Jukka Jokinen

Joint Regression and Association Models for Repeated Categorical Responses

Publications of the National Public Health Institute A 21/2006

Department of Vaccines
National Public Health Institute Helsinki, Finland
and
Division of Biometry
Department of Mathematics and Statistics
Faculty of Science
University of Helsinki, Finland

Helsinki 2006

JOINT REGRESSION AND ASSOCIATION MODELS FOR REPEATED CATEGORICAL RESPONSES

JUKKA JOKINEN

Academic Dissertation for the Degree of Doctor of Philosophy

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium III, Porthania (Yliopistonkatu 3), on February 3rd 2007, at 10 am.

Department of Vaccines
National Public Health Institute
Helsinki, Finland

and

Division of Biometry
Department of Mathematics and Statistics
Faculty of Science
University of Helsinki, Finland

Helsinki 2006

Publications of the National Public Health Institute
KTL A21 / 2006

Julkaisija - Utgivare - Publisher

Kansanterveyslaitos (KTL)
Mannerheimintie 166
00300 Helsinki
Puh. vaihde (09) 474 41, telefax (09) 4744 8408

Folkhälsoinstitutet
Mannerheimvägen 166
00300 Helsingfors
Tel. växel (09) 474 41, telefax (09) 4744 8408

National Public Health Institute
Mannerheimintie 166
FIN-00300 Helsinki, Finland
Telephone +358 9 474 41, telefax +358 9 4744 8408

ISBN 951-740-677-0
ISSN 0359-3584
ISBN 951-740-678-9 (pdf)
ISSN 1458-6290 (pdf)

Edita Prima Oy
Helsinki 2006

Supervised by

Professor (emer.) Anders Ekholm
Division of Biometry
Department of Mathematics and Statistics
Faculty of Science
University of Helsinki
Helsinki, Finland

Reviewed by

Professor Chris Skinner
Division of Social Statistics
School of Social Sciences
University of Southampton
Southampton, United Kingdom

Professor Esa Uusipaikka
Department of Statistics
University of Turku
Turku, Finland

Opponent

Professor Philippe Lambert
Institut des Sciences Humaines et Sociales
Université de Liège
Liège, Belgium

Jukka Jokinen, Joint regression and association models for repeated categorical responses
Publications of the National Public Health Institute, A21/2006, 35 pages
ISBN 951-740-677-0; 951-740-678-9 (pdf-version)
ISSN 0359-3584; 1458-6290 (pdf-version)
http://www.ktl.fi/portal/4043

Abstract

The focus of this study is on statistical analysis of categorical responses, where the response values are dependent of each other. The most typical example of this kind of dependence is when repeated responses have been obtained from the same study unit. For example, in Paper I, the response of interest is the pneumococcal nasopharengyal carriage (yes/no) on 329 children. For each child, the carriage is measured nine times during the first 18 months of life, and thus repeated respones on each child cannot be assumed independent of each other.

In the case of the above example, the interest typically lies in the carriage prevalence, and whether different risk factors affect the prevalence. Regression analysis is the established method for studying the effects of risk factors. In order to make correct inferences from the regression model, the associations between repeated responses need to be taken into account. The analysis of repeated categorical responses typically focus on regression modelling. However, further insights can also be gained by investigating the structure of the association. The central theme in this study is on the development of joint regression and association models.

The analysis of repeated, or otherwise clustered, categorical responses is computationally difficult. Likelihood-based inference is often feasible only when the number of repeated responses for each study unit is small. In Paper IV, an algorithm is presented, which substantially facilitates maximum likelihood fitting, especially when the number of repeated responses increase. In addition, a notable result arising from this work is the freely available software for likelihood-based estimation of clustered categorical responses.

Jukka Jokinen, Joint regression and association models for repeated categorical responses
Kansanterveyslaitoksen julkaisuja, A21/2006, 35 sivua
ISBN 951-740-677-0; 951-740-678-9 (pdf-versio)
ISSN 0359-3584; 1458-6290 (pdf-versio)
http://www.ktl.fi/portal/4043

Tiivistelmä

Tutkimus käsittelee kategorisen vasteen tilastollista analyysiä tilanteessa, jossa vastearvojen välillä on riippuvuutta. Tyypillisimmillään tällaista riippuvuutta esiintyy silloin, kun samalta tutkimuskohteelta on havaittu vaste useana ajankohtana. Esimerkiksi tämän työn ensimmäisessä artikkelissa tutkimuskohteena on 329 lasta, ja tutkittavana vasteena on pneumokokkibakteerin nielukantajuus (kyllä/ei). Kantajuus on mitattu kultakin lapselta yhdeksän kertaa ensimmäisen 18 ikäkuukauden aikana, jolloin saman lapsen toistuvien mittausten ei voida olettaa olevan riippumattomia toisistaan.

Esimerkin kaltaisessa tilanteessa ollaan tyypillisesti kiinnostuneita kantajuuden yleisyydestä, sekä siitä, onko tietyillä riskitekijöillä vaikutusta yleisyyteen. Riskitekijöiden vaikutusta tarkastellaan regressiomallilla. Jotta regressiomallista tehtävät päätelmät eivät olisi virheellisiä, on analyysissä otettava huomioon toistettujen mittausten välinen riipuvuus. Analyysin pääpaino on tavallisesti virheettömässä regressiomallinnuksessa. Kuitenkin vastearvojen välisen riippuvuuden tutkimuksella voidaan saavuttaa arvokasta lisäinformaatiota. Tämän työn keskeisenä teemana on regression ja vastearvojen riippuvuuden samanaikainen tilastollinen mallinnus.

Toistetun, tai muuten ryhmitellyn, kategorisen vasteen analyysi on laskennallisesti haastavaa. Uskottavuusperusteinen päättely on tyypillisesti mahdollista vain, jos toistettuja mittauksia on kultakin tutkimuskohteelta vain muutama. Tämän työn neljännessä artikkelissa esitellään laskenta-algoritmi, joka helpottaa huomattavasti suurimman uskottavuuden estimointia, eritoten kun toistojen lukumäärä kasvaa. Olennainen osa tutkimuksen tuloksia on myös vapaasti saatavilla oleva ohjelmisto ryhmitellyn kategorisen vastemuuttujan uskottavuusanalyysiin.

Acknowledgements

I want to express my sincere gratitude to my supervisor, Anders Ekholm, for his guidance and overwhelming support throughout my doctoral studies. His impeccable attitude toward scientific research has taught me a lot, most of all patience, at the expense of his own at times. I am also very grateful for the extensive and fruitful collaboration with Peter Smith and Mac McDonald in Southampton, which has played a pivotal role in my studies. However, this collaboration would not have been possible without the head of the Department of Vaccines, Terhi Kilpi. Her astute and continuous encouragement was often as important as the opportunities she provided me with for pursuing my studies. In addition, I warmly thank the head of the Division of Biometry, Elja Arjas, for his help and guidance, especially during the final stages of my studies.

I also thank my friends and my colleagues in Southampton and in Helsinki for healthy and not so healthy distractions throughout the years. Most sincere thanks to my family, for providing the most important distraction, and for never bothering me with awkward questions related to my studies.

Grateful acknowledgements to the National Public Health Institute Department of Vaccines, the Rolf Nevanlinna Institute Research Foundation, the Yrjö Jahnsson Foundation, the Academy of Finland and the Southampton Statistical Sciences Research Institute for their financial support.

Helsinki, December 2006

Jukka Jokinen

List of original publications

This thesis is based on the following original articles which are referred to in the text by their corresponding Roman numerals.

I Ekholm A, Jokinen J, Kilpi T. 2002. Combining regression and association modelling for longitudinal data on bacterial carriage. Statistics in Medicine, 21, 773-791.

II Ekholm A, Jokinen J, McDonald JW, Smith PWF. 2003. Joint regression and association modelling for longitudinal ordinal data. Biometrics, 59, 795-803.

III Jokinen J, McDonald JW, Smith PWF. 2006. Meaningful regression and association models for clustered ordinal data. Sociological Methodology, 36, 173-199.

IV Jokinen J. 2006. Fast estimation algorithm for likelihood-based analysis of repeated categorical responses. Computational Statistics and Data Analysis, 51, 1509-1522.

Contents

1. Introduction

In almost all fields of science where statistical models are applied, we encounter situations where the response variable of interest is clustered in some way. By clustering we mean that, within the levels of some observed factor, the responses are unlikely to be independent of each other. The most typical example of this type of clustering is a set of repeated measurements in time for the same unit, often known as a longitudinal response. Other examples of clustering include a series of responses to similar questions on the same individual, and family studies, where the same respose has been obtained from several family members. These examples demonstrate that the clustering is often included in the study by design. However, for the statistical analysis of these types of responses, the assumption of independece is often erroneously applied. Part of this may be due to lack of proper tools for the analysis, especially when the response is measured on a categorical scale.

