 0.40, 0.40, 0.73$, and 0.21), a larger sample size giving rise to more correct identification results. When misidentification occurs, it is mostly Frank copula that is preferred above Cook-Johnson copula. The $\chi^2$ deviation criterion was dropped for further analysis as it again underperforms with an average of 13 correct identifications out of 20.
Table 4.3: Correct model identification of the true dimensional increasing IID generating process.

<table>
<thead>
<tr>
<th>Condition</th>
<th>MDL1</th>
<th>MDL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>mid</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>high</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>MDL1</th>
<th>MDL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>mid</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>high</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>MDL1</th>
<th>MDL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>mid</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>high</td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

20 replications per condition.
4.4.4 Results: Dimensional decreasing LID

Table 4.4 shows correct identification results when applying the deviance, MDL1, MDL2, and C$\nu$ as model selection criterion on 20 replicated datasets for each condition of the simulation design. Their overall correct identification is on average 16, 16, 9, and 17, respectively. The $\chi^2$ deviation criterion was dropped for further analysis as it again underperforms with an average of 8 correct identifications out of 20. As in the dimensional increasing LID situation, selecting on goodness-of-fit alone in the dimensional decreasing LID situation already results in a correct identification rate of 80%, whereas only the model weights of C$\nu$ can slightly improve on these results. The factor with the largest contribution to the explained variance in correct identification of the true underlying model was for all model selection criteria, the sample size ($\hat{\omega}_P^2 = 0.53, 0.53, 0.47, \text{ and } 0.33$), a larger sample size giving rise to more correct identification results. Furthermore, correct identification was also effected by the interaction between subset item difficulty and degree of LID except for the MDL2 criterion ($\hat{\omega}_{B \times D}^2 = 0.11, 0.11, 0.03$ and 0.07). Increasing degree of LID and having an equal difficulty LID subset improves correct identification of the underlying dimensional decreasing LID. Notice that the MDL2 criterion underperforms in this situation compared to the other criteria. This is due to a difficulty in distinguishing the Frank copula model from the Gumbel-Hougaard copula model. In general, when misidentification occurs, it is mostly Frank copula that is preferred above Gumbel-Hougaard copula for all model selection criteria.

4.4.5 Results: Model complexity

Without knowing how much a model constrains possible outcomes, one can not know how impressed to be when the observed data and model are consistent. The underlying idea is that the more specific a model is, the narrower its predictions, and the more impressive a confirmation
Table 4.4: Correct model identification of the true dimensional decreasing IID generating process.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Model selection criterion</th>
<th>-2LL</th>
<th>MDL1</th>
<th>MDL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td></td>
<td>12</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>mid</td>
<td></td>
<td>12</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td>12</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

*Note: 20 replications per condition.*
of this model (see e.g., Meehl, 1997; Roberts & Pashler, 2000). Hence, interest goes to the extent to which the four copula IRT models under consideration differ in model complexity, with the more complex model, being the more flexible and as such less constrained. The functional complexity term involving the Fisher information in the MDL1 and MDL2 criterion should be able to give a clearer picture whether or not the copula IRT models actually differ in functional complexity. The general trend is that the difference in functional complexity within the set of copulas is rather small, with a noticeable larger difference for comparisons to independence (see Table 4.5). This adds to the general picture of the model selection results, in which the log likelihood appeared to be already highly diagnostic when selecting the true copula within the set of LID copulas (correct identification rates of 70% and more). Note, that there was also a huge congruency between model selection decisions based on the different model selection criteria in the LID situations. Thus, it appears that the set of copulas each are able to fit a roughly equal, but well-separated, amount of the possible-data space.

<table>
<thead>
<tr>
<th>Model</th>
<th>MDL1</th>
<th>MDL1</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>LSI</td>
<td>F</td>
</tr>
<tr>
<td>LSI</td>
<td>0.00</td>
<td>2.40</td>
</tr>
<tr>
<td>F</td>
<td>0.00</td>
<td>3.07</td>
</tr>
<tr>
<td>C-J</td>
<td>0.00</td>
<td>3.10</td>
</tr>
<tr>
<td>G-H</td>
<td>0.00</td>
<td>3.11</td>
</tr>
</tbody>
</table>

### 4.5 Discussion

The motivating problem in this paper was the need to build realistic measurement models that are able to adequately account for the presence of local item dependencies in a test. Furthermore, the exact nature of these local item dependencies, i.e., the complexity of the association
structure, can have potential diagnostic value for test improvement and an understanding of the construct one wants to measure. The copula IRT approach is able to model these local item dependencies, and allows for the specification of a variety of LID association structures. This flexibility in defining the association structure also implies an inherent model (i.e., copula) selection issue, because a priori information on the nature of the LID and theoretical guidelines are mostly lacking when having to choose between competing models for the LID association structure. To make this choice in a proper way, a model selection approach is needed that adequately offsets model complexity and the effect of overfitting to prevent poor generalization of the chosen model.

The results of our simulation study that was setup to assess the feasibility of choosing the ‘true’ underlying copula IRT model, indicated in general a good recovery of the true model for each of the model selection criteria. When selecting between the copula LID models and the regular LSI model, traditional likelihood tools (likelihood ratio test, Wald test for the copula association parameter $\delta$) sufficiently accomplished the goal of balancing goodness-of-fit and generalizability. However, a badness-of-fit measure alone was totally ignorant when selecting between a copula and the independence case, but generalizability could still be warranted by offsetting this measure by means of a complexity penalty in terms of model dimension, as is done in traditional model selection criteria as AIC and BIC.

When considering selection between the different copula LID models, we expressed concern about possible complexity differences between copulas that are overlooked by regular model selection approaches as AIC and BIC, which define complexity merely in terms of model dimension (i.e., number of parameters), while functional differences might play as well. However, the high recovery performance and congruent choices of the different model selection criteria lead to greater confidence in the fact the three copulas under consideration demarcate essentially different LID association patterns that result in highly diagnostic badness-of-fit
differences as shown in the likelihood. Discongruent model selection decisions would have pointed at differences in model aspects or properties that are magnified by specific model selection approaches. The functional complexity differences between the three LID copula models under consideration, as indicated by the MDL model selection approach were also assessed to be rather small, dismissing our initial concern.

The two alternative and new model selection approaches that were introduced, Minimum Description Length and model weights $C^w$, and that we thought are more suitable for non-nested models with an equal amount of parameters, do not essentially outperform the traditional model selection criteria, but perform equally well.

MDL2 outperformed all other model selection criteria in the dimensional stable LID situation, but unfortunately it also disperformed when the underlying LID was dimensional decreasing. This points out a possible less stable estimation of the model parameters in case of an underlying dimensional decreasing LID, which might require more information in the data to lead to equally-stable parameter estimates. Furthermore the implementation of the MDL approach required several ad hoc simplifications and approximations to make the approach practically feasible. In the end one might wonder about making the effort to implement MDL for applied model selection issues in complex statistical models, as long as the gap between its practical implementation and theory is not bridged.

A similar inconsistency in model selection results holds for $C^w$, that performed nicely in the dimensional changing LID situations, but disperformed when the underlying LID was dimensional stable. This is also visable in the weights of the copulas (which are on average 0.51, 0.17, and 0.32, for Frank, Cook-Johnson and Gumbel-Hougaard copula in the convex sum $C^w$ given true underlying stable LID), and are quite likely another expression of the balancing issue in estimating this convex sum of parametric copulas.

In conclusion, despite their different mathematical form, the three
copula functions are very comparable to each other regarding model complexity. For this reasons, it is hard to beat a simple criterion like the likelihood in chosing a good model.
References


Frank, M. J. (1979). On the simultaneous associativity of \( F(x, y) \) and \( x + y - F(x, y) \). *Aequationes Mathematica*, **19**, 194–226.


Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp.


Rissanen, J. (2001). Strong optimality of the normalized ml models as universal codes and information in data. *IEEE Transactions on...
Model selection in copula IRT models

Information Theory, 47, 1712–1717.


Appendix: Data simulation

In this appendix it will be explained how data can be simulated under the various copula IRT models discussed in the paper. As a start, generate latent proficiencies $\theta_p$ from a distribution (e.g., a standard normal), and fix a set of $J$ item parameters, being item discrimination $\alpha$ and difficulty $\beta$ for a 2PL model. Compute the cumulative probability of a zero response given these model parameters,

$$F(Y_{pi} = 0|\theta_p) = (1 + \exp(\alpha_i(\theta_p - \beta_i)))^{-1}. $$

Next, formulate the copula IRT model. Let us assume for simplicity that LID only occurs between items 1 and 2, and that LSI holds for the others:

$$\Pr(Y_p|\theta_p) = \Pr_C(Y_{p1}, Y_{p2}|\theta_p) \prod_{i=3}^{J} \Pr(Y_{pj}|\theta_p),$$

For each independent item, $j = 3, \ldots, J$, generate a uniform random variable $U_{pj}$, and evaluate it against $F(Y_{pj} = 0|\theta_p)$; when $U_{pj} > F(Y_{pj} = 0|\theta_p)$ this results in an item response $Y_{pj} = 1$, and $Y_{pj} = 0$ otherwise.

For a locally dependent item set $s$, in this case $\{1,2\}$, generate a set of dependent random variables $U_p = (U_{p1}, U_{p2})$ (see next paragraph on how to do this) that are jointly distributed following a selected copula function $C$ with uniformly distributed margins and association parameter $\delta$, and evaluate them against $F(Y_{pj} = 0|\theta_p)$; when $U_{pj} > F(Y_{pj} = 0|\theta_p)$ this results in an item response $Y_{pj} = 1$, and $Y_{pj} = 0$ otherwise.

Making use of the mixture of power distributions formulation (Oakes, 1982; Marshall & Olkin, 1988) we can simulate dependent uniform random variables $U_p = (U_{p1}, U_{p2})$ as arising from the 3 copulas under consideration by means of well-known univariate distributions. See Table 4.6 for the necessary simulation steps. Note that this works for the whole range of positive dependency and the subset can be larger than a pair of items.
Table 4.6: Simulation of dependent uniform random variables $U$ by means of a copula $C$.

