POLITECNICO DI TORINO

SCUOLA DI DOTTORATO

Dottorato in Matematica per le Scienze dell'Ingegneria - XXVI Ciclo

Settore disciplinare: STATISTICA SECS-S/01

Tesi di Dottorato

Ordinal data: a new model with applications

Moreno Ursino

Tutore
Prof. Mauro Gasparini

Coordinatore del corso di Dottorato
Prof. Lamberto Rondoni

Marzo 2014

to my family

I can prove anything by statistics
except the truth.
George Canning

Preface

This thesis is divided in two parts.

The first part concerns the main work and the title of the thesis. We present a new model to analyse ordinal data and extend it to longitudinal case. In the first chapter, an overview of the analysis of ordered categorical data is presented. We focused on regression models, in particular on *cumulative logistic regression* and on CUB model. Clustered and repeated ordinal response data are also discussed.

In the second chapter our new discrete-beta distribution is introduced. We postulate the existence of a beta latent random variable and, with fixed cut-points, we find a discrete probability distribution that can well fit the distribution of a response variable Y which has ordinal outcomes. We investigate its shapes and its properties. Comparisons with respect to binomial, to beta-binomial and to CUB models are discussed.

The third chapter concerns the investigation of the problem of parameters estimation in both cases of a single population and in presence of covariates. We do not follow the approach of generalized linear models, in which it is allowed for an arbitrary function (the link function) of the mean of response variable, $g(\mathbb{E}[Y])$, to vary linearly with the predicted values, rather than assuming that the response itself must vary linearly. In a natural way, according to CUB model framework, we prefer to directly introduce covariates for the parameters of the latent variable, without a direct reference to the expectation of this random variable. Then, a case study of gastroesophageal reflux disease is shown.

In the fourth chapter an extension to longitudinal data is presented. In order to take into account the correlation between the repeated responses, we add *random effects* to the model presented in the previous chapter. An EM algorithm is proposed in order to obtain parameter estimates.

In the last chapter conclusions is summarised and in the appendices the R-scripts

used are shown.

The second part consists of an article on a proximity-based method to identify genomic regions correlated with a continuously varying environmental variable. One strategy for the detection of spatial selection signatures is the outlier approach. Our idea is to reinforce the outlier approach by considering a criterion of *proximity* between significant variants. We therefore adopt a search-and-confirm approach which integrates the outlier approach by identifying regions of the genome where not just one, but a significant number of SNPs (single nucleotide polymorphisms) are located in the tails of the distribution of the relevant statistic, when compared to the number of SNPs originally genotyped in the same region.

Acknowledgements

I would like to express my special appreciation and thanks to my advisor Professor Mauro Gasparini for his encouragement and for his valuable suggestions during the planning and development of this work. My grateful thanks are also extended to Drs Edda Battaglia for giving me the possibility to work with her clinical data, and to Dr Mario Grassini. I would also like to thank my referees, Professor Domenico Piccolo and Professor Kai-Sheng Song for their constructive suggestions.

Finally, special thanks to my family, to my friends and, above all, to my colleagues Anna, Chiara and Stefano, with whom I shared, even outside of work, these three wonderful years.

Part I
Ordinal data: a new model with applications

Contents

Chapter 1

State of the art in ordinal data analysis

Ordered categorical data are ubiquitous in the medical and health sciences, and research in this field has been vigorous over the past 30 or so years. An observed ordered categorical variable may rise as a result of *categorizing a continuous variable*, or it may be an inherently categorical measurement. In the medical field it is usual to classify patients as being at various degrees of risk for developing a disease or condition according to specified ranges of risk factors. For example, serum cholesterol and blood pressure levels are important risk factors in cardiovascular disease. These variables are intrinsically continuous, but it is useful to analyse them as ordered categorical data. Survey data, in which respondents are asked to characterize their opinions on a scale ranging from "strongly disagree" to "strongly agree", are another common example of such data.

On the other hand, in classifying levels of pain relief attained with a treatment it is often only possible to arrive at subjective categorizations, such as "no relief", "some relief", "considerable relief" and "complete relief". For our purposes, the defining property of ordinal data is that there exist a clear ordering of the response categories, but no underlying interval scale between them. For example, it is generally reasonable to assume an ordering on the form

$$\text{no relief} < \text{some relief} < \text{considerable relief} < \text{complete relief},$$

but it is usually does not make sense to write operations between them, such as

$$\text{“considerable relief”} - \text{“no relief”} = ...$$

Even when the observed data arises from the subjective assignment of an observed response into a category on an ordered scale, it might be reasonable to assume that there indeed exists an underlying continuous random variable for which the discrete classification is an imperfect measure.

It should be noted that methods developed for categorical data in general can be applied to the analysis of ordered categorical data. There are, however, important advantages to using models and methods developed explicitly to take account of the ordinal structures of categories. In particular, models for ordered categorical data tend to be more parsimonious than their more general counterparts, thus resulting in more efficient inferences and facilitating the interpretation of parameters.

The models and methodology for ordinal data focus on two general areas of statistical analysis, *association* and *regression*. For the first area, see [5, 49, 2]. In this chapter we outline the regression models. The most popular models apply a link function to the cumulative probabilities, most commonly the logit o probit.