The typical research hypothesis addresses the question whether the responses differ in observed subgroups of the population under study. In order to answer this, the univariate means need to be regressed on explanatory variables. Standard regression analysis assumes that the responses are independent of each other. In order not to abuse this assumption of independence, the whole set of responses within a cluster can be regarded as the response variable, which points us towards multivariate probability distributions. There exists a well-established theory on the regression analysis of repeated, or otherwise clustered, responses that are normally distributed [1, 2]. This is because the multivariate normal distribution has some convenient properties, namely that the joint distribution is fully specified by the first two moments, that is, by (i) the univariate means, and (ii) the variance-covariance matrix. However, the joint distribution of a multivariate categorical response is more complex. Take, for example, the response of interest in Paper I: a binary variable that measures the presence or abscence of bacterial carriage at 9 repeated time-points. Unlike for a multivariate normal response, the 9 first and $\binom{9}{2} = 36$ second order moments do not fully specify the joint distribution: from $\binom{9}{3} = 84$ moments of order three, up to $\binom{9}{9} = 1$ moment of order nine are also required. While 9 moments of order one are regressed on explanatory variables, altogether $2^9 - 9 - 1 = 502$ higher order moments, describing the associations between the responses, also need to be specified. It is apparent that an unstructured analysis of the associations, where the 502 parameters are estimated from the data, is unfeasible. This example illustrates how the number of association measures increase geometrically with increasing number of repetitions, which poses notable computational challenges for estimation.

In this thesis, we focus on a specific likelihood-based method for the analysis of repeated, or otherwise clustered, categorical responses. This work applies and extends the work by Ekholm and co-authors [3, 4], who specified the joint distribution of a multivariate binary response by utilizing the moment parameterisation, and proposed a set of association models that reduce the number of parameters needed to describe the associations. Rather than just achieving parsimony for the association parameters, the proposed association models aim to describe the mechanisms that generate the dependence between the responses. In other words, equal emphasis is put on regression modelling of the marginal means, and on modelling of the associations between repeated responses.

In what follows, we give an overview of the method, which serves as an introduction to four original research papers, referred to as Paper I - Paper IV. In Paper I, the method is applied, using a novel association model, to a longitudinal dataset reporting pneumococcal carriage of children in the FinOM Cohort Study [5]. Paper II extends the method to handle multivariate ordinal responses, and also to allow modelling of the dropout mechanism within the proposed framework. In Paper III, the applicability and interpretability of the method is discussed in relation to other approaches to the analysis of clustered categorical responses. Paper IV concentrates on the computational aspects of the method, and extends further the modelling of the associations by including explanatory variables. In addition to Papers I - IV, an essential part of this work is also the package called `drm` for statistical software R [6], that can be applied to fit models to clustered categorical datasets within the proposed framework. This package is freely available from the author's website at: `http://www.helsinki.fi/~jtjokine/drm`.

2. Approaches to the analysis of repeated categorical responses

Proposed methods for the analysis of repeated categorical responses are often categorized in three exclusive classes [2, 7, 8]:

- marginal or population-averaged models
- random-effects or cluster-specific models
- transition or conditional models

There is some variability in the literature how the differences of these approaches are perceived, but the main argument in distinguishing these approaches is in the interpretations of the regression model parameters. For a more detailed discussion, see Section 6 in Paper III. As was outlined in the Introduction, here we concentrate on methods that provide population-averaged regression coefficients for the explanatory variables. To avoid confusion with the definitions in the literature, these models will be subsequently referred to as population-averaged models. Note that the standard analysis of univariate responses is population-averaged, which is a particularly useful approach in clinical trial and epidemiological settings, where the effects of treatment or certain risk factors are of interest.

2.1. **Likelihood-based population-averaged models.** For likelihood-based inference, the joint probability of repeated categorical responses need to be specified. The common feature of the methods with population-averaged interpretation is that the first order moments, that is, the univariate means, are regressed on explanatory variables. The methods differ in the way the second and higher order moments are parameterised. As an introduction to different association measures, consider the simplest multivariate case; the bivariate binary response $\boldsymbol{Y} = (Y_1, Y_2)$. Denote the first two moments by

$$\begin{array}{lcll} E(Y_j) & = & \mathrm{pr}(Y_j = 1) & = \mu_j, \quad j = 1, 2, \\ E(Y_1 Y_2) & = & \mathrm{pr}(Y_1 = 1, Y_2 = 1) & = \mu_{12}. \end{array}$$

There are four possible realisations of the response profile: $(1,1), (1,0), (0,1), (0,0)$. Table 1 portrays these possible realisations as cells of a 2×2 -table. It is apparent

TABLE 1. 2×2 -table for a bivariate binary case

	$Y_2 = 1$	$Y_2 = 0$	*sum*
$Y_1 = 1$	μ_{12}		μ_1
$Y_1 = 0$			
sum	μ_2		1

that all four cell probabilities can be expressed as a function of the first and second order moments. Therefore, the joint probability of a bivariate binary response is fully specified with $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_{12})$. The second order moment contains the information regarding the association between Y_1 and Y_2. However, in itself, it is not a useful measure of the association, since it holds no comparison to the case of independence. Therefore, different parameterisations for the associations have been proposed.

2.1.1. *Correlation coefficient.* Bahadur [9] proposed the correlation coefficient for the associations. For bivariate binary responses, it is defined as

$$\rho_{12} = \frac{\mu_{12} - \mu_1 \mu_2}{\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)}}. \tag{2.1}$$

In order to express the four cell probabilities using $(\mu_1, \mu_2, \rho_{12})$, the following transformation is required: $\mu_{12} = \rho_{12}\sqrt{\mu_1(1-\mu_1)\mu_2(1-\mu_2)} + \mu_1\mu_2$. Correlation coefficient is an attractive parameterisation since it is familiar from the analysis of normally distributed responses. However, this familiarity may be misleading when analysing repeated categorical responses: the customary range of ρ, that is, $[-1, 1]$, applies only if $\mu_1 = \mu_2$. For example, if $\mu_1 = 0.1$ and $\mu_2 = 0.5$, the range of correlation coefficient is from -1/3 to 1/3.

2.1.2. *Odds ratio.* Several authors have proposed different variations of odds ratios for the associations [10, 11, 12, 13, 14]. These variations imply different parameterisations for higher than second order moments and/or multicategorical responses. However, all odds ratio parameterisations are analogous in the bivariate binary case, which we are considering here. The odds, or cross-product, ratio is defined as

$$\chi_{12} = \frac{\mu_{12}(1-\mu_1-\mu_2+\mu_{12})}{(\mu_2-\mu_{12})(\mu_1-\mu_{12})}. \tag{2.2}$$

The range of odds ratio is from zero to infinity, with 1 corresponding to independence, and values greater or less than 1 implying positive or negative associations respectively. In order to express the four cell probabilities using $(\mu_1, \mu_2, \chi_{12})$, a quadratic equation needs to be solved. It follows that, for $\chi_{12} \neq 1$,

$$\mu_{12} = \frac{\chi_{12}(\mu_1+\mu_2) + (1-\mu_1-\mu_2) - \sqrt{\{\chi_{12}(\mu_1+\mu_2) + (1-\mu_1-\mu_2)\}^2 - 4(\chi_{12}-1)\mu_1\mu_2\chi_{12}}}{2(\chi_{12}-1)},$$

while, for $\chi_{12} = 1$, $\mu_{12} = \mu_1\mu_2$. No explicit solution is available for higher than second order moments, irrespective of the variants of the odds ratio parameterisations. Therefore, iterative procedures are generally required for specifying the joint distribution.

2.1.3. *Dependence ratio.* Ekholm and co-authors [3] described the joint probability in terms of moment parameters, and proposed the dependence ratio for the associations. For the bivariate binary response, the dependence ratio, τ_{12}, is defined as

$$\tau_{12} = \frac{\mu_{12}}{\mu_1\mu_2}, \tag{2.3}$$

that is, the joint success probability divided by the joint success probability assuming independence. Similarly to odds ratio, 1 corresponds to independence, whereas values greater or less than 1 imply positive or negative associations respectively. In contrast to the odds ratio, the dependence ratio is constrained by the marginal means. For example, if $\mu_1 = 0.1$ and $\mu_2 = 0.5$, the maximum dependence ratio is $\min(1/\mu_1, 1/\mu_2) = 2$. This property in turn is more akin to the relative risk, or risk ratio: if the baseline risk is 0.5, the maximum risk ratio is similary 1/0.5=2. In order to express the four cell probabilities using $(\mu_1, \mu_2, \tau_{12})$, the apparent transformation is $\mu_{12} = \mu_1\mu_2\tau_{12}$. This transformation generalizes to moments of all orders and therefore a closed-form solution is available for specifying the joint distribution.

In this work, the focus is on the dependence ratio parameterisation approach. The properties of the dependence ratio as a measure of the association are discussed in each of the Papers I-IV, and compared with the odds ratio in [15]. The dependence ratio approach provides a unified framework for joint regression and association modelling, and in Section 3 we elaborate on how the dependence ratios are utilized for modelling the associations. If, however, the associations are considered to be of only secondary interest, approaches that focus primarily on the marginal regression may be applied. These types of population-averaged models are subsequently referred to as marginal models [16, 17]. The most well-known method for marginal

modelling of repeated categorical responses is the Generalized Estimation Equations (GEE) approach, which we briefly consider next.

