Using monotonicity constraints for the treatment of ordinal data in regression analysis

Javier Alejandro Espinosa Brito

Thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

Department of Statistical Science
Faculty of Mathematical & Physical Sciences
University College London

September 2020

Declaration

I, Javier Alejandro Espinosa Brito, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Javier Alejandro Espinosa Brito

Abstract

A regression model is proposed for the analysis of an ordinal response variable depending on a set of multiple covariates containing ordinal and potentially other types of variables. The ordinal predictors are not treated as nominal-scaled variables, and neither transformed into interval-scaled variables. Therefore, the information provided by the order of their categories is neither ignored nor overstated. The proportional odds cumulative logit model (POCLM, see McCullagh (1980)) is used for the ordinal response, and constrained maximum likelihood estimation is used to account for the ordinality of covariates. Ordinal predictors are coded by dummy variables. The parameters associated with the categories of the ordinal predictor(s) are constrained, enforcing them to be monotonic (isotonic or antitonic).

A monotonicity direction classification procedure (MDCP) is proposed for classifying the monotonicity direction of the coefficients of the ordinal predictors, also providing information whether observations are compatible with both or no monotonicity direction. The MDCP consists of three steps, which offers two instances of decisions to be made by the researcher.

Asymptotic theory of the constrained MLE (CMLE) for the POCLM is discussed. Some results of the asymptotic theory of the unconstrained MLE developed by Fahrmeir and Kaufmann (1985) are made explicit for the POCLM. These results are further adapted to extend the analysis of asymptotic theory to the constrained case. Asymptotic existence and strong consistency of the CMLE for the POCLM are proved. Asymptotic normality is also discussed. Different scenarios are identified in the analysis of confidence regions of the CMLE for the POCLM, which leads to the definition of three alternative confidence regions. Their results are compared through simulations in terms of their coverage probability. Similarly, different scenarios are identified in the analysis of confidence intervals of the CMLE and alternative definitions are provided. However, the fact that monotonicity is a feature of a parameter vector rather than of a singular parameter value becomes a problem for their computation, which is also discussed.

Two monotonicity tests for the set of parameters of an ordinal predictor are proposed. One of them is based on a Bonferroni correction of the confidence intervals associated with the parameters of an ordinal predictor, and the other uses the analysis of confidence regions.

Six constrained estimation methods are proposed depending on different approaches for making the decision of imposing the monotonicity constraints to the parameters of an ordinal predictor or not. Each one of them uses the steps of the MDCP or one of the two monotonicity tests. The constrained estimation methods are compared to the unconstrained proportional odds cumulative logit model through simulations under several settings.

The results of using different scoring systems that transform ordinal variables into interval-scaled variables in regression analysis are compared to the ones obtained when using the proposed constrained regression methods based on simulations.

The constrained model is applied to real data explaining a 10-Points Likert scale quality of life self-assessment variable by ordinal and other predictors.

Impact statement

The contributions in this thesis have an impact on both inside and outside academia. In academia, this work makes a contribution to the literature about the analysis and treatment of ordinal variables in regression models. In the regression analysis literature there are methods for the treatment of either ordinal responses or ordinal predictors. However, there is no literature about the treatment of an ordinal response and ordinal predictors simultaneously. This thesis proposes, among other statistical tools, a method for the treatment of an ordinal response in presence of ordinal and possibly other types of predictors in regression analysis. This method and all of its by-products can be extended to any type of response variable, increasing the value of its potential impact.

The undertaken approach requires assuming monotonic effects of ordinal predictors, which is also of interest to investigate in its own right. There is also a surprising lack of literature associated with the analysis of monotonic effects of ordinal predictors in multiple regression analysis. In this regard, a monotonicity direction classification procedure is proposed as well as two monotonicity tests. The former serves as a tool to make informed decisions about the direction to be used when imposing monotonic effects or even to drop the monotonicity assumption. With respect to the latter, and to my knowledge, the monotonicity test approaches that have concentrated the main interest of the research community are related to testing regression monotonicity in non-parametric models, considering a single (non-ordinal) predictor and/or taking into account only one monotonicity direction, which this thesis attempts to tackle by proposing two different monotonicity tests for ordinal predictors where others of any type are allowed in a parametric model. This helps researchers assess the validity of imposing monotonicity constraints or make inference about the association between some ordinal predictor and the response variable, being a novel contribution.

In general, this thesis demonstrates the application of a constrained regression model, the monotonicity direction classification procedure and the monotonicity tests as an effective set of tools to bridge the existing regression methods and

the need of specific treatment of ordinal data that includes ordinal predictors in the regression models framework, leading to a more informed regression modelling process. This will probably drive the advancement in the treatment of ordinal data in models belonging to other frameworks.

Regarding applications of the proposed methods outside academia, they are useful for regression analyses where monotonic effects of ordinal predictors are assumed, when these are the main point of interest, or when imposing monotonic associations between an ordinal predictor and the response variable improves interpretability. For example, in the finance industry they can be used to analyse the credit rating of bonds, in the public administration industry to study the country risk, in the education industry to identify the factors affecting the level of education attained by a given group of interest, etc. The proposed methods can be applied in any industry where regression models are used to conduct analyses of an ordinal response with ordinal predictors.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Christian Hennig who gave me the opportunity to study my PhD in the Department of Statistical Science at UCL, where I received his continuous support and guidance since I wrote my MSc thesis and then through all the years of my PhD. All of his comments, suggestions and detailed criticism played a crucial role to complete this project and were an important part of my learning process as a researcher. Working with Christian has been a great and enriching experience. I would also like to thank Professor Gianluca Baio, my second supervisor, and Professor Richard E. Chandler, Head of the Department of Statistical Science at UCL, whose comments about some parts of this thesis were very helpful to improve its quality.

I am also deeply grateful to my wife, Daniela Orellana. As expected, we spent almost all of the time during my PhD together with our two children, Constanza and David, far away from our home country. We knew that this would be a challenge. However, this project was much more than that and it turned out to be a life-changing experience, which we did not expect. We lived several types of moments, many will remain in our happy memories and some others will not, but her generosity, commitment and love for our family made her able to transform all of those moments into a unifying experience. Undoubtedly, without her help and support none of the good moments would have even existed. Many thanks to my most loved ones too, Constanza and David, who every day encourage me to go ahead and, without even knowing it, make me feel happy.

Last but not least, I also want to express my greatest gratitude to my parents and brother for all their love, encouragement and support throughout my whole life.

Contents

List of Figures

List of Tables

Chapter 1

Background

1.1 Introduction

The treatment of ordinal data has been subject of a long-standing controversy, mainly because of the partial agreement on the use of a wider set of permissible statistics for ordinal-scaled variables that are transformed into variables of interval scale type. In Section 1.2, this controversy is examined starting from the association proposed by Stevens (1946) between different scale types and the statistics that are allowed for each one of them (the so-called *permissible statistics*), which includes the ordinal scale type. This is referred to as the *measurement-statistics* association. Thereafter, many other authors have argued for or against it. For example, Lord (1953), Labovitz (1967, 1970, 1971) argued against the measurement-statistics association proposed by Stevens (1946), considering it too restrictive. However, others discredited the analyses with which these authors' arguments against the measurement-statistics association were built, such as Mayer (1970) and Vargo (1971), although it did not necessarily mean that they supported Stevens proposal without special considerations. For instance, Mayer (1970) specified the type of transformations and permissible statistics that could be used for variables of ordinal scale type.

In general, there are two common ways to address the treatment of ordinal variables. These approaches are classified in two cases. The first case is to transform an ordinal variable into one of interval scale type, which is a common practice that

is done to gain access to a wider range of permissible statistics. The second case is to use a transformation that does not take into account the order of categories, then the range of permissible statistics is reduced. In the first case the information provided by the order of categories is overstated, i.e. it is assumed that it allows to compute the distance between categories, but one of the main features of ordinal variables is that these distances are undetermined. In the second case, the information provided by the order of categories is ignored. The implications of these transformations are discussed in Section 1.3.

The methods that transform an ordinal variable into an interval-scaled variable are called scoring systems. These assign values to the ordinal categories according to different criteria, of which a subset is presented in Section 1.4. Some of these systems use information from the ordinal variable itself (see, for instance, Edwards and Kenney (1946), Veenhoven et al. (1993), Bross (1958), Brockett (1981), Van der Waerden (1952), Lehmann and D'abrera (1975), Blom (1958), Tukey (1962), Brockett (1981), Agresti (2010), and Sections 1.4.1 to 1.4.6), whereas others incorporate information from other variables also (see Hensler and Stipak (1979), Martin and Maes (1979), Young (1981), and Sections 1.4.7 to 1.4.9).

Given that these approaches overstate the information provided by the categories of an ordinal variable, there are some methods that treat the ordinal variable as it is, without the need of any previous transformation. This is the case of some regression models that were specially defined for ordinal responses, which is discussed in Section 1.5.

There is not a unique way of accounting for the order of categories of a response variable (see Agresti (2007) and Section 1.5.2). Among all the options of regression models for ordinal responses, one of the most popular ones is the proportional odds cumulative logit model (POCLM, see Section 1.5.3). The POCLM is part of the family of generalised linear models proposed by McCullagh and Nelder (1989) (see Section 1.5.1) and is the regression model to be used throughout this thesis to deal with the modelling of an ordinal response.

Most of the attention in the literature about ordinal variables has been focused on the response variable. However, little has been said about the treatment of or-

dinal predictors (OPs). This is discussed in Section 1.6, where some methods are examined. For instance, the penalised maximum likelihood approach proposed by Tutz and Gertheiss (2014) is studied in Section 1.6.1. This approach penalises the coefficients associated with the categories of an ordinal predictor (OP) to make them closer when they violate monotonicity of effects. Given that it uses penalisation rather than constraints, an analysis of its implications on the parameter estimates associated with an ordinal predictor is included as part of Section 1.6.1, which shows that non-monotonic effects can be penalised but they never get to be monotonic. Hence, getting monotonic effects for the ordinal predictors turned to be one of the characteristics to be considered in the proposed models of this thesis.

An alternative approach to deal with ordinal predictors is isotonic regression, also known as monotonic regression (see, for example, Dykstra et al. (1982), de Leeuw et al. (2009), Stout (2015), and Section 1.6.2). This uses order restrictions on the parameter estimates assuming a particular direction and the number of parameters is equal to the number of observations. An important restriction is that all of the predictors are restricted to be associated with monotonic effects, which is a characteristic that is not required in the proposed models of this thesis. The isotonic regression approach will be discussed in Section 1.6.2.

Another contribution to the subject was made by Rufibach (2010), who proposed to use constrained maximum likelihood to estimate the effects of ordinal predictors with response variables that could be continuous, binary, or represent censored survival times. This method is discussed in Section 1.6.3. With this approach, isotonic effects are achieved, but antitonic effects are not allowed. Therefore, this issue was also considered in the proposed models of this thesis.

It is also possible to assume that an unobserved continuous variable underlies an ordinal variable. This unobserved continuous variable is called latent variable. A latent variable model (LVM) for ordinal variables is presented in Section 1.7 (see Bartholomew et al. (2011)). Moustaki (2000) proposed a class of latent variable models for ordinal manifest (observed) variables, which is presented in Section 1.7.2. This method will be used later as a dimensionality reduction technique in

the regression model context to transform a set of ordinal predictors into a single latent variable representing their information (Section 6.3).

The final section of this chapter, Section 1.9, will outline the structure of the thesis.

1.2 The controversy on permissible statistics for the ordinal scale measurement

One of the broadest definitions of measurement is the numeric assignment to objects or events based on rules (Stevens, 1946). Therefore, different rules may lead to different scales of measurement. According to Stevens (1946), when defining a scale of measurement, it is necessary to specify the rules with which the numeric assignment can be done, identifying the mathematical transformations after which the scale type remains the same and the statistical operations that could be applied to it.

Stevens (1946) proposes four types of scales of measurement. They were defined considering what Stevens called the *basic empirical operations*, which are (1) determination of equality, (2) determination of greater or less, (3) determination of equality of intervals or differences, and (4) determination of equality of ratios. In addition, for each scale type, Stevens assigned a *mathematical group structure* and a set of *permissible statistics*. The mathematical group structure describes the mathematical transformations that make the scale type to remain unchanged, which are referred to as *invariant* transformations. The permissible statistics indicate those statistics that are compatible with an scale type according to its basic empirical operations.

The scale types are listed in Table 1.1 in ascending order according to the field "Basic Empirical Operations." This field associates the operations listed in it with each scale type in a cumulative way, i.e., for a particular scale type of interest, the determination of basic empirical operations includes not only the one(s) associated with it, but also all those corresponding to preceding scale types. Therefore, the wider the range of basic empirical operations can be determined, the *higher* the scale type. For instance, the basic empirical operations of the

ordinal scale type allow to determine equality and greater or smaller, whereas the latter two operations are not possible for the nominal scale type. Therefore, in this sense, the nominal scale type is at a lower level than the one of ordinal-scaled variables. Correspondingly, interval-scaled variables are at a higher level than ordinal-scaled variables.

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
Nominal	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
Interval	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Table 1.1: Stevens' classification of types of scales of measurement.

According to Stevens' proposal, each type of scale allows different sets of statistical operations. The case of the ordinal scale type is a good example to illustrate the meaning of permissible statistics. Ordinal variables provide information on the order of their categories only, not on the magnitude of their distances. This means that statistics like the median and percentiles are suitable for this kind of scale type because they rely on frequencies and rank order only, whereas other statistics, such as the mean and standard deviation, are not suitable for ordinal variables because they do rely on knowledge that is absent in the ordinal scale type. However, this measurement-statistics association is a source of controversy as will be discussed later on in this section. Statistics associated with a particular scale type are considered as "permissible" by Stevens when they fulfil the criterion of *invariance* under transformations in the corresponding mathematical

group, where each mathematical group describes a set of functions associated with the scale type, decreasing the range of possible transformations for higher scale types. Note that Stevens uses the term *invariant* in a wider sense compared to the same concept defined in some other textbooks, such as in Young and Smith (2005), where "...the statistic t is invariant to the action of the group G if *its value* does not depend on whether x or $g(x)$ was observed, for any $g \in G : t(x) = t(g(x))$." However, Stevens associates the term *invariant* to the relative location of every item instead. For instance, elements at (and lower to) the median of an ordinal variable x keep their relative location after any transformation belonging to the isotonic group defined as $x' = f(x)$ with $f(x)$ being any monotonic transformation; a data point at the mean of an interval-scaled variable will be located at the mean after any transformation belonging to the general linear group and not to the groups of transformations of lower level scale types such as the isotonic and permutation groups. Once again, these statistics are shown in Table 1.1 in a cumulative way, meaning that each scale type allows its corresponding set of statistics together with all those of the lower ones. Conversely, each additional set of statistics considered as "permissible" for each higher scale type should not be used for variables classified in any lower scale type.

Based on this classification, the additional set of statistics that is made "permissible" when increasing the type of the scale from ordinal to interval or ratio should not be used when analysing ordinal variables (or nominal). This comes from the fact that the basic empirical operations for ordinal variables are the *determination of equality, and greater or less*, which implies that the distance between any pair of ordinal categories is not determined. When dealing with ordinal variables, the use of any of the permissible statistics in the additional set of statistics for interval and ratio-scaled variables necessarily forces the researcher to assume that something that is not originally present in the ordinal variable is being exploited, i.e., it takes features from the ordinal scale type as information despite the fact that it is not, leading to statistical results that are an artefact. However, using these types of statistics is a common practice in many applications, where, for instance, the labels of the ordinal categories are used as numeric values

(for example, see Pasta (2009), Frederick and Jenkins (2015), Santina and Perez (2003), and Hren et al. (2004)). By doing so, it would imply that the level of information contained in ordinal variables is overstated because more knowledge than the relative rank-order is needed to perform these statistics.

Restricting the number of permissible statistical methods in presence of ordinal variables is a practice that has been criticised by many authors. In fact, one of the first opponents to this restriction went even further by supporting the idea of using statistics for interval data on nominal data. Lord (1953) used an example where nominal data represented by the numbers that football players had to wear on their uniform were treated as an interval-scaled variable, making his statement "Since the numbers don't remember where they came from, they always behave just the same way, regardless." one of the important arguments to be used by subsequent antagonists to the measurement-statistics association.

Another important opponent to the restrictive use of statistical methods on ordinal-scaled variables is Labovitz. In Labovitz (1967), the author analysed the effect on statistical results produced by transforming ordinal data into interval data. Data from hypothetical subjective responses to two types of therapy were used (see Table 1.2).

	Subjective Response to Therapy				
	-	0	+	++	Total
Therapy A	12	18	18	9	57
Therapy B	18	12	12	3	45
Total	30	30	30	12	102

Table 1.2: Labovitz (1967) example: Hypothetical subjective responses to two types of therapy.

Seven different ways to transform the ordinal responses to therapy ('−', '0', '+', and '++') into numbers were used to treat the ordinal data as interval data by arbitrarily assigning numbers to the ordinal categories under the restriction of being consistent with the original rank order. All of them transformed '-' into 0, '++' into 10, and the remaining categories '0' and '+' in different arbitrary ways such as 1 and 2 correspondingly, or 1 and 9, 6 and 8, among others. Several

statistical results coming from these different transformations were compared, such as different measurements of correlation and some statistical tests. For example, one of them was the Point-biserial r_{pb} correlation, which is defined as

$$r_{pb} = \frac{M_A - M_B}{s_n}\sqrt{\frac{n_A n_B}{n^2}},$$

where M_A and M_B are the means of the scoring system for therapies A and B correspondingly, n_A and n_B are their sample sizes, $n = n_A + n_B$, and s_n is the overall standard deviation. The Point-biserial r_{pb} correlation between the two therapies ranged from 0.161 to 0.221 depending on the transformation that was used, which was considered as a slight difference and used as evidence to support the use of scoring systems to gain access to a wider range of statistical methods.

In Labovitz (1967), the impact of using an arbitrary monotonic assignment of numbers was assessed. If one of the seven monotonic transformations is assumed to be the 'true' one, then the correlation between each transformed variables and the 'true' monotonic transformation can be understood as the degree in which each transformation method is correct. In addition, significance tests for the difference between means were performed. Labovitz showed these results to conclude that no matter the monotonic transformation the researcher choose, it will always be highly correlated to the unknown but true one. Therefore, statistical analyses should not be affected too much by any arbitrary monotonic transformation, making it feasible and advisable in order to reach a wider range of "permissible" statistics.

One of the main Labovitz's conclusions is that an arbitrary monotonic number assignment to ordinal categories does not produce an important degree of alteration in statistical analyses results. This supports the use of parametric statistics after transforming ordinal data to interval data even though strict adherence to Stevens' scale types of measurement is not met. Furthermore, it is argued that being inflexible to scale type transformations may lead to an important waste of information based on a highly restricted choice of statistical methods and tests. However, he made such general conclusions based on just one applied example, making his methodology not particularly convincing.

Few years later, in 1970, Labovitz published the article "The Assignment of Numbers to Rank Order Categories", in which he reinforced the idea of that there is little difference in the statistical analyses results after transforming ordinal variables into interval ones by doing arbitrary but monotonic assignments of numbers to rank order categories (Labovitz, 1970). The relation between the ordinal-scaled variable *occupational prestige* and the ratio-scaled variable *suicide rates* was analysed to assess to what extent statistical results change when treating the ordinal variable as interval in several arbitrary ways. The number of occupations was 36 and their prestige scores indicate the rank of each occupation relative to the others, ranging from 7 to 97. This score was considered as one of the possible transformations, another one was the assignment of consecutive integers (1-36) and 18 extra monotonic transformations with (constrained) random values ranging from 1 to 10,000 were generated completing a total of 20 options for the ordinal to interval arbitrary transformation. The analysis was based on the same methodology as in Labovitz (1967) explained before. Nevertheless, Mayer (1970) criticised Labovitz's analysis by arguing that "all he has shown is that the Pearson r is fairly stable with respect to non-linear monotone transformations on the numbers assigned to ranks." In addition, the sample correlation coefficient is invariant under changes in scale and location, making unclear how this statistic could serve as a measure to assess the positive and negative effects of ordinal to interval scale transformation. Mayer agreed with the fact that the way in which ordinal data is transformed into interval data is not important as long as it is monotonic and the statistics to be used are invariant under changes in scale and location. However, when they are not invariant, arbitrary assignment of numbers to ranks is not adequate for some statistic analyses, such as multiple regression and discriminant analysis, where the results are clearly affected by variable transformations. Moreover, Vargo (1971) discredited Labovitz's methodology by stating that he did not demonstrate the benefits of ordinal to interval transformation but proved a strong association between monotonicity and correlation, which is expected because the latter results from the construct of the analysis. The high correlations between scoring systems in Labovitz's example respond to the combination of (i) monotonicity constraints,

(ii) their large range differences, and (iii) the large number of ordinal categories. Therefore, the numbers generation process was forced to be a quasilinear transformation method. In the paper "In Defense of Assigning Numbers to Ranks" (Labovitz, 1971), the discussion continued by emphasising the virtues of gaining access to a wider range of permissible statistical methods and by stating that Mayer and Vargo's counterarguments were based either on misinterpretations or extreme counterexamples.

A different approach against Stevens' proposal is that of stressing the distinction between measurement theory and statistical theory, questioning the association of permissible statistics to scale types. Gaito (1980) discusses the importance of measurement scales aspects for the use of some statistical methods, emphasising that the assumptions of a particular statistical method follow from the corresponding mathematical model and not from the measurement scales aspects. In addition, he argues that the psychological meaning of a variable is not a matter of statistics, therefore, the link between measurement scales and statistical procedures should not be regarded as crucial. However, the measurement-statistics relationship has been used in textbooks and other literature, which, in Gaito's words, has led to a *misconception*. The author concludes totally against this association by stating that "Statistical procedures do not require specific scale properties."

Velleman and Wilkinson (1993) discussed the association of measurement scales to statistics from a conceptual point of view by making an extensive literature review. They claimed that it is problematic to apply Stevens' rules when determining the type of statistics that are suitable for the variables at hand, because they do not always fall exactly into the scale types defined in Stevens (1946). Depending on the context, a single variable can be treated as ordinal or interval. However, classifying a variable into only one of the four scale types simplifies the concept of measurement level so far as to be insufficient. In addition, the authors are in favour of transforming data because the proper use of statistics depends on model assumptions and not on variable scale features, and it has proven to be irrelevant to statistical inference.