1.1 Logistic regression

The logistic regression and loglinear models were primarily developed in 1960s and 1970s. Although ordinal data received some attention in those years [70, 7], a stronger focus on this field was inspired later by articles by Mc-Cullagh [52] on logit modeling of cumulative probabilities and by Goodman [28] on loglinear modeling relating the odd ratios, that are natural for ordinal variables. The logistic regression is a regression model for binary outcomes typically called “success” or “failure”, in which the mean response is linked linearly to a set of predictor variables or covariates via the log-odds for success versus failure. That is, let $\pi(\mathbf{x}_i)$ denote the probability of success for subject i, with covariate vector $\mathbf{x}_i$. Let

$$\text{logit}\left[\pi(\mathbf{x}_i)\right] = \ln \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}, \tag{1.1}$$

denote the so-called logit for observation i. Then the standard logit regression is

$$\text{logit}\left[\pi(\mathbf{x}_i)\right] = \beta_0 + \beta\mathbf{x}_i, \tag{1.2}$$

where β_0 is the intercept term.

Moving to ordinal data there are multiple approaches to defining logit for the response. The three logits that figure most prominently in the biostatistics literature are adjacent-categories logits, continuation-ratio logits, and cumulative logits. We take the number of category for the response to be m, let $\pi_j(\mathbf{x}_i)$ denote the probability of response category j given the covariate vector $\mathbf{x}_i$ and let $\gamma_j(\mathbf{x}_i) = \sum_{k=1}^{j} \pi_k(\mathbf{x}_i)$.

The *adiacent-categories logit* model [69, 29] is defined as

$$\ln\left[\frac{\pi_j(\mathbf{x}_i)}{\pi_{j+1}(\mathbf{x}_i)}\right] = \beta_{0j} + \beta_j \mathbf{x}_i, \quad j = 1, \cdots, m-1. \tag{1.3}$$

The model is a special case of the baseline-category logit model commonly used for nominal response variables, with reduction in the number of parameters by using the natural ordering to obtain a common effect.

The so called *continuation-ratio* logit model uses logits of form

$$\text{logit}\left[\pi_j(\mathbf{x}_i)\right] = \ln\frac{\pi_j(\mathbf{x}_i)}{1-\gamma_j(\mathbf{x}_i)} \quad \text{or} \quad \text{logit}\left[\pi_j(\mathbf{x}_i)\right] = \ln\frac{\pi_{j+1}(\mathbf{x}_i)}{\gamma_j(\mathbf{x}_i)}. \tag{1.4}$$

Tutz [75, 76] showed that this model is useful when a sequential mechanism determines the response outcome.

Currently, the most popular model for ordinal responses is *cumulative logit*. Early work with this approach includes [82] and [69]. The model has form

$$\text{logit}\left[\gamma_j(\mathbf{x}_i)\right] = \ln\frac{\gamma_j(\mathbf{x}_i)}{1-\gamma_j(\mathbf{x}_i)} = \alpha_j - \boldsymbol{\beta}\mathbf{x}_i\ , \quad j = 1, \cdots, m-1. \tag{1.5}$$

The minus sign in $\boldsymbol{\beta}$ makes the sign of each component of the vector have the usual interpretation, that is a positive or negative value means a positive or negative effect. The parameters $\{\alpha_j\}$ are called *cut-points*. This model applies simultaneously to all $m-1$ cumulative probabilities and it assumes an identical effect of the predictors for each cumulative probability. More precisely, models where the vector of the regression coefficients $\boldsymbol{\beta}$ is constrained to be equal for all $j = 1, \cdots, m-1$ are referred to as **proportional-odds models**. McCullagh [52] has been highly influential in promoting the use of proportional-odds cumulative logit models. To fit this models, it is unnecessary to assign scores to the response categories. It is possible to build the model from the postulation of the existence of an underlying latent variable Y^* [3] associated with each