<table>
<thead>
<tr>
<th>Copula $C$</th>
<th>$U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>$V_1$ and $V_2 \sim \text{Uniform}(0,1)$</td>
</tr>
<tr>
<td></td>
<td>$W_1$ and $W_2 \sim \text{Uniform}(0,1)$</td>
</tr>
<tr>
<td></td>
<td>$S = \text{round}(1 + \frac{\log(V_1)}{\log(1 - \exp(-\delta V_2))})$</td>
</tr>
<tr>
<td></td>
<td>$T = W^{(1/S)}$</td>
</tr>
<tr>
<td></td>
<td>$(-1/\delta) \log(1 - (1 - \exp(-\delta T))$</td>
</tr>
<tr>
<td>Cook-Johnson</td>
<td>$S \sim \text{Gamma}(1/\delta, 1)$</td>
</tr>
<tr>
<td></td>
<td>$V_1$ and $V_2 \sim \text{Exp}(1)$</td>
</tr>
<tr>
<td></td>
<td>$(1 + V/S)^{-1/\delta}$</td>
</tr>
<tr>
<td>Gumbel-Hougaard</td>
<td>$S \sim \text{Uniform}(0, \pi)$</td>
</tr>
<tr>
<td></td>
<td>$W, T_1, &amp; T_2 \sim \text{Exp}(1)$</td>
</tr>
<tr>
<td></td>
<td>$\delta_1 = 1/\delta$ and $\delta_2 = (1 - \delta_1)/\delta_1$</td>
</tr>
<tr>
<td></td>
<td>$V_1 = \sin(\delta_1 S)(\sin(1 - \delta_1)S)^{\delta_2}$</td>
</tr>
<tr>
<td></td>
<td>$V_2 = \sin(S)^{1/\delta_2}W^{\delta_2}$</td>
</tr>
<tr>
<td></td>
<td>$exp(-(T/(V_1/V_2))^{\delta_1})$</td>
</tr>
</tbody>
</table>
CHAPTER 5

Fitting explanatory IRT models:

a Matlab IRTm toolbox

Abstract: Item response theory models are the central tools in modern measurement and advanced psychometrics. We offer a MATLAB IRT modeling (IRTm) toolbox that is freely available, and can fit a large variety of IRT models, including the recent copula IRT models to handle local item dependencies. A design-matrix approach is followed, giving it possibilities for end-user control and for flexibility in building a model that goes beyond standard models like the Rasch or 2PL model. This approach is suitable for a measurement course to acquire more insight in, and demystify the structure and construction of common (and less common) IRT models.
5.1 Introduction

Item response theory (IRT) models are the central tools in modern measurement and advanced psychometrics. They differ from classical test theory in the sense that IRT posits a mathematical model to explain the pattern of observed item responses on a test (see e.g., Lord & Novick, 1968), while classical test theory does not. The key idea is that tests are designed to measure an unobservable variable of interest, a latent trait. Hence, multiple measurements have to be made of the same person in order to obtain a reliable estimate of a person’s latent proficiency. IRT assumes that both persons as well as test items have a position on this same underlying dimension the test is measuring. In relating both persons and items to this underlying dimension or latent trait, restrictions are imposed on the probability model for the observed item responses. This is a constructive theory-driven scientific strategy, as one can revise the test to fit the model restrictions (like for instance is done in strict Rasch measurement) or, alternatively, one can adapt the model to better fit the data.

Over the last years, it has been shown that models from item response theory can be conceptualized as generalized linear or nonlinear mixed models (see e.g., Agresti, Booth, Hobert, & Caffo, 2000; De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003), if one considers the latent trait to be a person-specific random effect with a given distribution in the population. This allows the estimation of such models by means of standard statistical software as SAS and Mplus. Well established software specific for IRT does exist and adds a multitude of specialized features and graphics. The most popular packages are Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2006), Multilog (Thissen, Chen, & Bock, 2003), and Conquest (Wu, Adams, & Wilson, 1998). Besides this commercial software, there a couple of less extensive, but free, packages available in the R computing environment (see e.g., Rizopoulos, 2006; Johnson, 2007).
We will present an IRT modeling (IRTm) toolbox for MATLAB that is freely available, and can fit a large variety of IRT models for binary responses, including the recent copula IRT models to handle local item dependencies (Braeken, Tuerlinckx, & De Boeck, 2007). This latter class of models is not implemented elsewhere in any software package. MATLAB is a general-purpose, matrix programming language with a large user base in research and business. To our knowledge no other IRT specific toolboxes are available, and as such this IRT toolbox may be useful to some of these users. A design-matrix approach is followed, giving it the advantages of end-user control and flexibility in building a model that goes beyond standard models like the Rasch or 2PL model. A marginal maximum likelihood approach will be followed for estimation purposes.

In what follows some practical information is given about the IRTm toolbox for MATLAB and its general use will be laid out. In the next sections some didactical examples are used to build up from a basic Rasch model to a more general copula IRT model. These examples, including experimental design, cross-validation, latent regression, differential item functioning, hidden person groups, local item dependency, and data simulation, will be used to demonstrate some of the most interesting modeling possibilities in practice and will be accompanied by explicit key references that can direct the reader when they are interested in more technical background and extensive discussion of the topic at hand. For ease of recognition, all MATLAB expressions, variables and functions will be typeset in a different font and in italic. The main text functions as software manual and hence, can best be combined with a hands-on approach on the computer. An appendix is included with some more statistical details about the general copula IRT model that the toolbox is able to fit. A short discussion section will point out possible extensions, further optimization possibilities, and the main objective underlying the current toolbox.
5.1.1 IRTm toolbox

**Requirements:** The IRTm toolbox requires that you have MATLAB 7.2 (Version R2006a) or a more recent version installed (7.6, Version 2008a is preferable), and the Optimization toolbox also needs to be installed. The IRTm toolbox was developed and tested on Windows and Linux platforms.

**Installation:** The IRTm toolbox for MATLAB can be freely downloaded via [http://ppw.kuleuven.be/okp/software/IRTm](http://ppw.kuleuven.be/okp/software/IRTm). Upon filling out a form, you will be emailed a link to where a ZIP archive is available. Unpackage the file to the folder in which you want to install the toolbox and then execute the included installer, `IRTmInstall`, from the MATLAB command window. The installer will guide you from there. Note that during installation, you will be asked to accept a license agreement prior to completion of the installation. When the installation is finished, you can test the toolbox by running through some example case studies of which the code is provided in the m-file `IRTmCase` (also accessible through the MATLAB demo feature). Since IRTm is constantly under development, its most recent version, bug reports, and fixes can be found on the web site above.

**License agreement:** Please note that, although IRTm and its source code may be freely downloaded, it is not permitted to distribute the code or derived code without the authors’ consent. Cooperation from third parties in further developing the IRTm toolbox is welcomed, but in order to maintain transparency regarding exactly which methods an end user has implemented, we want the distribution of this toolbox to remain centralized. This will also allow us to keep an eye on the quality and correctness of the toolbox at all times and to have a bit of an overview of the user base which will allow better support and communication. We would appreciate acknowledgement if you would use the IRTm toolbox in your own work and papers. Note, however, that as usual, no warranty whatsoever is given for the results and outcomes obtained by means of the application of any function in this toolbox.
5.1.2 General use

The general use of the toolbox involves three main actions: (1) Specifying the data and algorithm to be used \( [\text{SETTINGS}] \), (2) building the IRT model \( [\text{MODEL}] \), (3) and estimating the model, such that one can acquire the necessary output and evaluate the model results \( [\text{OUTPUT}] \). These three main actions correspond to defining two structures \( \text{SETTINGS} \) and \( \text{MODEL} \) in MATLAB, and feeding these to the main function \( \text{IRTm} \) to obtain the resulting \( \text{OUTPUT} \).

We will start with a concrete and simple example in which we guide the reader through the different steps, and then go on and explore more of the functionalities of the IRTm toolbox in subsequent sections and examples. The code used in these examples can also be retrieved in the m-file \( \text{IRTmCase.m} \), which allows you to access the code by opening it in the editor and easily go through the examples running the commands in this file from the MATLAB command line. The full model functionality of the toolbox is laid out in a separate slightly technical appendix at the end of the manuscript. Readers who prefer a top-down approach and are more familiar with IRT models might be more inclined to start with the contents of this appendix and then return to the practical examples in the current section.

**Data requirements and preprocessing:** A preliminary step to the procedure is making sure that your data are in the right format. The data to be analyzed with the IRTm toolbox are expected to be complete-case (no missing data), numeric and of a rectangular format, persons (rows) -by- repeated measurements (columns). Item responses are expected to be binary (0 and 1) coded. Additional information on the persons can be given under the form of continuous or categorical variables. The latter need to be recoded into binary dummy variables, in order to avoid that they are considered to be continuous.
Example 1: A simple Rasch model

In this first example we will analyze the dataset ‘example_rasch.txt’ (included in the IRTm toolbox) following the Rasch model (Rasch, 1960), a well-known standard model of item response theory. The data consist of a variable with an unique number to identify the person (first column) and binary responses on 6 test items (columns 2 to 7) for all 500 persons. Note that there is no additional information present and that the data are already in the required format so that no preprocessing is needed.

Data and algorithm setup [ SETTINGS ]: Our first action is to define the structure SETTINGS which contains fields with setup information about the data as well as specifications for the optimization algorithm to be used (see Table 5.1). A default SETTINGS structure can be acquired by the function call

\[
\text{SETTINGS = IRTmSet()}
\]

after which one can fill in and adapt the necessary fields. Specifying the dataset to be used can be done by the statement

\[
\text{SETTINGS.input = 'example_rasch.txt'},
\]

after which you only need to indicate the columns that correspond to the items in the dataset

\[
\text{SETTINGS.I = [2 3 4 5 6 7]},
\]

and the delimiter that separates these columns \text{SETTINGS.delim=' '} (here a space, which is also the default). If your current MATLAB working directory is not the directory in which you installed the IRTm toolbox, you will be required to give up an absolute, instead of relative, address to locate the data file ( \text{SETTINGS.input = 'c:\dir\example_rasch.txt'}). Note that you can use regular MATLAB tricks to define the items field. For instance, we could also have simply stated \text{SETTINGS.I=[2:7]} to indicate the item sequence. The defaults settings for the estimation algorithm will suffice for our example.
Model building [MODEL]: In a Rasch model the characteristic function of $Y_{pi}$, the binary item response of person $p$ ($p = 1, 2, \ldots, P$) on item $i$ ($i = 1, 2, \ldots, I$) is defined as

$$\Pr(Y_{pi} = 1|\theta_p) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)},$$

with $\theta_p$ a model parameter representing the proficiency of person $p$, and $\beta_i$ indicating the difficulty of item $i$. Thus, the probability of answering correctly is an S-shaped (logistic) function of the positions, of both the item and the person, on the latent trait the test is measuring. Note that the value of $\beta_i$, the difficulty of an item, corresponds exactly to the location on the latent trait where a person would have a chance of a half to answer the item correctly (i.e., the point of inflection of the characteristic curve).

The joint probability of the item responses on the full test ($Y_p = [Y_{p1}, \ldots, Y_{pI}]$) for a person $p$ is then

$$\Pr(Y_p = y_p|\theta_p) = \prod_{i=1}^{I} \frac{\exp(y_{pi}[\theta_p - \beta_i])}{1 + \exp(\theta_p - \beta_i)}.$$

Notice that in this model a person’s item responses are assumed to be independent given his/her proficiency on the latent trait. This assumption is commonly referred to as conditional independence or local stochastic independence, and implies that the latent trait accounts for all the dependency between a person’s responses on the test. Thus, the latent proficiency explains why there are differences between persons in performance and why the item responses of a given person interrelate.

The IRTm toolbox follows a marginal maximum likelihood approach to model estimation (see Appendix for technical details and references), and hence it is assumed that these latent proficiencies follow a distribution in the population (see e.g., Rijmen et al., 2003). Here we will assume $\theta_p \sim N(\mu, \sigma^2)$, a normal distribution with mean $\mu$ set at zero (a convention to fix the scale of the model) and variance $\sigma^2$. The larger
the estimated value of this variance, the larger the individual differences between persons on the latent trait targeted by the test are assessed to be.