This is just part of a long-standing unsettled controversy.

1.3 Changing the scale of measurement: some implications

When facing the task of analysing ordinal variables, there is a certain tendency to either decrease or increase the level of measurement of these types of variables by transforming them into nominal or interval-scaled variables correspondingly. However, the proper use of the information provided by ordinality has called little attention in some areas of statistics such as in regression analysis.

On the one hand, using statistical methods for nominal data when dealing with ordinal variables leads to waste of information. This becomes palpable by the fact that results are invariant to permutation of the categories, which means that the order of the categories is not being exploited as it should. For example, using the Pearson $\mathcal{X}^2$ test of independence provides results that may be quite different from those obtained using methods for ordinal variables (see Agresti (2010)).

On the other hand, using statistical methods for interval data when dealing with ordinal variables is not advisable in some cases. This kind of applications of statistical methods demands an ordinal to interval transformation of data. These transformations are referred to as *scoring systems*, and defined as systematic methods for assigning numerical values to ordinal categories (Golden and Brockett, 1987). The ambiguity in the determination of the scores leads to the existence of several scoring systems, which implies that it is necessary to make a choice that is not straightforward for cautious analysts, because different scoring systems could lead to different results (Casacci and Pareto, 2015). Despite the fact that scoring systems are permissible by Stevens, when they are used in the data pre-processing step to transform the scale type reaching a higher one (ordinal to interval transformation), subsequent statistical analyses rely on information that is an artefact of the selected method. Then, it should be of interest to analyse to what extent the chosen scoring system could affect final results, which in practical applications is a problem with unclear implications because it could actually produce a significant effect or not. However, there are some problems with clear implications when using scoring systems on a response variable. For example, when using Or-

dinary Least Squares (OLS) regression on an interval response variable resulting from the transformation of an ordinal response into a sequence of integers going from 1 to its number of categories, then, at a given set of values for the predictors, the prediction will not be a set of estimated probabilities for the original response categories, and probably it will not even be one of the scores. In addition, other problems in the same context are associated with the so called "ceiling and floor effects", from which one of the most evident is that predicted values from the OLS regression may fall out of the range of possible category scores (above or below), which would imply the use of an additional method if the researcher wants to adjust their range. Despite these limitations, OLS can be useful in this context depending on the purpose of the analysis, such as the case of identifying statistically significant effects, or making simple descriptions (see Agresti (2010)). Therefore, using a scoring system as a way of getting access to statistical methods that are suitable for interval data is not an indisputable approach when dealing with ordinal variables.

1.4 Scoring systems

There are several scoring systems for transforming scale types from ordinal to interval. In this section some of them will be presented, for which common notation is used when it is possible. Some of these scoring systems will be used later in sections 6.3 and 6.4.2 in order to compare their results against the ones of the methods that will be proposed in the following chapters.

Consider an ordinal response variable Y with k categories, a sample size of n with frequencies $n_1, n_2, ..., n_k$ in the categories of Y. The total number of observations can be computed as $n = \sum_{j=1}^{k} n_j$, and the sample proportions are denoted as $\{p_j = n_j/n\}$. Assuming that there is at least one observation for each one of the k categories, then $p_j > 0$ for all j.

The probability of response in category j is denoted as π_j, and the cumulative probabilities are

$$F_j = P(Y \leq j) = \pi_1 + \cdots + \pi_j, \quad j = 1, 2, ..., k.$$

As Y is an ordinal variable and assuming $\pi_j > 0$, then $0 < F_1 < F_2 < \cdots < F_k = 1$,

or

$$F_{j-1} < F_j, \quad j = 1, 2, ..., k,$$

with $F_0 = 0$ and $F_k = 1$. The corresponding cumulative sample proportions are denoted by $\{\hat{F}_j\}$, where

$$\hat{F}_j = p_1 + \cdots + p_j, \quad j = 1, 2, ..., k. \tag{1.4.1}$$

The general purpose of a scoring system is to assign numerical values to ordinal categories, producing scores $v_1 < v_2 < \cdots < v_k$.

1.4.1 Linear score

The most obvious and simplest transformation is to use consecutive natural integers:

$$v_1 = 1, v_2 = 2, \ldots, v_k = k, \tag{1.4.2}$$

as the ones used in Likert scales (see for example, Edwards and Kenney (1946)).

The linear score has also been modified in different simple ways. One of them is the one presented in Kalmijn (2013), where the linear scale is transformed into a 0-10 scale to meet what Veenhoven et al. (1993) proposed in order to get one common scale as a measure of happiness. Therefore, a k-points scale going from 1 to k as in (1.4.2) is transformed according to:

$$v_j = 10\frac{j-1}{k-1}, \quad j = 1, 2, ..., k, \tag{1.4.3}$$

which is here referred to as the "Veenhoven" scoring system.

1.4.2 Ridit score

Another simple scoring system that allows to introduce the main one of this section is based on the distribution of the observed Y. It uses the cumulative sample proportions defined in Equation (1.4.1), with which the scoring system is:

$$v_1 = \hat{F}_1, v_2 = \hat{F}_2, \ldots, v_k = \hat{F}_k. \tag{1.4.4}$$

A usual modification of (1.4.4) introduced by Bross (1958) is the average cumulative proportion, also known as ridit. According to Bross (1958), the term

"ridit" was chosen by analogy with "probit" and "logit", and its first three letters stand for Relative to an Identified Distribution. The ridit for category j is the proportion of individuals or observations in categories 1 to $j-1$ plus one half of the proportion in category j, more formally,

$$ridit_j = \sum_{c=1}^{j-1} p_c + \frac{1}{2}p_j, \quad j = 1, 2, ..., k. \tag{1.4.5}$$

In terms of the sample cumulative proportions

$$ridit_j = \frac{\hat{F}_{j-1} + \hat{F}_j}{2},$$

with $\hat{F}_j = p_1 + \cdots + p_j$ and $\hat{F}_0 = 0$.

By definition, $ridit_j$ for $j = 1, \ldots, k$ results with the same order as the ordinal categories and their weighted average with respect to the sample distribution is

$$\begin{aligned}\sum_{j=1}^{k} p_j ridit_j &= \sum_{j=1}^{k} p_j \left(\sum_{c=1}^{j-1} p_c + \frac{1}{2}p_j \right) = \sum_{j=1}^{k} p_j \sum_{c=1}^{j-1} p_c + \frac{1}{2}\sum_{j=1}^{k} p_j^2 \\ &= \frac{2\sum\sum_{c<j} p_j p_c + \sum_{j=1}^{k} p_j^2}{2} = \frac{(\sum_{j=1}^{k} p_j)^2}{2} \\ &= 0.5,\end{aligned}$$

which guarantees that their weighted average with respect to the sample distribution will always be the same, 0.5, meaning that the transformation affects the dispersion of the resulting ridits in order to assign numbers to the original categories.

1.4.3 Midranks

Midranks are averages of the ranks that would be assigned to the observations in a category if they could be overall ranked without ties. The midrank for category j is the average between the number of individuals or observations in categories 1 to $j-1$ increased in one unit and the number of those falling in categories 1 to j, more formally,

$$midrank_j = \frac{[(\sum_{c=1}^{j-1} n_c) + 1] + \sum_{c=1}^{j} n_c}{2}, \quad j = 1, 2, ..., k.$$

The addition of 1 in the numerator ensures that any $midrank_j$ ranges between 1 and n. Whereas midrank scores fall between 1 and n, ridit scores fall between 0 and 1. Their close relation from the conceptual point of view is also true in mathematical terms,

$$\begin{aligned} midrank_j &= \frac{[(\sum_{c=1}^{j-1} n_c) + 1] + \sum_{c=1}^{j} n_c}{2} = \frac{n\sum_{c=1}^{j-1} n_c/n + n\sum_{c=1}^{j} n_c/n}{2} + 0.5 \\ &= n\left(\frac{2\sum_{c=1}^{j-1} n_c/n + n_j/n}{2}\right) + 0.5 \\ &= n \times ridit_j + 0.5 \end{aligned}$$

and, therefore

$$ridit_j = \frac{midrank_j - 0.5}{n} \tag{1.4.6}$$

which shows a linear relationship between the two scoring systems (Agresti, 2010), making ridits and midranks equivalent for most tasks.

1.4.4 Normal scores and conditional median scoring

Consider an unobserved continuous latent variable that follows the standard normal distribution with cumulative distribution function denoted by $\Phi(\cdot)$, and assume this variable to underlie the ordinal variable Y. Then, normal scores based on ridits are

$$\upsilon_j = \Phi^{-1}(ridit_j), \tag{1.4.7}$$

where $ridit_j$ is the ridit score for category j defined in equation (1.4.5). This is also known as the Conditional Median Scoring. It can be thought of as an approximation of the median of the standard normal distribution for the range delimited by $\Phi^{-1}(p_{j-1})$ and $\Phi^{-1}(p_j)$, with $\Phi^{-1}(p_0) = -\infty$ and $\Phi^{-1}(p_k) = \infty$.

Originally, Brockett (1981) proposed this scoring system in a more general fashion, where the researcher could assume any particular cumulative distribution G of an unobserved continuous latent variable assumed to underlie Y. The author showed that the score υ_j is equal to the conditional median of $G^{-1}(x)$ given category j under the assumed cumulative distribution G.

For example, consider the continuous uniform distribution $U(a, b)$ with cumulative density function (CDF)

$$G(x) = \begin{cases} \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise,} \end{cases} \tag{1.4.8}$$

and its inverse

$$G^{-1}(x) = a + (b-a)x. \tag{1.4.9}$$

If (1.4.9) is used in Equation (1.4.7) instead of $\Phi^{-1}(x)$ with $x = ridit_j$, then

$$\begin{aligned} \upsilon_j &= G^{-1}(ridit_j) \\ &= a + (b-a)ridit_j. \end{aligned} \tag{1.4.10}$$

Therefore, for $a = 0$ and $b = 1$, $\upsilon_j = ridit_j$, i.e., ridits can be considered as a conditional median scoring using a continuous uniform distribution $U(0, 1)$.

1.4.5 Rank-based normalisation procedures

Several scoring systems are based on different ways of determining the probabilities where $\Phi^{-1}(\cdot)$ is assessed (Agresti, 2010). For instance, normal scores based on midranks $\{r_j\}$ (defined in Section 1.4.3) are

$$\upsilon_j = \Phi^{-1}\left(\frac{r_j}{n+1}\right), \tag{1.4.11}$$

which are a rank-based normalisation using the Van der Waerden's formula (see Van der Waerden (1952) and Lehmann and D'abrera (1975)).

Another scoring system that approximates the percentiles of the standard normal distribution is

$$\upsilon_j = \Phi^{-1}\left(\frac{r_j - 0.375}{n + 0.25}\right), \tag{1.4.12}$$

which is known as the Blom scoring system (see Blom (1958) and Harter (1961)).

Equation (1.4.7) is also known as the rankit scoring system (see Ipsen and Jerne (1944)), usually expressed as

$$\upsilon_j = \Phi^{-1}\left(\frac{r_j - 0.5}{n}\right). \tag{1.4.13}$$

Another similar version of rank-based normalisation procedures was proposed by Tukey (1962), which was declared as a "simple and surely an adequate approximation to what is claimed to be optimum,":

$$v_j = \Phi^{-1}\left(\frac{r_j - 1/3}{n + 1/3}\right). \tag{1.4.14}$$

All these scoring systems use the inverse of the cumulative standard normal distribution. However, they could be based on the inverse of another cumulative distribution function.

1.4.6 Conditional mean based on density

Hensler and Stipak (1979) proposed the use of some scoring systems to estimate interval scale values for survey item response categories. They presented two types of systems, one based on observed frequencies and distributional assumptions, and another one based on estimations from *criterion variables*. Among the scoring systems of the first type, two were discussed: the ridit score, which was already presented in Section 1.4.2, and a second one that will be explained in this section. The scoring system based on criterion variables will be summarised in the next section.

The second scoring system based on observed frequencies and distributional assumptions uses the observed category proportions as estimates of areas under the probability density function of an underlying variable associated with the observed ordinal variable, which is assumed to be distributed under the standard normal distribution.

Each cumulative proportion $\hat{F}_j$ determines two boundaries assuming the inverse of the cumulative standard normal distribution $\Phi^{-1}(\cdot)$, being $\Phi^{-1}(\hat{F}_j)$ and $\Phi^{-1}(\hat{F}_{j-1})$, with $\hat{F}_0 = 0$ and $\hat{F}_k = 1$, and therefore $\Phi^{-1}(\hat{F}_0) = -\infty$ and $\Phi^{-1}(\hat{F}_k) = \infty$ correspondingly. The idea is to compute an average for each of these paired values. However, as it is not possible to compute the average for the first and last pair, e.g., between $-\infty$ and -1.96 for a category with a proportion of 5%, the method uses the ordinates of the probability density function of the standard

normal distribution, $\phi(\cdot)$. Then, the scores are computed as follows:

$$\upsilon_j = \frac{\phi(\Phi^{-1}(\hat{F}_{j-1})) - \phi(\Phi^{-1}(\hat{F}_j))}{p_j}. \tag{1.4.15}$$

This method assigns the mean value for the segment of the density function corresponding to category j as its category value. This is why this method is known as the conditional mean scoring system.

Brockett (1981) showed that the conditional mean scoring system can be generalised to assume any distribution function using

$$\upsilon_j = \frac{1}{p_j} \int_{G^{-1}(\hat{F}_{j-1})}^{G^{-1}(\hat{F}_j)} x g(x) dx \tag{1.4.16}$$

$$= \frac{1}{p_j} \int_{\hat{F}_{j-1}}^{\hat{F}_j} G^{-1}(u) du, \tag{1.4.17}$$

where G is an assumed particular cumulative distribution of an unobserved continuous latent variable that is assumed to underlie the ordinal variable Y, $G^{-1}(\cdot)$ its inverse, and $g(x)$ its corresponding density function.

Both (1.4.16) and (1.4.17) can be thought of as the conditional mean of $G^{-1}(\cdot)$ for a given category j under the assumed cumulative distribution G.

1.4.7 Estimation from criterion variables

Hensler and Stipak (1979) also proposed the use of other scoring systems under the assumption that some observed variable(s) provide information about the values to be assigned to the categories of an ordinal variable. For example, the observed values of the variable "temperature in celsius degrees" could give some information about the values to be assigned to the ordinal variable called "individual thermal sensation" with categories "low", "medium", and "high". In this context, the variable that is used to assign values to the ordinal categories is called the "criterion variable" (temperature) and the ordinal variable is called the "target variable" (thermal sensation). More than one criterion variable is also possible. The estimation from criterion variables uses the relationship between the criterion and target variables.

In the single criterion variable case, the categories of the target variable simply take the conditional mean value of the criterion variable as their scale given an

ordinal category. This is the result of fitting a linear regression model with the criterion variable as the dependent variable and the target variable as the independent one in the form of a set of dummy variables as usual. Another example of this is to treat a "feeling thermometer" used to assess the performance of a president ranging from 0 to 100 as interval-scaled and use it as the criterion variable to define the interval scale of a categorical target variable representing different levels of assessment: "very good job", "good job", "fair job", or "poor job."

When the single criterion (dependent variable) is determined by a target (independent) variable and other independent variables, it is advisable to use multiple regression (or some other method) to estimate the effects of several independent variables upon the criterion variable, including the target variable represented by a set of dummy variables. Thus, the dummy variables' regression coefficients are treated as category value estimates relative to the zero point defined by the omitted category.

In the multiple criterion variable case, a criterion index is calculated based on the weighted sum of the criterion variables. There are two options to compute the weights: (i) they are arbitrarily determined (e.g., equal weights), or (ii) they are estimated using canonical correlation analysis (CCA) to maximize the strength of the relationship of the criterion index with the target variable. Canonical correlation analysis can be seen as a generalisation of multiple regression analysis in the sense that it uses more than one independent variable, i.e., CCA aims to find a linear association between two sets of variables (Martin and Maes, 1979).

Sometimes the criterion variables are conceptually homogeneous and highly correlated, or the researcher decides to equally weight them, then a weighting procedure is not absolutely necessary. However, in general, the researcher might be unsure about the conceptual homogeneity of the criterion variables, and therefore a weighting technique is necessary. Then, CCA is used to simultaneously estimate weights for the criterion variables and category values for the target variable.

In CCA, two sets of variables are used as input data. In this case, the criterion variables form one set of input variables, and the other set is formed by $k - 1$ dummies representing the target variable of k categories. CCA estimates a

canonical variate for each of the two sets. In this application, those are the *target canonical variate* and the *criterion canonical variate*, each of which is a linear weighted sum of their corresponding set of variables. These weights are chosen to maximise the correlation between the canonical variates. Therefore, CCA finds the set of weights for the category dummies that maximise the correlation of the target canonical variate with the criterion canonical variate, where the weights for the category dummies are the category values.

1.4.8 An overview of optimal scaling

Optimal scaling (OS) is another technique for scaling ordinal (or nominal) variables from information provided by interval-scaled variables. This technique aims to represent each observation of a categorical variable, either nominal or ordinal, by a parameter. The measurement scale of the variable to be transformed implies the use of constraints on the estimation of each parameter, e.g., for ordinal categorical variables, order constraints should be imposed (Young, 1981).

The following overview of the optimal scaling procedure is based on Jacoby (2015) mainly.

OS is a procedure for obtaining $\mathbf{x}^*$ from $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}$ is a qualitative vector of observations $x_1, x_2, \ldots, x_n$ to be transformed, $\mathbf{y}$ is a quantitative vector with elements $y_1, y_2, \ldots, y_n$, and $\mathbf{x}^*$ is the vector of optimally scaled values of $\mathbf{x}$, to be estimated considering the observed correspondence between $\mathbf{x}$ and $\mathbf{y}$.

The vector $\mathbf{x}^*$ is defined to be maximally correlated with the entries in $\mathbf{y}$, while taking into consideration the *measurement characteristics* that are assumed for $\mathbf{x}$. The term *measurement characteristics* is composed by three key concepts: (i) measurement level, (ii) measurement process, and (iii) measurement conditionality, which are described as follows:

measurement *level*: is associated with the level of $\mathbf{x}$. As OS works for categorical $\mathbf{x}$, this vector can be either nominal or ordinal.

measurement *process*: is associated with $\mathbf{x}^*$. It indicates how the observations within a given category of $\mathbf{x}$ must be assigned to $\mathbf{x}^*$. These:

- must be assigned to the same optimally scaled value in $\mathbf{x}^*$ if the process is **discrete**, or
- can be assigned to different values within a closed interval, if the process is **continuous**.

measurement *conditionality*: separates a data set into partitions. Within each partition it is possible to make meaningful comparisons among observations and also among scores.

Let's assume that the measurement *conditionality* is considered "fixed" because the vector $\mathbf{x}$ comprises a single partition. Therefore, there are only four combinations of measurement characteristics between measurement *level* and measurement *process*.

The nominal level - discrete process case.

The only measurement restriction on the entries in the vector of optimally scaled values $\mathbf{x}^*$ is:

$$x_i = x_j \implies x_i^* = x_j^*. \tag{1.4.18}$$

Here, the OS procedure assigns the conditional means of the y_i's given each observational category of $\mathbf{x}$ to the entries in $\mathbf{x}^*$.

The ordinal level - discrete process case.

The measurement restrictions on the entries in the vector of optimally scaled values $\mathbf{x}^*$ are:

$$x_i = x_j \implies x_i^* = x_j^* \tag{1.4.19}$$

$$x_i \prec x_j \implies x_i^* \leq x_j^*. \tag{1.4.20}$$

The procedure computes the conditional means as in the previous case and then applies a monotonic transformation to the conditional means by using a numerical method proposed by Kruskal with its secondary approach for the treatment of tied values (see Kruskal (1964)).

The nominal level - continuous process case.

The measurement restrictions on the entries in the vector of optimally scaled values $\mathbf{x}^*$ are:

$$x_i = x_j \implies x^*_{max,lower} \leq x^*_i, x^*_j \leq x^*_{min,higher}. \tag{1.4.21}$$

Where $x^*_{max,lower}$ is the largest entry in $\mathbf{x}^*$ that is assigned to any $x_k \neq x_i, x_j$, but that is still smaller than either x^*_i or x^*_j. Conversely, $x^*_{min,higher}$ is the smallest entry in $\mathbf{x}^*$ that is assigned to any $x_k \neq x_i, x_j$, but that is still greater than either x^*_i or x^*_j. In other words, categories from $\mathbf{x}$ now correspond to intervals of real numbers in $\mathbf{x}^*$.

One option to perform OS in this case, the nominal-continuous, is the two-step "pseudo-ordinal" procedure (see De Leeuw et al. (1976)):

Step 1 The data are treated as nominal-discrete to obtain x^*_i, with which an ordering of the categories in $\mathbf{x}$ is established.

Step 2 The categories are treated as ordinal-continuous (see next section) to obtain the interval of optimally scaled values for each of the x_i's.

The next section describes the the general treatment for the ordinal-continuous case.

The ordinal level - continuous process case.

Like in the ordinal-discrete case, the OS transformation for ordinal-continuous data is performed by carrying out Kruskal's monotonic transformation, but now with the primary approach for the treatment of tied values (instead of the second one), which requires that if $y_i \leq y'_i$ then $x^*_i \leq x^{*'}_i$ for the tie $x_i = x'_i$. The transformation is applied to the individual y_i's, rather than to the conditional means.

According to Young (1981), OS can be seen as a numerical assignment to observation categories based on the maximisation of the relation between the observations and the statistical model taking into account the measurement levels of the data. In fact, it is common to see optimal scaling applications using $\mathbf{y}$

as a vector of predicted values from a statistical model, for example, as part of the alternating least square (ALS) algorithm (see, for example, Mair and Leeuw (2008) for some applications).

Details about these optimal scaling procedures, monotonic transformations and treatments of tied values can be found in, for example, Kruskal (1964), De Leeuw et al. (1976), and Jacoby (2015). The OS techniques are presented in general terms only in this section to include them as part of the methods used to transform categorical variables into interval scaled-variables. In addition, OS techniques are also used as part of algorithms such as "alternating least square with optimal scaling" (see next Section 1.4.9), which extends the use of OS to cases where more than one categorical variable is being transformed using a multiple regression model (see the example of Section 1.4.9 using Multiple Optimal Regression via Alternating Least Squares (MORALS)).