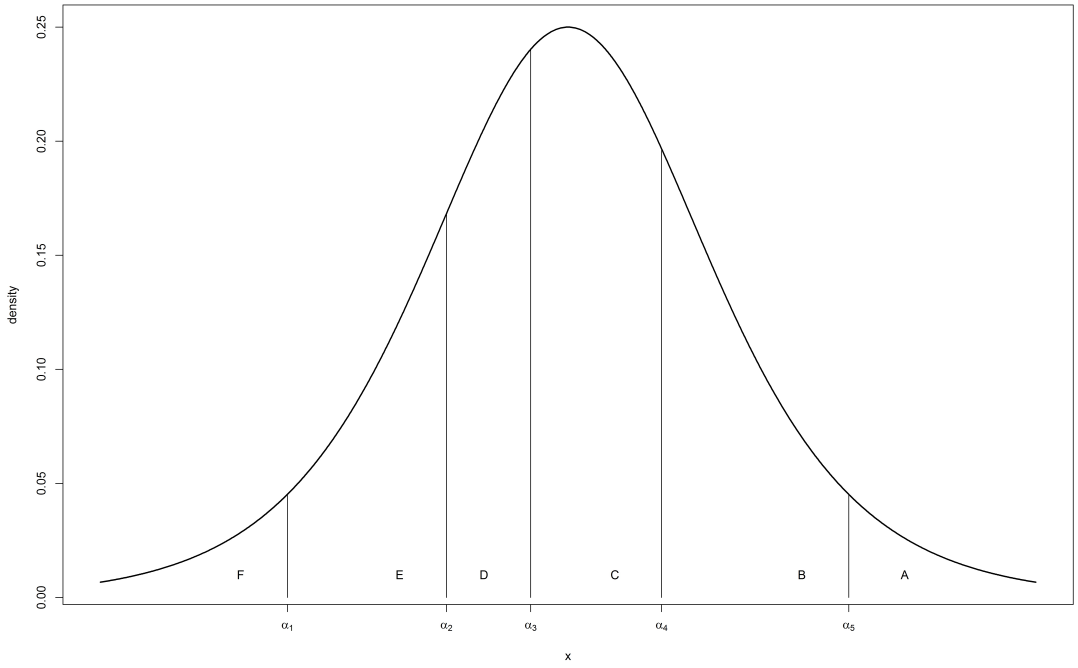

Figure 1.1: An example of a logistic distribution with five cut-points in order to describe the probabilities on the discrete set $\{A, B, C, D, E, F\}$ with $A > B > C > D > E > F$, such as grades in an exam.

response. Such variables are often assumed to be drawn from a continuous distribution centered on a mean value that is modeled as a linear function of the covariates vector. Then, if we choose Y^* with logistic conditional distribution with constant variance and we chop it with cut-points $\{\alpha_j\}$, we obtain the discrete distribution of Y and the model described above [42].

The definitions of the adjacent-categories logits and the cumulative logits are invariant with respect to reversing the ordering of the response categories, but the high to low ordering of the categories gives a different set of continuation ratio logits than are obtained with the low-to-high ordering. The decision of which set of logits to use, and which ordering in the case of the continuation-ratio logits, should be based on the substantive questions of interest that are to be addressed in the statistical analysis.

When proportional-odds structure seems inadequate, alternative possible strategies to improve the fit include

1. trying different link functions, such as the log-log or the probit link (where Y^* are

assumed to have a standard normal distribution);

2. adding additional terms, such as interactions, to the linear predictor;

3. adding dispersion parameters [52, 16];

4. permitting separate effects for each logit for some but not all predictors (i.e., *partial proportional odds* [59]).

1.2 CUB model

D'Elia and Piccolo [19] have proposed a mixture model (CUB model) that has been proven to be a useful tool for analysing preference data sets in several contexts. Their approach is motivated by a direct investigation of the psychological process that generates the choice mechanism among m alternatives. The idea is that the response is the result both of *feeling* and *uncertainty* components, which are different random variables to be combined in a mixture [63]. Both feeling and uncertainty, F^* and U^*, are continuous and latent random variables, whereas ordinal data are expressed as discrete responses, Y, thus, they are required efficient transformations of continuous variables into a discrete set:

- *feeling*: if it is assumed that the records are generated by an unobserved continuous random variable Normally distributed, then its discrete form obtained by varying cut-points is well fitted by a *shifted Binomial* random variable F

$$\mathcal{P}(F = r) = \binom{m-1}{r-1}(1-\xi)^{r-1}(\xi)^{m-r} \quad r = 1, \cdots, m;$$

- *uncertainty*: under the circumstance that the subject shows a complete indifference towards a given item, then it seems appropriate to model ranks by means of a discrete uniform random variable U

$$\mathcal{P}(U = r) = \frac{1}{m} \quad r = 1, \cdots, m.$$

Finally, taking into account the composite nature of the elicitation process by means of a mixture model, where the feeling and uncertainty components are adequately weighted, the authors proposed the (discrete) MUB random variable Y with parameters ξ and π,

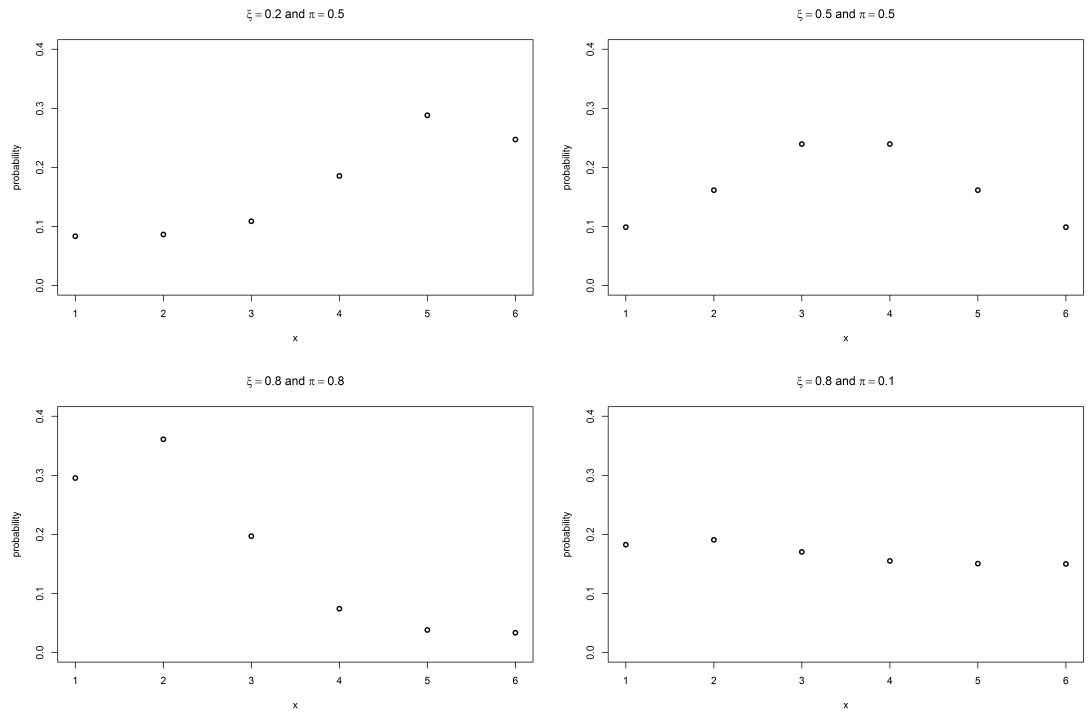

Figure 1.2: Examples of shapes assumed by a random variable with a MUB distribution.