Given the model specifications above, a default Rasch model can be built with the function call

$$\text{MODEL} = \text{IRTmModel}(6, 'Rasch');$$

where the first argument of the function takes the number of items in the test, the second argument defines the requested type of model structure.

**Fitting the model and acquiring results [ OUTPUT ]:** To fit this Rasch model on our dataset we feed the two structures `SETTINGS` and `MODEL` to the main function of the IRTm toolbox by means of the call

$$\text{OUTPUT} = \text{IRTm(SSETTINGS,MODEL)};$$

The model likelihood $l(\beta, \sigma)$ that is optimized contains the fixed parameter vector $\beta$, gathering the 6 item difficulties $\beta_i$, and the fixed parameter $\sigma$, whereas the random parameters $\theta_p$ will be integrated out and estimated by means of empirical Bayes methodology (see Appendix for technical details). The results of the procedure are gathered into the structure `OUTPUT`. It contains compressed data information stored in a format using only the unique response patterns and their frequency of occurrence, an index to keep track of which person has which response pattern, information on the algorithmic optimization outcome and model estimates (for a description of the fields of this structure, see Table 5.2).

The information and results inside this structure can easily be accessed. Model convergence can for instance be checked by looking at `OUTPUT.optim`, a zero indicating non-convergence (and some related fields as `OUTPUT.message` and `OUTPUT.HEScheck`). The compilation of some summary tables has also been preprogrammed, but customized displays of `OUTPUT` features are easily implemented either within the MATLAB environment or a word processing software.
A table with the different unique response patterns, item sum score, frequency of occurrence, predicted frequency, and corresponding latent trait estimates can be requested by the call

\[ \text{IRTmSummary} \left( \text{OUTPUT}, 'pred' \right) \] 

A basic table of the fixed model parameters, with their estimates, standard errors and a \( p \)-value from the corresponding Wald test (indicating whether the parameter deviates significantly from zero or not), can be displayed by means of the call

\[ \text{IRTmSummary} \left( \text{OUTPUT}, 'param', \text{MODEL} \right) \] 

Note that not providing the third argument \( \text{MODEL} \) in the function call will leave out the parameter labels in the table. A third table, with most common goodness-of-fit measures can be displayed by

\[ \text{IRTmSummary} \left( \text{OUTPUT}, 'gof' \right) \] 

**Comparison with commercial software:** To have an idea of the performance of our toolbox, results were compared to two gold standards, being the commercial software packages SAS (SAS Institute, 2002) and MPLUS (Muthén & Muthén, 1998). The simple Rasch model was fitted on this dataset under similar conditions in each of the packages, and the results are shown in Table 5.3.

Parameter estimates for both the fixed parameters, being the vector of difficulty parameters \( \beta \) and the variance \( \sigma^2 \) of the latent trait, as well as the empirical Bayes estimates for the latent trait \( \theta_p \) are comparable across programs. Note however, that SAS appears to follow a slightly different and more time-intensive way to estimate the latent trait, and that for more complicated models, MPLUS follows a factor analytical reformulation of common IRT models. We did not find an option to access the standard errors for the latent trait estimates in MPLUS, and the time spent on the algorithm was not reported accurately, as it was displayed to be zero seconds. To represent this in the table, we used ‘(na)’, short for not available.
Post-processing of model results: The fact that the toolbox is integrated in the MATLAB environment allows convenient and easy post-processing of the model output. We will illustrate this with some examples. The command \( \text{theta} = \text{OUTPUT.ebe}(\text{OUTPUT.Index}) \) gives each person its corresponding empirical Bayes estimate of the latent trait following his/her recorded response pattern and stores it in the variable \( \text{theta} \). To plot a histogram of the resulting latent trait estimates the following command is used: \( \text{hist(theta)} \). These figures can then be further adapted, annotated and exported by means of the graphic functions and editor included in MATLAB. To get an impression of the precision of measurement over the latent trait scale we can use the following commands:

\[
\text{thetaSE} = \text{OUTPUT.ebeSE}(\text{OUTPUT.Index});
\]

\[
[s \text{ order}]=\text{sort(theta)};
\]

\[
\text{plot(theta(order),thetaSE(order)});
\]

This will result in a plot of the ordered estimated latent trait values on the X-axis and their corresponding standard error on the Y-axis. The test is most precise for areas on the latent trait dimension where standard errors are the smallest.

5.2 Model building by means of design matrices

The IRTm toolbox supports non-standard models as well and this by means of design matrices (see e.g., De Boeck & Wilson, 2004). The purpose of the design matrices is to allow very general models as well as models that further constrain parameter sets. First, the main idea will be illustrated by putting the previously introduced Rasch model within the proposed framework of design matrices. Next, it is shown how this framework can be used, for instance, to fix the latent distribution to be a standard normal, to take into account an experimental design on the item responses, and to crossvalidate a model.
**General idea:** Each parameter vector in the likelihood, like for instance the latent trait variance $\sigma^2$ and the item difficulties $\beta_i$ in the Rasch model, will have a model specification part summarized in a corresponding MATLAB substructure (see Appendix and Table 5.4) consisting of fields that allow to control and construct the design in detail. The following fields are included: a general design matrix $D$, giving the exact parameterization of the model, a restriction matrix $R$, indicating which of the model parameters are to be freely estimated or fixed at a given value, and an offset matrix $O$, containing values that need to be added to the model parameter. The latter is a real valued matrix, whereas the two others are of the logical class. To compile the actual model the following expression is used:

$$\text{MODEL} = \text{Design} \times [\text{diag(Restrictions)}] \times \text{Parameters} + \text{Offset}.$$  

To make it more clear how this combination of matrices specifies the model parameterization, the previously introduced Rasch model will be used.

**Illustration:** If we take a closer look at $\text{MODEL.TS}$, the substructure for $\sigma^2$ the variance parameter of the latent trait distribution, we can see the following: the design $\text{MODEL.TS.D} = \text{logical(1)}$ indicates that there is a variance for the latent trait distribution in our model that is parameterized by one parameter. The restriction $\text{MODEL.TS.R} = \text{logical(1)}$ indicates that this parameter is to be freely estimated, and $\text{MODEL.TS.O} = 0$ indicates that its offset is 0. Thus, we have

$$\text{MODEL} = \text{Design} \times [\text{diag(Restrictions)}] \times \text{Parameters} + \text{Offset}$$

$$\sigma^2 = 1 \times [1 \times \sigma^2_{\text{est}} + 0] = \sigma^2_{\text{est}},$$

resulting in one free model parameter $\sigma^2_{\text{est}}$ defining the variance of the latent trait $\sigma^2$ in our Rasch model.
The Rasch model also requires an unique difficulty parameter $\beta_i$ for each item. This is also reflected in the design for this parameter set $\beta$ given in $MODEL.B$:

$$MODEL = \text{Design} \times [\text{diag}(\text{Restrictions}) \times \text{Parameters} + \text{Offset}]$$

$$\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6
\end{bmatrix} = \text{diag}(\begin{bmatrix}
1 \\
1 \\
1 \\
1 \\
1 \\
1
\end{bmatrix}) \times \begin{bmatrix}
\beta_1^{est} \\
\beta_2^{est} \\
\beta_3^{est} \\
\beta_4^{est} \\
\beta_5^{est} \\
\beta_6^{est}
\end{bmatrix} + \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix} = \begin{bmatrix}
\beta_1^{est} \\
\beta_2^{est} \\
\beta_3^{est} \\
\beta_4^{est} \\
\beta_5^{est} \\
\beta_6^{est}
\end{bmatrix}. $$

The main design $MODEL.B.D$ takes the form of a logical matrix, number of items $I$ by number of item parameters $K$ for that vector, with a one indicating that the parameter $k$ needs to be included for item $i$. Restrictions on the parameters are imposed by a logical vector $MODEL.B.R$, that contains the diagonal elements of a matrix, and indicate which of the parameters are free (one) and which are fixed (zero). The offset $MODEL.B.O$ is a vector with $K$ rows, consisting of values that need to be added to the model parameters (here the offsets are all set to 0). This results in each of the item difficulties $\beta_i$ in the Rasch model being parameterized by an unique difficulty parameter $\beta_{est,i}$.

**Fixing a parameter at a given value:** Consider that we want to restrict this model by requiring that the latent trait follows a standard normal distribution. This can be done by setting the restriction $MODEL.TS.R=\text{logical}(0)$, which indicates that the variance $\sigma^2$ of our latent trait is a fixed parameter now, and setting the offset of this parameter at 1 by $MODEL.TS.O=1$. All this together results in

$$MODEL = \text{Design} \times [\text{diag}(\text{Restrictions}) \times \text{Parameters} + \text{Offset}]$$

$$\sigma^2 = 1 \times [0 \times \sigma_{est}^2 + 1] = 1,$$

Thus, this constrains the variance of the latent distribution to be equal to one.
We can fit this alternative model and store the results in a second set of outcome structures:

\[
\text{OUTPUT}(2) = \text{IRTm}(\text{SETTINGS}, \text{MODEL});
\]

Comparing the two alternative models can be done by looking at the goodness-of-fit criteria,

\[
\text{IRTmSummary(OUTPUT, 'gof');}
\]

Note, that \text{OUTPUT} now consists of two structures \text{OUTPUT}(1) and \text{OUTPUT}(2). Given that the two models are nested with a difference of one free parameter, we can also opt for a likelihood ratio test with one degree of freedom:

\[
1 - \text{chi2cdf}(2 \ast (\text{OUTPUT}(2).LL - \text{OUTPUT}(1).LL), 1).
\]

Both the likelihood ratio test and the information criteria reject the more restrictive standard normal model in favor of our original Rasch model.

**Experimental design:** Let us assume that the six items are in fact realizations from an experiment in which the accuracy of hitting a target is measured in three conditions, and each person has to run through 2 trials per condition. In the first condition (item 1 and 4), there is loud noise added to the environment, in the second condition (item 2 and 5) the light in the room switches off and on, in the third condition (item 3 and 6) both disturbing effects are combined. Hence for each person 6 recordings (item responses) are made whether he/she missed the target or not. This experimental design can be reflected in the model design of the difficulty parameters using a Linear Logistic Test Model formulation (Fischer, 1973):
MODEL = Design × \[\text{diag}(\text{Restrictions}) \times \text{Parameters} + \text{Offset}\]

\[
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6
\end{bmatrix} = \begin{bmatrix}
1 \\
1 \\
1 \\
1 \\
1 \\
1
\end{bmatrix} \times \begin{bmatrix}
\beta_1^{\text{est}} \\
\beta_2^{\text{est}} \\
\beta_3^{\text{est}} \\
\beta_4^{\text{est}} \\
0 \\
0
\end{bmatrix} + \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

The first parameter \(\beta_1^{\text{est}}\) is the item difficulty due to the loud noise, \(\beta_2^{\text{est}}\) is the difficulty corresponding to the light flashes, \(\beta_3^{\text{est}}\) is the added difficulty when these effects are jointly present, and \(\beta_4^{\text{est}}\) is the training effect of having a rerun in trial 2. Thus in fact, the difficulty of the individual items is decomposed into common difficulties due to the experimental design, leading to a more parsimonious, but also more restrictive model.