1.4.9 Alternating least square with optimal scaling (ALSOS)

The alternating least square algorithm works with two mutually exclusive and exhaustive subsets of parameters: (i) the parameters of the model; and (ii) the parameters of the data (or optimal scaling parameters). The general procedure of the ALS algorithm is to alternate between two steps (Young, 1981):

Step 1 Obtain the least square estimates of the parameters in one subset while assuming that the parameters in all other subsets are constant. This is called the *conditional least squares estimate*.

Step 2 Replace the old estimates of these parameters by the new estimates and then switch to another subset of parameters to apply step 1 again.

When the first step is performed on a subset of model parameters, it is referred to as the *model estimation phase*, and when it is performed on a subset of parameters associated with the data, it is called the *optimal scaling phase*.

ALS also allows to perform multivariate analysis by noting that each subset could be formed of several mutually exclusive and exhaustive subsets, where each

variable is associated with one of them.

A more detailed description of the algorithm (based on Jacoby (2015)) uses the following steps:

1. The variables are assigned initial optimal scale values, and the measurement characteristics are set.

2. Least-squares estimates are obtained for the parameters of the statistical model (*model estimation phase*).

3. If model fit has not improved over the previous iteration, terminate the procedure; otherwise proceed with the following steps.

4. The predicted values from the statistical model are used to generate new optimal scale values for the variables (*optimal scaling phase*).

5. Return to Step 2 and re-estimate the model using the updated optimally-scaled variable values.

One of the main advantages of this algorithm is that the OS phase does not depend on the type of statistical model used for the model estimation phase. Therefore, given a specific data analysis situation, it is possible to use any suitable statistical model in step 2.

An application of ALSOS is the Multiple Optimal Regression via Alternating Least Squares (MORALS) algorithm, which can be used to fit a multiple regression model to variables measured at a variety of levels (Young et al., 1976). There are also many other applications such as ALSOS for nonlinear canonical analysis and for nonlinear principal component analysis (see De Leeuw et al. (2009) for an R package to handle these models).

An example of MORALS applied on politics.

An example in Jacoby (2015) has been replicated to show how the ALS algorithm can be applied to multiple regression analysis with ordinal dependent and ordinal independent variables. The example uses data from the Center for Political Studies' 1992 National Election Study (NES), the number of observations is 1,653.

Consider the dependent variable "Choice" (a measure of relative candidate preference ranging from −100 to +100) and the following three independent variables: "Party identification" (from 0 Strong Democrat, through 3 Independent, to 6 Strong Republican), "Ideological self-placement" (from 1 Extremely Liberal, through 4 Moderate, to 7 Extremely Conservative), and "Did nation's economy become better or worse over past four years?" (from 1 Much Better, trough 3 Stayed the same, to 5 Much Worse).

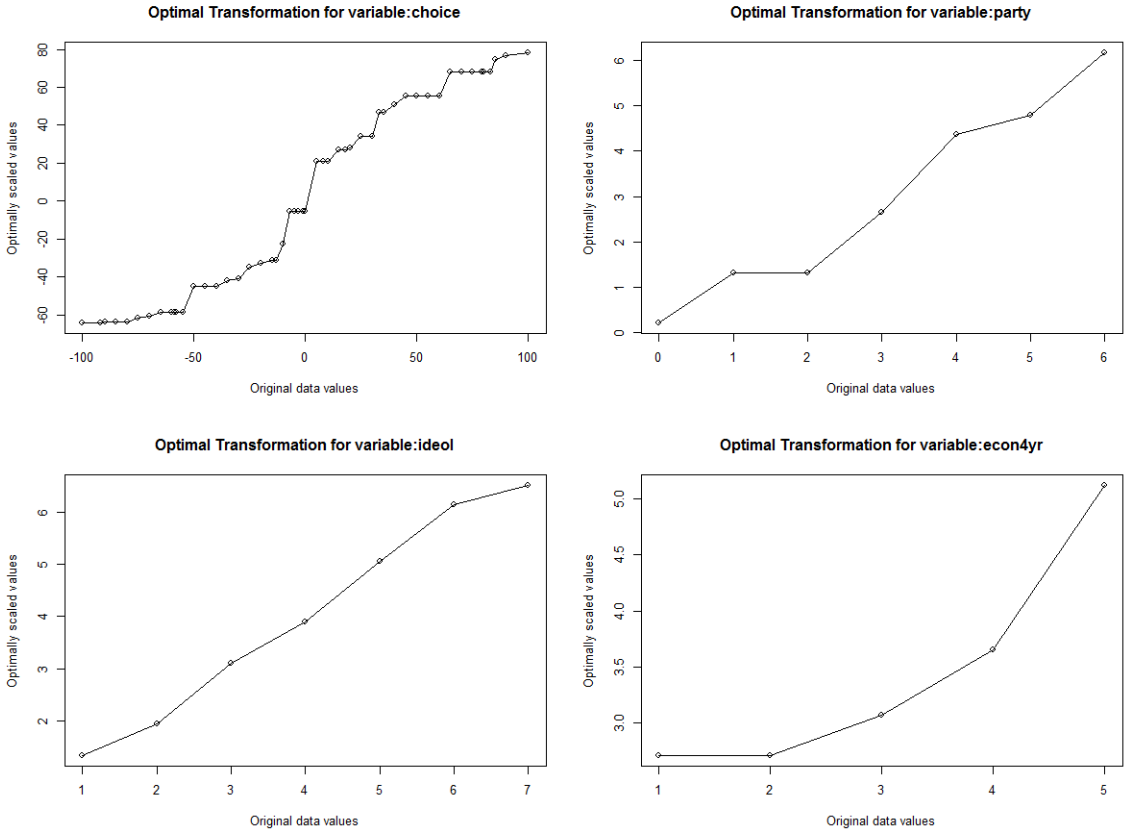

Figure 1.1: Example of Optimal Scores resulting from MORALS.

An ALSOS analysis is used in this illustration as a diagnostic test to determine whether it is reasonable to treat these variables as of interval scale type or not, which indeed are routinely treated like this arguing "practical purposes." Therefore, all of the variables are considered as measured at the ordinal scale type. Then, the MORALS algorithm is used to estimate the regression parameters and optimal scores for the dependent and independent ordinal variables. If the resulting optimal scores for a variable are linearly related to the original scores (see

Figure 1.1), then nothing is lost if they are treated as interval-level measurement in this statistical analysis.

The upper-left plot shows the relationship between the original values of the dependent variable and the ones resulting from its optimally scaled transformation. Their relationship is clearly non-linear, because the differences in the transformed values associated with extreme original values are much less strong than the ones associated with the original values at the middle of their range.

Party identification (see the upper-right plot) shows a non-linear relationship between its original coding and its optimally scaled values. Original values 1 and 2 are assigned to the same optimally scaled value (1.31), meaning that those who classify themselves at intermediate categories between '0 Strong Democrat' and '3 Independent' are quite similar to each other in terms of their answers to the other questions (variables in the model). Something similar happens with original values 4 and 5, where those between '3 Independent' and '6 Strong Republican' are relatively similar to each other as well.

Economic assessments (lower-right) seems to produce big differences in the second half of the scale but not much difference in the first. This means that people assessing the economy as '5 Much Worse' considerably differs from other in the second half. In general, differences between adjacent categories decrease while decreasing in the original coding of the variable (better assessment of economy). In fact, those assessing the economy as '1 Much Better' do not produce any difference in its optimal scale value compared to '2 Better'.

Ideological self-placement is a special case. It is the only variable that shows a nearly linear relationship between the optimally scaled scores and its original coding. Meaning that the differences between the optimally scales values corresponding to adjacent original categories are almost the same regardless their position in the original scale.

Jacoby (2015) concluded this analysis with the statement "these results suggest that the usual practice of treating feeling thermometers, party identification, and judgements about the American economy as interval level variables may, in fact, be problematic." A missing analysis in Jacoby (2015) is related to testing linearity.

In some cases it is not that clear whether linearity is not met by the association between the original scores and the resulting optimally scaled ones. For example, for party identification (see the upper-right plot of Figure 1.1), and despite the fact that the sample size is relatively large (n =1653), it could be the case that its deviation from linearity is not significant at certain significant level, and therefore linearity would not be rejected.

1.5 Some regression models for ordinal responses

Regression models for ordinal response variables can be considered as extensions of one of the most popular regressions models for binary responses, the logistic regression. These extensions account for $k \geq 2$ ordinal response categories rather than only $k = 2$. The general structure of the link function remains the same, the logit link function. However, there are several ways of defining the logits depending on how the category order is taken when determining probabilities. As logistic regression is a particular model within a broader class of models, the Generalized Linear Models (GLM), its general framework is the starting point in this section.

1.5.1 Generalised linear models

Consider a vector of observations $\mathbf{y}$ having n components and a matrix $\mathbf{X}$ of dimensions $n \times p$ representing the observed values of p explanatory variables for the n observations. The vector $\mathbf{y}$ is assumed to be a realization of a random response variable $\mathbf{Y}$ whose components are independently distributed with means $\boldsymbol{\mu}$.

Under this setting, linear models can be defined as

$$E(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{with } \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{1.5.1}$$

where the elements of $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2)$, and $\boldsymbol{\beta}$ is a vector of p parameters to be estimated from the data.

Generalised linear models (GLM) are used to study the relationship between a response variable and a set of explanatory variables relaxing the restrictive assumption of normality for the response. GLMs are regarded as an extension of the

theory behind linear models to the more general case where the response variable $\mathbf{Y}$ follows a distribution belonging to the exponential family of distributions. The most known member of this family is the Normal distribution, but there are many others such as the Poisson, Binomial and Multinomial (the last two with fixed number of trials but unknown probability parameter(s)).

Each element of $\mathbf{Y}$ is assumed to have a distribution that belongs to the exponential family, which depends on parameters θ and ϕ, and is defined in McCullagh and Nelder (1989) as

$$f_Y(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{1.5.2}$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are known functions. The parameter ϕ is the dispersion parameter or scale parameter, which defines a one-parameter exponential family of distributions when it is known and a two-parameter exponential family if it is not.

The binomial distribution is of interest when studying the logistic regression. Therefore, it will be represented in the exponential family form as follows. Let the random variable Y be the number of "successes" in n independent trials in which the probability of success, π, is the same in all trials. Then Y has the Binomial distribution with probability mass function

$$f_Y(y;\pi) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y} = \binom{n}{y}\pi^y(1-\pi)^{n-y}. \tag{1.5.3}$$

The logarithm of (1.5.3) is

$$\begin{aligned}
\log f_Y(y;\pi) &= \log\binom{n}{y}\pi^y(1-\pi)^{n-y} \\
&= \log\binom{n}{y} + y\log\pi + (n-y)\log(1-\pi) \\
&= \log\binom{n}{y} + y\log\pi + n\log(1-\pi) - y\log(1-\pi).
\end{aligned}$$

Then, applying the exponential function and rearranging to reach an equivalent

form to (1.5.2)

$$\begin{aligned} f_Y(y;\pi) &= \exp\left[y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right] \\ &= \exp\left[y(\log\pi - \log(1-\pi)) + n\log(1-\pi) + \log\binom{n}{y}\right] \\ &= \exp\left[y\log\frac{\pi}{(1-\pi)} + n\log(1-\pi) + \log\binom{n}{y}\right], \end{aligned} \tag{1.5.4}$$

so that comparing with the structure of equation (1.5.2) we find the parameter of interest

$$\theta = \log\frac{\pi}{(1-\pi)}, \tag{1.5.5}$$

the well-known logit function.

A GLM consists of three parts or components:

1. The *random component*: independent observations $Y_1, \ldots, Y_n$ with distribution in the exponential family;

2. The *systematic component*: the linear predictor of Y_i, denoted by η_i for the i^{th} observation $(i = 1, \ldots, n)$, defined as

$$\boldsymbol{\eta} = \sum_{h=1}^{p} \mathbf{x}_h \beta_h, \tag{1.5.6}$$

   with $\mathbf{x}_h$ the vector of n observations of the h^{th} variable;

3. The *link* between the random and systematic components through the use of a link function g, i.e $g(\mu_i) = \eta_i$ where $\mu_i = E(Y_i)$.

Then, for the i-th observation, Y_i has some probability distribution belonging to the exponential family with mean μ_i, such that

$$g(\mu_i) = \sum_{h=1}^{p} \beta_h x_{i,h} = \eta_i. \tag{1.5.7}$$

The function $g(\cdot)$ is assumed to be a monotonic and differentiable function, and is called the *link function*. It describes how the expected response is linked to the explanatory variables. When $g(\mu) = \theta$ in the exponential family, it is called to be a *canonical link*. Therefore, it is sensitive to use (1.5.5), a canonical link function,

when modelling a response variable that follows the Binomial distribution, as it is in the well-known logistic regression.

In the following section, several ways of defining π of Equation (1.5.5) will be presented, which allows to deal with more than two categories in an ordinal response variable.

1.5.2 Logits and ordinal information

When the dependent variable has two categories, one of them is classified as the event of success and usually the choice of the regression model to be used reduces to the logistic or probit regression models depending on the type of link function to be chosen, the logit or probit link functions correspondingly, as shown by (1.5.8) and (1.5.9) below:

$$P(Y = 1|X = x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}, \tag{1.5.8}$$

which results from using the logit function (1.5.5), and

$$P(Y = 1|X = x) = \Phi(\alpha + \beta x), \tag{1.5.9}$$

with $\Phi(\cdot)$ being the cumulative density function of the standard normal distribution.

In both cases, once the category associated with the event of success has been set, there is no need of making any further decision regarding the definition of the probability of success to be used in either (1.5.8) or (1.5.9). However, when there are more than two categories, there are different ways of determining the probabilities to be used in the logit function.

For the case when the dependent variable has three or more unique values, such as *Christianity*, *Islam*, *Hinduism*, *Nonreligious* or *Other Religion*, then the researcher should be interested in analysing results from some suitable multicategory models, such as the *multinomial logistic regression*. Agresti (2007) provides a thorough review of such models and others for categorical data, including the logistic regression model, being the main source of literature for this section.

Given that the categories of an ordinal variable are contiguous on the ordinal

scale type, it is possible to group them without affecting the general order. This leads to different definitions for the logits.

In general, define k as the number of ordinal categories and π_j as the probability of the j^{th} category of the ordinal variable Y in the population, with $j = 1, \ldots, k$.

Cumulative logits

The cumulative logits group the k category responses into two sets, $Y \leq j$ and $Y > j$. Then, the cumulative logit function is

$$\begin{aligned} \text{logit}[P(Y \leq j)] = \log \frac{P(Y \leq j)}{1 - P(Y \leq j)} &= \log \frac{\pi_1 + \ldots + \pi_j}{1 - (\pi_1 + \ldots + \pi_j)} \\ &= \log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_k}, \quad j = 1, \ldots, k-1. \end{aligned} \tag{1.5.10}$$

This can be seen as an ordinary binary logit with exhaustive and contiguous cumulative probabilities.

Adjacent-categories logits.

Unlike the cumulative logits, the adjacent-categories logits use the information provided by two adjacent categories only,

$$\log \frac{\pi_j}{\pi_{j+1}}, \quad j = 1, \ldots, k-1.$$

This logit is equal to an ordinary logit that reduces the possible response outcomes to categories j or $j+1$:

$$\begin{aligned} \text{logit}[P(Y = j | Y = j \text{ or } Y = j+1)] &= \log \frac{P(Y = j | Y = j \text{ or } Y = j+1)}{1 - P(Y = j | Y = j \text{ or } Y = j+1)} \\ &= \log \frac{P(Y = j, (Y = j \text{ or } Y = j+1))/P(Y = j \text{ or } Y = j+1)}{P(Y = j+1, (Y = j \text{ or } Y = j+1))/P(Y = j \text{ or } Y = j+1)} \\ &= \log \frac{P(Y = j, (Y = j \text{ or } Y = j+1))}{P(Y = j+1, (Y = j \text{ or } Y = j+1))} \\ &= \log \frac{P((Y = j \text{ or } Y = j+1) | Y = j) P(Y = j)}{P((Y = j \text{ or } Y = j+1) | Y = j+1) P(Y = j+1)} \\ &= \log \frac{P(Y = j)}{P(Y = j+1)}. \end{aligned}$$

The adjacent-categories logits can also be seen as the difference between two adjacent *baseline category logits*, which are commonly used to model nominal response variables:

$$\log \frac{\pi_j}{\pi_{j+1}} = \log \frac{\pi_j}{\pi_k} - \log \frac{\pi_{j+1}}{\pi_k}, \quad j = 1, \ldots, k-1.$$

Continuation-ratio logits.

To some extent, the continuation-ratio logits combine the previous two ways of defining logits by introducing one single probability in the numerator and a cumulative probability in the denominator. The continuation-ratio logits are defined as

$$\log \frac{\pi_j}{\pi_{j+1} + \ldots + \pi_k}, \quad j = 1 \ldots, k-1.$$

Continuation-ratio logits are generally used when categories of the response variable are such that they can be reached only successively step by step, which is covered by sequential models (Tutz, 1991).

In general, if we define

$$\omega_j = P(Y = j | Y \geq j) = \frac{\pi_j}{\pi_j + \ldots + \pi_k}, \quad j = 1, \ldots k-1,$$

then the ordinary logits of these conditional probabilities are the continuation-ratio logits:

$$\log \frac{\omega_j}{1-\omega_j} = \log \frac{\frac{\pi_j}{\pi_j+\ldots+\pi_k}}{1 - \frac{\pi_j}{\pi_j+\ldots+\pi_k}} = \log \frac{\frac{\pi_j}{\pi_j+\ldots+\pi_k}}{\frac{\pi_{j+1}+\ldots+\pi_k}{\pi_j+\ldots+\pi_k}} = \log \frac{\pi_j}{\pi_{j+1} + \ldots + \pi_k}.$$

The same rationale should be applied when modelling the inverse sequential mechanism to find alternative continuation-ratio logits:

$$\log \frac{\pi_{j+1}}{\pi_1 + \ldots + \pi_j}, \quad j = 1, \ldots, k-1.$$

In this case, the focus is on the probability associated with observing the response category $j+1$ with respect to the one of observing lower categories.

1.5.3 Proportional odds cumulative logit models (POCLM)

Instead of fitting $k-1$ models separately, ordinal models incorporate the $k-1$ logits into a single model simultaneously. A POCLM is one example of this.

Let y_i be the outcome category for the ordinal response variable Y for the i-th subject, and $\mathbf{x}_i$ its corresponding column vector of explanatory variables. The

model simultaneously uses all $k-1$ cumulative logits, and is defined as

$$\begin{aligned}\text{logit}[P(Y_i \leq j)] &= \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i \\ &= \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots, \end{aligned} \tag{1.5.11}$$

for $j = 1, \ldots, k-1$, and with $\boldsymbol{\beta}$ being a column vector of parameters associated with the explanatory variables.

The logit for cumulative probability j has its own intercept, α_j, which increases with j because $P(Y \leq j)$ are cumulative probabilities. Also, the effects $\boldsymbol{\beta}$ are the same for each cumulative logit. This is why (1.5.11) is called the "Proportional Odds" model. The model can also be expressed as

$$\begin{aligned}\text{logit}[P(Y_i \leq j)] &= \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i \\ \log \frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} &= \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i \\ \frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} &= \exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\} \\ P(Y_i \leq j) &= \exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\} - P(Y_i \leq j)\exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\} \\ P(Y_i \leq j) &= \frac{\exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\}}{1 + \exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\}}, \end{aligned} \tag{1.5.12}$$

and for each cell probability,

$$\begin{aligned}P(Y_i = j) &= P(Y_i \leq j) - P(Y_i \leq j-1) \\ &= \frac{\exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\}}{1 + \exp\{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\}} - \frac{\exp\{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i\}}{1 + \exp\{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i\}}, \end{aligned} \tag{1.5.13}$$

with $\alpha_0 = -\infty$ and $\alpha_k = +\infty$.

POCLM: Continuous predictor

Consider the case of a single continuous predictor x and an ordinal response variable with four possible outcomes. Then, the model is

$$\text{logit}[P(Y_i \leq j)] = \alpha_j + \beta_1 x_i, \quad j = 1, \ldots, 3.$$

The parameter effect, β_1, is the same for the three cumulative logits. Therefore, there are three different response curves when plotting $(x, P(Y \leq j))$, but sharing the same shape. They only differ in their location. The size of $|\beta|$ determines how

quickly the curves react over changes in x. When Y is statistically independent of x, then $\beta = 0$ and $P(Y \leq j)$ as a function of x is depicted by a horizontal line.

Since the curves have the same shape, it is possible to get one from another by adjusting the value of x, translating one of the curves towards the other.

$$P[Y \leq c|X = x] = P\Big[Y \leq j|X = x + \frac{\alpha_c - \alpha_j}{\beta}\Big] \quad \text{for } j < c \leq k. \tag{1.5.14}$$

Hence, the difference between two parameters in $\{\alpha_j\}$ gives a notion of the distance between their corresponding cumulative distributions, and the parameter β describes the effect of x.

POCLM: The proportional odds property

In a multivariate setting, consider two different subjects with explanatory variables vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ that have been used to fit the model (1.5.11). Therefore,

$$\begin{aligned}\text{logit}[P(Y \leq j|\mathbf{x}_1)] - \text{logit}[P(Y \leq j|\mathbf{x}_2)] &= \log \frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} \\ &= \alpha_j + \boldsymbol{\beta}'\mathbf{x}_1 - \alpha_j - \boldsymbol{\beta}'\mathbf{x}_2 \\ &= \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2),\end{aligned}$$

which can be rewritten as

$$\begin{aligned}\log \frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} &= \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2) \qquad (1.5.15) \\ \frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} &= \exp\{\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)\} \\ P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1) &= \exp\{\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)\}P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2).\end{aligned}$$

The odds of $Y \leq j$ given $\mathbf{x}_1$ are proportional to the ones of $Y \leq j$ given $\mathbf{x}_2$ in a magnitude of $\exp\{\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)\}$. Alternatively, the log cumulative odds ratio in (1.5.15) is proportional to the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ in a magnitude of $\boldsymbol{\beta}$.