on the finite support $\{r : r = 1, \cdots, m\}$

$$\mathcal{P}(Y = r) = \pi \left[\binom{m-1}{r-1} (1-\xi)^{r-1} (\xi)^{m-r} \right] + (1-\pi) \frac{1}{m} \quad r = 1, \cdots, m, \tag{1.6}$$

with $\xi \in [0, 1]$ and $\pi \in [0, 1)$ [61]. Therefore Y can be seen as a point in the parametric space $\xi \times \pi = [0, 1] \times [0, 1)$. Moreover, the requirement $m > 3$ allows the identifiability of the models as it avoids the cases of a degenerate random variable (if $m = 1$), an indeterminate model (if $m = 2$) and a saturated model (if $m = 3$) [39]. With only two parameters they are able to model extreme and intermediate modes, positive and negative skewness, flat and peaked behaviours, and so on. In particular, π is related to uncertainty, since each subject has a propensity to adhere to a resolute choice and a totally uncertain choice, weighted by π and $1 - \pi$ respectively. On the other hand, the meaning of ξ changes with the setting of the analysis and it may be a degree of perception, a measure of closeness, a rating of concern and so on.

Since different sets of (π, ξ) give the same expectation of Y, it seems preferable to look for a link among model parameters and subjects' covariates, $\mathbf{y}_i$ and $\mathbf{w}_i$, without a direct reference to the expectation of this random variable. Thus, in [19] it is proposed a logistic mapping among the covariates and the parameters π and ξ defined by

$$\pi \mid \mathbf{y}_i = \frac{1}{1 + e^{-\mathbf{y}_i \boldsymbol{\beta}}} \tag{1.7}$$

$$\xi \mid \mathbf{w}_i = \frac{1}{1 + e^{-\mathbf{w}_i \boldsymbol{\gamma}}} \tag{1.8}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of regression parameters. CUB models are MUB random variable where we assume that π and ξ are functions of subjects covariates. This approach is logically related to GLM since the parameters of a model are related to subjects' covariates. However, they have not introducing a link function between the expectation and the covariates, and the probability distribution does not belong to the exponential family.

Three extensions are represented by the models that take into account the presence of a *shelter choice* [37], the CUBE models, where a beta-binomial is used in place of binomial in order to capture overdispersion [40, 62, 38], and the hierarchical CUB models [36]. However, all these models have more than two parameters.

Other multinomial response models

Tutz [77] extended the form of generalized additive models [31] by considering semi-parametrically structured models. These models contains linear parts ($\mathbf{X}_i\boldsymbol{\beta}$), additive parts of covariates with an unspecified functional form, and interactions.

1.3 Clustered or repeated ordinal response data

A more general area in which there has been much recent activity is the analysis of repeated measures data in the form of an ordered categorical response. Such data arise, for example, in longitudinal studies, crossover experiments, and studies of families or sibship. All these data have some sort of clustering. General approach to the analysis of repeated measures data have been applied to such problems, including methodology based on *generalized estimating equations* [32, 45] and methodology based on incorporating *random effects* [23, 33, 45] into the models. The primary emphasis is on three

classes of models:

- *marginal models* describe the so-called *population-averaged* effects which refer to an averaging over clusters at particular level of predictors. The model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses;
- *conditional models* (*subject-specific models*) describe effects at the cluster level. The mean response depends not only on covariates but also on a vector of random effects;
- *transitional models* describe a response conditional on past responses and other covariates.

1.3.1 Marginal models

Let N denote the number of clusters, for example the number of patients, Y_{it} the response variable for the ith subject on the t-th measurement and T_i the number of repeated measurements of the response on the ith subject. The response for the ith subject can be grouped into a $T_i \times 1$ vector

$$\mathbf{y}_i = \left(y_{i1}, y_{i2}, \cdots, y_{iT_i}\right)^T = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{bmatrix} \quad i = 1, \cdots, N; \tag{1.9}$$

and $\mathbf{y}_i$'s are assumed to be independent of one another. Associated with each response y_{it} is a $p \times 1$ vector of covariates $\mathbf{x}_{it}$

$$\mathbf{x}_{it} = \begin{bmatrix} x_{it1} \\ x_{it2} \\ \vdots \\ x_{itp} \end{bmatrix} \quad i = 1, \cdots, N; \quad t = 1, \cdots, T_i. \tag{1.10}$$

The marginal expectation of the response, $\mathbb{E}(Y_{it}) = \mu_{it}$ depends on the covariates, $\mathbf{x}_{it}$, through a known link function $g(\mu_{it}) = \mathbf{x}'_{it}\beta$. The marginal model with cumulative logit

link has the form

$$\text{logit}\left[\mathcal{P}\left(Y_{it} \leq j|\mathbf{x}_{it}\right)\right] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}_{it} \qquad (1.11)$$
$$j = 1, \cdots, m-1; \qquad t = 1, \cdots, T_i.$$

Marginal models describe the relationship between the mean of the outcome Y_{it} and the risk factors $\mathbf{x}_{it}$ across all observations simultaneously regardless of whether the observations come from a common individual or not.