The following commands setup and fit this model on our example data

\[
\text{MODEL.TS.R} = \text{logical}(1);
\]

\[
\text{MODEL.TS.O} = 0;
\]

\[
\text{MODEL.B.D} = \text{logical}([1 \ 0 \ 0 \ 0; 1 \ 0 \ 0 \ 1; 1 \ 1 \ 1 \ 1; 1 \ 0 \ 0 \ 1; 0 \ 1 \ 0 \ 1; 1 \ 1 \ 1 \ 1]);
\]

\[
\text{MODEL.B.R} = \text{true}(4,1);
\]

\[
\text{MODEL.B.O} = \text{zeros}(4,1);
\]

\[
\text{OUTPUT}(2) = \text{IRTm(SETTINGS,MODEL)};
\]

To interpret the model estimates we request the table with parameter estimates by a call to \(\text{IRTmSummary(OUTPUT(2),\ 'param',MODEL)}\).

Note that again the parameters are given in order of appearance in the likelihood, meaning that for instance the second parameter corresponds to the second column in the design \(\text{MODEL.B.D}\) of the item difficulties.

At first glance all effects seem to be significant, which should be good news to our experimenter. The effect of \(\beta_4^{\text{est}}\) is 0.801 on the latent trait scale, which means that there is a negative training effect, the
difficulty of a condition increases on the second trial compared to the first. An effect size measure interpretation is that the odds of being accurate at a rerun are twice (i.e., $\exp(0.801) = 2.228$) as low compared to the first trial (perhaps, the individuals get tired or less concentrated). Comparing this model to the original less restrictive Rasch model, one would conclude based upon the statistical criteria that the Rasch model does better. However, the interests and research questions of the experimenter are better answered by the latter LTTM model.

**Crossvalidation:** Another advantage of a design matrices approach is the possibility to crossvalidate a model (see e.g., Browne, 2000). Consider our simple example dataset, and randomly divide the sample of 500 persons in two halves. The first half will be used to estimate the model, the second half will be used to crossvalidate the estimated model.

The command `[ign randomorder] = sort(rand(500,1))` results in a random shuffle of the numbers 1 to 500 stored in the variable `randomorder`. By adapting the `SETTINGS` structure, we can select the set of persons that are to be included in the analysis, by assigning a vector of row numbers, here the first 250 numbers in the `randomorder` variable, to `SETTINGS.Sel=[randomorder(1:250)]`. A standard 2PL model is requested for the example dataset by means of `MODEL = IRTmModel(6, '2PL')`, leading to a likelihood $l(\alpha, \beta)$ with two parameter sets, item discrimination and difficulty parameters. Note that for identification reasons, the distribution of the latent trait is set to be a standard normal. The model is fitted by means of

```
OUTPUT = IRTm(SETTINGS,MODEL)
```

The set of estimated parameters reports first the set of 6 item discrimination parameters followed by the set of item difficulty parameters. For instance, in our random sample the item parameters of the third item were $\alpha_3 = 1.256$ and $\beta_3 = 0.050$.

We can crossvalidate this estimated model on the last 250 persons in the `randomorder` variable that were previously not used in the analysis. To do this we first need to select again the appropriate rows in
the dataset, restrict all the previously free parameters to be fixed, filling in the estimated parameters of the previous model in the corresponding offset fields, and of course fit this new model.

\[
\text{SETTINGS.Sel=} \{\text{randomorder}(251:500)\};
\]

\[
\text{MODEL.A.R=logical(zeros(6,1))};
\]

\[
\text{MODEL.A.O=OUTPUT.param(1:6)};
\]

\[
\text{MODEL.B.R=logical(zeros(6,1))};
\]

\[
\text{MODEL.B.O=OUTPUT.param(7:12)};
\]

\[
\text{OUTPUT(2)=IRTm(SETTINGS,MODEL)};
\]

You can then calculate self-defined crossvalidation indexes within the MATLAB environment. For instance, a simple index can be the ratio of the crossvalidated model log likelihood and the estimated model log likelihood

\[
\text{CVI = OUTPUT(2).LL/OUTPUT(1).LL}
\]

5.2.1 Accounting for covariate information

Example 2: An international study on geometry

Consider a mathematics test, included in the IRTm toolbox as MATLAB dataset ‘covtest.mat’, that is designed for an international study on geometric knowledge and that consists out of 10 items (columns 1 to 10), each representing a problem case that needs to be solved correctly. Suppose there is some background information about the 1000 pupils participating in the study, being whether they live in country E or in country F (column 11, value 1 indicating that the pupil belongs to country F, and 0 referring to country E) and a general mathematics end score (column 12) on a scale of 0 up to 100 from their last year in primary school.
We will show how this potentially interesting information can be taken into account in the IRT model for the test, including concerns about differential item functioning (DIF) and hidden person groups.

**Overall person covariate effects:** In our first example we assumed that the latent trait followed a normal distribution in the population with a given mean and variance. It is not unreasonable to think that pupils with a higher score on mathematics in primary school will on average also have a higher geometric knowledge, i.e., the latent trait the test is measuring. This reasoning can be reflected in our model formulation as \( \theta_p \sim N(\mu, Z_{pj} \lambda_j, \sigma^2) \), where \( \lambda_j \) is a person covariate parameter indicating this expected increase in geometry knowledge \( \theta_p \) depending on a person’s \( p \) mathematics score in primary school \( Z_{pj} \), and \( \mu \) is the mean parameter (set by convention at zero to fix the scale of the model) and \( \sigma^2 \) is the variance. Equivalently, this can be seen as the decomposition of the geometric knowledge \( \theta_p \) in a fixed part \( Z_{pj} \lambda_j \) due to the mathematics score and a random residual part \( \theta_p^* \), which leads to the following item characteristic function

\[
\Pr(Y_{pi} = 1|\theta_p^*, Z_p) = \frac{\exp(\theta_p^* + z_{pj} \lambda_j - \beta_i)}{1 + \exp(\theta_p^* + z_{pj} \lambda_j - \beta_i)},
\]

with \( \theta_p^* \sim N(0, \sigma^2) \).

This type of model is often referred to as a Latent Regression Model (see Andersen & Madsen, 1977) and can be regarded as the person-side analogue of the LLTM with its item decomposition in the previous experimental design example (see e.g., De Boeck & Wilson, 2004).

The new model part \( \lambda_j \) has a corresponding set of model design matrices that are situated in the `MODEL.LM` substructure. The same principles as in the previous section apply here as well. The size of the main design matrix is \( J \) covariates by \( K \) parameters. We will use the Rasch form of the Latent Regression Rasch model as a starting point in our example.
Reading in the data and setting up the algorithm can be done by first requesting and adapting the standard `DATA` structure by

```matlab
SETTINGS = IRTmSet();
load('covtest.mat')
dataset(:,12) = dataset(:,12) - mean(dataset(:,12));
SETTINGS.input = dataset;
SETTINGS.I = [1:10];
SETTINGS.J = [11:12];
```

For ease of interpretation, we did recode the continuous end score variable by centering around the sample mean.

We can request the default `MODEL` structure for a latent regression model by `MODEL = IRTmModel(10,'Rasch',2)` , in which the third argument indicates the number of person covariates, and then continue with fitting the model. However, the second covariate, the end score, is continuous and more or less out of the range of the other variables as well as the latent trait, and this might lead to estimation issues which are technically due to the global step size used in the optimization algorithms. There are two options to solve this kind of problem.

One is manually setting the initial parameter estimates for the covariate parameters $\lambda$ to appropriate values, instead of choosing for the default random start, by filling in the initial value field of the relevant model substructure. For instance set

```matlab
MODEL.LM.I = [0.5;-0.5]
```

and try fitting the model again (this can take some time!)

```matlab
OUTPUT=IRTm(SETTINGS,MODEL);
```
The second option consists of rescaling the covariate to a more reasonable range compared to the other variables and parameters. We will set the end score on a scale from -0.5 to 0.5 by dividing it by 100.

\[ \text{dataset(:,12)} = \text{dataset(:,12)} ./ 100; \]

\[ \text{SETTINGS.input} = \text{dataset}; \]

\[ \text{MODEL.LM.I} = []; \]

\[ \text{OUTPUT(2)} = \text{IRTm(SETTINGS,MODEL)}; \]

Compare the times to estimate these models, one can conclude that the second option is more efficient ( \([\text{OUTPUT.time}]\) ). The model likelihood and estimates however, do not change. Only the scale of the covariate parameter is different, i.e., instead of \( \lambda_{\text{endscore}} = 0.0227 \), it will be \( \lambda_{\text{endscore}} = 2.2763 \) (see e.g., \([\text{OUTPUT.param}]\) ), because the value of the parameter corresponds to the increase in geometric knowledge \( \theta_p \) per unit of the covariate \( Z_{pj} \). The country effect \( \lambda_{\text{country}} \) has a similar interpretation; the pupils in country F have on average 2.5 points more on the geometric knowledge scale \( \theta_p \), compared to the reference group, being the pupils in country E. The item difficulty parameters have to be interpreted acknowledging the known covariates, and as such they represent the difficulty of an item for a pupil with an average score on \( \theta_p \).

**Differential item functioning:** In practice, country effects are a sensitive issue and a topic of much discussion. Perhaps pupils with the same ability, but from a different country, do not have an equally fair chance of answering on a particular item due to some contextual or culture-specific information in the given item. This type of phenomenon is commonly referred to as DIF, because the item functions in a different way depending on the group it is presented to. DIF has an important application area in the development of culture-fair tests and can also be investigated within IRT models (Holland & Wainer, 1993).
A way of doing this is to sequentially check whether an item functions differently by adding an item-covariate interaction parameter $\zeta_{ji}$ to the latent regression model. This leads to the following item characteristic curve for the item under investigation

$$\Pr(Y_{pi} = 1 | \theta_p^*, Z_p) = \frac{\exp(\theta_p + z_{pj}\lambda_j - \beta_i - z_{pj}\zeta_{ji})}{1 + \exp(\theta_p + z_{pj}\lambda_j - \beta_i - z_{pj}\zeta_{ji})},$$

while the other items retain their original formulation. The interaction parameter $\zeta_{ji}$ represents the added difficulty for an item $i$ given a person with a value $z_{pj}$ on person covariate $j$ and given proficiency $\theta_p + z_{pj}\lambda_j$.

The introduction of this new model parameter in the likelihood again coincides with a new set of model matrices given in the substructure `MODEL.BD`. The model structure of the item-covariate interaction parameters has one added field `.P`, a row vector of length $J$, that indicates how many columns in the main design matrix `MODEL.BD.D` relate to which covariate (order corresponding to the order in `SETTINGS.J`). An example may clarify what is meant by this.