2.2. **Generalized estimation equations, GEE.** Diggle et al. [2, p.146] argued that it is difficult to specify simple and sensible models for third and higher order moments, regardless of which association parameterisation is adopted. Lesaffre et al. [18] also pointed out that likelihood-based estimation of population-averaged models using odds ratios is generally unfeasible for cluster sizes greater than five. These difficulties have lead to the development of quasi-likelihood methods, namely the Generalized Estimation Equations (GEE) approach, first proposed by Liang and Zeger [19].

The idea in the GEE is to focus primarily on marginal means and treat the associations as of secondary interest. In order to gain efficiency in estimating the regression parameters $\boldsymbol{\beta}$, the standard score equations under the assumption of independence are generalized by introducing a working correlation matrix for the second order associations. Liang and Zeger [19] show that, by solving the resulting quasi-score equations, the $\boldsymbol{\beta}$ are consistent and asymptotically normal if the model for the univariate means is correctly specified. They also propose the so-called sandwich estimator for the asymptotic variance of $\boldsymbol{\beta}$, and show that this estimate is consistent, even if the working correlations are wrongly specified. For this reason, the result from the sandwich estimator is often known as the robust variance estimate. Prentice [20] proposed a further extension, commonly referred to as GEE2, where the first two moments are estimated jointly. This method leads to more efficient $\boldsymbol{\beta}$, but the trade-off is that both the first and second order moments need to be correctly specified in order to retain consistency. Other variations of the GEE differ either in the parameterisation of the second order moments, or in their computational solutions [21, 22, 23, 24].

The GEE approach has arguably become the most popular approach to the analysis of repeated categorical responses during the last decade. One of the attractions of the GEE is that it corresponds to the more familiar analysis of multivariate normal response, where the two first moments specify the joint distribution. Even the proposed working correlation matrix structures, typically included in statistical softwares (such as independence, AR1, exchangeable, and unstructured), are analogous to the standard analyses of the correlations for multinormal responses. However, like the example in the Introduction demonstrated, it is important to bare in mind that the correspondence with the multivariate normal response is only for convenience, not on substantive grounds.

3. Joint regression and association modelling

In GEE estimation, higher than second order moments are left unspecified, and therefore likelihood-based inference is generally unavailable. If likelihood-based marginal modelling is preferred, the second and higher order association measures need to be treated as nuisance parameters in the analysis. In this case, it is desirable that

the regression and association parameters are orthogonal to each other [25]. The so-called mixed parameterisation approach [10, 16, 26] satisfies this requirement, where the univariate means are complemented with conditional log odds ratios, the canonical parameters of the log-linear model. However, Lindsey and Lambert [17] argue that marginal modelling is not a reasonable approach for the analysis of repeated responses. They note that models should be derived from multivariate or conditional models representing plausible, if simplistic, physiological mechanisms. In order to achieve such a comprehensive analysis of repeated responses, further efforts need to be put on modelling the associations. Since in Papers I-IV, the proposed dependence ratio modelling approach is advocated as a method that allows meaningful modelling of the associations, it is appropriate to give a brief overview what we mean by a 'meaningful model' in this context.

3.1. **Meaningful models.** Of course, it can be argued that all models are meaningful, since they are based on (hopefully) relevant choices of probability distributions for the responses. Note, however, that this requirement already leaves the GEE approach out of scope. The role and proper specification of statistical models is a broad issue that has been addressed by eminent statisticians [27, 28], so we do not plan to bring any new insights to this complex theme. The following brief discussion only reiterates some of the points relevant in our context.

The considerations regarding the role of models in statistical analysis typically distinguish two different types of models: (i) empirical, and (ii) substantive. Empirical models are the most common type of statistical models: rather than building the model on any specific subject-matter considerations, these models aim to represent the form and assess the strength of the way the responses depend on the explanatory variables at hand. Standard regression and ANOVA models are of this type, where the primary goal is to estimate unknown parameters of interest by e.g. confidence intervals. In contrast, substantive models aim to describe the mechanisms that have generated the observed data, and the parameters of the model typically describe quantities that are not directly observed. Thus the formulation of the model requires some theoretical notions about the underlying mechanism. Substantive models are typically based on subject-matter considerations independently of the data at hand. However, another form of a substantive model is where the lack of fit calls for retrospective formulation of substantive issues. In Sections 6.1 and 6.3, we present two examples of this type of models.

Admittedly, a 'meaningful model' is a problematic phrase, since models may be meaningful for some but not for others. For example, it may be that certain models are meaningful only for statisticians. According to Cox [28], a meaningful model needs to fullfill the following two criteria: (i) it has a substantive interpretation and (ii) it can be used fairly directly to simulate data: "the essential idea is that if the investigator cannot use the model directly to simulate artificial data, how can 'Nature' have used anything like that method to generate real data?"

In the following section, we explore different ways of modelling the associations in the light of these notions on model specification.

3.2. **Association mechanisms.** Consider the bacterial carriage response described in the Introduction: although the number of repeated measures is no more than 9, altogether $2^9 - 9 - 1 = 502$ association measures are required to specify the joint distribution. This example illustrates how the number of measures needed to describe the associations increase geometrically with increasing cluster size. Therefore, already for cluster sizes greater than three, it is necessary to impose some type of structure on the association measures to reduce the number of parameters to be estimated. Two different ways of achieving parsimonious modelling of the associations can be distinguished:

(i) Imposing constraints on the measures of the association.
(ii) Deriving the measures from underlying association mechanisms.

Arguably, the first approach is an empirical way of modelling, whereas the latter one is substantive. The methods based on marginal, local or global odds ratio parameterisations [11, 13, 14] generally use the first approach, where symmetry constraints, or other type of equalities of odds ratios, are used for reducing the number of parameters to be estimated. For the dependence ratio approach, the general way of proceeding is that the underlying mechanism generating the associations is first postulated, and the dependence ratios are derived from it. To illustrate this, we next consider a latent class association mechanism [4] as an example.

3.2.1. *Example 1: Latent binary factor.* Consider a situation where we have observed q repeated binary responses on a subject. Suppose that the subject belongs to one of the two latent groups, $L = 0$ or $L = 1$. By utilizing the rule of total probability, the marginal univariate probability can be expressed as a weighted sum of two conditional probabilitites: for $j = 1, \ldots, q$,

$$\mu_j = \text{pr}(L = 1)\text{pr}(Y_j = 1|L = 1) + \text{pr}(L = 0)\text{pr}(Y_j = 1|L = 0). \tag{3.1}$$

Following the same logic, for $j, k = 1, \ldots, q$, $j \neq k$, the second order moment can be expressed as

$$\mu_{jk} = \text{pr}(L = 1)\text{pr}(Y_j = 1, Y_k = 1|L = 1) + \text{pr}(L = 0)\text{pr}(Y_j = 1, Y_k = 1|L = 0), \tag{3.2}$$

and similarly for moments of order $3, \ldots, q$.

For substantive and also for computational purposes, it is useful to have a understandable interpretation of the parameters. Therefore, denote by

$$\begin{array}{lcl} \text{pr}(L = 1) & = & \nu \\ \text{pr}(Y_j = 1|L = 1) & = & \psi_j \\ \text{pr}(Y_j = 1|L = 0)/\text{pr}(Y_j = 1|L = 1) & = & \kappa \end{array}$$

Thus (3.1) can now be expressed as $\mu_j = \nu\psi_j + (1 - \nu)\psi_j\kappa$. Next postulate that this unobserved factor L is the source of the observed association. In other words, the dependence between the responses within a cluster are independent given L. It follows that, for $l = 0, 1$,

$$\text{pr}(Y_j = 1, Y_k = 1|L = l) = \text{pr}(Y_j = 1|L = l)\text{pr}(Y_k = 1|L = l). \tag{3.3}$$

Furthermore, (3.2) can now be expressed as $\mu_{jk} = \nu\psi_j\psi_k + (1-\nu)\psi_j\kappa\psi_k\kappa$. The second order dependence ratio, assuming the latent binary factor mechanism, is thus:

$$\tau_{jk} = \frac{\mu_{jk}}{\mu_j\mu_k} = \frac{\nu\psi_j\psi_k + (1-\nu)\psi_j\psi_k\kappa^2}{\{\nu\psi_j + (1-\nu)\psi_j\kappa\}\{\nu\psi_k + (1-\nu)\psi_k\kappa\}} = \frac{\nu + (1-\nu)\kappa^2}{\{\nu + (1-\nu)\kappa\}^2}. \tag{3.4}$$

Higher than second order dependence ratios are straightforward generalisations of (3.4), derived by using the multivariate form of the local independence assumption (3.3).

The latent binary association mechanism thus has two parameters, where ν is the proportion in the population with $L = 1$, and κ is the ratio of probabilities in groups $L = 0$ and $L = 1$ respectively. An extension to more than two latent classes can be found in Paper III. However, if a latent continuous mechanism seems more plausible for the phenomenon under study, one can fit a model with a latent continuous, Beta-distributed variable [4]. Furthermore, in Papers II-III, similar mechanisms for ordinal and nominal responses are introduced. Note that, rather than imposing constraints directly on the association measures, the dependence ratio parameterisation is mainly used here as a convenient tool for specifying the latent association mechanisms.