Maximum likelihood estimation for these models is computationally challenging, but it is easier to apply a generalized estimating equations (GEE) method. GEE is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes. Parameter estimates from the GEE are consistent even when the covariance structure is misspecified, under mild regularity conditions. The focus of the GEE is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual. The GEE methodology, originally specified for marginal models with univariate distributions such as the binomial and the Poisson, extends to cumulative logit models [48] and cumulative probit models [74] for repeated ordinal response.

Alternatively, [44] proposed the classical marginal models with cumulative logit [52] where covariate effects are allowed to vary with time. In other words, the parameters $\boldsymbol{\beta}(t)$ depends on time t.

1.3.2 Generalized Linear Mixed Effects Models

Mixed effects models focus initially on the regression relationship restricted to observations on a single individual. The model is then extended to multiple individuals by allowing some pieces of the model to vary from individual to individual in a prescribed manner, while other components remain the same. The models have conditional interpretations with cluster-specific effects. When the response has distribution in the exponential family, generalized linear mixed models (GLMMs) add random effects to generalized linear models (GLMs).

For example, in `R` software, library `lme4`, that we use in our applications in the last chapter, adds random effects to models implemented in library `glm`. In general,

random effects in models can account for a variety of situations, including subject heterogeneity, unobserved covariates and other form of over-dispersion. Building a mixed effects model focuses on the introduction of *random effects*, these are the pieces of the model that vary across individuals, in addition to *fixed effects*, the relationships that are assumed to be identical for every subject. This approach indirectly describes and interprets the covariance structure for longitudinal observations. The repeated measures are typically assumed to be independent, given the random effects, but variability in the random effects induces a marginal non-negative association between pairs of response after averaging over the random effects. The model with cumulative logit link and random effects has the form

$$\text{logit}\left[\mathcal{P}\left(Y_{it} \leq j | \mathbf{x}_{it}, \mathbf{z}_{it}\right)\right] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}_{it} - \mathbf{u}_i'\mathbf{z}_{it} \tag{1.12}$$
$$j = 1, \cdots, m-1; \quad t = 1, \cdots, T_i; \quad i = 1, \cdots, N;$$

where $\mathbf{z}_{it}$ refers to a vector of explanatory variables for the random effects and $\mathbf{u}_i$ are *iid* from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ [78]. In the simplest case $\mathbf{z}_{it}$ is a vector of 1's and $\mathbf{u}_i$ to be a single random effects from a univariate normal distribution $\mathcal{N}\left(0, \sigma^2\right)$. The form (1.12) can be extended to models that use different link function, such as in [30, 15, 58] or to models with scalar random effects with different variance among individuals [66].

When $\mathbf{z}_{it}$ is a vector of size greater than two, that is the vector of random effects contains more than a couple of terms, model fitting with maximum likelihood estimation can be challenging. The best linear unbiased prediction of parameters of an underlying continuous distribution was used in [18]. The classical approach is based on computing the joint likelihood function as the usual products of multinomials, such as in the case in which random effects were known. Since they are unobservable, in order to obtain a likelihood to maximize, we integrate out the random effects from joint likelihood [47]. In other words, we compute the maximum likelihood from the marginal distribution of $\mathbf{Y}_i$. Except in rare cases, such as the complementary log-log link with the log of a gamma or inverse Gaussian distribution for random effects [17, 73], this integral does not have closed form and it is necessary to use some approximations.

For simple models, such as random effects intercept models, Gauss-Hermite quadrature is usually adequate [33, 34]. An adaptive version of Gauss-Hermite quadrature can be found in [50, 64]. For models with higher-dimensional integrals, more feasible meth-

ods use Monte Carlo methods, which use the randomly sampled nodes to approximate integrals. In [9], the authors proposed an automated Monte Carlo EM algorithm for generalized linear mixed models which assesses the Monte Carlo error in current parameter estimates and increases the number of nodes if the error exceeds the change in the estimates from the previous iteration.

Pseudo-likelihood method can avoid the problems associated with the integration, making computations simpler [10]. Unfortunately, these methods are biased for highly non-normal cases such as Bernoulli response data [10, 22], with the bias increasing as variance components increase.

If the distribution of the response variable Y, which is the discrete form of the distribution of the latent variable Y^*, can be assumed to be well approximated by a simpler discrete distribution such as binomial or beta-binomial distribution, then it is possible to use directly the theory of GLMMs. This approach has been described by authors in [56] and the GLMM has been extended by adding non-parametric random effects (GAMLSS)[71].

1.3.3 Transition models

Another type of model to analyse repeated measurement data, called *transition model*, describes the distribution of a response conditional on past responses and explanatory variables. In transition models, the probability of the outcome of individual i at time t, Y_{it}, is a function of the individual's covariates at time t, $\mathbf{x}_{it}$, and the individual's outcome history $Y_{i1}, \cdots, Y_{i(t-1)}$, with $t > 1$ [25]. Such models are appropriate when there is a natural sequencing of the response, as in longitudinal studies. This approach has received substantial attention for binary data [8]. Many transitional models have Markov chain structure taking into account the time ordering [55, 43].