Given 10 items and two covariates, in theory a maximum of $(10+10) = 20$ item-covariate interaction parameters $\zeta_{ji}$ exist. However, most of the times a more parsimonious parameterization is more appropriate or necessary due to identification issues. Consider, that we want to investigate potential DIF on item 7 for the country effect, and on item 5 and 6 for the mathematics end score effect. Furthermore, assume that there are reasons to believe that this DIF would be the same for item 5 and 6. This can be reflected in the model design as follows. For each covariate we have a model for the interaction parameter, for instance $\zeta_{1i}$

MODEL = Design $\times [\text{diag}(\text{Restrictions}) \times \text{Parameters} + \text{Offset}]$
\[
\begin{bmatrix}
\zeta_{1,1} \\
\vdots \\
\zeta_{1,5} \\
\zeta_{1,6} \\
\zeta_{1,7} \\
\vdots \\
\zeta_{1,1}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0 \\
1 \\
0 \\
0
\end{bmatrix} \times \begin{bmatrix}
\zeta_{est}^{1,1} \\
\end{bmatrix} + \begin{bmatrix}
0 \\
0 \\
0 \\
\zeta_{est}^{1,1} \\
0 \\
0
\end{bmatrix}.
\]

and for \(\zeta_{2,i}\)

\[
\begin{bmatrix}
\zeta_{2,1} \\
\vdots \\
\zeta_{2,5} \\
\zeta_{2,6} \\
\zeta_{2,7} \\
\vdots \\
\zeta_{2,1}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
1 \\
1 \\
0 \\
0
\end{bmatrix} \times \begin{bmatrix}
\zeta_{est}^{2,1} \\
\end{bmatrix} + \begin{bmatrix}
0 \\
0 \\
\zeta_{est}^{2,1} \\
\zeta_{est}^{2,1} \\
0 \\
0
\end{bmatrix}.
\]

These matrices are consequently put next together within the \texttt{MODEL.BD} substructures, such that

\texttt{MODEL.BD.P=[1;1];}

\texttt{MODEL.BD.D=logical([zeros(6,1);1;zeros(3,1)])}

\texttt{[zeros(4,1);ones(2,1);zeros(4,1)])};

\texttt{MODEL.BD.R=logical(ones(2,1));}

\texttt{MODEL.BD.O=zeros(2,1);}

\texttt{OUTPUT(3)=IRTm(SETTINGS,MODEL);} 

\texttt{IRTmSummary(OUTPUT(3),'param',MODEL)}

The estimated parameters are again stored following their order in the likelihood which is \(l(\beta, \zeta, \lambda, \sigma)\), hence parameters 11 and 12 are the DIF parameters \((\zeta_{est}^{1,1} = 0.250(0.241), \zeta_{est}^{2,1} = 0.200(0.273))\). Looking at the p-values of the Wald tests for these item covariate interaction
parameters it appears that there is no DIF for each set of item-covariate combination we investigated so far. Note that there is less information available to estimate these interaction parameters, than there is information present in the data to estimate the other model parameters. Hence, a Wald test can not entirely be trusted, and it would be more appropriate to use for instance a likelihood ratio test and sequentially check each item-covariate combination separately by comparing a full model with the item-covariate interaction and a reduced model without that interaction (For more on DIF, see Holland & Wainer, 1993).

**Latent groups and finite mixtures:** Consider a situation in which the country information is not available in the data, yet does have a significant impact on the pupil’s results. We can create this situation by removing the country variable from the dataset. A possible way to retrieve this (now hidden) grouping on country, is to make use of a finite mixture formulation of the latent trait.

Instead of assuming that your latent trait is normally distributed with a given mean and variance, you assume that the latent trait is actually a combination of two component distributions

\[ \theta_p \sim \pi_1 N(\mu_1, \sigma^2_1) + \pi_2 N(\mu_2, \sigma^2_2), \]

in which the component weights \( \pi_g \) necessarily sum to one and are each restricted to the interval \([0, 1]\).

This type of finite mixture formulation (McLachlan & Peel, 2000) has two advantages. First, from a substantive point of view it implies that you consider having a hidden (latent) grouping on the persons in your data which might be reflected in the posterior allocation of persons among the components of the mixture and in the average latent trait values for these hidden groups. Second, from a technical point of view, this is a way to lose the restrictive assumption of a normally distributed
latent trait, and to consider non-normal distributions instead, because mixtures can mimic skewed or multimodal distributions.

To illustrate this approach, store the country covariate in the variable $Z$ for future reference, $Z = dataset(:,11)$, and then delete it from the data $dataset(:,11) = []$. We first fit a latent regression Rasch model on the ‘new’ data (with only one covariate left, the mathematics end score) to have a model to compare the finite mixture alternative to.

```matlab
SETTINGS.input=dataset;
SETTINGS.J=[11];
MODEL=IRTMModel(10,'Rasch',1);
OUTPUT=IRTMSETTINGS,MODEL);
```

Fitting the finite mixture formulation of this model, involves setting the correct algorithm, `SETTINGS.alg=1`, which is a generalized expectation maximization algorithm instead of the default Quasi-Newton implementation for the non-mixture models. For this illustration, we also limit the number of iterations to 50 and make the stopping criterion less strict to avoid having to wait too long. We put on the display option to be able to monitor the progress.

```matlab
SETTINGS.alg=1;
SETTINGS.iter=50;
SETTINGS.crit=0.01;
SETTINGS.display='on';
```

The latent trait part of the model has to be adapted as well to reflect the mixture distribution. We specify the model substructure for the means of the mixture distribution, `MODEL.TM`, and fix the mean of the first component to zero, hereby functioning as reference and setting the scale.
MODEL.TM.D=logical(eye(2));
MODEL.TM.R=logical([0;1]);
MODEL.TM.O=zeros(2,1);

For identification reasons the variances of the components have to be restricted to be the same and remain fixed at 1.

MODEL.TS.D=logical(ones(2,1));
MODEL.TS.R=logical([0]);
MODEL.TS.O=1;

Next the component weights design structure is specified, with one weight for each component.

MODEL.TW.D=logical(eye(2));
MODEL.TW.R=logical(ones(2,1));
MODEL.TW.O=zeros(2,1);

Note that the combination of restrictions on the component weights are arranged internally in the algorithm and as such the restriction matrix \( MODEL.TW.R \) can be kept free as shown above.

Having completed all necessary steps, the mixture model fitting can be started.

\[ OUTPUT(2)=IRTm(SETTINGS,MODEL); \]

Because we specified the \( SETTINGS.display \) to be on, the algorithm will display a set of numbers at each iteration. These are respectively, the iteration number, the maximum absolute difference between parameter values of consecutive iterations, the difference in log likelihood between consecutive iterations, the value of the objective function in the maximization step, and the log likelihood at the given iteration. Note that the algorithm stops when the given maximum number
of iterations is exceeded or when both the maximum absolute difference between parameter values and the difference in log likelihood between consecutive iterations, are below the stopping criterion defined in SETTINGS.crit.

Compare the fit of both models and acquire the latent group division based upon the maximum posterior component probabilities:

\[ \text{[OUTPUT.LL;OUTPUT.BIC]} \]

\[ [a \ postG] = \max (\text{OUTPUT(2).l(OUTPUT(2).Index,:)},[],2); \]

To check how good the group division is retrieved, we compare the two classifications (known and latent) in a confusion matrix \( M \) with columns components and rows observed groups (cfr. variable \( Z \)). Note that the third row and columns, contain the respective marginal total proportions:

\[ Z1=\text{sum}(Z(Z==1)); \]
\[ Z0=\text{OUTPUT(2).lN}-Z1; \]
\[ C1=\text{sum}(\text{postG}(\text{postG==1})); \]
\[ C2=\text{sum}(\text{postG}(\text{postG==2}))./2; \]
\[ Z1C1=\text{sum}(Z(\text{postG==1})); \]
\[ Z1C2=\text{sum}(Z(\text{postG==2})); \]
\[ Z0C1=C1-Z1C1; \]
\[ Z0C2=C2-Z1C2; \]
\[ M=[Z0C1 \ Z0C2 \ Z0;Z1C1 \ Z1C2 \ Z1;C1 \ \text{C2} \] \]
\[ \text{OUTPUT(2).lN}.\text{OUTPUT(2).lN} \]
If the main diagonal compares to the marginal row totals you obtained a good classification based upon your mixture components. Note that it is good practice to rerun the optimization a couple of times when applying the EM algorithm (multistart procedure), as it only guarantees to find at least a local maximum. So do not be surprised if you end up with a bad result in a first attempt. Thus unfortunately, the advantages of the mixture approach are counterbalanced a bit by its lesser convergence behavior (speed and local optima).

5.3 Conditional independence and copulas

In the previous sections, the latent trait is considered to account for all the dependency between a person’s responses on the test, referred to as the conditional independence assumption which allows the convenient formulation of the joint probability of the item responses on the full test for a person \( p \) as the product of the marginal probabilities

\[
\Pr(Y_p = y_p|\theta_p) = \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi}|\theta_p)
\]

However in practice, some subsets of items might be more interrelated than can be explained by the general latent trait \( \theta_p \) underlying the test. For instance, think of items that appeal to the same specific background knowledge, are subquestions of the same problematic case, or perhaps just follow the same question format. These item subsets are sometimes called item bundles or testlets in the literature. The prototypical example are items relating to the same reading-passage.

Residual dependencies between the items imply that conditional independence does not hold for all items anymore and that the model estimates and inferences will be distorted if you ignore it. In other words, a more adequate formulation of \( \Pr(Y_p = y_p|\theta_p) \), the joint probability of someone’s response vector is needed.
Copula functions: A convenient way of accounting for residual dependencies within an item subset are copula functions (Braeken et al., 2007). A copula is a type of function that is able to connect sets of marginal distributions to form a multivariate distribution that preserves these margins (for reference works on copula theory see Nelsen, 1999; Joe, 1997). The main idea here is that instead of coupling the marginal distributions $F(Y_{pi} = y_{pi}|\theta_p) = \Pr(Y_{pi} \leq y_{pi}|\theta_p)$ by taking their product to form the multivariate distribution under conditional independence

$$F(Y_p = y_p|\theta_p) = f(F(Y_{p1} = y_{p1}|\theta_p), \ldots, F(Y_{pI} = y_{pI}|\theta_p))$$

$$= \prod_{i=1}^{I} F(Y_{pi} = y_{pi}|\theta_p)$$

another function than the product is taken to couple these margins to form a multivariate distribution that does allow for a residual dependency structure between the different margins (i.e., items)

$$F(Y_p = y_p|\theta_p) = f(F(Y_{p1} = y_{p1}|\theta_p), \ldots, F(Y_{pI} = y_{pI}|\theta_p))$$

This function $f$ is exactly the copula function, which makes already clear that the regular independence case is a specific instance of the larger class of copula functions.

In practice, this means that we will consider the total set of items $J = \{1, 2, \ldots, I\}$ to consist of mutually exclusive item subsets $J_s (s = 1, \ldots, S)$, for which conditional independence holds between the different subsets, but where the joint probability of the responses in subset $s$ $\Pr(Y_p^{(s)} = y_p^{(s)}|\theta_p)$ is evaluated from a copula function $C_s$ (when subset size $I_s > 1$) allowing for local item dependency within the subset.

$$\Pr(Y_p = y_p|\theta_p) = \prod_{s=1}^{S} \Pr(Y_p^{(s)} = y_p^{(s)}|\theta_p)$$
Notice, that when \( S = I \) (i.e., \( I_s = 1 \forall s \)), thus each item is its own subset, this formulation reduces to the regular model formulation under conditional independence.