POCLM: Interpretation

There are several ways to interpret the results of the POCLM. One approach is to compare predictive values for cumulative probabilities given different values of the explanatory variable(s). Another approach is to find the maximum and minimum for $\hat{P}(Y = 1)$ and $\hat{P}(Y = k)$ using (1.5.13) over the set of predictor values to

find those that provide these extremes. Finally, another approach is to use an approximation for the rate of change of a probability with respect to a change in an explanatory variable x_h. For the latter, using the expression (1.5.12) and taking the derivative with respect to x_h,

$$\begin{aligned}\frac{\partial P(Y \leq j|\mathbf{x})}{\partial x_h} &= \frac{\beta_h \exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}}(1+\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}}) - \beta_h(\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})^2}{(1+\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})^2} \\ &= \frac{\beta_h \exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}} + \beta_h(\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})^2 - \beta_h(\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})^2}{(1+\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})^2} \\ &= \beta_h \frac{\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}}}{(1+\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})} \frac{1}{(1+\exp^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}})} \\ &= \beta_h P(Y \leq j|\mathbf{x})[1-P(Y \leq j|\mathbf{x})]. \end{aligned} \tag{1.5.16}$$

Therefore, the effect on the cumulative probability depends on both the parameter value β_h and the level of $P(Y \leq j|\mathbf{x})$. For example, suppose that $\hat{\beta}_h = 0.20$ for the effect of x_h and that $\hat{P}(Y \leq j) = 0.60$. This means that an increase of one unit in x_h while keeping fixed the other predictors corresponds to approximately a $0.20(0.60)(0.40) = 0.048$ estimated increase in $\hat{P}(Y \leq j)$. Note that for $\beta_h > 0$, an increase in x_h implies a greater probability of falling in response category Y_j or lower. This could be considered as counter-intuitive because positive parameters are associated with higher probabilities for lower response categories. For this reason, some researchers use an alternative representation of the model (1.5.11) by imposing a negative sign to the parameter vector $\boldsymbol{\beta}$ as follows

$$\text{logit}[P(Y_i \leq j)] = \alpha_j - \boldsymbol{\beta}'\mathbf{x}_i. \tag{1.5.17}$$

However, this alternative version will not be used here, keeping the model (1.5.11) as the one to be analysed.

POCLM: Model fitting

In order to fit the model, let $y_{i1}, \ldots, y_{ik}$ be the binary indicators of the response, where $y_{ij} = 1$ if the response of subject i falls in category j and 0 otherwise. Define $\pi_j(\mathbf{x}_i)$ as the probability of the response of subject i to fall in category j, $P(Y_i = j|\mathbf{x}_i)$. Under the usual assumption of independent observations, the likelihood function is based on the product of the multinomial mass functions for the n subjects:

$$
\begin{aligned}
L(\{\alpha_j\}, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} \pi_j(\mathbf{x}_i)^{y_{ij}} \right\} = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} P(Y_{ij} = j|\mathbf{x}_i)^{y_{ij}} \right\} \\
&= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} [P(Y_{ij} \leq j|\mathbf{x}_i) - P(Y_{ij} \leq j-1|\mathbf{x}_i)]^{y_{ij}} \right\} \\
&= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} \left[\frac{e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}} \right]^{y_{ij}} \right\}. \qquad (1.5.18)
\end{aligned}
$$

To obtain the likelihood equation for an effect parameter β_h, differentiate the logarithm of $L(\{\alpha_j\}, \boldsymbol{\beta})$, denoted as $\ell(\{\alpha_j\}, \boldsymbol{\beta})$, with respect to a particular parameter and equate the derivative to zero, which is

$$
\sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} x_{ih} \frac{g(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - g(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{G(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i) - G(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} = 0,
$$

where $G(z) = \frac{e^z}{1+e^z}$ and $g(z) = \frac{e^z}{(1+e^z)^2}$.

Iterative methods, such as the Fisher scoring algorithm, are used to solve the likelihood equations and obtain the ML estimates of the model parameters.

POCLM: Inference

Statistical inference on the parameters is based on the Maximum Likelihood parameter estimates and their standard errors.

Estimating Standard Errors.

In order to estimate the standard errors of the parameters in the POCLM the information matrix is used (see Agresti (2010)). The information matrix describes the curvature of the log-likelihood function. The more highly curved the log likelihood function at the ML estimates, the smaller the standard errors and more precise the ML estimates of the model parameters. The information matrix contains the negative second partial derivatives of $\ell(\{\alpha_j\}, \boldsymbol{\beta})$ with respect to the model parameters. There are two versions of the information matrix that can be used:

- The *observed* information matrix: It uses the actual second partial derivatives. The element in row a and column b of the observed information matrix

is

$$-\partial^2 \ell(\{\alpha_j\}, \boldsymbol{\beta})/\partial\beta_a \partial\beta_b.$$

The Newton-Raphson algorithm uses the observed information matrix.

- The *expected* information matrix: It uses the expected values of the second partial derivatives. The element in row a and column b of the expected information matrix is

$$E\{-\partial^2 \ell(\boldsymbol{\beta})/\partial\beta_a \partial\beta_b\}.$$

The inverse of the expected information matrix is used in the Fisher scoring algorithm for obtaining the ML model fit.

Either the *observed* or *expected* information matrix can be used to estimate the information matrix by substituting $(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})$. The estimated standard errors are the square roots of the main-diagonal entries of the inverted estimated information matrix.

Inference About Model Parameters

With the ML estimates, their standard errors, and the maximised likelihood function, it is possible to perform statistical inference on the model parameters in the traditional way by using tools such as the Wald confidence interval, the z-test, the Likelihood-Ratio test, and others.

A 95% Wald confidence interval for a parameter β_h is

$$\hat{\beta}_h \pm 1.96(SE_{\hat{\beta}_h}),$$

where $SE_{\hat{\beta}_h}$ is the standard error of the parameter estimate $\hat{\beta}_h$, which relies on the Central Limit Theorem (CLT).

According to Agresti (2010), when it is of interest to test the significance of a parameter, then $H_0 : \beta_h = 0$ can be tested by using

$$z = \hat{\beta}_h / SE_{\hat{\beta}_h},$$

with the usual interpretations and implications for the resulting z statistic and its corresponding p-value. More generally, when testing whether β_h is equal to any other value $(H_0 : \beta_h = \beta_{h0})$,

$$z = \frac{\hat{\beta}_h - \beta_{h0}}{SE_{\hat{\beta}_h}}.$$

1.6 Some regression models for non-ordinal responses and ordinal predictors

In the regression model framework there are some models that give special treatment to ordinal predictors. Some of these are discussed in this section as an overview of available approaches, although they were designed for non-ordinal responses.

1.6.1 Penalised maximum likelihood

Tutz and Gertheiss (2014), noted that "in most advanced books on statistical modelling, ordinal responses are treated but ordinal predictors are, if at all, just mentioned; and proper treatment is hardly considered." Therefore, they extended existing methodology to the framework of generalized linear models to propose a penalisation method giving ordinal predictors a special treatment other than transforming its measurement scale. In addition, they presented their penalisation approach as a method for the selection of predictors and clustering of their categories.

Apart from using statistical models specifically developed for ordinal data, these authors also state that there are two common approaches for statistical modelling when using rating scales as predictors. One of these approaches is to treat ordinal predictors as metrically scaled variables using a scoring system, and the other one is to treat them as nominal-scaled variables by transforming each ordinal predictor in a set of dummy variables.

The first approach leads to the choice of a scoring system. For example, a model for one ordinal predictor under this approach is

$$y = \beta_0 + A\beta_1 + \varepsilon, \tag{1.6.1}$$

where A is a rating-scaled variable with $p+1$ categories, $A \in \{0, \ldots, p\}$, but assumed to be measured on metric scale level, and $\varepsilon \sim N(0, \sigma^2)$.

The second approach uses p dummy variables to represent the $p+1$ ordered categories of the predictor variable, usually being the first category the baseline,

which is omitted in the model. For example, a model for one ordinal predictor under this approach is

$$y = \alpha_0 + x_{A(1)}\beta_1 + \ldots + x_{A(p)}\beta_p + \varepsilon, \tag{1.6.2}$$

where $x_{A(j)} = 1$ if $A = j$ and $x_{A(j)} = 0$ otherwise for all $j = 1, \ldots, p$.

Tutz and Gertheiss (2014) allow their model for smooth and monotonic effects across categories using a penalised version of the model (1.6.2) with their corresponding tuning parameters λ and λ_m accordingly.

Smooth effects

Consider the case of one ordinal predictor. The data is composed by n observations of y and A, where y_i is the response for the ith observation and $A_i \in \{0, \ldots, p\}$ is its corresponding observed ordinal predictor, (y_i, A_i). It is assumed that the distribution of the response variable is a member of the exponential family, just as it is in the setting for GLMs. The linear predictor is defined as

$$\eta_i = \alpha_0 + x_{A(1),i}\beta_1 + \cdots + x_{A(p),i}\beta_p, \tag{1.6.3}$$

where $x_{A(h),i}$ are dummy variables for the category h and observation i as defined for the model (1.6.2). The parameter β_0 that corresponds to $x_{A(0),i}$ is omitted in equation (1.6.3) as it has been set to 0, indicating the reference category. This model does not take into account the categories order yet.

Tutz and Gertheiss (2014) proposed to maximise the penalised log-likelihood for the estimation of the parameters $\beta_1, \ldots, \beta_p$, recall that $\beta_0 = 0$, then we maximise

$$l_{penal}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda \mathbf{J}(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ is the unpenalised log-likelihood, λ is a tuning parameter, and $\mathbf{J}(\boldsymbol{\beta})$ is a penalty term. For any $\lambda > 0$, then the fitted model penalises the likelihood if $\mathbf{J}(\boldsymbol{\beta}) > 0$. Therefore, if the penalty term is defined as:

$$\mathbf{J}(\boldsymbol{\beta}) = \sum_{h=1}^{p} (\beta_h - \beta_{h-1})^2, \tag{1.6.4}$$

greater differences between pairs of parameters for adjacent categories, β_h and β_{h-1}, produce greater penalisation.

Setting a larger λ and using the penalty term (1.6.4) forces the maximisation procedure to choose a set of parameters in such a way that differences between the parameters associated with adjacent categories are smaller, resulting in estimating smoother effects. Also, a larger λ leads parameter estimates to be closer to zero because $\beta_0 = 0$. In the limit $\lambda \to \infty$ and provided that $\beta_0 = 0$, all the parameters tend to be equal, then $\hat{\beta}_h \to 0$. The authors state that "the parameter λ establishes how important the order information in the predictor is." However, this would mean that the more important the order information is, the closer to zero the parameter estimates are. Therefore, their statement implies that using the smoothing approach when the order information is highly important increases the probability of estimating effects $\hat{\beta}_h$ that are close to zero, which can be considered as a disadvantage of the penalisation approach based on the penalty term (1.6.4). This is why they propose a more general one.

A penalty term that allows to avoid estimating $\hat{\beta}_h \to 0 \ \forall h$ when $\lambda \to \infty$ is

$$\mathbf{J}(\boldsymbol{\beta}) = \sum_{h=d}^{p} (\Delta^d \beta_h)^2, \tag{1.6.5}$$

where Δ is the difference operator and d is the number of times this difference operator is used, that is, for $d = 1$ then $\Delta^1 \beta_h = \beta_h - \beta_{h-1}$, for $d = 2$ then $\Delta^2 \beta_h = \Delta^1(\beta_h - \beta_{h-1}) = \beta_h - 2\beta_{h-1} + \beta_{h-2}$, and so on.

Consider the special case when the parameters are linear, $\beta_h = \gamma h$. The penalty term (1.6.5) is $\sum_{h=1}^{p} \gamma^2$ if $d = 1$, since $\beta_h - \beta_{h-1} = \gamma h - \gamma(h-1) = \gamma$. Then it is still true that $\hat{\beta}_h \to 0$ when $\lambda \to \infty$ for $d = 1$. However, if $d = 2$ the penalty term is

$$\begin{aligned}
\mathbf{J}(\boldsymbol{\beta}) &= \sum_{h=2}^{p} (\Delta^2 \beta_h)^2 = \sum_{h=2}^{p} (\beta_h - 2\beta_{h-1} + \beta_{h-2})^2 \\
&= \sum_{h=2}^{p} (\gamma h - 2\gamma(h-1) + \gamma(h-2))^2 \\
&= \sum_{h=2}^{p} (\gamma h - 2\gamma h + 2\gamma + \gamma h - 2\gamma)^2 \\
&= 0,
\end{aligned}$$

meaning that there is no penalisation for the linear form. Therefore, this approach

produces unpenalised linear parameter estimates $\hat{\beta}_h = \hat{\gamma}h$ in this case, even if $\lambda \to \infty$.

Monotonicity of effects

In Tutz and Gertheiss (2014) the case of increasing effects of the ordered predictors is considered only, that is, $\beta_0 \leq \beta_1 \cdots \leq \beta_p$. This is incorporated as an extra penalty term based on the concept of asymmetric difference penalties, yielding to the total penalty of the form

$$\lambda \sum_{h=d}^{p} (\Delta^d \beta_h)^2 + \lambda_m \sum_{h=1}^{p} \upsilon_h (\Delta \beta_h)^2, \tag{1.6.6}$$

where υ_h are weights defined as $\upsilon_h = 1$ if $\Delta\beta_h < 0$ and $\upsilon_h = 0$ if $\Delta\beta_h \geq 0$.

Notice that the weights υ_h depend on the parameters, which is a problem because they need to be estimated. However, during the iterative procedure, it is possible to compute the weights from previous estimates, then the weights are treated as known.

Given that penalising the differences between parameters associated with adjacent ordered categories decreases $\Delta\hat{\beta}_h \ \forall h$ in each step of the maximisation procedure, there is no reason to think that the monotonicity will hold as a final result, reaching $\Delta\hat{\beta}_h \geq 0$. In the following sub-section the penalisation approach will be discussed in more detail, analysing to what extent the parameter estimates get close to be monotonic.

Despite the fact that the authors make reference to the tuning parameters λ and λ_m, the order of differences d in the penalty terms (1.6.5) and (1.6.6), is not referred to as a tuning parameter in Tutz and Gertheiss (2014). Given that d must be chosen, preferably using an integer greater than one, then it is considered here as a tuning parameter.

Some remarks on penalised estimates

Consider the model (1.5.11) for an ordinal response (with three categories) and one ordinal predictor only,

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] &= \alpha_j + \sum_{h_1=2}^{p_1} \beta_{1,h_1} x_{i,1,h_1} \\ &= \alpha_j + \boldsymbol{\beta}\mathbf{x}_i, \end{aligned} \tag{1.6.7}$$

where $p_1 = 8$, $h_1 = 1, \ldots, 8$, and $k = 3$, i.e., $j = 1, 2$. As the aim of the following analysis is to highlight some aspects of the penalisation approach given a certain set of unpenalised and unconstrained parameter estimates (UPE), it will be assumed that the true pattern of the ordinal predictor parameters is isotonic and an example data set for which its parameter estimates violate this assumption will be analysed. The data were simulated according to the following setting: $\alpha_1 = -0.5$, $\alpha_2 = 0.1$, $\boldsymbol{\beta}' = (0, 0.5, 1.3, 1.27, 1.24, 1.18, 1.13, 1.14)$, the values of the ordinal predictor were randomly drawn from its assumed population distribution of 5% for each of the first two categories and 15% for the remaining six, the sample size was 2,000 observations. The parameters of the OP were chosen in such a way to get decreasing monotonicity for an adjacent subset of the UPEs as shown by those for $\beta_{1,4}, \ldots, \beta_{1,8}$ in Figure 1.2(a), which violates monotonicity.

In order to fit this model, Tutz and Gertheiss (2014) proposed to maximise

$$\ell_{penal}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \left(\lambda \sum_{h_1=d}^{p_1} (\Delta^d \beta_{1,h_1})^2 + \lambda_m \sum_{h_1=2}^{p_1} v_{1,h_1} (\Delta \beta_{1,h_1})^2 \right) \tag{1.6.8}$$

where $\lambda \sum_{h_1=d}^{p_1} (\Delta^d \beta_{1,h_1})^2 + \lambda_m \sum_{h_1=2}^{p_1} v_{1,h_1} (\Delta \beta_{1,h_1})^2$ is the total penalty term, which is decomposed into the two sums. Its first part aims to obtain smooth effects, with λ being the tuning parameter and Δ^d a difference operator for adjacent parameters, $\Delta \beta_{1,h_1} = \beta_{1,h_1} - \beta_{1,h_1-1}$, $\Delta^2 \beta_{1,h_1} = \Delta(\Delta \beta_{1,h_1}) = \Delta(\beta_{1,h_1} - \beta_{1,h_1-1})$, and so on. The second part is related to monotonicity, with λ_m being its tuning parameter and, for the isotonic case, each $v_{1,h_1} = 1$ if $\Delta \beta_{1,h_1} < 0$ and $v_{1,h_1} = 0$ otherwise.

An unconstrained and unpenalised version of the model (1.6.7) was fitted to obtain the UPEs and their corresponding 95% confidence intervals (CIs) as shown in Figure 1.2(a). The first two parameter estimates, $\hat{\beta}_{1,2}$ and $\hat{\beta}_{1,3}$, suggest an

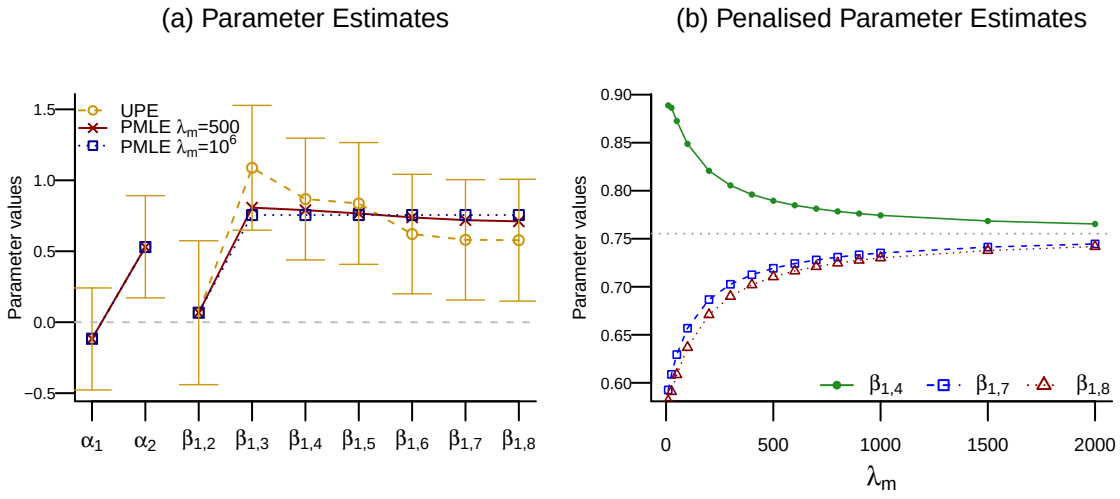

Figure 1.2: (a) Unpenalised Parameter Estimates and Penalised Maximum Likelihood Estimates with $\lambda_m = 500$ and $\lambda_m = 10^6$. (b) Penalised MLE for different parameters and the tuning parameter for monotonicity λ_m.

isotonic pattern. However, the UPEs for $\beta_{1,4}, \ldots, \beta_{1,8}$ are antitonic. Based on the analysis of individual CIs, the UPEs are compatible with an isotonic pattern, because the CI of $\hat{\beta}_{1,2}$ is fully above the one of $\hat{\beta}_{1,3}$ and the remaining ones overlap with the latter, so it seems to be valid to penalise deviations from an isotonic pattern.

The method proposed by Tutz and Gertheiss (2014) was applied to obtain the Penalised Maximum Likelihood Estimates (PMLE) for the model (1.6.7) with no smoothness penalty, $\lambda = 0$. Two solutions regarding monotonicity are shown in Figure 1.2(a). The one with $\lambda_m = 10^6$ provides an almost monotonic solution (blue dotted line), whereas the violation of monotonicity by the one with $\lambda_m = 500$ (red solid line) is much clearer. For $\lambda_m = 10^6$, the resulting parameter estimates of $\beta_{1,3}, \ldots, \beta_{1,8}$ produce a horizontal line. However, this is just a visual effect because the PMLE for $\beta_{1,3}$ is 0.7552680 and 0.7552072 for $\beta_{1,8}$, which is not compatible with monotonicity neither. In fact, this solution is asymptotic with respect to monotonicity, as shown in Figure 1.2(b).

If we consider the solution with $\lambda_m = 10^6$, then the value of the PMLE for $\beta_{1,3}$, 0.7552680, should be considered as a reference point by the following penalised

parameter estimates because they belong to the same ordinal predictor and are assumed to be monotonic (isotonic for this example). Let's consider the case of the PMLE for $\beta_{1,4}$. It starts at 0.868 and decreases as λ_m increases converging to a point around the one depicted by the grey dotted line (see Figure 1.2(b)). This solution converges to monotonicity but it does not reach it for any level of $\lambda_m \neq \infty$.

Under the current setting of consecutive monotonicity violations, the PMLE solution produces a cumulative deviation from monotonicity. Note that the last three UPEs are also violating monotonicity and they all are below the almost horizontal line describing the final solution with $\lambda_m = 10^6$ (blue dotted line in Figure 1.2(a)). In this case, the solution of the PMLE for $\beta_{1,8}$ is asymptotic with respect to the one of the PMLE for $\beta_{1,7}$ as shown in Figure 1.2(b). Therefore, if we set a not big enough λ_m, not only monotonicity is not achieved but also the higher the ordinal category, the higher the magnitude of the violation for those cases were consecutive monotonicity violations are present.

1.6.2 Isotonic regression

The isotonic regression fits an increasing function to a set of observations. In general, there is no information regarding the true regression function. Therefore, if the analysis is based on order restrictions, it is necessary to assume a particular direction of the ordering, ascending or descending. For example, that the value of the response variable increases as the value of the explanatory variable does. The isotonic regression is a particular and typical case belonging to a broader class, the monotonic regression. When a monotonic regression fits a decreasing function to a set of observations, then it is referred to as an antitonic regression.

A general isotonic regression model was described in de Leeuw et al. (2009) as follows. A simple linear regression estimates the parameters α and β describing a linear relationship between a predictor $x = (x_1, \ldots, x_n)$ and a response $y = (y_1, \ldots, y_n)$ by minimising the loss function

$$L(\alpha, \beta) = \sum_{i=1}^{n} w_i (y_i - \alpha - \beta x_i)^2, \tag{1.6.9}$$

over α and β, where w_i are some observation weights.