The choice among marginal, conditional and transitional models depends on whether one prefers interpretations to apply at the population or the subject level and on whether it is sensible to describe effects of explanatory variables conditional on previous responses. Conditional models are especially useful for describe within-cluster effects, such as within-subject comparisons in a crossover study. Marginal models may be adequate if one is interested in summarizing responses for different groups (e.g., gender and race) without spending much effort on modeling the dependence structure.

Chapter 2

The discrete-beta distribution

In this chapter we introduce a new distribution that will be useful in the field of ordinal data analysis. The idea is similar to those described in [19]: finding a discrete probability distribution that can well fit the distribution of a response variable Y, that has ordinal outcomes. We want a distribution depending on the smallest number of parameters, but that describes properly the choice of an individual. For example, the binomial distribution has only one parameter, p, but it is not very flexible, since its mean and variance are both functions of p (note that the number of Bernoulli trials n is fixed and is not a parameter). Thus, we focus on a distribution with two parameters, where the first describes the level, and the second is linked to variability.

2.1 From latent variable to discrete variable

As in [42], we suppose the existence of an underlying latent distribution. A normal distribution, $\mathcal{N}(\mu, \sigma^2)$, could be a candidate, since it has two parameters with simple interpretation, i.e. mean and variance. Really, it hides many other parameters, since, moving to the discrete case, we need cut-points, that should be estimated. In other words, they are also parameters. In order to overcome this increasing in the number of parameters, our idea consists of assuming a latent variable with compact support and fixed cut-points.

We choose a *beta* distribution, because it has a highly flexible shape, including bell-shapes (symmetric or skewed), U-shapes and J-shapes. In the following, we assume that

Y^*, the latent variable, follows a beta distribution, i.e its density is

$$f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1) \tag{2.1}$$

where $\mu \in (0,1)$, $\phi > 0$ and $\Gamma(t)$ is the gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\, dx.$$

This is not the usual parametrization of the beta law, where

$$\begin{cases} \alpha = \mu\phi \\ \beta = (1-\mu)\phi \end{cases} \tag{2.2}$$

but it is convenient for modeling purposes and it is often used in beta regression analysis, as we can see in [24]. Since

$$\mathbb{E}\left(Y^*\right) = \mu$$

$$\text{Var}\left(Y^*\right) = \frac{\mu(1-\mu)}{1+\phi}$$

μ is really the mean of the latent variable and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ, the larger the value of ϕ, the smaller the variance of Y^*. Figure 2.1 shows a few different beta densities along with the corresponding values of (μ, ϕ). It is noteworthy that the densities can display quite different shapes depending on the values of the two parameters. In particular, when $\mu = 0.5$ it is symmetric and otherwise it is asymmetric. It is also interesting to note that in the upper panels, two densities have J shapes and two others have inverted J shapes. When $\mu = 0.5$ and $\phi = 2$ the density reduces to a standard uniform distribution. The beta density can also be U shaped (skewed or not), and this situation is also displayed in Figure 2.1.

In order to move to the discrete case, we choose fixed cut-points, and in such a way they are not parameters. Similarly to what Morrison describes in [54], where he divides the stated intention of purchase by n in order to get values on $(0, 1/n, 2/n, \ldots, 1)$, we set uniformly spaced cut-points in the interval $[0, 1]$ (the domain of Y^*). Thus, if Y is the ordinal variable that assumes values in $[0, 1, \ldots, n]$, with $n \geq 2$, the set of cut-points is $\left\{0, \frac{1}{n+1}, \ldots, \frac{n}{n+1}, 1\right\}$. Then, we imposed that $Y = k$ if and only if $Y^* \in \left(\frac{k}{n+1}, \frac{k+1}{n+1}\right)$. In other words, the probability that Y is equal to k is the same of the probability that

the latent variable Y^* falls in the interval $\left(\frac{k}{n+1}, \frac{k+1}{n+1}\right)$. In this framework, we obtain that

$$\begin{aligned} P\left(Y=k\right) &= \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}\, dx \\ &= I_{\frac{k+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \end{aligned} \tag{2.3}$$

where $I_x(p,q)$ is called the *incomplete beta function ratio*[41] (but we will called it simply