To implement this method, the IRTm toolbox again makes use of a small variant of the previously introduced model structures. The design matrix \( \text{MODEL.Copula.D} \) is of size number of copula functions by copula properties, with a row \( s \) looking like

\[
\begin{bmatrix}
C_s & I_s & item_1^{(s)} & \ldots & item_{I_s}^{(s)} & J_s
\end{bmatrix}
\]

The first column (and copula property) indicates the type of copula function \( C_s \), the second column indicates the number of items \( I_s \) in this copula, the third column until the \((I_s + 3)\)th column provide the item indexes in the subset \( s \), and the last column is \( J_s \) a subset identification number.

In Table 5.5 the implemented copula functions in IRTm are listed with their corresponding type number (an overview on copula families can be found in both Nelsen, 1998 and Joe, 1997). Note that a new copula \( C_s \) can also be composed as a convex linear combination of copula functions:

\[
F(Y_p = y_p | \theta_p) = C_s(F(Y_{p1} = y_{p1} | \theta_p), \ldots, F(Y_{pI} = y_{pI} | \theta_p); \delta_s)
\]

\[
= \delta_{1,2} f_1(F(Y_{p1} = y_{p1} | \theta_p), \ldots, F(Y_{pI} = y_{pI} | \theta_p); \delta_{1,1})
\]

\[
+ \delta_{2,2} f_2(F(Y_{p1} = y_{p1} | \theta_p), \ldots, F(Y_{pI} = y_{pI} | \theta_p); \delta_{2,1})
\]

with the usual restrictions for mixture-like models

\[
\sum_{k=1}^{K} \delta_{k,2}, \delta_{k,2} \in [0, 1].
\]

For this reason, one needs to specify in IRTm for each (component) copula function two parameters: \( \delta_{k,1} \), a parameter specific to the \( k \)th copula in the linear combination of copula functions, and \( \delta_{k,1} \), a param-
eter indicating the weight of the $k^{th}$ component within this combination. The restriction matrix $\mathbf{R}$ and offset matrix $\mathbf{O}$ are therefore matrices of size number of copulas by two. The parameter order in the output again follows their order in the likelihood $l(\mathbf{beta}, \delta, \sigma)$, where $\delta$ is the matrix of copula parameters, and the first column of copula parameters is reported first and then the second column (of weights for the copula functions). A few examples are needed to further clarify this.

**Example 3: A reading-comprehension test**

Consider a 10 items reading-comprehension tests, taken by 1000 students applying to a university. Some of the questions refer to similar parts in the main text that the students had to read, hence there may be some concern regarding the potential presence of local item dependency issues.

**Model 1:** To have a base of reference, a regular Rasch model is formulated under conditional independence and fitted

\[
\text{SETTINGS}=\text{IRTmSet};
\]
\[
\text{SETTINGS.input}=\text{‘example_lid.txt’};
\]
\[
\text{SETTINGS.I}=[1:10];
\]
\[
\text{MODEL}=\text{IRTmModel}(10, \text{‘Rasch’});
\]
\[
\text{OUTPUT}=\text{IRTm} (\text{SETTINGS,MODEL});
\]

**Exploration:** The IRTm toolbox offers a rough tool to assist in exploring the structure of the test, based upon the comparison of the Mantel-Haenszel (MH) statistics (Mantel & Haenszel, 1959) for each pair of items. This MH statistic essentially measures whether the odds ratio for two items is equal over the different levels of the latent trait, which is implied by the conditional independence assumption. Hence, a large value of this statistic is an indication of local item dependency for the specific item pair. A call to the function \text{IRTmLIDscan} computes
these statistics and displays two figures. A figure containing the matrix of MH statistics for all item pair combinations, and a dendrogram of a rough hierarchical clustering based upon these MH statistics. Both of these can assist in determining whether and where one has to account for residual dependency issues in the dataset. We will use the sum score $\text{Sum}$ over items for each person as a temporary proxy for the division in latent trait level groups to be able to calculate the MH statistics.

Applying this for our example, gives the following

$$Y = \text{OUTPUT}.Y(\text{OUTPUT}.\text{Index},:);$$

$$\text{Sum} = \text{OUTPUT}.S(\text{OUTPUT}.\text{Index});$$

$$\text{IRTmLIDscan}(Y, \text{Sum});$$

Based upon the MH matrix one can distinguish two subsets that have a LID issue, $J_1 = \{4, 5\}$ and $J_2 = \{6, 7, 8\}$, both of them also pop up as the lowest level clusters in the dendrogram. Thus, the items within each of these subsets will be coupled by means of a copula, hereby correcting for the extra dependency they show (beyond the dependency due to the latent trait), which should result in a more appropriate joint probability of the item responses and a better model for our test.

**Model 2:** Based upon the previous exploratorions, a copula Rasch model with copulas to couple the items within the respective subsets $J_1$ and $J_2$ is constructed. Adding the copula part to the model can be done as follows

$$[C_s \ I_s \ item_1^{(s)} \ ... \ item_I^{(s)} \ J_s]$$

$$\text{MODEL}(2) = \text{MODEL};$$

$$\text{MODEL}(2).\text{Copula.D} = [3 \ 2 \ 4 \ 5 \ 0 \ 1; \ 2 \ 3 \ 6 \ 7 \ 8 \ 2];$$

The first copula is of type 3, which is a Gumbel-Hougaard copula, contains 2 items (4 and 5) and is labeled as subset 1. The second copula is
of type 2, which is a Cook-Johnson copula, contains 3 items (6, 7, and 8) and is labeled as subset 2. Notice that the first copula row is padded with a zero to obtain the same length for the design row of each copula function.

\[ \text{MODEL(2).Copula.R} = \text{logical([1 0;1 0])}; \]

The parameter specific to each copula (first column) are set free to be estimated and the component weight parameters (second column) are fixed to one, because each subset is only modeled by one copula function.

\[ \text{MODEL(2).Copula.O} = [1 1;0 1]; \]

The parameter for the first copula (first row, first column), which is a Gumbel-Hougaard copula is offset by one to improve estimation. The parameters for the copula mixture weights (second column) are both offset at 1, because there is only one copula involved for each subset. This copula Rasch model is then estimated by calling

\[ \text{OUTPUT(2) = IRTm(SETTINGS,MODEL(2));} \]

Looking at the parameter estimates in the summary table, \( \text{IRTmSummary(OUTPUT(2), 'param',MODEL(2))} \), both copula parameters appear to be significant (\( p < 0.0001 \)), indicating that there is indeed a LID issue in the dataset. Furthermore, when comparing the regular conditional independence Rasch model and the copula Rasch alternative, one can see that the latter model outperforms the former by far, \( \text{IRTmSummary(OUTPUT,'gof')} \).

**Model 3:** Alternatively, we could have chosen other copula functions, and for instance opt for a mixture of the Fréchet-Hoeffding upper bound and the independence case instead of Gumbel-Hougaard copula for the first subset

\[ \text{MODEL(3)=MODEL(1);} \]

\[ \text{MODEL(3).Copula.D} = [0 2 4 5 0 1]; \]
The last column in the design \( \text{MODEL(3).Copula.D} \) indicates that the first two copula functions relate to the same subset and should be put together in a linear combination. Both component weights are set free and the copula parameters are fixed as both the independence copula and Fréchet-Hoeffding upper bound copula are non-parametric limiting cases. Note that for estimation purposes the constrained optimization has to be chosen in the \( \text{SETTINGS} \) structure as the mixture parameters in the linear convex combination have to sum up to one and should each be within the interval \([0,1]\). To help the constrained optimization algorithm, the initial values of the item difficulty parameters are set at their parameter estimates under the regular Rasch model.

Comparing the three models, \( \text{IRTmSummary(OUTPUT,’gof’) } \), shows that our first copula Rasch model was the best one, but that the alternative copula Rasch model is a close second. Constrained optimization is more demanding than regular unconstrained optimization, so that the possibility should be taken into account of ending up with a local maximum or a non-existing solution as a result of a bad run of the algorithm. Furthermore for identification reasons, parametric copulas can only be put in a linear combination with their copula parameters as fixed and known (e.g., by estimating them separately in a first step).
In our case, running this sequence of three models took about 2 minutes in time on a personal computer running Windows XP and MATLAB7.4.0(R2007a) with a Pentium 2.8Ghz processor and 512MB RAM. This can be checked by vectorizing the time field of the different OUTPUT structures, and dividing the sum of this vector by 60 to obtain the total time spend in minutes( \( \text{sum([OUTPUT.time])}/60 \)).

5.4 Data simulation

The IRTm toolbox can also be used to for data simulation. For instance simulation studies can be set up to investigate what the consequences are of for instance ignoring certain model assumption violations or putting constraints on specific model parameters. Data simulation is also needed for parametric bootstraps (Efron & Tibshirani, 1993), in which new data are simulated following the fitted model, and compared to the original dataset, might be useful to further assess model fit.

With the IRTm toolbox, data can be simulated from a MODEL structure built up by means of the following function call \([Y \ t h] = \text{IRTm Sim}(\text{MODEL}, P, Z)\). The matrix \( Y \) will contain the simulated item responses, while the vector \( th \) will contain the latent trait values of the \( P \) simulated persons. Person covariate information can be added using an optional matrix \( Z \), of size \( P \times J \), to be specified when calling the function. Note that the offset field will be taken as parameter values, and hence the model you will be simulating data from is

\[
\text{MODEL} = \text{Design} \times [\text{Parameters} = \text{Offset}]
\]

Illustration: For instance, to simulate item responses for 500 persons following a Rasch model with 10 items generated from a standard normal distribution and a latent trait with mean 0 and variance 1.5, you can do the following:

\[
\text{MODEL} = \text{IRTmModel}(10, \text{’Rasch’});
\]
\[ MODEL.B.O = \text{randn}(10, 1); \]
\[ MODEL.TS.O = 1.5; \]
\[ [Y \ th] = \text{IRTmSim}(MODEL, 500); \]

To save all variables in the workspace in a MATLAB datafile, use

\[ \text{save}('\text{nameoffile.mat}'); \]

If you want to save only specific variables use the following format

\[ \text{save}('\text{nameoffile.mat}', '\text{nameofMATLABvariable}'); \]

To save the data in a space-delimited text file use the following

\[ \text{dlmwrite}('\text{nameoffile.txt}', Y, ' '); \]

In order to save the file in another directory than the working directory (which is the default), one has to specify the filename using an absolute address.