However, in monotonic regression the setting is different because there is no linear relationship to be fitted between the response and the predictor, but monotonic. Now, let X be the set of the ordered observations of the predictor with no ties, i.e., $\{x_1, x_2, \ldots, x_n\}$ with $x_1 < x_2 < \cdots < x_n$. The observed response vector is y and the vector of unknown response values to be fitted is $z = (z_1, \ldots, z_i, \ldots, z_n)'$ (commonly defined as $\hat{y}$ under the usual linear regression framework). The ordinal predictor does not provide more information than the ordering, then the loss function (1.6.9) does not hold and y_i follows the order of x_i after sorting all the observed pairs (y_i, x_i) based on x_i. The least squares function in monotonic regression has to be minimised over z and can be stated as

$$L(z) = \sum_{i=1}^{n} w_i (y_i - z_i)^2, \tag{1.6.10}$$

which means that the number of parameters is equal to the number of observations and that, for the isotonic regression, the minimisation has to be done under the inequality restrictions $z_1 \leq z_2 \leq \cdots \leq z_n$. The predictor x_i is not explicitly expressed in (1.6.10). However, it is implicitly playing a role in (1.6.10) because its ordering affects the one of y_i.

The fact that in monotone regression the number of parameters is equal to the number of observations might suggest that it is possible to reach the perfect fit. However, the order restrictions on z is an impediment of getting a perfect solution. Under the isotonic regression scenario, the best fitting monotone function computation is based on the fact that if $y_i \geq y_{i+1}$, then $\hat{z}_i := \hat{z}_{i+1}$. This means that if one value of y is in descendent order with respect to the consecutive previous one, then the two corresponding consecutive values for the solution $\hat{z}$ will be equal.

To fit this model, the up-and-down-blocks algorithm used for the Nonmetric Multidimensional Scaling method developed by Kruskal (see Kruskal (1964)) provides a numerical method as a solution.

For a non-strict partial order of the predictors, i.e., $x_1 \leq x_2 \leq \ldots \leq x_n$, several algorithms for the treatment of ties can be considered. Different approaches are grouped in three classes. They all partition the index set $\{1, 2, \ldots, n\}$ into a

number of tie blocks $I_1, I_2, \ldots, I_k$ with $k \leq n$, where a *block* is defined as a set of consecutive points with the same value. However, the approaches differ in the way they treat ties mainly. Consider the tied observations i and i', where $x_i = x_{i'}$:

1. **Primary Approach**. Implies that z_i does not necessarily equal $z_{i'}$. In turn, it does imply that if $y_i \leq y_{i'}$ then $z_i \leq z_{i'}$, forcing to hold the monotonicity condition.

2. **Secondary Approach**. Requires $z_i = z_{i'}$ for the tie $x_i = x_{i'}$, regardless which y-values were observed.

3. **Tertiary Approach**. For each tie block $I_1, I_2, \ldots, I_k$ the unit of analysis are the weighted means $\bar{z}_{I_1}, \bar{z}_{I_2}, \ldots, \bar{z}_{I_K}$. Therefore, the tertiary approach requires only that these means are monotonic across tie blocks, not individual values. As a consequence, it abandons the monotonicity condition because $y_i \leq y_{i'}$ does not imply $z_i \leq z_{i'}$.

Theory and algorithms to solve this kind of isotonic regression problems are discussed in de Leeuw et al. (2009), in particular, the Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. However, their proposed solutions are restricted to one predictor of any scale type that allows "determination of greater or less" according to Steven's measurements scales (Stevens, 1946). Beran and Dümbgen (2010) described active set methods to fit regression models with two predictors via least squares or least absolute deviations, and some other authors have proposed isotonic regression for multiple predictors, including an extension of the work in de Leeuw et al. (2009) developed by Burdakov et al. (2004) (for other contributions, see Dykstra et al. (1982) and Stout (2015)). Nevertheless, none of these approaches allows a non-monotonic association between a particular predictor and the response variable.

1.6.3 Constrained maximum likelihood

Rufibach (2010) proposed an active set algorithm to estimate parameters in generalised linear models with ordered predictors. Consider an ordinal predictor variable w with ordinal categories, $1, \ldots, k$, and a response variable y that may be

continuous, binary, or represent censored survival times. If we take the first possible type of y as an example to illustrate what is proposed in Rufibach (2010), then its corresponding linear model for one continuous response and one ordinal predictor is

$$\begin{aligned} y_i &= \beta_2 w_{i,2} + \cdots + \beta_k w_{i,k} + \epsilon_i \\ \mathbf{y} &= \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \end{aligned} \tag{1.6.11}$$

where the ordinal predictor w introduces $k-1$ dummy variables in the model defining $w_{i,j} = 1$ if the ith observed value falls in the jth category and 0 otherwise for all $j = 2, \ldots, k$ and $i = 1, \ldots, n$. The author proposed an algorithm to solve the problem of maximising the log-likelihood of the model (1.6.11) assuming $\beta_1 = 0$ and isotonic effects, i.e.,

$$0 \leq \beta_2 \leq \ldots \leq \beta_k. \tag{1.6.12}$$

The constrained maximum likelihood approach proposed by Rufibach (2010) is based on using "the available knowledge (or our 'prior belief')" in order to apply the monotonicity constraints in certain direction, i.e., in the isotonic form as shown in (1.6.12).

In general, $\mathbf{y}$ and $\mathbf{W}$ are given observations. Then it is required to maximise a criterion function L (e.g. a likelihood function) over the possible values of $\boldsymbol{\beta} \in \mathbb{R}^p$ to estimate the parameter vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$:

$$L = L(\mathbf{y}, \mathbf{W}, \boldsymbol{\beta}) : \mathbb{R}^n \times \mathbb{R}^{n \times p} \times \mathbb{R}^p \to \mathbb{R}. \tag{1.6.13}$$

The matrix $\mathbf{W}$ contains the values associated with any type of predictors, but they require to be ordered in a specific way. This matrix contains two groups of columns, the first c columns represent quantitative variables and the second $f = (p - c)$ set of columns are associated with ordinal predictors. In addition to interval and ratio-scaled predictors, nominal-scaled predictors are assumed to be part of the c first columns of $\mathbf{W}$ in the form of dummy variables. The last f predictors $\mathbf{w}_{\cdot,j}$, $j = c+1, \ldots, p$, are ordered factors, each with k_j levels. Therefore, the total number of columns of $\mathbf{W}$ is $p = c + f$. The elements of each ordered

factor $\mathbf{w}_{\cdot,j} = (w_{ij})_{i=1}^{n}$ are assumed to be $w_{ij} \in \{1, \ldots, k_j\}$, $i = 1, \ldots, n$, where the higher the level of the ordered factor, the higher the number w_{ij}. Define two sets of indices, one denoting the indices of the ordinal predictors $\mathcal{F}_{c,p} = \{c+1, \ldots, p\}$, and another one denoting the indices of the ordinal categories for each ordinal predictor, $\mathcal{L}_j = \{2, \ldots, k_j\}$ for $j \in \mathcal{F}_{c,p}$. Based on $\mathbf{W}$, build a new data matrix $\mathbf{X} \in \mathbb{R}^d$ in such a way that the

$$\left(\sum_{j=c+1}^{p} k_j \right) - (p - c)$$

dummy variables represent the levels of the ordinal predictors excluding their first category.

The new criterion function has the form

$$L = L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) : \mathbb{R}^n \times \mathbb{R}^{n \times d} \times \mathbb{R}^d \to \mathbb{R} \tag{1.6.14}$$

and the constrained parameter space is

$$\mathcal{B}(c, p, \mathbf{k}) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_{j,2} \geq 0, \beta_{j,l+1} - \beta_{j,l} \geq 0, 2 \leq l \leq k_j - 1, j \in \mathcal{F}_{c,p}\}, \tag{1.6.15}$$

where $\mathbf{k} = ((0)_{i=1}^{c}, k_{c+1}, \ldots, k_p) \in \mathbb{R}^p$.

The optimisation procedure takes into account the constraints in $\mathcal{B}$, a criterion function $\ell(\boldsymbol{\beta})$ (e.g., a log-likelihood function or the negative of a loss function), and $\boldsymbol{\beta}$ as its maximiser. Therefore, the constrained maximisation problem is

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathcal{B}}{\text{maximise}}\ \ell(\boldsymbol{\beta}), \tag{1.6.16}$$

whereas the unconstrained one is

$$\hat{\boldsymbol{\eta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\text{maximise}}\ \ell(\boldsymbol{\beta}). \tag{1.6.17}$$

An active set algorithm was proposed in Rufibach (2010) to solve the constrained problem (1.6.16) in presence of isotonic relationships between the ordinal predictors and a response variable that could be continuous, binary, or represent censored survival times.

1.7 A latent variable approach for ordinal variables

One of the main objectives of many statistical models is to specify the joint distribution of a set of random variables. According to Bartholomew et al. (2011), latent variable models are used when some or all of these variables are unobservable, the so called latent variables.

There are two main reasons why latent variable models are used:

1. Dimensionality reduction. Latent variable models are used to see whether there are patterns in the interrelationships among a set of variables (e.g., answers to a set of questions) by constructing a smaller number of latent variables that can represent a large amount of the information provided by the observed variables.

2. Represent a complex construct by numbers. Specially in social sciences, it is often of interest to analyse certain constructs, for instance, the researcher could be interested in measuring Self-Assessment of Quality of Life, Business Confidence or Ideological Self-Placement in politics, but these constructs do not manifest themselves directly but rather indirectly through other variables. Each of these constructs can be analysed through the use of a variety of directly measurable variables that are related to the construct, such as yes/no answers to a set of questions, multiple choice questions, etc., for which their analysis is conducted with latent variable models. The latent variables that are obtained from the model refer to the underlying construct that is indirectly measured based on the information provided by the variables that are actually measured, the so-called manifest variables.

From the perspective of the treatment of ordinal covariates in a regression model context, LVMs can be used as a way of transforming a set of ordinal covariates into a single continuous latent variable that represents them, which works as a scoring system. The resulting score can be used in the regression model as a

continuous covariate replacing the set of ordinal covariates. This is the approach that will be used later and therefore a LVM for ordinal observed variables will be presented in Section 1.7.2, after an introduction to the general framework of LVMs is discussed in the next section.

1.7.1 Introduction to the general framework for LVMs

The main source of information for this introduction to the general framework for LVMs is Bartholomew et al. (2011).

The observed variables are referred to as manifest variables (MVs) and the ones representing their underlying joint distribution are called latent variables (LVs). The main interest is in the information generated by the latent variables after observing the manifest variables, which introduces a conditional probability approach. This approach is known as the "item response function approach", where the whole response pattern is specified by the LVM as will be seen in this section.

Consider a set of p MVs forming a vector $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ and q LVs defining the vector $\mathbf{z}$, all these variables are considered to be random. It is of interest to build q latent variables explaining the interrelationship among the p manifest variables with q as small as possible, certainly not greater than p. If the MVs are conditionally independent when conditioning on LVs held fixed, then the set of latent variables in $\mathbf{z}$ explains the dependencies among the manifest variables $\mathbf{y}$ and the set of LVs is said to be complete. Therefore, the number of latent variables q will correspond to the smallest q that fulfils

$$g(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{p} g_i(y_i|\mathbf{z}), \tag{1.7.1}$$

where q is the dimensionality of $\mathbf{z}$.

This is true when the density function $f(\mathbf{y})$ of the manifest variables $\mathbf{y}$ is

$$f(\mathbf{y}) = \int_{R_z} h(\mathbf{z}) g(\mathbf{y}|\mathbf{z}) d\mathbf{z} \tag{1.7.2}$$

$$= \int_{R_z} h(\mathbf{z}) \prod_{i=1}^{p} g_i(y_i|\mathbf{z}) d\mathbf{z}, \tag{1.7.3}$$

for some q, h and $\{g_i\}$. The main problem is to know about the conditional distribution of the latent variables in $\mathbf{z}$ given the manifest variables in $\mathbf{y}$ with q as small as possible according to (1.7.1). Based on Bayes' theorem this conditional probability can be expressed as

$$k(\mathbf{z}|\mathbf{y}) = \frac{h(\mathbf{z})g(\mathbf{y}|\mathbf{z})}{f(\mathbf{y})}, \tag{1.7.4}$$

where $k(\mathbf{z}|\mathbf{y})$ is also called the *posterior distribution* of the latent variables given the manifest variables. In (1.7.4), the only function that can be estimated is f, which requires to restrict the classes of functions to be considered for h and g. Given that the LVs in $\mathbf{z}$ are a construct, their true distribution is unobservable, and therefore h is chosen. This is why the density function $h(\mathbf{z})$ is called the *prior distribution*. On the other hand, the conditional distributions $g_i(y_i|\mathbf{z})$ are the ones that are modelled while the smallest q for which (1.7.3) holds is found.

Bartholomew (1983), based on the existing framework of latent variable models for categorical data, extended these types of models to take into account the order of categories of the variables. Moustaki (2000) proposed a class of LVMs for observed ordinal variables where the whole response pattern is the unit of analysis, making assumptions on the conditional joint distribution of the observed variables given a set of latent variables, as presented above. This is one of the main characteristics of methods belonging to the "response function approach" for LVMs. Another general approach is the "underlying variable approach", which assumes that each observed variable is generated by an underlying latent continuous variable that is usually assumed to be normally distributed. Both approaches use underlying variables. However, unlike the "underlying variable approach", the "response function approach" does not require to define an underlying variable for each ordinal observed variable and therefore it does not require to make assumptions about the distribution of those variables. The model proposed by Moustaki (2000) is presented in more detail in the following section.

1.7.2 LVMs for ordinal manifest variables

Consider the p manifest variables in $\mathbf{y}$, defined in Section 1.7.1, as ordinal-scaled variables. Define the number of ordered categories for each observed ordinal vari-

able as m_i, with $i = 1, 2, \ldots, p$, and their probabilities as $\pi_{i1}(\mathbf{z}), \pi_{i2}(\mathbf{z}), \ldots, \pi_{im_i}(\mathbf{z})$, which depend on q latent variables in $\mathbf{z}$. As in Moustaki (2000), all variables and their realisations will be denoted by lower-case letters.

The general form given in McCullagh (1980) of a linear model for the i-th ordinal variable (also denoted as i-th response) is:

$$\begin{aligned} link[\kappa_{is}(\mathbf{z})] &= link\ [P\{y_i \leq s|\mathbf{z}\}] \\ &= \alpha_{is} - \sum_{j=1}^{q} \tau_{ij} z_j, \quad i = 1, 2, \ldots, p; s = 1, 2, \ldots, m_i, \end{aligned} \tag{1.7.5}$$

where $link[\kappa_{is}(\mathbf{z})]$ is a function of $\kappa_{is}(\mathbf{z})$ that in the context of GLMs is known as a link function, and $\kappa_{is}(\mathbf{z})$ is the cumulative probability of the response falling in category s or lower of item y_i, written as $\kappa_{is}(\mathbf{z}) = \pi_{i1}(\mathbf{z}) + \pi_{i2}(\mathbf{z}) + \cdots + \pi_{is}(\mathbf{z})$, with $s \leq m_i$, where $\kappa_{is}(\mathbf{z})$ (or for easy of notation simply κ_{is}) is a function of the latent variables $\mathbf{z}$ in Moustaki (2000).

The probability of the i-th response in category s is:

$$\pi_{is} = \kappa_{i,s} - \kappa_{i,s-1}, \quad i = 1, 2, \ldots, p; \quad s = 2, \ldots, m_i. \tag{1.7.6}$$

The parameters α_{is} in (1.7.5) follow the restriction $-\infty = \alpha_{i0} < \alpha_{i1} \leq \alpha_{i2} \leq \cdots \leq \alpha_{i,m_i} = +\infty$, and there is one of them for each category of the MV y_i. The effect of the j-th latent variable in $\mathbf{z}$ on the link function of the cumulative probability for the i-th response is measured by the parameter τ_{ij} $\forall s$. Therefore, these are considered as factor loadings.

Within the current approach, the item response function approach, the unit of analysis is the complete response pattern. Consider $\mathbf{y}$ as the p-dimensional response pattern of a randomly selected individual. Its density function $f(\mathbf{y})$ is

$$f(\mathbf{y}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(\mathbf{y}|\mathbf{z}) h(\mathbf{z}) d\mathbf{z}. \tag{1.7.7}$$

The latent variables $\mathbf{z}$ are assumed to be independent, $g(\mathbf{y}|\mathbf{z})$ is the conditional density function and $h(\mathbf{z})$ is referred to as the *prior distribution*, which in practice is usually chosen to be the standard normal distribution.

As described in Section 1.7.1, the responses to the p items are conditionally

independent given q latent variables in $\mathbf{z}$, then Equation (1.7.1) holds, which is

$$g(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{p} g_i(y_i|\mathbf{z}). \tag{1.7.8}$$

For each term in the product of the right hand side of (1.7.8) the conditional probability is

$$\begin{aligned} g_i(y_i|\mathbf{z}) &= \prod_{s=1}^{m_i} \pi_{is}(\mathbf{z})^{y_{is}} \\ &= \prod_{s=1}^{m_i} (\kappa_{is} - \kappa_{i,s-1})^{y_{is}}, \end{aligned} \tag{1.7.9}$$

where $y_{is} = 1$ if the response of the i-th item is in category s and $y_{is} = 0$ otherwise.

Define $y^*_{i,s}$ as $y^*_{i,s} = 1$ if the response of the i-th item falls into category s or lower, and $y^*_{i,s} = 0$ otherwise. Therefore, $y^*_{i,s}$ can be thought of as a cumulative version of y_{is}. Now, using $y^*_{i,s}$, Equation (1.7.9) is equivalent to

$$g_i(y_i|\mathbf{z}) = \prod_{s=1}^{m_i-1} \left(\frac{\kappa_{is}}{\kappa_{i,s+1}}\right)^{y^*_{i,s}} \left(\frac{\kappa_{i,s+1} - \kappa_{i,s}}{\kappa_{i,s+1}}\right)^{y^*_{i,s+1} - y^*_{i,s}}. \tag{1.7.10}$$

Taking the log of (1.7.10) results into:

$$\begin{aligned} \log g_i(y_i|\mathbf{z}) &= \sum_{s=1}^{m_i-1} \left[y^*_{i,s} \log\left(\frac{\kappa_{is}}{\kappa_{i,s+1}}\right) - y^*_{i,s} \log\left(\frac{\kappa_{i,s+1} - \kappa_{i,s}}{\kappa_{i,s+1}}\right) + y^*_{i,s+1} \log\left(\frac{\kappa_{i,s+1} - \kappa_{i,s}}{\kappa_{i,s+1}}\right)\right] \\ &= \sum_{s=1}^{m_i-1} \left[y^*_{i,s} \log \frac{\kappa_{is}}{\kappa_{i,s+1} - \kappa_{i,s}} - y^*_{i,s+1} \log \frac{\kappa_{i,s+1}}{\kappa_{i,s+1} - \kappa_{i,s}}\right] \\ &= \sum_{s=1}^{m_i-1} \Big[y^*_{i,s}\theta_{is}(\mathbf{z}) - y^*_{i,s+1} b[\theta_{is}(\mathbf{z})]\Big], \end{aligned} \tag{1.7.11}$$

where each component of (1.7.11) is in the form of the definition of the exponential family distribution with parts:

$$\theta_{is}(\mathbf{z}) = \log \frac{\kappa_{is}}{\kappa_{i,s+1} - \kappa_{i,s}}, \quad s = 1, 2, \ldots, m_i - 1, \tag{1.7.12}$$

and

$$\begin{aligned} b(\theta_{is}(\mathbf{z})) &= \log \frac{\kappa_{i,s+1}}{\kappa_{i,s+1} - \kappa_{i,s}} \\ &= \log \frac{\kappa_{i,s+1} - \kappa_{i,s} + \kappa_{i,s}}{\kappa_{i,s+1} - \kappa_{i,s}} \\ &= \log \left(1 + \frac{\kappa_{i,s}}{\kappa_{i,s+1} - \kappa_{i,s}}\right) \\ &= \log \Big(1 + \exp\{\theta_{is}(\mathbf{z})\}\Big), \quad s = 1, 2, \ldots, m_i - 1. \end{aligned} \tag{1.7.13}$$

The parameter θ_{is} is not a linear function of the latent variable.

Equation (1.7.11) is in the form of a generalised linear model. The maximum likelihood results for a random sample of size n, are based on using the loglikelihood

$$L = \sum_{m=1}^{n} \log f(\mathbf{y}_m) = \sum_{m=1}^{n} \log \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} g(\mathbf{y}_m|\mathbf{z})h(\mathbf{z})d\mathbf{z}, \tag{1.7.14}$$

and the expectation maximisation algorithm.

To score an individual m on the latent dimensions identified by the analysis, Moustaki (2000) proposes to use the mean of the posterior distribution of the latent variable z_j given the individual's response pattern $E(z_j|\mathbf{y}_m)$. In the q-th factor model the posterior mean is given by

$$E(z_j|\mathbf{y}_m) = \int_{R_{z_1}} \cdots \int_{R_{z_q}} z_j k(\mathbf{z}|\mathbf{y}_m)d\mathbf{z}, \tag{1.7.15}$$

$$= \int_{R_{z_1}} \cdots \int_{R_{z_q}} z_j \frac{g(\mathbf{y}_m|\mathbf{z})h(\mathbf{z})}{f(\mathbf{y}_m)}d\mathbf{z}, \tag{1.7.16}$$

where R_{z_j}, with $j = 1, \ldots, q$, denotes the range of values for z_j and $k(\mathbf{z}|\mathbf{y}_m)$ is the posterior distribution of the latent variables given the observed variables for individual m.

An extension of these LVMs in order to include explanatory variables affecting the manifest and latent variables was proposed by Moustaki (2003).

1.8 Conclusion

The discussion about changing the scale of measurement of an ordinal variable in order to carry out statistical analysis has been a never ending controversy. Several scoring systems have been proposed to do so, including the computation of scores based on other variables (as seen in Section 1.4.7), which requires to use multivariate statistical methods. Some of these methods have been extended in order to take into account a special treatment for ordinal variables. In particular, there are regression models for ordinal responses, which are part of the framework of generalised linear models (Sections 1.5 and 1.5.1), from which the proportional odds cumulative logit model presented in Section 1.5.3 is of special interest because it serves as starting point for the proposed model in the next chapter. This

model will be used to deal with the information provided by the order of categories of the response variable, however the treatment of ordinal predictors still needs to be explored. There are some regression models for non-ordinal responses that deal with ordinal predictors (Section 1.6), such as those using penalised maximum likelihood, isotonic regression, and regression models based on constrained maximum likelihood estimation as discussed in Sections 1.6, 1.6.2, and 1.6.3. However, they do not offer a special treatment of ordinal variables at both sides of the regression formula simultaneously and allow one monotonicity direction only.