These latent variable association models are exchangeable, that is, independent of the ordering of the responses. However, if there is a natural ordering of the responses, like in longitudinal studies, temporal structures are generally more natural choices for the associations. Next we explore how these temporal association models can be expressed via the dependence ratio parameterisation.

3.2.2. *Example 2: First order Markov assumption.* Consider again q binary responses observed in time for the same subject. The joint probability of the response profile $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ can be decomposed as

$$\text{pr}(Y_1 = y_1, \ldots, Y_q = y_q) = \text{pr}(Y_1 = y_1)\text{pr}(Y_2 = y_2|Y_1) \cdots \text{pr}(Y_q = y_q|Y_{q-1}, \ldots, Y_1).$$

Further assume that the future response is conditionally independent of the previous responses, given the current response, that is, the first order Markov property. The joint probability can now be factorised into $q-1$ adjacent univariate and $q-1$ overlapping bivariate probabilities; see Equations (6) and (7) in Paper I. The computational advantages of the Markov assumption are considerable: in addition to univariate means, only $q-1$ adjacent dependence ratios $(\tau_{12}, \tau_{23}, \ldots, \tau_{(q-1)q})$ are required to specify the joint distribution.

Note also that in longitudinal studies, the probabilities of moving from one state to another are often of direct interest. The dependence ratio parameterisation provides the following connection to these transition probabilities:

$$\tau_{12} = \frac{\text{pr}(Y_2 = 1|Y_1 = 1)}{\text{pr}(Y_2 = 1)}.$$

It is arguable whether the Markov association model can be regarded as a substantive model. Although interpretable, there are also aspects that correspond more

closely to empirical models. For example, when making the Markov assumption, the possible constraints are imposed directly on the observed second order association measures. However, there is an example in Paper I where the estimated second order dependence ratios are further utilized to estimate the unobserved duration of carriage. In addition, it is common in longitudinal studies that the temporal association does not account for all the dependence between responses, and latent variable models are also required in order to describe adequately the underlying association mechanism. Applications to combined temporal and exchangeable association modelling can be found in each of the Papers I-IV.

4. Computational aspects

Thus far, we have considered the dependence ratio approach from two aspects: first, how to specify the joint probability in terms of univariate means and dependence ratios of all orders, and secondly, how to impose a structure on the dependence ratios in order to reduce the number of parameters to be estimated. In order to make the method useful in applied work, feasible computational solutions are imperative, as the popularity of the GEE approach has shown. A novel computational solution for finding the maximum likelihood (ML) estimates was developed for fitting the models in Papers I-IV. This estimation algorithm is described in detail in Paper IV. Next we present some of the key properties of this estimation technique.

4.1. **Score equations.** Consider a categorical response profile, with f alternatives and q repetitions. There are altogether $f^q = d$ possible realisations of the profile; for example, for the carriage response profile in the Introduction, the number of realisations is $2^9 = 512$. A natural choice is to assume that the response profiles follow a multinomial distribution, with d probabilities summing to one, $\sum_{k=1}^{d} \pi_k = 1$. The log likelihood for multinomially distributed response profiles is

$$l(\boldsymbol{\pi}) = \sum_{k=1}^{d} n_k \log(\pi_k), \tag{4.1}$$

where n_k is the observed count of the profile k, and $n = \sum n_k$. If $l(\boldsymbol{\pi})$ is smooth enough, its maximum satisfies $\partial l(\boldsymbol{\pi}) = 0$, commonly known as the score equation. To demonstrate the use of score equations, consider the case of multinomial probabilities: following from the unit-sum constraint, one of the probabilities is a linear combination of the others, for example, $\pi_d = 1 - (\pi_1 + \cdots + \pi_{d-1})$. Thus $\partial \pi_d / \partial \pi_k = -1$, $k = 1, \ldots, d-1$, and

$$\frac{\partial \log(\pi_d)}{\partial \pi_k} = \frac{1}{\pi_d} \frac{\partial \pi_d}{\partial \pi_k} = -\frac{1}{\pi_d}.$$

To eliminate redundancies, we treat the likelihood as a function of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{d-1})$, and differentiating $l(\boldsymbol{\pi})$ with respect to π_k gives the score equation

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_k} = \frac{n_k}{\pi_k} - \frac{n_d}{\pi_d} = 0.$$

The solution to the score equation satisfies $\hat{\pi}_k/\hat{\pi}_d = n_k/n_d$, and it follows that $\sum \hat{\pi}_k = 1 = \hat{\pi}_d \sum n_k/n_d = \hat{\pi}_d n/n_d$, and thus, $\hat{\pi}_k = n_k/n$, for $k = 1, \ldots, d$.

Note that closed-form solutions for the maximum likelihood estimates of $\boldsymbol{\pi}$ are available. However, in our case we aspire to model the probabilities as a function of regression and association model parameters, denoted by $\boldsymbol{\theta}$. In that case the score equations are nonlinear functions of the model parameters, and need to be solved iteratively. One of the most commonly used iterative methods for solving nonlinear equations is the Newton method [29, 30, 31]: consider a $p \times 1$ vector $\boldsymbol{\theta}^r$ at iteration round $r = 1, 2, \ldots$. The first order Taylor series expansion of $\partial l(\boldsymbol{\theta})$ at point $\boldsymbol{\theta}^r$ is

$$\partial l(\boldsymbol{\theta}) \approx \partial l(\boldsymbol{\theta}^r) + \partial^2 l(\boldsymbol{\theta}^r)(\boldsymbol{\theta} - \boldsymbol{\theta}^r), \tag{4.2}$$

where $\partial l(\boldsymbol{\theta}^r)$ is a $p \times 1$ gradient vector and $\partial^2 l(\boldsymbol{\theta}^r)$ is a $p \times p$ matrix of second derivatives, called the Hessian. By equating the right hand side of (4.2) with zero, the solution to the first order approximation of $\partial l(\boldsymbol{\theta})$ satisfies $\boldsymbol{\theta}^{r+1} = \boldsymbol{\theta}^r - \{\partial^2 l(\boldsymbol{\theta}^r)\}^{-1} \partial l(\boldsymbol{\theta}^r)$. This, the so-called Newton step, is repeated until convergence.

The proposed estimation techniques for the odds ratio parameterisations, reviewed in Section 2.1.2, generally utilize the Newton method for finding ML estimates. However, for cluster sizes greater than two, the estimation procedure is more cumbersome. For $r = 1, 2, \ldots$, consider the following mappings:

$$\boldsymbol{\theta}^r \rightarrow \boldsymbol{\pi}^r \rightarrow l(\boldsymbol{\pi}^r) \rightarrow \boldsymbol{\theta}^{r+1} \rightarrow \ldots$$

As we noted in Section 2.1.2, no closed-form solution is generally available for expressing the profile probabilities in terms of univariate means and the odds ratios. Thus in order to obtain ML estimates of the model parameters, the iterative procedure is required to have two steps within each round for updating the parameter values. However, note from Section 2.1.3 that when expressing the profile probabilities in terms of univariate means and dependence ratios, general explicit expression exists for mapping $\boldsymbol{\theta}^r \rightarrow \boldsymbol{\pi}^r$. Therefore, for the dependence ratio parameterisation, only one step is required for updating the parameter values.

In addition to reducing the number of steps in the iteration process, the closed-form solution also has other important implications for ML estimation, which we consider next.

4.2. **Nonderivative methods.** In previous applications of the dependence ratio approach [3, 4], the solution for fitting the models was to present the dataset in the form of contingency tables for each combination of the explanatory variables, and to model the cell frequencies n_k as independent Poisson variates. The estimation was performed by using the GLIM macro NLIN4 [32], which is a modification to the iteratively reweighted least squares (IRLS) algorithm used, e.g., for generalized linear models. Since the logarithms of the cell frequencies are nonlinear functions of $\boldsymbol{\theta}$, the design matrix, needed for IRLS, does not stay constant from one iteration to the next. NLIN4 macro recalculates the design matrix at each iteration by using numerical derivatives. In addition to fitting the models in [3, 4], this macro can be used more generally for fitting any nonlinear exponential family models. However,

this estimation technique has its limitations: if continuous or time-dependent covariates are present, the number of different covariate patterns is usually equal to the number of response profiles. In that case presenting the dataset as a contingency table for each combination of the explanatory variables rapidly becomes prohibitive. For example, for the carriage response profiles of 329 children in Paper I, the size of the contingency table is $329 \times 2^9 = 168\ 448$, with 329 cell counts of one, and 168 119 zero cells.

For the Papers I-IV, the computational solution was to use a numerical optimiser for the multinomial likelihood. We refer to this method as nonderivative maximisation of the likelihood since the derivatives for the Newton step are approximated numerically from the likelihood function. In Paper I, the likelihood function specific to the dataset was written using S-language, and the parameters were estimated using the numerical function minimisation routine in S-PLUS. For Papers II-IV, R software [6] and the package `drm` were applied. The `drm`-package includes a generic likelihood function that utilizes the explicit solution for the profile probabilities. Maximisation of the likelihood is performed using function `nlm` in R software, which is a Newton-type numerical iteration algorithm [29, 33]. See Section 4.2.2 for further details and Appendix for an example.