5.5 Discussion

In this manuscript, we presented the IRTm toolbox for fitting IRT models for binary item responses in the MATLAB environment. We briefly illustrated some applications of a quite general, liberal and explanatory IRT model and illustrated the potential of the toolbox using some didactical examples. However, advances can still be made, and we would definitely welcome cooperation. Some potential further developments are the following: further optimize speed and memory use, for instance by accelerating the implemented EM algorithm, allow for data missing completely at random, provide other prototype MODEL structures for standard IRT models, add extra utility functions for standard IRT test statistics and plotting. Some ongoing developments include polytomous item responses and alternative LID models like the conditional interaction models (Hoskens & De Boeck, 1997) and testlet models (Wainer, Bradlow, & Wang, 2007).
When developing this toolbox, our focus was not primarily on offering an alternative for existing commercial software, the focus was rather on model building in itself. The adopted design-matrix approach allows for control and flexibility and requires you to think during the model buildup stage, this in contrast to mindlessly applying one of the available and widely known standard models. This property might make the proposed tool suitable to be used in a measurement course to acquire more insight in, and to demystify common IRT models in practice. MATLAB is widely used by a large variety of users, both in research as well as in business, and as such this IRT toolbox may be useful to some of them. Furthermore, it is currently the only software that offers an implementation of the copula approach to LID modeling.
References


Cancer Institute, 22, 719–748.


5.6 Appendix: The copula IRT model

Over the last years, it has been shown that models from item response theory can be conceptualized as generalized linear or nonlinear mixed models (see e.g., Agresti et al., 2000; De Boeck & Wilson, 2004; Rijmen et al., 2003), if one considers the latent trait to be a person-specific random effect with a given distribution in the population. Within this mixed model framework, we will very concisely lay out the pieces that together compose the joint marginal likelihood of the general copula IRT model the IRTm toolbox can accommodate:

\[
l(\alpha, \zeta, \beta, \xi, \omega, \delta, \lambda, \psi, \mu, \sigma, \pi) = \prod_{p=1}^{P} \sum_{g=1}^{G} \pi_g \int_{\theta_{pg}} \prod_{s=1}^{S} \Pr \left( Y_{ps}^{(s)} | \theta_p \right) \phi(\theta_{pg}) d\theta_{pg}.
\]

The density of the component distribution \( g \) of the latent variable \( \theta_{pg} \) is

\[
\phi(\theta_{pg}) \sim N \left( \mu_g + \sum_{j=1}^{J} Z_{pj} \lambda_j, \sigma_g^2 + \sum_{j=1}^{J} Z_{pj} \psi_j^2 \right),
\]

in which both mean and variance can be decomposed in a fixed part due to the effect of a person covariate \( Z_{pj} \) (\( \lambda_j \) and \( \psi_j^2 \), respectively) and a latent part (\( \mu_g \) and \( \sigma_g^2 \), respectively). Each component of the latent trait distribution is further parameterized by means of a component weight \( \pi_g \), with \( \sum_{g=1}^{G} \pi_g = 1, \pi_g \in [0, 1] \). \( \Pr \left( Y_{ps}^{(s)} | \theta_{pg} \right) \) is then the joint probability of the subset \( s \) under the copula \( C_s \left( F(Y_{p1}^{(s)} | \theta_{pg}), \ldots, F(Y_{pIs}^{(s)} | \theta_{pg}); \delta_s \right) \) given \( \theta_{pg} \), the component latent trait estimate of person \( p \). The conditional distribution function of an item \( i \) is defined as

\[
F(Y_{pi} = y_{pi} | \theta_{pg}) = \begin{cases} 
0 & y_{pi} < 0 \\
\frac{(1-\omega_i)}{1+\exp(\eta_{pi})} & y_{pi} = 0 \\
1 & y_{pi} > 0 
\end{cases}
\]
with predictor component $\eta_{pi} = (\alpha_i - z_{pj}\varsigma_{ji})(\theta_{pg} - \beta_i - z_{pj}\zeta_{ji})$, where $\omega_i$ functions as lower-asymptote of the logistic function and is also known as the pseudo-guessing parameter in the 3-parameter logistic model, $\alpha_i$ is the item discrimination parameter, $\beta_i$ is the item difficulty parameter, and $\varsigma_{ji}$ and $\zeta_{ji}$ are corresponding item-covariate interaction parameters.

This model formulation allows the fitting of most known binary item response models (Rasch, 2PL, 3PL), but also accommodates the linear logistic test model (Fischer, 1973), latent regression models through person covariates, differential item functioning (Holland & Wainer, 1993), and copula models to deal with local item dependencies (LID; Braeken et al., 2007). Estimation of the models will be done following full information marginal maximum likelihood (Bock & Lieberman, 1970) by means of a quasi-Newton algorithm or a generalized Expectation-Maximization algorithm (McLachlan & Krishnan, 1997). The IRTm toolbox takes a marginal maximum likelihood approach to model estimation, and hence, assumes that the proficiency of the persons follows a distribution in the population. In the MML approach this population distribution is integrated out to surpass the Neyman-Scott problem (1948) leading to consistent estimates for both fixed and random parameters. The distribution of the latent trait can either be (1) parametric, a Normal distribution or (2) semi-parametric, a finite mixture of Normals (both approximated by means of Gauss-Hermite quadrature).

The integral(s) in the likelihood expression has no closed-form solution and will be approximated using Gauss-Hermite quadrature. Making use of the Bayes theorem, empirical Bayes estimates of the latent proficiencies $\hat{\theta}_p$ can be computed using

$$\hat{\theta}_p = \prod_{p=1}^{P} \sum_{g=1}^{G} \pi_g \int_{\theta_{pg}} \prod_{s=1}^{S} \Pr_s \left( Y_{(s)}^{p} | \theta_{pg} \right) \phi(\theta_{pg}) d\theta_{pg}$$

The stopping criterion of the generalized EM algorithm uses two refer-
ences, the maximal absolute difference between parameter values of two successive iterations and the corresponding difference in log likelihood. When both are below \texttt{SETTINGS.crit}, the algorithm stops.

Some features that will increase general computation time are non-linear terms in the predictor component (e.g., discrimination and pseudoguessing parameters), an increasing complexity of the latent trait distribution on the model part (e.g., multiple components), a large DIF part in the model, an increasing amount of items and unique response patterns on the data part, an increasing number of quadrature points, and decreasing the stopping criterion for the generalized EM technique on the algorithmic part.

Where possible, the performance of the algorithms was crosschecked with SAS Proc Nlmixed and was found to be quite satisfactory. It should be noted that SAS biggest handicap is time, especially when Proc Nlmixed is requested to provide estimates of the latent trait.
### Table 5.1: Field names of the IRTm SETTINGS structure and a brief description of their contents.

<table>
<thead>
<tr>
<th>Field name</th>
<th>Type</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Data</td>
<td>string containing path to data file or alternatively a MATLAB matrix</td>
</tr>
<tr>
<td>I</td>
<td>Data</td>
<td>vector with item column numbers</td>
</tr>
<tr>
<td>J</td>
<td>Data</td>
<td>vector with covariate column numbers (empty vector [] when not applicable)</td>
</tr>
<tr>
<td>Sel</td>
<td>Data</td>
<td>vector with persons to be included in the analysis (empty vector [] for whole sample)</td>
</tr>
<tr>
<td>delim</td>
<td>Data</td>
<td>delimiter between columns when reading in data file (default is space)</td>
</tr>
<tr>
<td>alg</td>
<td>Algorithm</td>
<td>algorithm to be used: 0 for quasi-Newton and 1 for generalized EM</td>
</tr>
<tr>
<td>qp</td>
<td>Algorithm</td>
<td>number of quadrature points to be used when integrating out the latent trait</td>
</tr>
<tr>
<td>iter</td>
<td>Algorithm</td>
<td>maximum number of iterations the algorithm is allowed to run</td>
</tr>
<tr>
<td>func</td>
<td>Algorithm</td>
<td>maximum number of likelihood evaluations an algorithm may run for one iteration</td>
</tr>
<tr>
<td>crit</td>
<td>Algorithm</td>
<td>stopping criterion value of generalized EM algorithm (no influence on quasi-Newton)</td>
</tr>
<tr>
<td>display</td>
<td>Algorithm</td>
<td>string (‘on’ or ‘off’) regulating display of the optimization procedure’s progress</td>
</tr>
<tr>
<td>con</td>
<td>Algorithm</td>
<td>optimization type: 0 for regular unconstrained and 1 for constrained</td>
</tr>
<tr>
<td>Field name</td>
<td>Type</td>
<td>Contents</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>AIC</td>
<td>Fit</td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Fit</td>
<td>Bayesian information criterion</td>
</tr>
<tr>
<td>Chi2</td>
<td>Fit</td>
<td>pseudo chi-square statistic based upon the response patterns present in the data</td>
</tr>
<tr>
<td>Freq</td>
<td>Data</td>
<td>frequencies of unique response patterns in the data</td>
</tr>
<tr>
<td>HES</td>
<td>Optim</td>
<td>Hessian matrix</td>
</tr>
<tr>
<td>HEScheck</td>
<td>optim</td>
<td>outcome of positive definiteness check of the Hessian</td>
</tr>
<tr>
<td>Index</td>
<td>Data</td>
<td>index towards the persons with corresponding unique response patterns</td>
</tr>
<tr>
<td>LL</td>
<td>Fit</td>
<td>optimized value of the log likelihood</td>
</tr>
<tr>
<td>Npred</td>
<td>Pred</td>
<td>predicted number of persons for each of the response patterns present in the data</td>
</tr>
<tr>
<td>RMSE</td>
<td>Trait</td>
<td>root mean squared error of the latent trait estimates</td>
</tr>
<tr>
<td>S</td>
<td>Data</td>
<td>sum score over items for each of the response patterns present in the data</td>
</tr>
<tr>
<td>Y</td>
<td>Data</td>
<td>item part of the unique response patterns in the data</td>
</tr>
<tr>
<td>Z</td>
<td>Data</td>
<td>covariate part of these unique response patterns</td>
</tr>
<tr>
<td>algorithm</td>
<td>Optim</td>
<td>name of applied optimization algorithm</td>
</tr>
<tr>
<td>ebe</td>
<td>Trait</td>
<td>empirical Bayes estimate of the latent trait</td>
</tr>
<tr>
<td>ebeSE</td>
<td>Trait</td>
<td>standard error of ebe</td>
</tr>
<tr>
<td>funCount</td>
<td>Optim</td>
<td>number of likelihood evaluations in the last quasi-Newton iteration</td>
</tr>
<tr>
<td>init</td>
<td>Optim</td>
<td>initial values for the model parameters</td>
</tr>
<tr>
<td>iterations</td>
<td>Optim</td>
<td>number of iterations run by the algorithm</td>
</tr>
<tr>
<td>l</td>
<td>Pred</td>
<td>posterior probability (one column per component)</td>
</tr>
<tr>
<td>II</td>
<td>Data</td>
<td>number of items</td>
</tr>
<tr>
<td>lJ</td>
<td>Data</td>
<td>number of covariates</td>
</tr>
<tr>
<td>lN</td>
<td>Data</td>
<td>number of persons</td>
</tr>
<tr>
<td>lP</td>
<td>Data</td>
<td>number of parameters</td>
</tr>
<tr>
<td>lX</td>
<td>Data</td>
<td>number of unique response patterns</td>
</tr>
<tr>
<td>message</td>
<td>Optim</td>
<td>general information on optimization outcome</td>
</tr>
<tr>
<td>muY</td>
<td>Data</td>
<td>proportions of items</td>
</tr>
<tr>
<td>muX</td>
<td>Data</td>
<td>means of covariates</td>
</tr>
<tr>
<td>optim</td>
<td>Optim</td>
<td>outcome code of the optimization procedure (0 indicating non-convergence)</td>
</tr>
<tr>
<td>p</td>
<td>Pred</td>
<td>p-values of Wald test for estimated model parameters</td>
</tr>
<tr>
<td>param</td>
<td>Pred</td>
<td>estimated model parameters in order of appearance in the model likelihood</td>
</tr>
<tr>
<td>paramSE</td>
<td>Pred</td>
<td>standard errors of estimated model parameters</td>
</tr>
<tr>
<td>stepsize</td>
<td>Optim</td>
<td>last step size in quasi-Newton iteration</td>
</tr>
<tr>
<td>time</td>
<td>Optim</td>
<td>total time spend by the procedure in seconds</td>
</tr>
<tr>
<td>trait</td>
<td>Trait</td>
<td>trait estimate being the sum of the ebe estimate of the residual latent trait and the covariate effect for each unique response pattern</td>
</tr>
</tbody>
</table>
Table 5.3: Rasch estimates for IRTm, SAS, and MPLUS.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IRTm</th>
<th>SAS</th>
<th>MPLUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-0.244 (0.119)</td>
<td>-0.244 (0.119)</td>
<td>-0.244 (0.119)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-1.762 (0.142)</td>
<td>-1.762 (0.142)</td>
<td>-1.762 (0.142)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.063 (0.118)</td>
<td>0.063 (0.118)</td>
<td>0.063 (0.118)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.286 (0.119)</td>
<td>0.286 (0.119)</td>
<td>0.286 (0.119)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-1.049 (0.126)</td>
<td>-1.049 (0.126)</td>
<td>-1.049 (0.127)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>1.234 (0.130)</td>
<td>1.234 (0.130)</td>
<td>1.234 (0.130)</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>1.649 (0.216)</td>
<td>1.649 (0.216)</td>
<td>1.649 (0.208)</td>
</tr>
</tbody>
</table>