In these regression models, ordinal variables play the role of being either the response or the predictor(s), i.e., models for both ordinal response and ordinal predictor(s) have not been explicitly discussed. Therefore, in the next chapter, a regression model is proposed for an ordinal response, ordinal predictors, and possibly other types of predictors.

1.9 Thesis structure

The following chapters start with the proposal of a regression model to deal with an ordinal response and ordinal predictors in Chapter 2. In that chapter, not only the use of monotonicity constraints on the parameters of ordinal predictors in a proportional odds cumulative logit model is proposed (see Section 2.3), but also a monotonicity direction classification procedure is developed (see Section 2.4). Given that the parameters associated with an ordinal predictor are estimated under monotonicity constraints, one of the inputs of the model is the pre-specified monotonicity direction of the parameters of every ordinal predictor (isotonic or antitonic). The monotonicity direction classification procedure serves as a tool that allows the researcher to make an informed decision on the direction to be assigned to the parameters of each ordinal predictor when fitting the constrained POCLM.

In Chapter 3, the asymptotic theory of the MLE for the constrained POCLM is developed. Some of the results of the asymptotic theory of the MLE for the unconstrained POCLM are presented in Section 3.2, particularly the log-likelihood ratio test (see Section 3.2.1) and confidence regions (see Section 3.2.2), which will

be used later on. As imposing monotonicity constraints on the parameters associated with the ordinal predictors affects the parameter space, this is analysed in Sections 3.3 and 3.4, which allows further analysis on asymptotic monotonicity direction detection and consistency of the constrained POCLM in Section 3.5. Asymptotic normality of the constrained POCLM is also discussed in Section 3.6. All of these results provide the foundations to analyse asymptotic confidence regions (see Section 3.7) for the parameters of the constrained model. For finite n, the use of approximate confidence regions could be considered as problematic depending on the results of the constrained MLE. Four cases are distinguished and three definitions of confidence regions are provided. These alternative definitions are compared against each other based on a simulation study of coverage probabilities in Section 3.7.1. Asymptotic confidence intervals are discussed in Section 3.8.

In Chapter 4, two monotonicity tests are proposed as a complementary tool to assess the validity of the monotonicity assumption for each ordinal predictor. They both allow to test whether the parameters associated with an ordinal predictor follow a monotonic pattern in the population. One is based on the Bonferroni correction (see Section 4.2) and the other is based on the analysis of confidence regions (see Section 4.3).

Both the MDC procedure proposed in Section 2.4 and the monotonicity tests proposed in Chapter 4 provide statistical evidence on the validity of the monotonicity assumption. This can be incorporated in the estimation procedure. In Chapter 5, different steps of the MDC procedure, together with the monotonicity tests are used to define five estimation methods. They make the decision about the ordinal predictors for which their parameter estimates will not be constrained to be monotonic and then estimate the constrained model. They differ in the way this decision is made, some being more restrictive than others. Two of them use different monotonicity tests and the remaining three use the steps of the MDC procedure in different ways. In addition, the same procedures may also detect that the data are consistent with zero influence of a variable, in which case the variable may be dropped, this is treated in Section 5.4.

In Chapter 6, the model proposed in Chapter 2 and its less restrictive versions proposed in Chapter 5 are compared against each other, against the unconstrained POCLM and against some other approaches based on scoring systems for ordinal predictors presented in Section 1.4. These comparisons will be conducted through simulations and a real data application.

The six constrained methods and the unconstrained one are compared through the analysis of the mean-squared error decomposition in Section 6.2 based on simulations. Furthermore, in Section 6.3 the constrained methods are compared against several models resulting from the unconstrained POCLM using different scoring systems as the treatment of ordinal predictors. Regarding the real data application, the proposed constrained approach is applied to real data from the Chilean National Socio-Economic Characterisation in Section 6.4 to analyse a quality of life self-assessment variable using a 10-Points Likert scale considering ordinal and other predictors. In addition, despite the fact that the transformation of ordinal predictors into interval-scaled variables overstates the information provided by the order of categories of OPs, the results of using the constrained approach are compared against methods using scoring systems for the treatment of ordinal predictors.

Finally, the concluding remarks are presented in Chapter 7, where the main contributions of this thesis are listed in Section 7.1 together with future work in Section 7.2.

Chapter 2

A constrained regression model for an ordinal response with ordinal predictors

2.1 Introduction

In many situations where regression models are suitable, the relationship between ordinal responses and ordinal predictors is of interest. However, statistical modelling for this type of relationship has received little attention. Even literature for ordinal predictors with any other type of scale of the response variable is scarce (see, for example, Tutz and Gertheiss (2014), and Rufibach (2010)).

One usual approach to the treatment of ordinal predictors is to treat them as if they were of nominal scale type, ignoring the information provided by the order of their categories, and another one is to assign numbers to the ordinal categories in order to transform an ordinal predictor into an interval-scaled one, assuming that the categories ordering provides more information than the one that it actually offers. These two common approaches are discussed in Section 2.2 and a constrained regression model for an ordinal response with ordinal predictors and possibly other types of predictors is proposed.

In Section 2.3, the proposed model is developed in detail to obtain both constrained parameter estimates for multiple ordinal predictors and unconstrained

estimates for other types of covariates. As the monotonic estimates can be either increasing (isotonic) or decreasing (antitonic) as the categories of the ordinal predictor increase, it is necessary to specify this relation while defining the constraints. Also, investigating possible directions of monotonicity for all ordinal predictors is of interest in its own right. Therefore, a monotonicity direction classification (MDC) procedure is introduced in Section 2.4 that determines the best possible combination of isotonic and/or antitonic associations as a way of assisting the estimation method of the constrained model introduced in Section 2.3. The way in which the MDC procedure works is discussed in detail through an illustration based on simulated data sets in Section 2.5, where the true parameters' pattern of different OPs will represent different degrees of monotonicity in order to explore the results of the MDC procedure for clear and unclear monotonicity directions. Further analyses of the MDC procedure's performance are left for Chapter 6, where patterns representing none and both monotonicity directions are incorporated in a new set of simulations. The contents of the current chapter were published already in Espinosa and Hennig (2019).

2.2 Ordinal response with ordinal predictors and possibly others

A regression model for an ordinal response with ordinal predictors and possibly others is proposed. In order to account for an ordinal response variable, proportional odds cumulative logit models (McCullagh, 1980) are used here in presence of multiple predictors allowing for different measurement scales. Special attention is paid to the treatment of ordinal-scaled predictors. Their parameter estimates are restricted to be monotonic through constrained maximum likelihood estimation (CMLE). To begin with, consider for simplicity one ordinal response variable y with k categories and one ordinal predictor x with p categories. The corresponding model for this setup is

$$\text{logit}[P(y_i \leq j|x_i)] = \alpha_j + \sum_{h=2}^{p} \beta_h x_{i,h}, \tag{2.2.1}$$

$j = 1, \ldots, k-1$. α_j and β_h for $h = 2, \ldots p$ are real parameters. The observations are $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. The vector $\mathbf{x}_i$ contains the $x_{i,h}$, which are dummy variables defined as $x_{i,h} = 1$ if x_i falls in the h-th category of the ordinal predictor and 0 otherwise, with $h = 2, \ldots, p$. Category number one is treated as the baseline category with $\beta_1 = 0$; therefore the dummy variable $x_{i,1} = 1 - \sum_{h=2}^{p} x_{i,h}$ is omitted and the sum in model (2.2.1) starts at $h = 2$. Monotonicity on $\{\beta_h\}$ is obtained by using CMLE. The general model is defined in Section 2.3, which allows for multiple ordinal predictors and other covariates of different measurement scales.

The monotonic effects approach to the ordinal predictors treatment is conceived here as an intermediate point between two general and common approaches within the context of regression analysis on observed variables. One of these common approaches corresponds to an unconstrained version of (2.2.1), treating the ordinal predictor as if it were nominal. This ignores the ordinal information. The other common approach treats an ordinal predictor as if it were of interval scale type, replacing it by a single transformed variable after applying some scoring method, f. More formally,

$$\text{logit}[P(y_i \leq j | x_i)] = \alpha_j + \beta \tilde{x}_i, \tag{2.2.2}$$

with $\tilde{x} = f(x)$. This treats $f(x)$ as interval-scaled. Numerous data-based methods for scaling of ordinal variables have been proposed in the literature, on top of using plain equidistant Likert scaling (see, e.g., Section 1.4), but ultimately in most situations the data do not carry conclusive information about the appropriateness of any scaling f.

The intermediate approach proposed here is defined to achieve a set of linear estimates described by multiple magnitudes, as in the nominal scale type approach, but allowing one direction only, as in the interval scale type approach. The latter is attained by restricting the effects of the model (2.2.1) to be monotonic in either direction. The monotonicity assumption should not necessarily be taken for granted in regression with ordinal predictor and response. But it has a special status, similarly to linearity between interval-scaled variables. According to Stevens (1946) the interval scale type is defined by the equality in the meaning

of differences between values regardless of the location of these differences on the measurement range. A linear relationship between interval-scaled variables means that the impact of a change in the predictor on the response is proportional to the meaning of the change of measurement at all locations of the measurement scale. For the ordinal measurement scale, only the order of measured values is meaningful. In this case, monotonic relationships are those that imply that a change in the predictor of the same meaning (i.e. changing to a value that is higher, or lower, respectively) at all locations of the measurement scale has an effect of the same meaning on the response.

Some other regression models for ordinal predictors are also based on the monotonic effects assumption. However, models for ordinal responses have not been explicitly discussed in this context. Tutz and Gertheiss (2014) used penalisation methods for modelling rating-scaled variables as predictors (see Section 1.6.1), and an active set algorithm was proposed by Rufibach (2010) to incorporate ordinal predictors in some regression models considering the response variable to be continuous, binary, or represent censored survival times, and assuming isotonic effects of the ordinal predictors' categories (see Section 1.6.3). Another related method is isotonic regression, mostly applied to continuous data (see, for example, Barlow and Brunk (1972), Dykstra et al. (1982), and Stout (2015), also Section 1.6.2). In a broader context, there are some other types of statistical models that deal with ordinal data, such as those in item response theory (IRT) (e.g., Tutz (1990), Bacci et al. (2014)), latent class models such as the one presented in Section 1.7.2 (see also, e.g., Moustaki (2000), Moustaki (2003), Vasdekis et al. (2012)), nonlinear principal components analysis (NLPCA) (e.g., De Leeuw et al. (2009), Linting and van der Kooij (2012) and Mori et al. (2016)), and nonlinear canonical correlation analysis (NLCCA) (e.g., Mardia et al. (1979) and De Leeuw et al. (2009)). However, their settings are somewhat different compared to the one corresponding to modelling an ordinal response with ordinal predictors (and others) in classical regression. For instance, unlike IRT models and latent class models, classical regression models do not assume latent variables; and in contrast to NLPCA and NLCCA, classical regression models are not used as a dimensionality reduction

technique and need a single dependent variable, respectively.

The monotonicity constrained regression model discussed here can be used for several purposes. When the unconstrained parameter estimates associated with the ordinal predictor are monotonic, then clearly there is no need of a constrained model. However, when these unconstrained estimates are not monotonic, then there are some reasons why the constrained model could be useful. It is often of interest to compare unconstrained and constrained fits in order to decide whether there is evidence for not monotonic relationship. In case that the unconstrained version does not provide a clearly better fit, the monotonic fit may be superior regarding interpretability, and may also lead to a smaller mean-squared error, as will be shown by simulations and a real data application in Chapter 6.

2.3 Proportional odds with monotonicity constraints

2.3.1 Model setting

Consider the unconstrained proportional odds cumulative logit model (POCLM). Define $\pi_j(\mathbf{x}_i) = P_{\boldsymbol{\gamma}}(y_i = j|\mathbf{x}_i)$ as the probability of the response of subject i to fall in category j, where y_i and j denote the number of the ordinal category of the response variable $y_i, j \in \{1, 2, \ldots, k\}$ and $\boldsymbol{\gamma}$ is the parameter vector for the probability distribution that will be defined in the end of this section. Let $y_{i1}, \ldots, y_{ik}$ be the binary indicators of the response for subject i, where $y_{ij} = 1$ if its response falls in category j and 0 otherwise. The response vector with these k binary components for the ith subject is $\mathbf{y}_i$. The POCLM for this probability is

$$\pi_j(\mathbf{x}_i) = F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}), \quad j = 1. \ldots, k, \quad i = 1, \ldots, n, \quad (2.3.1)$$

with $F(\zeta) = (1 + e^{-\zeta})^{-1}$, known as the cumulative logistic distribution function. In order that these k probabilities are strictly positive and sum up to one, it is assumed that the intercepts are restricted to be

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} < \alpha_k = \infty. \quad (2.3.2)$$

Model (2.3.1) can also be written in terms of the cumulative probability for the j-th category of the response variable as

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i, \quad j = 1, \ldots, k-1, \ i = 1, \ldots, n, \tag{2.3.3}$$

keeping the restrictions described in (2.3.2).

The predictors are assumed to be fixed and they can be ordinal, for which their parameter estimates will be constrained to account for monotonicity as explained later, and/or non-ordinal. The parameters of the t ordinal and v non-ordinal fixed predictors are contained in $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}'_{(ord)}, \boldsymbol{\beta}'_{(nonord)}\right)$ allocating their parameters in either $\boldsymbol{\beta}_{(ord)}$ or $\boldsymbol{\beta}_{(nonord)}$ correspondingly.

When model (2.3.1) includes t ordinal predictors (OPs), each ordinal predictor is denoted by the subindex s, with $s = 1, \ldots, t$, and contributes $p_s - 1$ dummy variables to the model representing its ordinal categories $\{1, \ldots, p_s\}$ assuming the first one as the baseline category, i.e., $\beta_{s,1} = 0$. The $(\sum_{s=1}^{t}(p_s - 1))$-dimensional parameter vector $\boldsymbol{\beta}_{(ord)}$ contains t vectors $\boldsymbol{\beta}_s$, $s = 1, \ldots, t$, each of which has $p_s - 1$ components representing the parameters associated with the ordinal categories of ordinal predictor s. Note that differences between the regression parameters belonging to the ordinal categories are independent of the baseline category. Confidence intervals (CIs) will be used for these parameters, the widths of which can depend on the baseline category. For ordinal variables, the beginning or end point of the scale seem elementary choices. Each dummy variable is defined as $x_{i,s,h_s} = 1$ if the i-th observation falls in the category h_s of the ordinal predictor s and 0 otherwise, with $h_s = 1, \ldots, p_s$. The model also allows to include v non-ordinal predictors. Therefore, $\mathbf{x}'_i = (x_{i,1,2}, \ldots, x_{i,1,p_1}, x_{i,2,2}, \ldots, x_{i,2,p_2}, \ldots, x_{i,t,2}, \ldots, x_{i,t,p_t}, x_{i,1}, \ldots, x_{i,v})$, where those variables with three indexes correspond to the observation of an ordinal predictor category and those with two are observations of other types of covariates. Hence, the model with t OPs and v non-ordinal predictors is represented as

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \sum_{s=1}^{t}\sum_{p_s=2}^{q_s}\beta_{s,h_s}x_{i,s,h_s} + \sum_{u=1}^{v}\beta_u x_{i,u}, \ j = 1, \ldots, k-1, \ i = 1, \ldots, n, \tag{2.3.4}$$

with restrictions given by (2.3.2). Therefore, the dimensionality of the parameter space is $p = (k-1) + \sum_{s=1}^{t}(p_s - 1) + v$, and putting all the parameters together

the p-dimensional parameter vector is defined as

$$\begin{aligned}\boldsymbol{\gamma}' &= (\boldsymbol{\alpha}', \boldsymbol{\beta}') \\ &= (\boldsymbol{\alpha}', \boldsymbol{\beta}'_{(ord)}, \boldsymbol{\beta}'_{(nonord)}) \\ &= (\boldsymbol{\alpha}', \boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_t, \beta_1, \ldots, \beta_v).\end{aligned}$$

2.3.2 Likelihood model fitting

For independent observations, the likelihood function is based on the product of the multinomial mass functions for the n subjects:

$$\begin{aligned}L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} \pi_j(\mathbf{x}_i)^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} P(y_i = j|\mathbf{x}_i)^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} [P(y_i \leq j|\mathbf{x}_i) - P(y_i \leq j-1|\mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} \left[\frac{e^{\alpha_j + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}}{1 + e^{\alpha_j + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}} \right.\right. \\ &\quad \left.\left. - \frac{e^{\alpha_{j-1} + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}}{1 + e^{\alpha_{j-1} + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}} \right]^{y_{ij}} \right\}. \end{aligned} \tag{2.3.5}$$

Hence,

$$\begin{aligned}\pi_j(\mathbf{x}_i) &= \frac{e^{\alpha_j + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}}{1 + e^{\alpha_j + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}} \\ &\quad - \frac{e^{\alpha_{j-1} + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}}{1 + e^{\alpha_{j-1} + \sum_{s=1}^{t} \sum_{h_s=2}^{p_s} \beta_{s,h_s} x_{i,s,h_s} + \sum_{u=1}^{v} \beta_u x_{i,u}}},\end{aligned} \tag{2.3.6}$$

and the log-likelihood function for the model is

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i). \tag{2.3.7}$$

As a constrained version of model (2.3.4) is of interest with the aim of getting monotonic increasing/decreasing effects for the ordinal predictors, it is necessary to define the set of constraints to be applied on the t sets of p_s coefficients. The isotonic constraints are

$$0 \leq \beta_{s,2} \leq \cdots \leq \beta_{s,p_s}, \quad \forall s \in \mathcal{I}, \tag{2.3.8}$$

where $\mathcal{I} \subseteq \mathcal{S}$, with $\mathcal{S} = \{1, 2, \ldots, t\}$, and $\beta_{s,1} = 0$. The antitonic constraints are

$$0 \geq \beta_{s,2} \geq \cdots \geq \beta_{s,p_s}, \quad \forall s \in \mathcal{A}, \tag{2.3.9}$$

where $\mathcal{A} \subseteq \mathcal{S}$, and $\beta_{s,1} = 0$. An estimation method based on a monotonicity direction classification (MDC) procedure will be discussed in Section 2.4, allocating the ordinal predictors in either of these two subsets, achieving $\mathcal{I} \cup \mathcal{A} = \mathcal{S}$.

These constraints can be expressed in matrix form as $\mathbf{C}\boldsymbol{\beta}_{(ord)} \geq \mathbf{0}$. The vector $\boldsymbol{\beta}_{(ord)}$ is part of the vector $\boldsymbol{\beta}$. The latter contains all the parameters associated with the t ordinal predictors and their $p_s - 1$ categories together with the v non-ordinal predictors, $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}'_{(ord)}, \boldsymbol{\beta}'_{(nonord)}\right)$, with $\boldsymbol{\beta}'_{(ord)} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_t)$ with $s = 1, \ldots, t$, and $\boldsymbol{\beta}'_{(nonord)} = (\beta_1, \ldots, \beta_v)$ with $u = 1, \ldots, v$, where each vector $\boldsymbol{\beta}'_s = (\beta_{s,2}, \ldots, \beta_{s,p_s})$ with $h_s = 2, \ldots, p_s$. The matrix $\mathbf{C}$ is a square block diagonal matrix of $\sum_{s=1}^{t}(p_s - 1)$ dimensions composed of t square submatrices $\mathbf{C}_s$ in its diagonal structure and zeros in its off-diagonal blocks as follows,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{C}_t \end{bmatrix}, \text{ with } s = 1, \ldots, t,$$

where

$$\mathbf{C}_s = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & -1 & 1 \end{bmatrix} \quad \forall s \in \mathcal{I},$$

$$\mathbf{C}_s = \begin{bmatrix} -1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 1 & -1 \end{bmatrix} \quad \forall s \in \mathcal{A},$$

and each square submatrix $\mathbf{C}_s$ has $p_s - 1$ dimensions.

Then, the maximisation problem is

$$\begin{aligned} &\text{maximise } \ell(\{\alpha_j\}, \boldsymbol{\beta}) \\ &\text{subject to } \mathbf{C}\boldsymbol{\beta}_{(ord)} \geq \mathbf{0}, \end{aligned} \tag{2.3.10}$$

where $\mathbf{0}$ is a vector of $\sum_{s=1}^{t}(p_s - 1)$ elements. Now, (2.3.10) can be expressed as the Lagrangian

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\gamma}) - \boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta}_{(ord)}, \tag{2.3.11}$$

where $\boldsymbol{\lambda}$ is the vector of $\sum_{s=1}^{t}(p_s - 1)$ Lagrange multipliers denoted by λ_{s,h_s}.

The set of equations to be solved is obtained by differentiating $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ with respect to its parameters and equating the derivatives to zero. In order to solve this in `R` (R Core Team, 2018), the package `maxLik` (Henningsen and Toomet, 2011) offers the `maxLik` function which refers to `constrOptim2`. This function uses an adaptive barrier algorithm to find the optimal solution of a function subject to linear inequality constraints such as in (2.3.10) (Lange, 2010).

2.4 Monotonicity direction classification procedure

Under the monotonicity assumption for all OPs, an important decision to be made involves their monotonicity direction, i.e., whether the association between the values of each ordinal predictor's set of effects and the ordered categories (also referred to as pattern), is either isotonic, namely $s \in \mathcal{I}$, or antitonic, $s \in \mathcal{A}$. Also outside the context of parameter estimation, it may be of interest whether a predictor is connected to the response in an isotonic or antitonic way, or potentially whether monotonicity may not hold or whether both directions are compatible with the data.

One possible way to deal with this decision is to just maximise the likelihood, i.e., to fit 2^t models, one for each possible combination of monotonicity directions for the t ordinal predictors, and then choose the one with the highest likelihood. However, as the number of ordinal predictors t increases, the number of possible combinations of monotonicity directions becomes greater, which could lead to a considerable number of models to be fitted, each involving a large number of covariates.

Another possible estimation method uses a monotonicity direction classifier to find the monotonicity direction for each ordinal predictor and then fits only

one model. This will be based on CIs for the parameters and on checking which monotonicity direction is compatible with these. In some cases, e.g. when there is an OP for which its parameter estimates are compatible with both monotonicity directions, this estimation method may choose a particular set of monotonicity directions that is not the same as the one of the model with the highest maximum likelihood among models with different combinations of monotonicity directions, but in some situations it may be desirable to take into account fewer than 2^t but more than a single model.