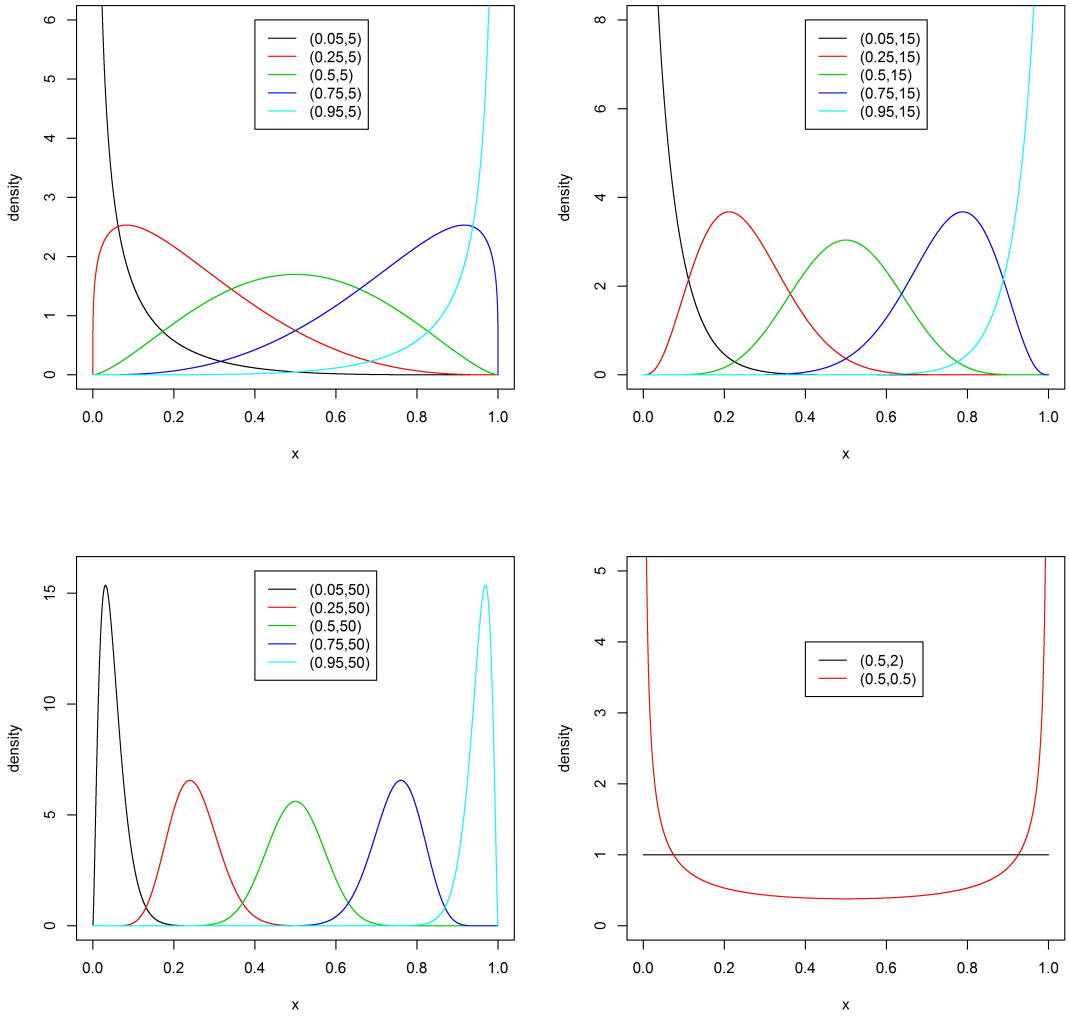

Figure 2.1: Beta densities plotted for different combinations of (μ, ϕ).

incomplete beta function from now on) and it is defined as

$$I_x(p,q) = \frac{1}{B(p,q)} \int_0^x t^{p-1}(1-t)^{q-1}\, dt \tag{2.4}$$

with

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{2.5}$$

We said that Y follows a *discrete-beta* distribution if its probabilities are equal to those described in eq. (2.3). We can also say that $Y = \lfloor (n+1)Y^* \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Note that it is different from the distribution defined in [65], that is defined, without re-parametrization, as follows

$$P(Y = y) = \frac{(1+y)^{\alpha-1}(m+1+y)^{\beta-1}}{\sum_{j=0}^{m}(1+j)^{\alpha-1}(m+1+j)^{\beta-1}}, \quad y = 0, 1, \ldots, m, \tag{2.6}$$

with $\alpha, \beta \in \mathbb{R}$ and $m \geq 1$.

This approach is based on the opposite idea found, for example, in [42]: usually, the latent distribution has a fixed form, i.e. a logistic with known variance, and the value of the mean of such distribution joint with the estimated cut-points give us the probabilities of the discrete random variable Y; on the contrary, here we have that the flexible shape of the latent variable defines the probabilities of Y, since cut-points are fixed. In Figure 2.2 different discrete-beta distributions are shown along with the corresponding values of (μ, ϕ), which are the same values used in Figure 2.1. The constrain of the uniform subdivision of the finite support of the beta distribution allows us to obtain the same trend of the latent variable. Thus, except in the limit case of the U shape, the only bimodal shape that a beta distribution can assume, the other shapes are uni-modal fitting well the usual scenario encountered in a study of ordinal outcomes. For example, in survey data in which a respondent is asked to characterize his opinions on a scale ranging from "strongly disagree" to "strongly agree", he may be doubtful between two or three near values, but rarely he may be doubtful between two distant values, such as "strongly disagree" and "strongly agree". Similarly, in medical field, when a pathologist must decide the magnitude of a disease in a patient, for example categorizing the result from 0 to 10, he may be doubtful between 5, 6 or 7, but rarely between 1 or 9 (note that in this example, natural numbers are treated such as ordinal categorical data). Thus, it makes sense to use unimodal distributions.