$\theta_{p|\text{sumscore}} = 0$ -2.019 (0.859) -1.927 (0.846) -2.019 (na)
$\theta_{p|\text{sumscore}} = 1$ -1.337 (0.797) -1.281 (0.781) -1.337 (na)
$\theta_{p|\text{sumscore}} = 2$ -0.731 (0.765) -0.701 (0.748) -0.731 (na)
$\theta_{p|\text{sumscore}} = 3$ -0.156 (0.755) -0.159 (0.738) -0.156 (na)
$\theta_{p|\text{sumscore}} = 4$ 0.422 (0.770) 0.389 (0.750) 0.422 (na)
$\theta_{p|\text{sumscore}} = 5$ 1.042 (0.809) 0.975 (0.790) 1.042 (na)
$\theta_{p|\text{sumscore}} = 6$ 1.752 (0.881) 1.647 (0.866) 1.752 (na)

Time in seconds 0.51 8.03 0.00 (na)

Table 5.4: Field names of the IRTm MODEL structure and model parts they refer to.

<table>
<thead>
<tr>
<th>Substructures</th>
<th>Model Part</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>item discrimination $\alpha_i$</td>
<td>Item</td>
</tr>
<tr>
<td>AD</td>
<td>item discrimination-covariate interaction $\beta_i$</td>
<td>Item</td>
</tr>
<tr>
<td>B</td>
<td>item difficulty $\beta_i$</td>
<td>Item</td>
</tr>
<tr>
<td>BD</td>
<td>item difficulty-covariate interaction $\zeta_{ji}$</td>
<td>Item</td>
</tr>
<tr>
<td>C</td>
<td>item pseudo-guessing $\omega_i$</td>
<td>Item</td>
</tr>
<tr>
<td>LM</td>
<td>covariate effect on the mean $\lambda_j$</td>
<td>Covariate</td>
</tr>
<tr>
<td>LS</td>
<td>covariate effect on the variance $\psi_{ji}$</td>
<td>Covariate</td>
</tr>
<tr>
<td>TM</td>
<td>component latent trait mean $\mu_g$</td>
<td>Trait</td>
</tr>
<tr>
<td>TS</td>
<td>component latent trait variance $\sigma_g^2$</td>
<td>Trait</td>
</tr>
<tr>
<td>TW</td>
<td>Component latent trait weight $\pi_g$</td>
<td>Trait</td>
</tr>
<tr>
<td>Copula</td>
<td>Copula LID subset $\delta$</td>
<td>Copula</td>
</tr>
</tbody>
</table>

Table 5.5: Field names of the IRTm MODEL structure and model parts they refer to.

<table>
<thead>
<tr>
<th>Type</th>
<th>Copula function</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Independence (non-parametric)</td>
</tr>
<tr>
<td>1</td>
<td>Frank copula</td>
</tr>
<tr>
<td>2</td>
<td>Cook-Johnson copula</td>
</tr>
<tr>
<td>3</td>
<td>Gumbel-Hougaard copula</td>
</tr>
<tr>
<td>4</td>
<td>Placket copula (only for $I_s = 2$)</td>
</tr>
<tr>
<td>5</td>
<td>Fréchet-Hoeffding lower bound (non-parametric)</td>
</tr>
<tr>
<td>6</td>
<td>Fréchet-Hoeffding upper bound (non-parametric)</td>
</tr>
</tbody>
</table>

*non-parametric in the sense that no additional parameters are involved.
EPilogue

Reflection

The primary focus of the dissertation was to introduce copula functions into the domain of latent variable modeling, to deal with violations of the conditional independence assumption. The first four chapters construct the basic foundations of this copula approach to residual dependencies, and a software tool is described in the fifth chapter. It is illustrated that the copula approach can be applied throughout latent variable models that have an increasing degree of model complexity and that model estimation can be implemented in a rather straightforward fashion. The copula approach is shown to have nice theoretical and analytical properties, mainly due to what McCullagh (1989) calls “upward compatibility”. The marginal model parts are not changed and retain their original format even after the introduction of the association structure (i.e. the copula part). This is useful for straightforward interpretations of the marginal model and for flexible model building. There is a large flexibility in defining the marginal probability model for individual indicators, but also a large flexibility regarding the kind of dependence structure that is implemented (by means of different copula functions) for the joint behavior of a set of manifest variables.

However, theoretical and analytical advantages do not guarantee usefulness in the applied field. Therefore, a next step should be to evaluate the performance of the copula approach on a more practical level, for instance in the context of the typical large scale assessments in educational measurement. This way the copula approach can be better
situated and its practical value determined in comparison to competing modeling approaches (see e.g., Hoskens & De Boeck, 1997; Bradlow, Wainer, & Wang, 1999; Thissen, Steinberg, & Mooney, 1989). It would be interesting to study in detail the influence on general diagnostics and other measures and practices (e.g., classification of masters and non-masters) within the measurement domain. The application of the copula approach in substantive contexts will also foster the need for and development of new insights in the model structure, results, and guidelines for specific modeling choices. The first steps are made, but many challenges still lay ahead, some of which we will explore in the following.

**Future Directions**

**Copulas for non-exchangeable and/or overlapping dependence structures**

The implemented copula approach assumes exchangeability within subset and exchangeability between the different mutually exclusive subsets. The mutual exclusive subsets imply that there is no overlap of indicators between subsets. Exchangeability is used here in a sense that the exchangeable entities provide information that is independent of the order in which they are collected. This results in an interpretative attractive simple structure for the dependency between the manifest indicators conditional upon the set of latent variables. However, some situations like test speededness, the measurement of change, or overlapping subdesigns within a set of indicators might require a more complicated proposed model structure. Hence, one might want to look into further generalizations of multivariate copula functions to handle these types of dependency structures. Construction methods to derive a non-exchangeable bivariate copula from a exchangeable bivariate copula have recently become available. However, this is not readily extended to the multivariate case, because the theory is quite complex as it is difficult to verify whether the proposed constructions lead to valid multivariate cop-
ula distributions is not straightforward (Joe, 1997). It should be noted though that there is some controversy about the usefulness of copula functions to model non-exchangeable asymmetrical association as for instance is the case in time-dependent processes (see for a lively discussion, Mikosch, 2006). The main issue is with the presentation of copulas as a universal modeling tool for all types of dependence. This is of course a bridge too far, but when one has good ideas of univariate distributions for individual variables and want to model the dependence as a second step, copula functions might indeed be an appropriate option.

While there is an extensive collection of copula functions for the bivariate case, this is less so for the multivariate case. In other words, the curse of dimensionality is also present in this approach, adding more margins (i.e., dimensions) complicates matters to construct a valid multivariate copula distribution. This is also the reason why we limited ourselves in this dissertation to three of the best-studied multivariate archimedean copulas (Frank, Cook-Johnson, Gumbel-Hougaard) that each have a prototypical and different dependency profile. An obvious extension to our approach is therefore to explore, implement, and investigate other copula families that can accommodate different types of association. The possibility of more complicated dependence structures also complicates the interpretation of the model-implied dependence structure. After all, the dependency between categorical variables can not be fully described by means of the widely known concept of linear correlation that applies for continuous (normal) variables. A general challenge is to find an accepted, informative and parsimonious way of describing the model-implied dependency within copula-extended latent variable models.

Copulas for non-linear relations between latent variables

So far copula functions were introduced to model residual dependencies between categorical manifest variables. However, one should certainly not limit the copula approach to this case, because most applications in
the literature actually involve the association between continuous variables. Recently there has been some interest in models that allow for non-linear relations between the continuous latent variables. The motivation of this is to be found in the interest of substantive researchers in these type of relations, as not all things in the world are connected in a mere linear way. Instead of the mathematical attractive and convenient linear correlations or linear regression relations between different latent variables, cross-product interactions and polynomial relations have been explored. The most simple approach uses two steps in which first non-linear combinations of manifest indicators are added to the model as virtual indicators and afterwards the resulting estimates are transformed following non-linear constraints. However, this method may not be statistically sound, because usual maximum likelihood theory is not fully applicable here. Arminger and Muthén (1998), and Lee and Zhu (2002) have now developed more appropriate estimation methods.

An alternative approach might be to apply the copula functions also on the latent variable level. The amount of known and well-studied multivariate distributions is rather scarce, certainly when compared to their univariate counterparts. Instead of applying a multivariate normal distribution by default due to a lack of other options, a new multivariate distribution can be constructed based upon chosen marginal univariate distributions for the individual latent variables and a specific copula function to combine them. This copula constructed multivariate distribution will lead to non-linear (non-normal) relations between the latent variables, and especially the copulas with tail dependence properties might be interesting choices here. In the context of non-linear repeated measurement models, Lindsey and Lindsey (2006) already showed that adopting different multivariate distributions than the default multivariate normal, generally yields an improved fit, and can even lead to different conclusions. Hence, making use of the copula approach to construct alternative multivariate distributions for the set of latent variables might be more than a simple robustness check.
References