The two approaches are put together in a three steps monotonicity direction classification (MDC) procedure exploiting their best features. Each of the first two steps uses a decision rule with different confidence levels for the CIs, and the last step applies the multiple models fitting process described above over those patterns with no single monotonicity direction established in the previous steps. Before describing its steps, consider some remarks and definitions.

The parameters' CIs from an unconstrained model are the main input for the decision rule proposed here. It is possible to compute the CI defined in equation (2.4.1) for the parameters of an unconstrained version of the model (2.3.4) (Agresti, 2010). Denote $SE_{\hat{\beta}}$ as the standard error of the parameter estimate $\hat{\beta}$, then an approximate confidence interval for β with a $100(1-\tilde{\alpha})\%$ confidence level is

$$\hat{\beta} \pm z_{\tilde{\alpha}/2}(SE_{\hat{\beta}}), \tag{2.4.1}$$

where $z_{\tilde{\alpha}/2}$ denotes the standard normal percentile with probability $\tilde{\alpha}/2$. The values for $\hat{\beta}$ and $SE_{\hat{\beta}}$ are obtained by fitting the proportional odds model (McCullagh, 1980) using the unconstrained model (2.3.4). The `R` function `vglm` of the package `VGAM` was used here, see Yee (2018).

The first two steps of the MDC procedure provide four possible outcomes for each pattern of unconstrained parameter estimates associated with an ordinal predictor's categories: 'isotonic', 'antitonic', 'both', and 'none'. The first two correspond to a classification of monotonicity direction whereas the remaining two to the case where a single direction is not found because either both directions of monotonicity are possible or the parameter estimates' pattern is not compatible

with monotonicity, respectively. The idea is that the intersections of all CIs for the parameters of a single ordinal predictor together will either allow for isotonic but not antitonic parameters, or for antitonic but not isotonic parameters, or for both, or for neither. Formally, the MDC of the parameter estimates' pattern is defined as

$$d_{s,\tilde{c}} = \begin{cases} \text{isotonic} & \text{if } \mathcal{D}_{s,\tilde{c}} = \{0,1\} \text{ or } \mathcal{D}_{s,\tilde{c}} = \{1\} \\ \text{antitonic} & \text{if } \mathcal{D}_{s,\tilde{c}} = \{-1,0\} \text{ or } \mathcal{D}_{s,\tilde{c}} = \{-1\} \\ \text{both} & \text{if } \mathcal{D}_{s,\tilde{c}} = \{0\} \\ \text{none} & \text{if } \mathcal{D}_{s,\tilde{c}} \supseteq \{-1,1\}, \end{cases} \tag{2.4.2}$$

where $\mathcal{D}_{s,\tilde{c}} = \{d_{s,h_s,h'_s,\tilde{c}}\}$ is defined as the set of distinct values resulting from (2.4.3) for the ordinal predictor s considering confidence intervals with a $100\tilde{c}\%$ confidence level, and

$$d_{s,h_s,h'_s,\tilde{c}} = \begin{cases} 1 & \text{if } \tilde{L}_{s,h_s,\tilde{c}} \geq \tilde{U}_{s,h'_s,\tilde{c}} \\ -1 & \text{if } \tilde{U}_{s,h_s,\tilde{c}} \leq \tilde{L}_{s,h'_s,\tilde{c}} \\ 0 & \text{otherwise,} \end{cases} \tag{2.4.3}$$

$\forall h'_s < h_s$ and $\forall h_s \in \{2,3,\ldots,p_s\}$, where $\tilde{U}_{s,h_s,\tilde{c}}$ is the confidence interval's upper bound of the parameter β_{s,h_s} associated with the category h_s of the ordinal predictor s given a $100\tilde{c}\%$ confidence level, and $\tilde{L}_{s,h_s,\tilde{c}}$ is its corresponding lower bound. Note that, by definition, the first category of all ordinal predictors is set to zero, so $\tilde{L}_{s,1,\tilde{c}} = \tilde{U}_{s,1,\tilde{c}} = 0$, $\forall s$. (2.4.3) yields 1 when the CI of the parameter β_{s,h_s} is fully above the one of β_{s,h'_s} and consequently their CIs only allow an isotonic pattern; -1 when it is fully below pointing to an antitonic pattern; and 0 when there exists an overlap, meaning that both monotonicity directions are still possible.

Each result of (2.4.3), denoted as $d_{s,h_s,h'_s,\tilde{c}}$, can be understood as an indicator of the relative position of the confidence interval of the parameter β_{s,h_s} compared to the one of β_{s,h'_s}, $\forall h'_s < h_s$ and $h_s \in \{2,3,\ldots,p_s\}$, belonging to the same ordinal predictor s and given a $100\tilde{c}\%$ confidence level. As this is a pairwise comparison, there exist $p_s(p_s-1)/2$ indicators for each ordinal predictor s. Equation (2.4.2)

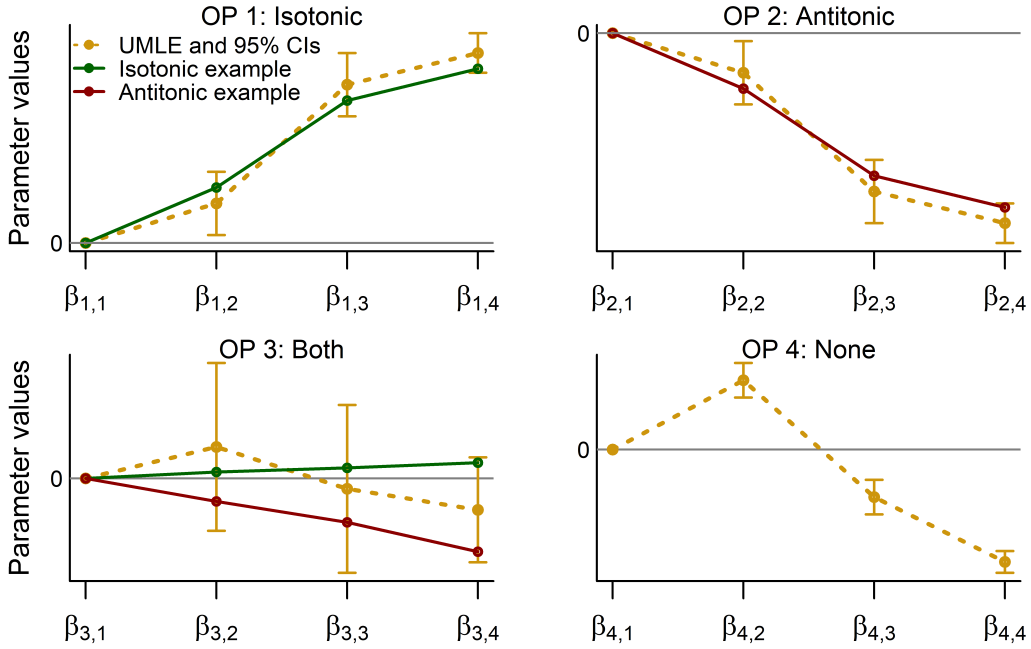

Figure 2.1: Illustration of particular examples for each possible monotonicity direction classification.

uses these indicators to classify the monotonicity direction of an ordinal predictor as a whole at a particular $\tilde{c}$.

As an illustration, Figure 2.1 shows some arbitrary patterns representing a particular example for each one of the possible results of (2.4.2). For instance, OP 1 is classified as 'isotonic' because all but one of the results of (2.4.3) are 1, where the only different is $d_{1,4,3,0.95} = 0$, and therefore $\mathcal{D}_{1,0.95} = \{0, 1\}$. The monotonicity direction of OP 2 is clear also, for which the results of (2.4.3) are -1 except for $d_{2,4,3,0.95} = 0$, with which (2.4.2) classifies this OP as 'antitonic'. All the individual confidence intervals of OP 3 jointly overlap and contain zero. Therefore, $d_{3,h_3,h_3',0.95} = 0\ \forall h_3' < h_3$ and thus $\mathcal{D}_{3,0.95} = \{0\}$, classifying OP 3 as 'both'. Finally, each individual confidence interval associated with the OP 4 is either fully above or fully below the ones of previous categories belonging to the same ordinal predictor. In particular, $\mathcal{D}_{4,0.95} = \{-1, 1\}$ because, for example, $d_{4,2,1,0.95} = 1$ and $d_{4,3,2,0.95} = -1$, which (2.4.2) classifies as 'none'.

The three steps MDC procedure has the following structure:

Step 1 Set $\tilde{c}$ at a relatively high $100\tilde{c}\%$ confidence level, say 0.99, 0.95 or 0.90,

and apply the MDC (2.4.2) to assign the subindexes s either to the set $\mathcal{I}$ or $\mathcal{A}$ defined in Section 2.3.2. Therefore, $\mathcal{I}_1 = \{s : d_{s,\tilde{c}} = \text{isotonic}\}$ and $\mathcal{A}_1 = \{s : d_{s,\tilde{c}} = \text{antitonic}\}$, where $\mathcal{I}_1$ and $\mathcal{A}_1$ denote the isotonic and antitonic sets resulting from the step 1 respectively. In addition, define $\mathcal{B}_1 = \{s : d_{s,\tilde{c}} = \text{both}\}$ and $\mathcal{N}_1 = \{s : d_{s,\tilde{c}} = \text{none}\}$. If $(\mathcal{I}_1 \cup \mathcal{A}_1) = \mathcal{S}$, then all the ordinal predictors' monotonicity directions have been decided, and there is no need to continue with the MDC procedure. Otherwise, the following step is used for the remaining cases only, $(\mathcal{B}_1 \cup \mathcal{N}_1)$.

Step 2 Consider the set of ordinal predictors $\{s : s \in (\mathcal{B}_1 \cup \mathcal{N}_1)\}$ and apply the MDC (2.4.2) in an iterative manner while varying the confidence level $100\tilde{c}\%$. A decrease/increase of $\tilde{c}$ reduces/enlarges the range of the CIs of the parameter β_{s,h_s} $\forall s \in (\mathcal{B}_1 \cup \mathcal{N}_1)$ and $h_s \in \{2, 3, \ldots, p_s\}$. These changes in $\tilde{c}$ produce different effects on the classification depending on whether $s \in \mathcal{B}_1$ or $s \in \mathcal{N}_1$, which must be used as follows:

(a) For each $s \in \mathcal{B}_1$, the second step is to gradually decrease $\tilde{c}$ while applying the decision rule (2.4.2) using a new confidence level $\tilde{c}'_s$ instead of $\tilde{c}$, obtaining $d_{s,\tilde{c}'_s}$. The level of $\tilde{c}'_s$ must be gradually decreased until either a pre-specified minimum confidence level referred to as tolerance level $\tilde{c}'^*_s$ is reached, with $0 < \tilde{c}'^*_s < \tilde{c}$, or the ordinal predictor s is classified as either isotonic or antitonic by $d_{s,\tilde{c}'_s}$.

(b) Conversely, for each $s \in \mathcal{N}_1$, gradually increase $\tilde{c}$ while applying the MDC (2.4.2) using a new confidence level $\tilde{c}''_s$ obtaining $d_{s,\tilde{c}''_s}$. The level of $\tilde{c}''_s$ must be gradually increased until either a pre-specified maximum confidence level referred to as tolerance level $\tilde{c}''^*_s$ is reached, with $\tilde{c} < \tilde{c}''^*_s < 1$, or the ordinal predictor s is classified as either isotonic or antitonic by $d_{s,\tilde{c}''_s}$.

Finally, $\mathcal{I}_2 = \mathcal{I}_1 \cup \{s : d_{s,\tilde{c}'_s} = \text{isotonic or } d_{s,\tilde{c}''_s} = \text{isotonic}\}$ and $\mathcal{A}_2 = \mathcal{A}_1 \cup \{s : d_{s,\tilde{c}'_s} = \text{antitonic or } d_{s,\tilde{c}''_s} = \text{antitonic}\}$, where the subindex of $\mathcal{I}_2$ and $\mathcal{A}_2$ denotes results from the second step. After completing the second step, if

$(\mathcal{I}_2 \cup \mathcal{A}_2) = \mathcal{S}$, then it is not necessary to continue with step 3 and the MDC procedure ends. If $(\mathcal{I}_2 \cup \mathcal{A}_2) \subset \mathcal{S}$, then the third and final step must be carried out.

Step 3 Fit $2^{\#\{s:s\notin(\mathcal{I}_2\cup\mathcal{A}_2)\}}$ models accounting for possible combinations of monotonicity directions of the ordinal predictors that were not classified as 'isotonic' or 'antitonic', i.e., those in the set $\{s : s \notin (\mathcal{I}_2 \cup \mathcal{A}_2)\}$, and choose the best model based on some optimality criterion, such as the maximum likelihood as used here.

In general, the MDC procedure describes two levels of decision. The first one is provided by step 1, where a confidence level is applied to all ordinal predictors by the use of a single parameter $\tilde{c}$. This step uses multiple confidence intervals, which does not necessarily assure that overall confidence levels will be kept. The analysis of multiple confidence intervals is used in the MDCP as a tool for decision making based on heuristic ideas. The second levels of decision is in step 2, where each ordinal predictor $s \in (\mathcal{B}_1 \cup \mathcal{N}_1)$ is classified based on its own confidence level. Step 2 allows to classify predictors that were not classified based on the fixed initial confidence level.

In step 2, classifying more parameter estimates' patterns with $s \in \mathcal{B}_1$ as either isotonic or antitonic requires a gradual reduction of the confidence level. The tolerance levels $\tilde{c}'^{*}_s$ and $\tilde{c}''^{*}_s$ determine the leeway allowed for the confidence levels in order to enforce a decision. The choice of these may depend on the number of ordinal variables; if the number is small, running step 3 may not be seen as a big computational problem, and it may not be necessary to enforce many decisions in step 2. The tolerance level $\tilde{c}'^{*}_s$ should not be too low, less than 0.8, say, because it is not desirable to make decisions based on a low probability of occurrence.

For those $s \in \mathcal{N}_1$ in step 2, the researcher does not face such a trade-off, because greater confidence levels could increase (not decrease) the number of new isotonic or antitonic classifications for those $s \in \mathcal{N}_1$.

It is important to reduce (or increase) the confidence level in step 2 in a gradual manner, by 0.01 or 0.005, say, for each iteration. The smaller the distance between

parameter estimates for adjacent categories of an OP, the smaller the size of the reduction (or increase). If the chosen intervals in the sequence of confidence levels to be assessed are too wide without assessing intermediate levels, then, for an ordinal predictor $s \in \mathcal{B}_1$, it is possible to switch its classification from 'both' to 'none' instead of updating it from 'both' to either 'isotonic' or 'antitonic'. Conversely, the class of an ordinal predictor $s \in \mathcal{N}_1$ could change from 'none' to 'both'. The thinner the intervals in the sequence of confidence levels to be assessed are, the less likely it is to switch from 'both' to 'none' or 'none' to 'both'. However, in some specific cases, there still is a probability of having such an undesired class change.

The researcher may also be interested in exploring other monotonicity directions rather than those resulting from the MDC procedure proposed here, although the maximum likelihood attained by the MDC procedure would not be reached. In this case, the correspondence of each ordinal predictor s to either $\mathcal{I}$ or $\mathcal{A}$ should simply be enforced when constructing $\mathbf{C}$, the matrix of constraints, as described in Section 2.3.2.

2.5 Illustration of the MDC procedure

In order to illustrate the MDC procedure, consider a particular example of model (2.3.4) with four ordinal predictors only ($t = 4$ and $v = 0$), where $p_1 = 3$, $p_2 = 4$, $p_3 = 5$, $p_4 = 6$, and $k = 4$, i.e., $j = 1, 2, 3$. The parameters are chosen to be $\alpha_1 = -1$, $\alpha_2 = -0.5$, and $\alpha_3 = -0.1$; and

$$\begin{aligned}
\boldsymbol{\beta}_1' &= (1.0, 1.5), \\
\boldsymbol{\beta}_2' &= (0.1, 0.2, 0.25), \\
\boldsymbol{\beta}_3' &= (-0.02, -0.04, -0.041, -0.05), \text{ and} \\
\boldsymbol{\beta}_4' &= (-0.2, -0.3, -0.31, -0.35, -0.36).
\end{aligned}$$

These parameters represent a situation in which all covariates are monotonic, with the elements of $\boldsymbol{\beta}_1'$ and $\boldsymbol{\beta}_2'$ being isotonic, and those of $\boldsymbol{\beta}_3'$ and $\boldsymbol{\beta}_4'$ antitonic patterns. Given monotonicity, the higher the distances between adjacent parameters are, the clearer the monotonicity direction is. In this illustration, these distances were

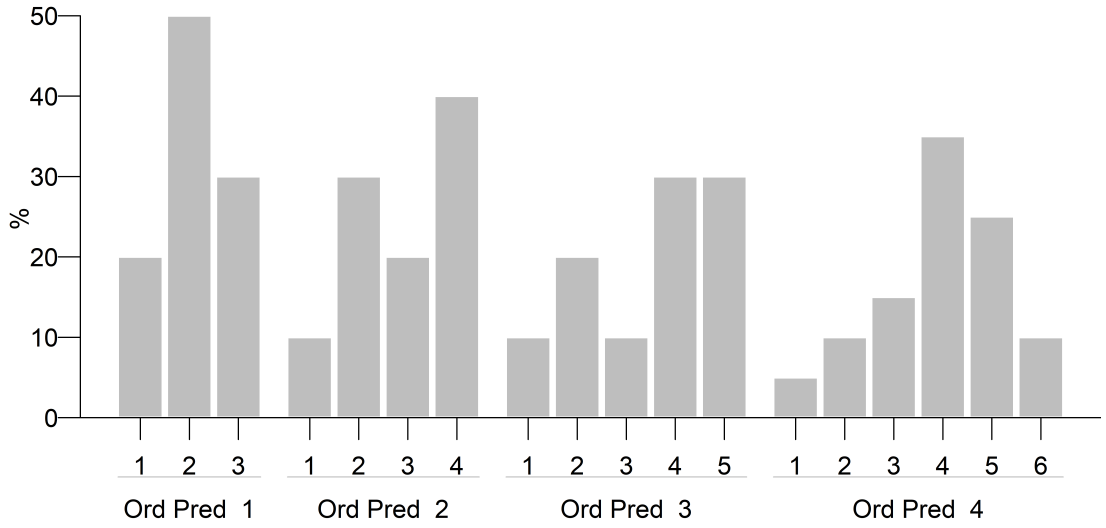

Figure 2.2: Distribution of ordinal categories for each simulated ordinal predictor (OP). OP 1 has 3 categories, OP 2 has 4, OP 3 has 5 and OP 4 has 6.

chosen to make the monotonicity direction clear for the first ordinal predictor only and less clear for the remaining ones, $s = 3$ being the most unclear and challenging case because all of its parameters show little distance between adjacent categories and consequently from zero.

The 2,000 simulated observations of the ordinal predictors were obtained from the population distributions shown in Figure 2.2.

Using this simulated data set, an unconstrained version of the model was fitted to obtain the parameter estimates and their standard errors, with which a confidence interval can be computed for any level of $\tilde{\alpha}$ using equation (2.4.1).

For the first step of the MDC procedure, the confidence level was set at a high $\tilde{c} = 0.99$. The resulting confidence intervals allowed to classify the first and second OP as 'isotonic', $\mathcal{I}_1 = \{1, 2\}$, and the remaining two patterns of parameter estimates as 'both', $\mathcal{B}_1 = \{3, 4\}$. Figure 2.3 shows that the latter two ordinal predictors allowed both directions of monotonicity, which is the reason why they were not classified as 'antitonic'. The second step was applied over each ordinal predictor $s \in \mathcal{B}_1 = \{3, 4\}$ using the same tolerance level, $\tilde{c}_3'^* = \tilde{c}_4'^* = 0.8$. For $s = 3$, it was not possible to classify its pattern as 'antitonic' before reaching the tolerance level. Therefore, it remained as 'both'. For $s = 4$, the procedure was

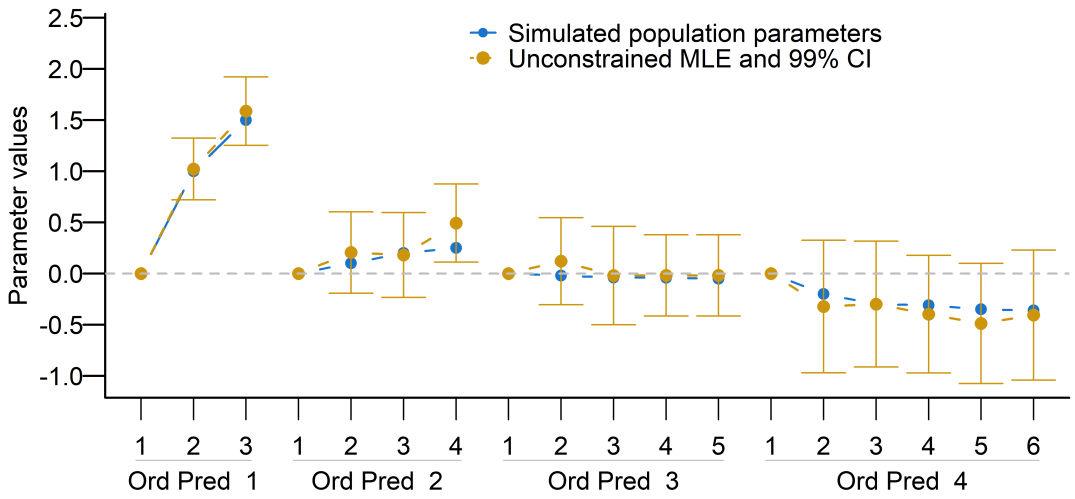

Figure 2.3: Simulated population parameters used in the data generation process (blue lines and dots) and unconstrained parameter estimates for ordinal predictors' categories and their 99% confidence intervals (golden lines and dots).

applied until reaching $\tilde{c}'_s = 0.96$, where the fourth OP was classified as 'antitonic'. Now, $\mathcal{I}_2 = \{1, 2\}$ and $\mathcal{A}_2 = \{4\}$. As no monotonicity direction was identified for the third OP, two models were fitted in step 3 of the MDC procedure, one treating the third OP as 'isotonic' and the other one as 'antitonic'. Finally, the model with the highest log-likelihood was selected as the final one.