4.2.1. *Inherent unit-sum constraint.* The function `nlm` is an unconstrained minimisation routine. Therefore, if the likelihood function has constraints, these need to be built in to the function. The apparent constraint for the multinomial likelihood is the unit-sum requirement. Note that although only the observed response profile probabilities are required for (4.1), all $f^q = d$ probabilities generally need to be calculated for each unit in order to express the redundant probability as a linear combination of the others: $\pi_d = 1-(\pi_1+\cdots+\pi_{d-1})$. In other words, the same curse of dimensionality than for the contingency table representation generally applies for constrained numerical optimisation of the multinomial likelihood.

However, note from Section 2.1.3 that the dependence ratio parameterisation is based on the moments of the distribution. Recall that the moment-generating function generates the moments of the probability distribution, and thus uniquely defines the distribution. All profile probabilities, including the redundant one, can therefore be expressed as a function of the moments. This implies that the unit-sum constraint is inherent in the moment parameterisation. The advantages of this inherent constraint are substantial, since we only need to calculate the observed profile probabilities as a function of the moments at the estimation stage: in the bacterial carriage dataset, instead of calculating $329 \times 2^9 = 168\ 448$ probabilities at each iteration, only 329 probabilities are required. The difference is even more notable in the government spending example in Paper IV: instead of $607 \times 3^9 = 11\ 947\ 581$ probabilities, only 607 are needed at each iteration. The probabilities of all the possible profiles need to be calculated only once, after the convergence of the estimates, in order to ensure that all profile probabilities are between 0 and 1. If this turns out not to be the case, the model is considered to be wrongly specified, and an alternative model formulation is required.

4.2.2. *Optimisation algorithms.* Most of the current statistical softwares include generic numerical optimisation routines. Therefore, if a closed-form solution exists for the likelihood, nonlinear models can often be estimated as easily as linear ones [34]. However, when choosing the appropriate optimisation method for a particular problem, it is valuable to have a coarse understanding how the major algorithms work. We finish our review of the computational aspects by summarising the reasons for choosing `nlm` as the optimiser for the software package `drm`.

The Newton method, discussed in Section 4.1, is a much used algorithm for solving nonlinear equations. However, the method has three drawbacks. Firstly, in order to approximate the score function, the Hessian need to be evaluated, which can be computationally demanding [30]. Secondly, if the starting values are at a far distance from the maximum, the first order approximation to the score function may not be adequate [31]. Thirdly, since the Hessian need not be positive definite at each iteration, the method does not ensure that the likelihood increases at each step; it may even converge to a local minimum [30, 31]. In order to overcome these drawbacks, modified Newton methods have been developed. These methods use finite-difference approximations to the gradient and/or the Hessian, perform backtracking routines for step lengths, and modify the Hessian for positive-definiteness. Two of this type of numerical optimisation routines in the base package of the R software [6] are `nlm`, and `optim` with an option `''BFGS''`. Our experience from fitting the models presented in this thesis is that `nlm` converges faster, sometimes in less than half the time that of `optim`, and tends to be more reliable in terms of finding the maximum.

Apparently, in addition to `drm`, the only other available softwares for likelihood-based population-averaged modelling are MAREG [35], based on the mixed parameterisation approach [10, 26] and a set of functions for R software [36], based on the local odds ratio representation [13], both of which are limited to cases where the responses profiles are short and essentially complete. Currently, however, `drm` is apparently the only general software package for repeated categorical responses that also includes the possibility of modelling the dropout mechanism, which we consider next.

5. Dropout in longitudinal studies

An ubiquitous problem in longitudinal studies is that some of the study subjects drop out before the end of the study. This adds another level of complexity to the analysis of repeated responses. Missing data problems have been a subject of considerable research during the last two decades. For example Little and Rubin [37] provide an extensive summary of the recent methodological developements for handling missing data.

Special attention is required when analysing longitudinal datasets with dropouts. Although the actual mechanism causing subjects to drop out is rarely the focus of the study, its possible influence on the regression and association parameters need to be investigated. Little and Rubin [37] distinguish three different missing data

mechanisms that each have different implications on inference regarding the model parameters: (i) missing completely at random (MCAR), (ii) missing at random (MAR), and (iii) missing not at random (MNAR). The following sections summarise heuristically some of the main features of these mechanims by using the dataset analysed in Paper II as an example.

5.1. **Missing completely at random, MCAR.** Consider the dataset, analysed in Paper II, where side-effects of a drug called Fluvoxamine were recorded at four different visits for 299 patients. For purposes of illustrating the missing data mechanisms, it is sufficient to consider a simplified case where the response variable of interest is dichotomised, that is, presence or abscence of side-effects, and where the regression analysis task is to compare the probability of side-effects between males and females. If the reason for dropping out from the study is independent of all the observed and unobserved information, that is, of observed and missing values of side-effects and of sex, the data are said to be missing completely at random, MCAR. An example of this type of missing data mechanism is a premature termination of the study, so that all subsequent follow-up visits are censored. Note that methods such as the GEE, for which the likelihood is not available, rely on the assumption that the observed data are a random sample of the population. Therefore inference is valid only if the data are MCAR.

5.2. **Missing at random, MAR.** Suppose that those who are inherently more prone to side-effects of the drug also tend to drop out easier than their counterparts. In this case the prevalence of side-effects is greater for those who drop out, and thus the reason for dropping out depends on the observed information. This type of mechanism is called missing at random, more specifically sequential MAR [38]. Another example of the MAR mechanism is when the dropout rate for males and females is different. In the presence of the MAR mechanism, the observed data no longer are a random sample of the population, and hence the GEE approach is invalid. However, likelihood-based inference allows the specification of the model for both the observed and unobserved parts of the data, and by summing over the possible realisations of the unobserved responses, the inference is valid even when the data are MAR [39]. Therefore, when analysing longitudinal data with missing values, a likelihood-based approach takes on added importance.

As an example, consider a patient with an observed response vector $\boldsymbol{y} = (0, 0, 1, \star)$, where $\star$ denotes a missing response. By marginalising over the possible realisations of the missing values, the contribution to the likelihood (4.1) can be expressed as

$$\log\{\pi(0,0,1,\star)\} = \log\{\pi(0,0,1,0) + \pi(0,0,1,1)\} = \log\{\pi(0,0,1)\}. \tag{5.1}$$

In words, the contribution carries no information regarding the fourth response. As long as the reason for dropping out depends only on the three observed values, or possibly on the corresponding covariate sex, likelihood-based inference is still valid. Note also that, in addition to dropouts, this equally applies to situations with intermitting missing values.

5.3. **Missing not at random, MNAR.** Finally, consider a situation where the current state of side-effects causes the patient not to attend the visit in question. The dropout is then dependent on the missing information and the responses are said to be missing not at random, MNAR. This dropout mechanism is called nonignorable, and it cannot be directly investigated from the data since that information is missing. If one adopts an analytical approach under MNAR, joint analysis of regression, association and dropout mechanism is required. Two of the most common analytical approaches are pattern mixture models and selection models [37]. We briefly review the latter approach, as specified by Diggle and Kenward [40].

In the context of our example, define the dropout indicator $D = 2, 3, 4, 5$ for the occasion when dropout occurs. For example, for our example case in (5.1), $d = 4$, and for completers, $d = 5$. Further define the conditional hazard for dropping out at time-point $t = 2, 3, 4$ as $\mathrm{pr}(D = t|D \geq t) = \phi(\delta)$. This dropout, or selection, model can be specified on top of the model for the response profiles. For example, the weighted profile probability for the patient in (5.1) can now be expressed as

$$\pi'(0, 0, 1, \star) = \{1 - \phi(\delta)\}\{1 - \phi(\delta)\}\{\phi(\delta)\pi(0, 0, 1, 0) + \phi(\delta)\pi(0, 0, 1, 1)\}. \tag{5.2}$$

Note that in (5.2) dropout is constant over time and over the two possible realisations of the profile. This is equivalent to assuming that the dropout mechanism is MCAR. The weighted probability factorises to $\pi'(0, 0, 1, \star) = f(\delta) \times \pi(0, 0, 1)$, and thus the inference regarding the regression and association parameters is unaffected by dropout. However, when we allow for dropout at time $t = 2, 3, 4$ to depend on the observed and unobserved data, a model for the dropout hazard [40] is specified as:

$$\mathrm{logit}\{\phi(\boldsymbol{\delta}; y_t, y_{t-1}, \mathrm{sex})\} = \delta_0 + \delta_1 y_t + \delta_2 y_{t-1} + \delta_3 \mathrm{sex}. \tag{5.3}$$

Consider our example patient at $t = 4$: if $\delta_1 = 0$, dropout does not depend on the missing response value. This is equivalent to assuming that the dropout mechanism is MAR, and the likelihood still factorises. However, if $\delta_1 \neq 0$, the probabilities of the two possible realisations of the profile in (5.2) will be assigned different weights, and the likelihood no longer factorises to two separate parts of dropout and the model for the profiles. In this case, it is advisable to investigate the influence of δ_1 on the results derived from the regression and association models.