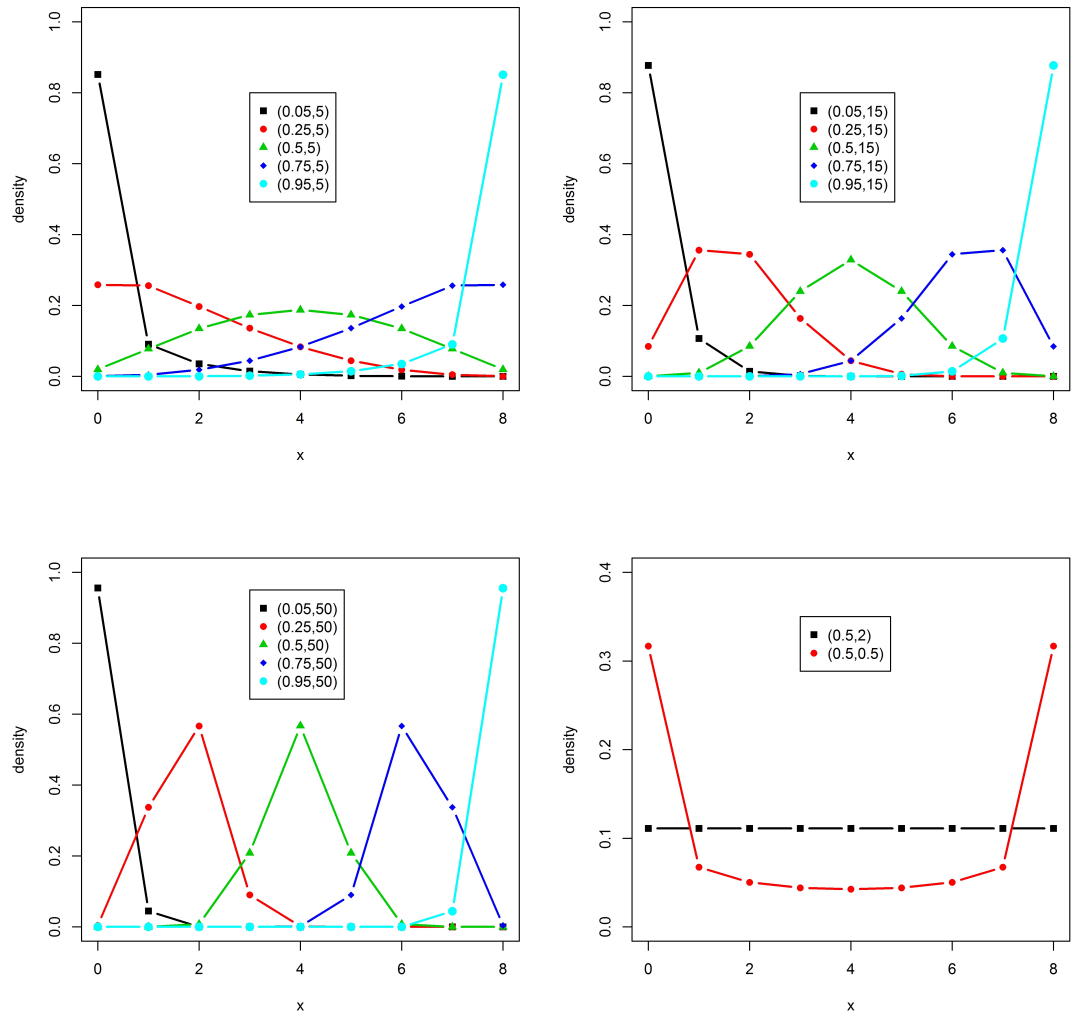

Figure 2.2: Discrete-beta densities plotted for different combinations of (μ, ϕ) and $n = 8$.

If Y follows a discrete-beta distribution, with n fixed and parameters μ and ϕ, $Y \backsim \text{Dbeta}(\mu, \phi, n)$, then

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{i=0}^{n} iP(Y = i) \\
&= \sum_{i=0}^{n} i \left(I_{\frac{i+1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) \right) \\
&= 0 + I_{\frac{2}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{1}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) + \cdots
\end{aligned}$$

$$\begin{aligned} &\cdots + nI_1\left(\mu\phi,(1-\mu)\phi\right) - nI_{\frac{n}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) \\ &= n - \sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) \end{aligned} \tag{2.7}$$

$$\begin{aligned} \mathbb{E}\left[Y^2\right] &= \sum_{i=0}^{n} i^2 P\left(Y=i\right) \\ &= \sum_{i=0}^{n} i^2 \left(I_{\frac{i+1}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) - I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right)\right) \\ &= 0 + I_{\frac{2}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) - I_{\frac{1}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) + \cdots \\ &\quad \cdots + n^2 I_1\left(\mu\phi,(1-\mu)\phi\right) - n^2 I_{\frac{n}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) \\ &= n^2 + \sum_{i=1}^{n} + \left((i-1)^2 - i^2\right) I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) \\ &= n^2 + \sum_{i=1}^{n} (1-2i)\, I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right). \end{aligned} \tag{2.8}$$

Thus, the variance can be written as

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left[Y\right] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[Y\right]^2 \\ &= n^2 + \sum_{i=1}^{n} (1-2i)\, I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) - \left(n - \sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right)\right)^2 \\ &= n^2 + \sum_{i=1}^{n} (1-2i)\, I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) - \left(n^2 - 2n \sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) + \right. \\ &\quad \left. + \left(\sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right)\right)^2 \right) \\ &= \sum_{i=1}^{n} (2n-2i+1)\, I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right) - \left(\sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi,(1-\mu)\phi\right)\right)^2. \end{aligned} \tag{2.9}$$

The impossibility to achieve a closed form for expectation and variance is a small limitation for this new class of models. Nevertheless Figure 2.3 shows their plots for $(\mu,\phi) \in (0,1) \times (0,20]$.