Although dropouts frequently occur in longitudinal studies, the available tools for analysing categorical longitudinal datasets with dropouts are limited. This is mainly because likelihood-based analysis is computationally challenging, even without missing data problems. However, for the dependence ratio parameterisation, this extension is relatively straightforward: by specifying the dropout model on top of the model for the profile probabilities as in (5.2) and (5.3), it follows that an explicit formula is available for the joint analysis of regression, association and dropout mechanisms; see Paper II. This also implies that the computational solution described in Section 4 equally applies to extensions with dropouts. The option for modelling the dropout probability, as specified in (5.3), is implemented in the package `drm`.

6. Summaries of the original papers

We now briefly summarise the papers in this thesis, paying particular attention to novel applications and how the theory and practise introduced by Ekholm and co-authors [3, 4] was developed further.

6.1. **Paper I: Application to longitudinal responses on bacterial carriage.** The application in Paper I is from the Finnish Otitis Media (FinOM) studies, which were conducted by the Department of Vaccines at the National Public Health Institute, Finland, between the years 1994 and 1999. The principal goal was to evaluate two pneumococcal conjugate vaccines in prevention of middle ear infections on children under two years of age [41]. Initially, a pilot study [5] for a subsequent vaccine trial was conducted, where one of the aims was to investigate the symptomless carriage of pneumococcal bacteria, and its potential development to disease. The dataset analysed is from the pilot study, where 329 children from Tampere, Finland, were enrolled at two months of age, and their pneumococcal carriage status was determined from nasopharyngeal swabs at scheduled visits at 2, 3, 4, 5, 6, 9, 12, 15 and 18 months of age. In addition, an extensive set of potential time-constant and time-dependent risk factors for carriage were collected at recruitment and throughout the follow-up.

The analysis was a novel application of an existing theory [3, 4]. Thus far, datasets with such large number of repetitions and so many explanatory variables had not been analysed using the dependence ratio approach. We fitted a marginal regression model, with five explanatory variables, combined with a model for temporal association. However, the lack of fit led us to modify the association model by including a latent factor mechanism, assuming that there is a proportion of children in the population protected against pneumococcal carriage during the follow-up in question. This approach can be viewed as retrospective formulation of a substantive model, as described in Section 3.1. The fitted association model was further utilized to obtain preliminary, although tentative, estimates regarding the acquisition rate and the median duration of carriage. Some selected results from the fitted model are that the carriage becomes steadily more prevalent throughout the first 18 months of life. In addition, the probability increases notably if the child attends daycare, but if the child has siblings at home, the additional effect of daycare is negligible. Furthermore, one of the interpretations from the association model is that approximately 10% of children are protected against pneumococcal carriage during their first 18 months of life.

6.2. **Paper II: Extension to longitudinal ordinal responses with dropout.** The dataset analysed in Paper II is from a psychiatric study, where therapeutic and side-effects of a drug called Fluvoxamine were recorded from a group of 299 patients at 4 time-points [14]. A reasonable concern in this type of studies is that some patients may drop out from the study because of no notable treatment effect, or worsening side-effects. In order to investigate the influence of the dropout on the inference regarding the response of interest, joint modelling of the regression,

association and dropout mechanisms are required. Previous applications to the Fluvoxamine data [14, 42] had been constrained to three repeated time-points because of computational difficulties.

Both responses are measured on a four-level ordinal scale, so extension of the dependence ratio theory to multicategorical case was required. In addition, a set of novel association mechanisms for repeated multicategorical responses were introduced. In order to investigate the influence of dropout on the regression and association parameters, a selection model [40] was specified on top of the regression and association models, resulting in an explicit solution for all three parts of the model. That allowed us to fit models to the entire dataset, that is, to all four time-points.

Temporal association models were fitted for both the therapeutic effect and side-effect responses. For therapeutic effect, the fitted dependence ratios for consecutive visits indicated that there was a general individual-level tendency towards weakening therapeutic effect through time. For the side-effect response, the probability of having the same side-effect at consecutive visits was significantly higher than under independence, whereas moving from one state to another was essentially as probable as under independence. These results were confirmed by comparing the fitted counts with the 20 most frequently observed profiles: for therapeutic effect, the majority of observed profiles had a decreasing trend, whereas for the side-effect, there was a notable dominance of constant or almost constant profiles. These findings highlight the usefulness of modelling the whole response profile rather than direct marginal modelling of the population means. The validity of these conclusions were further assessed by examining the potential influence of dropout on the results. The estimates from our selection model indicated that dropout in this study was non-ignorable but, fortunately, the results from the regression and association models were not much influenced by it.

6.3. **Paper III: Comparison with the other approaches for ordinal responses.** The aim in Paper III was to present more heuristically the applicability and the interpretability of the dependence ratio approach for readers that apply statistical methods in their empirical research. Four different datasets with clustered ordinal responses were analysed. By utilizing the examples, particular attention was put on comparisons with the other proposed approaches for clustered ordinal responses, especially with the random-effects modelling approach. The most apparent difference between the approaches is the parameter interpretation: the parameters of the random-effects model are subject-specific rather than population-averaged. For random-effects models, the unobserved heterogeneity is imposed in the linear predictor, and if the link function is nonlinear, the subject-specific and population-averaged effects are no longer equivalent. In addition, different link functions for the linear predictor imply different structure for the associations. For the dependence ratio approach, the unobserved heterogeneity, that is, the latent structures, are imposed directly on the univariate probabilities, independently of the link function used for the regression.

One of the datasets analysed is from the US National Youth Survey [43], where 237 teenagers (117 boys and 120 girls), aged 13 at the beginning of the study, filled in a questionnaire yearly for five consecutive years, answering questions regarding their use of marijuana during that year. The response is ordinal, with values never (non-user), less than once a month (occasional user) and more than once a month (frequent user). Although the conclusions derived from the regression model parameters were very similar, the patterns of the observed profiles were notably different for boys and girls. The dependence between the consecutive responses for boys was adequately explained by a temporal association model. The probabilities of staying as an occasional or a frequent user were, respectively, over two and three times higher than under independence. However the probabilities for moving from occasional use to frequent use and vice versa decreased with increasing age, change being even less probable than under independence at ages from 16 to 17. This can be interpreted as the boys who use marijuana, gradually develop a habit through the teenage years, which they are eventually reluctant to change.

For girls, a novel association model was introduced, through retrospective formulation of substantive issues; see Section 3.1. The associations between the responses for girls was found to be well-described by a model where, in addition to temporal association, only 63% of the teenage girls were categorised as susceptible to marijuana use. However, the lack of fit led us to modify the association model by formulating another latent subgroup for girls that may try marijuana, or use it occasionally, but categorically refuse to become frequent users. The estimated proportion of this type of girls among the susceptibles was 58%.

6.4. **Paper IV: Computational solutions for the likelihood estimation.** As was described in Section 4, a novel computational solution and the accompanying software was developed for the applications in this thesis. Paper IV describes in detail the algorithm for the maximum likelihood estimation and also consists of an unified representation of the dependence ratio approach simultaneously for binary, nominal and ordinal responses, along with all the proposed association mechanisms published so far. Paper IV is thus the general reference article for the software package `drm`.

For illustrative purposes of the computational capabilities of the method, two applications were presented. The first dataset is from the Madras psychiatric study [2, p.234], where the presence or abscence of six different schizophrenia symptoms of 86 patients were recorded at 12 repeated time-points. The response of interest is thought disorders, and the focus is to find out whether the longitudinal thought disorder prevalence differs according to age-at-onset and sex. We fitted a regression model combined with a model for temporal association. No significant interaction between time and the covariates of interest were found. Since altogether six symptoms were investigated in the study, we also postulated that there are patients whose symptom characteristics do not include susceptibility to thought disorders. This type of structure proved to fit the data well. Furthermore, the proportions of the susceptibles were found to vary between patient subgroups: according to

our model, practically all males older than 20 years were susceptible to thought disorders, whereas for females less than 20 years, the susceptibility was only 68%.

The second dataset is from a 1989 US General Social Survey, where 607 adults, aged over 18 years, were asked their opinion regarding government spending on nine different targets. Each of the responses is ordinal, with levels a) too little, b) about right, c) too much. Here the focus was to examine whether the opinions of subjects differ according to their political party affiliation. This dataset had been previously analysed with four government spending targets, resulting in $3^4 = 81$ possible realisations of the response profile [13]. To illustrate the computational potential resulting from the inherent unit-sum constraint, we fitted a model to a total of nine repeated responses, with altogether 3^9=19 683 possible realisations of the response profile. We also included an extensive set of explanatory variables for the regression. According to our regression model, the opinions of the Independents generally lie in between the Democrats and the Rebublicans, who in turn differ the most in opinions concerning Health and Assistance to the Poor. A latent binary mechanism for the associations, see Section 3.2.1, fitted the data clearly better than a model assuming independence of the responses. We further found that the proportions in the two latent classes, along with their corresponding conditional probabilities, were significantly different between the Democrats and the Rebublicans. In conclusion, when answering questions about government spending, the Democrats and the Rebublicans differ both in their marginal means as well as with regard to their response profiles.

7. Discussion

In this thesis we concentrate on a specific likelihood-based method for extending the analysis of univariate categorical responses to a multivariate case. However, our aim is not only to focus on the analysis of the univariate means but also to build a plausible model for the whole response profile. This approach is in contrast to the marginal modelling approaches, most notably the GEE, where the associations are of secondary interest. The GEE approach is often described as a robust method, since estimates of the regression parameters and their corresponding sandwich variance estimates are consistent, even when the correlation structure is misspecified. However, consistency relies on large sample properties; a luxury that we do not often have. As Drum and McCullagh [44] have noted: 'To advertise as robust a method whose only demonstrated property is consistency is to invite the wrath of SASA, the Statistical Advertising Standards Authority'. Arguably, the GEE approach is therefore useful only for datasets such as obtained from survey studies, where the sample sizes are typically large [17, 44], and the response profiles are short and essentially complete [2, p.80]. Since none of the datasets analyzed in Papers I-IV fulfill these requirements, a likelihood-based modelling approach for both the regression and the associations is more appropriate.

When choosing a method for modelling the whole response profile, several considerations need to be taken into account. For example Joe [45] distinguishes four properties that are desirable for any multivariate statistical model:

(i) Interpetability
(ii) Upwards compatibility
(iii) Flexible and wide range of dependence
(iv) Closed-form representation of the joint probability distribution

(i) Joe [45] concludes that the parameters of an interpretable model preferably relate to temporal or latent variable representations. Easily understandable model parameters are not only useful when interpreting the fitted model, but can also be helpful in computations, for example when specifying the starting values for the iterations. It is arguable whether the modelling approaches, where the constraints are imposed directly on the association measures, can be viewed as interpretable models. For example, Joe [45] concludes that the global odds ratio approach [14, 42] is only partly interpretable. In Papers I-IV, we present several association models with latent variable and temporal mechanims, or mixtures of these, with straightforward parameter interpretations such as proportions of the population, or ratios of probabilities.

(ii) In order to preserve upwards compatibility, the model need to specified in such a way that the association measures of order $2, \ldots, q-1$ are independent of q. In other words, it is desirable that the inference regarding, say, τ_{124}, is unaffected by the fact whether the third response is observed or not. This is a particulary important property in longitudinal studies, where often the lengths of the observed response profiles vary because of dropout and intermediate missing values. In general, likelihood-based approaches fullfill this property, although the mixed parameterisation approach [10, 26] is a notable exception.

(iii) The range of dependence corresponds to the constraints of the association measure imposed by the probability space. It is clear that the odds ratio parameterisation has much wider range of dependence than the dependence ratio. However, having constraints can also be a useful feature: our recent empirical findings suggest that finite bounds of the dependence ratio also imply smaller standard error and better identifiability of the association parameter compared to the odds ratio with infinite upper bound; see Figure 6 in [46]. Furthermore, rather than focusing on the properties of the association measure, the range of dependence should be viewed from a larger perspective, including the ability to specify plausible association mechanisms. A set of association models, summarised in Paper IV, cover a wide range of dependence from several latent categorical, to latent continuous, and temporal dependence models. These models are also flexible in the sense that they can include covariates.

(iv) For any available method, a closed-form representation of the model's probability function is not only a matter of theoretical elegance. The dependence ratio

approach demonstrates the potentials of an explicit solution: the extensions of the method are more straightforward, such as inclusion of the selection model specification for modelling the dropout mechanism. In addition, implementation is easier since the tools for maximum likelihood estimation are readily available in most statistical softwares. Furthermore, the explicit solution is a prerequisite for utilizing the inherent unit-sum constraint of the dependence ratio parameterisation, that provides an important stride in computational speed for large cluster sizes. As a final note, with regards to the definition of a meaningful model by Cox [28], a model with an explicit expression for the profile probabilities can be straightforwardly used to simulate artificial data.

The apparent criticism regarding the proposed computational solution for the dependence ratio approach is that no positivity constraints on the probabilities of the unobserved profiles are imposed in the estimation stage. What this means is that for some models, the converged parameter estimates may produce negative probabilities for some of the unobserved profiles. It is clear that in this case, the fitted model should be discarded outright, since it does not fall into the probability space. However, note from (4.1) that if all the possible profiles are observed, that is, each cell of the contingency table is positive, the positivity constraints are included also in this estimation algorithm. What this generally means is that the estimated negative probabilities can typically occur only in datasets with large cluster sizes, generally unfeasible to fit with any other proposed computational solutions other than the GEE. It is intuitively clear that, when the complexity and the number of unobserved profiles increase, there is more uncertainty regarding the underlying structure of the profiles. Since in this case, the data provide very few hints about the phenomenon under study, our view is that the estimated negative probabilities then serve as an additional tool for model validation; as an indication of model misspecification.

Another potential criticism regarding the validity of the proposed modelling approach relates to the latent variable specification. Since the parameters of the latent variable models refer to quantities that are not observed, the data generally provide no direct way to test the validity of the models. For example, consider the result presented in Section 6.1, with the interpretation that 10% of the children are protected against pneumococcal carriage up until the age of 1.5 years. The data cannot provide any confirmation whether this phenomenon is actually true, or possibly an artefact arising from the prolonged sampling intervals after the age of six months. Similar reservations apply to other latent variable models presented in this thesis. Our conclusion is that models with latent variable representation should always be interpreted with particular caution. Note, however, that this aspect of model validation is not specific to the dependence ratio approach, but applies more generally to all modelling approaches with latent variables. Equally, or even more so, these concerns also apply to the selection model specifications, which are similarly untestable because of the very nature of missing information.

To conclude, we have applied and extended the theory for the joint regression and association modelling of repeated, or otherwise clustered, categorical responses. The dependence ratio parameterisation provides a computational advantage over the other proposed likelihood-based approaches and therefore also presents a viable alternative to the GEE approach, which is dominating the applied field because it is implemented in many statistical softwares. In terms of applicability and extensibility of our proposed approach, a notable contribution of this thesis is thus the freely available package `drm` that hopefully will provide more insights and experience concerning the usefulness of this method in applied work.

References

[1] Hand D, Crowder M. 1996. Practical longitudinal data analysis. Chapman & Hall, London.

[2] Diggle PJ, Heagerty P, Liang KY, Zeger SL. 2002. The analysis of longitudinal data, second ed. Oxford University Press, Oxford.

[3] Ekholm A, Smith PWF, McDonald, JW. 1995. Marginal regression analysis of a multivariate binary response. Biometrika 82, 847–854.

[4] Ekholm A, McDonald JW, Smith PWF. 2000. Association models for a multivariate binary response. Biometrics 56, 712–718.

[5] Kilpi T, Herva E, Kaijalainen T, Syrjänen R, Takala A. 2001. Bacteriology of acute otitis media in Finnish children during the first two years of life. Pediatric Infectious Disease Journal 7, 654–662.

[6] Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J. Comput. Graph. Statist. 5, 299-314.

[7] Agresti A, Natarajan R. 2001. Modeling clustered ordered categorical data: a survey. Int. Statist. Rev. 69, 345-71.

[8] Lindsey JK 1999. Models for repeated measurements, second ed. Oxford University Press, Oxford.

[9] Bahadur RR. 1961. A representation of the joint distribution of responses to n dichotomous items. In studies on item analysis and prediction (ed. H. Solomon), pp 158-168. Stanford Mathematical Studies in the Social Sciences VI, Standford University Press, Standford, California.

[10] Fitzmaurice GM, Laird NM. 1993. A likelihood-based method for analysing longitudinal binary responses. Biometrika 80, 141-151.

[11] Glonek GFV, McCullagh P. 1995. Multivariate logistic models. J. R. Statist. Soc. B 57, 533-546.

[12] Glonek GFV. 1996. A class of regression models for multivariate categorical responses. Biometrika 83, 15-28.

[13] Lang JB, Agresti A. 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. J. Amer. Statist. Assoc. 89, 625-632.

[14] Molenberghs G, Lesaffre E. 1994. Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. J. Amer. Statist. Assoc. 89, 633-644.

[15] Ekholm A. 2003. Comparing the odds and the dependence ratios. Pp. 13-25 in Statistics, Econometrics and Society: Essays in honour of Leif Nordberg. R. Höglund, M. Jäntti, and G. Rosenqvist (eds). Helsinki: Statistics Finland.

[16] Fitzmaurice GM, Laird NM, Rotnitzky AG. 1993. Regression models for discrete longitudinal responses (with discussion). Statistical Science 8, 284-309.
[17] Lindsey JK, Lambert P. 1998. On the appropriateness of marginal models for repeated measurements in clinical trials. Statistics in Medicine 17, 447-469.
[18] Lesaffre E, Spiessens B. 2000. Discussion of the Paper by Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. Statistical Science 15, 1-26
[19] Liang K-Y, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.
[20] Prentice RL. 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics 44, 1033-1048.
[21] Zhao LP, Prentice RL. 1990. Correlated binary regression using a quadratic exponential model. Biometrika 77, 642-648.
[22] Lipsitz SR, Laird NM, Harrington DP. 1991. Generalized estimation equations for correlated binary data: using the odds ratio as measure of association. Biometrika 78, 153-160.
[23] Liang K-Y, Zeger SL, Qaqish B. 1992. Multivariate regression analyses for categorical data. J. R. Statist. Soc. B 54, 3-40.
[24] Carey V, Zeger SL, Diggle P. 1993. Modelling multivariate binary data with alternating logistic regressions. Biometrika 80, 517-526.
[25] Cox DR, Reid N. 1987. Parameter orthogonality and approximate conditional inference (with discussion). J. R. Statist. Soc. B 49, 1-39.
[26] Heumann C. 1996. Marginal regression modeling of correlated multicategorical response: a likelihood approach. SFB386 - Discussion Paper 19. Universität München.
[27] Lehmann EL. 1990. Model specification: The views of Fisher and Neyman, and later developments. Statistical Science 5, 160-168.
[28] Cox DR. 1990. Role of models in statistical analysis. Statistical Science 5, 169-174.
[29] Dennis JE, Schnabel RB. 1983. Numerical methods for unconstrained optimization and nonlinear equations. Prentice-Hall, Englewood Cliffs, NJ.
[30] Lange K. 1999. Numerical analysis for statisticians. Springer-Verlag, New York.
[31] Seber GAF, Wild CJ. 2003. Nonlinear regression. New York, John Wiley.
[32] Ekholm A, Green M. 1994. Fitting nonlinear models in GLIM4 using numerical derivatives. GLIM Newsletter 23, 12-20.
[33] Schnabel RB, Koontz JE, Weiss BE. 1985. A modular system of algorithms for unconstrained minimization. ACM Trans. Math. Software 11, 419-440.
[34] Lindsey JK, Lindsey PJ. 2006. Multivariate distributions with correlation matrices for nonlinear repeated measurements. Comput. Statist. Data Anal. 50, 720-732.
[35] Kastner C, Fieger A, Heumann C. 1997. MAREG and WinMAREG A tool for marginal regression models. Comput. Statist. Data Anal. 24, 237-241.
[36] Lang JB. 2004. Maximum likelihood fitting of multinomial-Poisson homogeneous (MPH) models for contingency tables using MPH.FIT. Online HTML-document, March 30, 2004. `http://www.cs.uiowa.edu/~jblang/mph.fitting/mph.fit.documentation.htm`
[37] Little RJA, Rubin DB. 2002. Statistical Analysis with Missing Data, second ed. New York, John Wiley.
[38] Robins JM, Rotnitzky A, Zhao LP. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Amer. Statist. Assoc. 90, 106-121.
[39] Rubin DB. 1976. Inference and missing data. Biometrika 63, 581-592.
[40] Diggle PJ, Kenward MJ. 1994. Informative dropout in longitudinal data analysis (with discussion). Applied Statistics 43, 49-94.

[41] Eskola J, Kilpi T, Palmu A, Jokinen J, Haapakoski J, Herva E, Takala A, Käyhty H, Karma P, Kohberger R, Siber G, Mäkelä PH and the Finnish Otitis Media Study Group. 2001. Efficacy of a pneumococcal conjugate vaccine against acute otitis media. New England Journal of Medicine 344, 403-409.

[42] Molenberghs G, Kenward MJ, Lesaffre E. 1997. The analysis of longitudinal ordinal data with nonrandom dropout. Biometrika 84, 33-44.

[43] Lang JB, McDonald JW, Smith PWF. 1999. Association-marginal modeling of multivariate categorical responses: a maximum likelihood approach. J. Amer. Statist. Assoc. 94, 1161-1171.

[44] Drum M, McCullagh P. 1993. Discussion of the paper by Fitzmaurice GM, Laird NM, Rotnitzky AG. Regression models for discrete longitudinal responses. Statistical Science 8, 284-309.

[45] Joe H. 1997. Multivariate models and dependence concepts. Chapman & Hall. London.

[46] Ekholm A, Jokinen J, McDonald JW, Smith PWF. 2006. A latent class model for bivariate binary responses from twins. S^3RI methodology working papers, M06/10. http://eprints.soton.ac.uk/archive/00039276

Appendix

R-function `drm` in package `drm` (`http://www.helsinki.fi/~jtjokine/drm/`) is the wrapper for the essential algorithms of the fast estimation procedure. In this Appendix, we summarise the essential R code how to calculate the probabilities of the observed profiles, with given regression and association parameters, and how to obtain the maximum likelihood estimates. For clearer presentability, these may differ to a small degree from the actual code in package `drm`, but the functionality remains.

The core of the estimation algorithm can be split into three parts:

(1) Transformation of the observed profiles
(2) Calculation of the probabilities of the observed profiles
(3) Numerical optimisation of the likelihood function

The part (1) transforms each of the observed profiles $\boldsymbol{y}_i$, $i = 1, \ldots, n$, into a sequence of values 0, 1, -1. This is required in order to obtain the correct linear combination of the moments that correspond to the probability of the observed profile. In other words, for each of the observed profiles, the corresponding row of the matrix W_q (see Eq. (7) in IV) is needed. This is calculated with function `kronecker.drm` before the numerical optimisation.

```
#####################################################################
### TRANSFORMATION OF THE OBSERVED PROFILES.
### Adapted from kronecker.drm (kroneckerd.drm for the dropout model)
## y = observed response profile, with 1 as the reference level
## nrep = number of repetitions within a profile
## nclass = number of possible categories for the response

## create vectors [1, -1, ..., -1] for the responses at level 1:
ymat <- matrix(0, ncol = nrep, nrow = nclass)
ymat[, y == 1] <- c(1, rep(-1, (nclass - 1)))

## and vectors [0,...,1,...] for the responses at levels 2,..,nclass
for (i in 2:nclass)
  ymat[, y == i] <- c(rep(0, (i - 1)), 1, rep(0, (nclass - i)))

## parse generic command that calculates vector products recursively
w <- paste(paste(rep("c(", nrep - 1), collapse = ""),
           paste("ymat[,1]",paste("%*% t(ymat[,",2:nrep,"]))",
                                  collapse = "")))
## evaluate the parsed command:
w <- eval(parse(text = w))
#####################################################################
```

Eventually, `kronecker.drm` is utilized to produce a matrix with n columns, where the columns are the transformed $(1 \times nclass^{nrep})$ vectors of the observed profiles.

The part (2) calculates the probabilities of the observed profiles as the product of the univariate means, the dependence ratios, and the matrix calculated in part (1). The example code below, adapted from function `logliks.drm`, shows the calculation for a binary response (`nclass` =2), assuming the structure $\mathcal{B}$ (see Sec. 2.3 in IV).

```
#####################################################################
### CALCULATION OF THE PROFILE PROBABILITIES (ASSOCIATION B)
## regr.parm/ass.parm = regression/association parameters
## X = design matrix for the regression
## w = (2^nrep x n) matrix, calculated in part(1)

## Number of repetitions of level 2 for each possible profile.
## (in other words, r corresponds to |w| in Sec.2.3 of IV).
r <- apply(expand.grid(rep(list(1:2),nrep)), 1,
           function(i) length(i[i==2]) )

### Adapted from logliks.drm (loglikd.drm for dropout model):
##  Fit a regression model and transform to univariate means
mu <- cbind(1, binomial()$linkinv(offset + X %*% regr.parm))

## Create (2^nrep x n) matrix of products of univariate means;
## calculate products of (1,mu_{i1}),..,(1,mu_{i nrep}) as in part(1)
muv <- paste(paste(rep("c(", nrep - 1), collapse = ""),
             paste("mu[(1+(", nrep, " * (i - 1))),]",
                   paste("%*% \nmu[(", 2:nrep, "+ (", nrep,
                         " * (i - 1))),drop=F,])", collapse = "")))
mu <- sapply(1:n, function(i, muv) eval(parse(text=muv)), muv=muv)

## Create corresp.(1 x 2^nrep) vector of tau's (see Sec.2.3 of IV)
tau <- c(1, sapply(2:(2^nrep), function(i, ass, r) {
  prod(c(ass[1] + (1:r[i]) - 1)/c(sum(ass) + (1:r[i]) - 1))*
    ((sum(ass)/ass[1])^r[i])}, ass = ass.parm, r = r))

## Assume here that the dependence ratios are constant over n
tau <- matrix(tau)[,rep(1,n)]

## profile probability vector with given regr.and assoc.parameters:
lik <- rep(1, 2^nrep) %*% (mu * tau * w)
#####################################################################
```

In part (3), initial parameter values and function `logliks.drm`, along with the arguments required by it, are given to the numerical optimisation function `nlm` (in R base package) to obtain the maximum likelihood estimates. For details of the optimisation algorithm, see the help-file of `nlm`, and the references therein [29, 33].