The procedure successfully classified the ordinal predictors $s = 1, 2, 3, 4$ as 'isotonic', 'isotonic', 'antitonic', and 'antitonic', respectively, despite the fact that the unconstrained parameter estimates of the last three are not monotonic. Furthermore, it reduced the number of possible models to be fitted from 17 (the unconstrained model and 16 constrained models) to 3 (the unconstrained and two models in step 3) while making decisions based on individual confidence levels of 96% or greater.

As shown in Figure 2.3, it is not easy to classify cases like $s = 3$ where all the parameter estimates are close to zero and their confidence intervals are big enough to make the monotonicity direction classification infeasible for any reasonable tolerance level. In this case, the tolerance level would have needed to be set at $\tilde{c}'^{*}_3 \leq 0.53$ had we wanted the MDC procedure to classify the third ordinal pre-

dictor as either 'isotonic' or 'antitonic'. In fact, when doing so, the MDC makes a mistake and classifies it as 'isotonic'. This relationship between low tolerance levels and misclassification is the main reason why the procedure needs to start with a relatively high confidence level $\tilde{c}_s$ and then gradually decrease it until reaching a reasonable tolerance level if necessary.

In cases like $s = 3$, one option is to remove this variable from the model because all of the CIs associated with it contain zero even if we choose a tolerance level lower than 0.80, which we consider too low. Removing this variable would have allowed us to fit just two models (the unconstrained and one constrained) instead of three in the whole procedure. This is a valid decision to be made from the point of view of a variable selection procedure based on the significance of the parameters associated with the variable. However, the objective of the analysis should also be considered, which could lead to making a different decision, for instance, keeping the variable because removing it may not be good if the aim is to obtain a model with optimal predictive power.

2.6 Conclusions

A constrained regression model for an ordinal response with ordinal predictors and possibly other type of predictors is proposed in Section 2.3. The ordinal response variable is properly treated by using an existing model for ordinal responses, the POCLM. The ordinal predictors are not treated as nominal-scaled variables neither they are treated as interval-scaled variables. Therefore, the information provided by the order of their categories is neither ignored nor overstated. A set of $(p_s - 1)$ dummy variables for each ordinal predictor and constrained maximum likelihood estimation are used, avoiding transformations of ordinal predictors and exploiting the information provided by their category ordering.

The proposed model allows the researcher to fit a regression model for an ordinal response imposing monotonicity constraints on the effects of OPs assuming any monotonicity direction. This means that the proposed regression model does not require a tuning parameter. However, the researcher could also be interested in some method that delivers the monotonicity directions to be imposed, for which

the MDC procedure defined in Section 2.4 is proposed.

The MDC procedure relies on the analysis of the confidence intervals of unconstrained parameters, for which the confidence levels must be set beforehand. The pattern of parameters associated with an ordinal predictor's categories is automatically classified as isotonic or antitonic by the MDC procedure after the execution of its three steps. Intermediate steps (steps 1 and 2), can classify a pattern not only as isotonic or antitonic, but also as compatible with both monotonicity directions or none, providing two levels of decision. This feature will be used to define part of a set of estimation methods proposed in Chapter 5, and also could be used as a variable selection method, which will be discussed in the same chapter.

Given monotonicity of effects, the MDC procedure delivers the model that maximises the likelihood among all possible combinations of monotonicity directions. The results of simulations discussed in Section 2.5 show that steps 1 and 2 of the MDC procedure reduce the number of possible combinations of monotonicity directions to be assumed to obtain the constrained maximum likelihood estimates, which turns out to be an increasingly valuable feature as the number of OPs in the model increases.

Chapter 3

Asymptotics of the MLE for the POCLM

3.1 Introduction

In order to study the asymptotic theory of the constrained MLEs for the POCLM, some of the unconstrained results are used. The log-likelihood ratio test and confidence regions for the unconstrained MLE (UMLE) of the POCLM are presented in Section 3.2 (see Sections 3.2.1 and 3.2.2 correspondingly). The latter will be considered in the discussion about confidence regions for constrained MLE (CMLE) in Section 3.7, which is part of the results of the analysis of asymptotic theory for the POCLM under monotonicity constraints on the coefficients of ordinal predictors.

When imposing monotonicity constraints on the effects associated with ordinal predictors, the parameter space of the UMLE is reduced to the one of the CMLE. This is analysed in Section 3.3, where, in addition, the likelihood function of the POCLM and its logarithm are proved to be continuous and differentiable in the constrained and unconstrained parameter spaces. Openness and convexity of the parameter space of the constrained POCLM are discussed in Section 3.4.

In Section 3.5, it will be shown that, asymptotically, the true monotonicity direction classification corresponds to the one of the MLEs with probability one (see Section 3.5.4) and that consistency holds for the constrained MLEs of the

POCLM (see Section 3.5.5). For this purpose, some other results must be discussed earlier. Consistency of GLMs with natural link function will be analysed based on Fahrmeir and Kaufmann (1985) (see Section 3.5.2). However, the link function of the POCLM is non-natural. Therefore, an extension to the case of the general link function (natural and non-natural) will be made explicit in Section 3.5.3. Given that some general setting of the generalised linear models will also be required, it is presented in Section 3.5.1. Asymptotic normality for the CMLE is discussed in Section 3.6, as a continuation of the results about consistency.

The definition of asymptotic confidence regions for the CMLE of the POCLM is analysed in Section 3.7, where four cases are distinguished and discussed depending on whether the UMLE are the same as the CMLE and on whether the confidence region is compatible with either one combination of monotonicity directions, multiple combinations of monotonicity directions, or even non-monotonicity. Three definitions of confidence regions are proposed taking into account the four cases. For some of them, using asymptotic theory could be problematic. The reasons why they could be considered as problematic and their implications will also be explored in terms of their confidence regions' coverage probability (see Section 3.7.1).

Finally, asymptotic confidence intervals for the CMLE of the POCLM are analysed in Section 3.8. For similar reasons to the ones of asymptotic confidence regions, approximate confidence intervals could also be problematic. Some or all of the values that are part of non-monotonic parameter vectors could belong to individual confidence intervals. Therefore, these values should be removed from confidence intervals associated with constrained parameters. However, given that monotonicity is a feature of a parameter vector (not of a single parameter), then each individual confidence interval does not provide information about monotonicity. This means that it is not possible to identify all those values that are part of non-monotonic parameter vectors by analysing individual confidence intervals. This problem is discussed in Section 3.8.

3.2 Unconstrained POCLM

Before analysing the asymptotic theory of the constrained MLEs for the POCLM, some unconstrained results are presented. In particular, these are the log-likelihood ratio test and confidence regions. These will be used in the discussion of one of the main results of this chapter, the definition of confidence regions for constrained MLEs for the POCLM in Section 3.7.

3.2.1 The log-likelihood ratio test

Recall the log-likelihood function defined in (2.3.7) for the unconstrained POCLM using $\boldsymbol{\gamma}' = (\{\alpha_j\}, \boldsymbol{\beta}')$,

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i), \tag{3.2.1}$$

with $\pi_j(\mathbf{x}_i)$ as defined in (2.3.6).

Let the parameter vector of model (2.3.4) belong to a p-dimensional space called U_{UM} and defined as

$$U_{UM} = \{\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_t', \boldsymbol{\beta}_{(nonord)}') \in \mathcal{R}^p : -\infty < \alpha_1 < \cdots < \alpha_{k-1} < \infty\}, \tag{3.2.2}$$

where $p = k - 1 + \sum_{s=1}^{t}(p_s - 1) + v$. The unconstrained maximum likelihood is

$$\ell(\hat{\boldsymbol{\gamma}}) = \max_{\boldsymbol{\gamma} \in U_{UM}} \ell(\boldsymbol{\gamma}), \tag{3.2.3}$$

then $\hat{\boldsymbol{\gamma}}$ is the vector of unconstrained maximum likelihood estimators belonging to the parameter space U_{UM}.

Define $r = p - q$, with $0 \leq q < p$, as the number of parameter values to be tested in the log-likelihood ratio test. When conducting a hypothesis test on one, some, or all of the parameter values, then $\boldsymbol{\gamma}$ is partitioned into two components so that $\boldsymbol{\gamma}' = (\boldsymbol{\beta}_r', \boldsymbol{\phi}')$. Denote as $U_{UM,0}$ the parameter space for which the r parameters of $\boldsymbol{\beta}_r$ will be tested to be $\boldsymbol{\beta}_{r,0}$, i.e., $\boldsymbol{\beta}_r = \boldsymbol{\beta}_{r,0}$, then the maximum likelihood is

$$\ell(\boldsymbol{\beta}_r, \tilde{\boldsymbol{\phi}}) = \max_{(\boldsymbol{\beta}_r, \boldsymbol{\phi}) \in U_{UM,0}, \boldsymbol{\beta}_r = \boldsymbol{\beta}_{r,0}} \ell(\boldsymbol{\beta}_r, \boldsymbol{\phi}). \tag{3.2.4}$$

and therefore $\tilde{\boldsymbol{\phi}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$ for fixed $\boldsymbol{\beta}_r = \boldsymbol{\beta}_{r,0}$, all of them in $U_{UM,0}$.

Given that $U_{UM,0}$ is a subspace of U_{UM}, then

$$\ell(\hat{\boldsymbol{\gamma}}) - \ell(\boldsymbol{\beta}_r, \tilde{\boldsymbol{\phi}}) \geq 0,$$

representing the amount by which the log-likelihood decreases as a result of testing r parameters at specific values. When this difference is large enough, $\ell(\hat{\boldsymbol{\gamma}}) - \ell(\boldsymbol{\beta}_r, \tilde{\boldsymbol{\phi}}) \geq w$ (with large enough positive w), then the impact of fixing some values in the parameter space suggests that the model under the null hypothesis that $(\{\alpha_j\}, \boldsymbol{\beta}') \in U_{UM,0}$ is a poor description of the data. The point now is how to determine w. Wilks (1938) found that:

$$W \equiv 2[\ell(\hat{\boldsymbol{\gamma}}) - \ell(\boldsymbol{\beta}_r, \tilde{\boldsymbol{\phi}})] \sim \chi^2_{(p-q)}, \tag{3.2.5}$$

where W is the Wilks' log-likelihood ratio statistic. Therefore, using a significance level α, the researcher can set $w = \chi^2_{(p-q);1-\alpha}$, where $\chi^2_{(p-q);1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2 distribution with $(p-q)$ degrees of freedom, and reject $H_0 : \boldsymbol{\gamma} \in U_{UM,0}$ if

$$2[\ell(\hat{\boldsymbol{\gamma}}) - \ell(\boldsymbol{\beta}_r, \tilde{\boldsymbol{\phi}})] > \chi^2_{(p-q);1-\alpha}. \tag{3.2.6}$$

These hypothesis tests can be used to construct confidence regions. In general, according to Lehmann and Romano (2005), Section 3.5, a confidence region is the totality of parameter values for which the null hypothesis is not rejected when the data are observed, which will be used in the following sections. In particular for the unconstrained version of the POCLM, McCullagh and Nelder (1989), p.473, includes a discussion of the likelihood ratio for GLMs and presents its use to construct confidence regions.

3.2.2 Confidence regions

Confidence region for all p parameters

The likelihood ratio test can be used to construct a confidence region for the p-dimensional parameter vector $(\{\alpha_j\}, \boldsymbol{\beta}')$. The confidence region is composed of all those p-dimensional points $(\{\alpha_{j0}\}, \boldsymbol{\beta}'_0)$ for which their log-likelihood ratio test

does not reject $H_0 : (\{\alpha_j\}, \boldsymbol{\beta}') = (\{\alpha_{j0}\}, \boldsymbol{\beta}_0')$, namely for those with a log-likelihood (3.2.3) that is not significantly greater than the maximum log-likelihood (3.2.4). Formally, a p-dimensional approximate $100(1-\alpha)\%$ confidence region is usually expressed in some of the following alternative ways

$$\begin{aligned}
\text{CR}_{\text{all}} &= \left\{(\{\alpha_{j0}\}, \boldsymbol{\beta}_0') : 2[\ell(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}}) - \ell(\{\alpha_{j0}\}, \boldsymbol{\beta}_0)] \leq \chi^2_{(p);1-\alpha}\right\} \\
&= \left\{(\{\alpha_{j0}\}, \boldsymbol{\beta}_0') : \log \frac{L(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})}{L(\{\alpha_{j0}\}, \boldsymbol{\beta}_0)} \leq \frac{1}{2}\chi^2_{(p);1-\alpha}\right\} \\
&= \left\{(\{\alpha_{j0}\}, \boldsymbol{\beta}_0') : \frac{L(\{\alpha_{j0}\}, \boldsymbol{\beta}_0)}{L(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})} > e^{-\frac{1}{2}\chi^2_{(p);1-\alpha}}\right\} \qquad (3.2.7)
\end{aligned}$$

where the degrees of freedom are p.

In particular, when $p = 2$

$$e^{-\frac{1}{2}\chi^2_{(p);1-\alpha}} = \alpha,$$

and then the border of the 2-dimensional approximate $100(1-\alpha)\%$ confidence region is depicted by the contour $\{(\{\alpha_{j0}\}, \boldsymbol{\beta}_0') : L(\{\alpha_{j0}\}, \boldsymbol{\beta}_0) = \alpha L(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})\}$.

Confidence region for a subset of parameters

The previous sub-section discussed the use of confidence regions where r, the number of parameters being set, is p. This section extends the method to construct a confidence region when $1 \leq r \leq p$.

Define a vector with r parameters of interest as $\boldsymbol{\beta}_r$, then the overall parameter vector $\boldsymbol{\gamma}' = (\{\alpha_j\}, \boldsymbol{\beta}')$ is partitioned as $\boldsymbol{\gamma}' = (\boldsymbol{\beta}_r', \boldsymbol{\phi}')$, where $\boldsymbol{\phi}'$ is the transpose of a vector $\boldsymbol{\phi}$ with the remaining $(p-r)$ parameters. The corresponding (unconstrained) MLEs of $\boldsymbol{\beta}_r$ and $\boldsymbol{\phi}$ are now denoted as $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\phi}}$ accordingly.

The confidence region for the parameter vector $\boldsymbol{\beta}_r$ can be constructed by:

$$\text{CR}_{\text{updated MLE}} = \left\{\boldsymbol{\beta}_{0r} : 2[\ell(\hat{\boldsymbol{\beta}}_r, \hat{\boldsymbol{\phi}}) - \ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}})] \leq \chi^2_{(r);1-\alpha}\right\} \qquad (3.2.8)$$

where the degrees of freedom are r because it is the number of parameter values that are being tested to be $\boldsymbol{\beta}_{0r}$, and $\tilde{\boldsymbol{\phi}}$ is the vector of maximum likelihood estimators as a function of the value of $\boldsymbol{\beta}_{0r}$, where $\tilde{\boldsymbol{\phi}}$ is defined by $\ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}}) = \max_{(\boldsymbol{\beta}_r, \boldsymbol{\phi}) \in U_{UM}, \boldsymbol{\beta}_r = \boldsymbol{\beta}_{0r}} \ell(\boldsymbol{\beta}_r, \boldsymbol{\phi})$ for each value of $\boldsymbol{\beta}_{0r}$. Therefore, $\tilde{\boldsymbol{\phi}}$ can be thought of as the updated MLE of $\boldsymbol{\phi}$ for each value of $\boldsymbol{\beta}_{0r}$. If $r = p$, $\boldsymbol{\beta}_r$ is p-dimensional, the terms $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\phi}}$ are omitted and in terms of notation $\boldsymbol{\beta}$ is replaced by $\boldsymbol{\gamma}$.

3.3 Monotonicity constraints and parameter space

When ordinal predictors are treated as of nominal scale type, the parameter space is the set U_{UM} defined in 3.2.2 without imposing monotonicity constraints on any of the t vectors $\boldsymbol{\beta}_s$, $s = 1, \ldots, t$. This is why the subscript UM is used in (3.2.2), which stands for unconstrained model.

The proposed treatment of an ordinal predictor s is to impose monotonicity constraints on its $p_s - 1$ parameters. The isotonic constraints for those parameters are

$$0 \leq \beta_{s,2} \leq \cdots \leq \beta_{s,p_s}, \quad \forall s \in \mathcal{I}, \tag{3.3.1}$$

where $\mathcal{I} \subseteq \mathcal{S}$, with $\mathcal{S} = \{1, 2, \ldots, t\}$, and the antitonic constraints are

$$0 \geq \beta_{s,2} \geq \cdots \geq \beta_{s,p_s}, \quad \forall s \in \mathcal{A}, \tag{3.3.2}$$

where $\mathcal{A} \subseteq \mathcal{S}$.

Despite the fact that model (2.3.1) is constrained on its intercepts with (2.3.2), it will be referred to as the unconstrained model, and the model under monotonicity constraints (3.3.1) and/or (3.3.2) will be regarded as the constrained model and its parameter space will be studied in this section. In this chapter the asymptotic theory of the MLEs for the POCLM will be addressed assuming that the true parameters of each ordinal predictor is strictly monotonic, dropping the equality sign of constraints (3.3.1) and (3.3.2), keeping their corresponding inequalities unaltered.

The parameter set of model (2.3.1) under monotonicity constraints on the parameters of its ordinal predictors with *undetermined* monotonicity directions is defined as

$$\begin{aligned} \tilde{U}_{CM} = \{\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_t', \boldsymbol{\beta}_{(nonord)}') \in \mathcal{R}^p : -\infty < \alpha_1 < \cdots < \alpha_{k-1} < \infty, \\ [(\beta_{s,2} > 0, \beta_{s,h_s} > \beta_{s,h_s-1}) \text{ or} \\ (\beta_{s,2} < 0, \beta_{s,h_s} < \beta_{s,h_s-1})] \quad \forall (s, h_s) \in \mathcal{S} \times \{3, \ldots, p_s\}\}. \end{aligned} \tag{3.3.3}$$

When the monotonicity directions are *established*, then the admissible set of the parameter space of model (2.3.1) under strict monotonicity constraints on its

ordinal predictors is defined as

$$U_{CM} = \{\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_t', \boldsymbol{\beta}_{(nonord)}') \in \mathcal{R}^p : -\infty < \alpha_1 < \cdots < \alpha_{k-1} < \infty,$$
$$(\beta_{s,2} > 0, \beta_{s,h_s} > \beta_{s,h_s-1}) \quad \forall (s, h_s) \in \mathcal{I} \times \{3, \ldots, p_s\},$$
$$(\beta_{s,2} < 0, \beta_{s,h_s} < \beta_{s,h_s-1}) \quad \forall (s, h_s) \in \mathcal{A} \times \{3, \ldots, p_s\}\}. \tag{3.3.4}$$

Therefore, the parameter space with established monotonicity directions, U_{CM}, is a subset of the one with undetermined monotonicity directions, $\tilde{U}_{CM}$, i.e., $U_{CM} \subset \tilde{U}_{CM}$.

The following two propositions analyse the likelihood function and its logarithmic version in terms of their continuity and differentiability at the values of the parameter vectors in the three different sets presented before: U_{UM}, $\tilde{U}_{CM}$ and U_{CM}.

Proposition 3.1. *$L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x})$ and its logarithm are continuous at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM}.* ■

The proof of Proposition 3.1 is based on the following continuity rules presented in Haggarty (1993):

- Sum rule: If f and g are continuous at x_0 then $f + g$ is continuous at x_0.
- Product rule: If f and g are continuous at x_0 then $f \cdot g$ is continuous at x_0.
- Reciprocal rule: If f is continuous at x_0 and $f(x_0) \neq 0$ then $1/f$ is continuous at x_0.

Proof. Since $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$, the likelihood function (2.3.5) can be rewritten as

$$L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} \left[\frac{e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}} \right]^{y_{ij}} \right\}. \tag{3.3.5}$$

Consider the function $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i) = \frac{e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}$. Given $\mathbf{x}_i$, the exponential function is continuous in $\boldsymbol{\gamma}$, therefore the numerator is continuous. By the reciprocal rule, the inverse of $(1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i})$ is also continuous. Therefore, by the product rule, $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i)$ is continuous. Analogously, so it is $z(\alpha_{j-1}, \boldsymbol{\beta}|\mathbf{x}_i)$. By the sum

rule $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i) - z(\alpha_{j-1}, \boldsymbol{\beta}|\mathbf{x}_i)$ is also continuous and the result of the remaining operations in (3.3.5) is continuous by the product rule.

Given that the logarithmic function is continuous, then $\log L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x})$ is also continuous at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM}. ■

Similarly, it can be shown that the likelihood function of the POCLM and its logarithm are differentiable at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM}.

Proposition 3.2. *$L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x})$ and its logarithm are differentiable at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM}.* ■

The proof of Proposition 3.2 is based on the following differentiability rules presented in Haggarty (1993):

- Sum rule: If f and g are differentiable at x_0 then $f + g$ is differentiable at x_0.
- Product rule: If f and g are differentiable at x_0 then $f \cdot g$ is differentiable at x_0.
- Reciprocal rule: If f is differentiable at x_0 and $f(x_0) \neq 0$ then $1/f$ is differentiable at x_0.

Proof. Since $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$, the likelihood function (2.3.5) can be rewritten as in Equation (3.3.5).

Consider the function $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i) = \frac{e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}$. Given $\mathbf{x}_i$, the exponential function is differentiable in $\boldsymbol{\gamma}$, therefore the numerator is differentiable. By the reciprocal rule, the inverse of $(1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i})$ is also differentiable. Therefore, by the product rule, $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i)$ is differentiable. Analogously, so it is $z(\alpha_{j-1}, \boldsymbol{\beta}|\mathbf{x}_i)$. By the sum rule $z(\alpha_j, \boldsymbol{\beta}|\mathbf{x}_i) - z(\alpha_{j-1}, \boldsymbol{\beta}|\mathbf{x}_i)$ is also differentiable and the result of the remaining operations in (3.3.5) is differentiable by the product rule.

Given that the logarithmic function is differentiable, then $\log L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x})$ is also differentiable in $\boldsymbol{\gamma}$. ■

A useful result about continuity is stated by the next theorem:

Theorem 3.3. *If f is differentiable at x_0 then f is continuous at x_0.*

See Haggarty (1993), p.169, for its proof. Theorem 3.3 will be referred later when statements about continuity of $L(\boldsymbol{\gamma}|\mathbf{y},\mathbf{x})$ or its logarithm at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM} are made. Given that Proposition 3.2 proves that $L(\boldsymbol{\gamma}|\mathbf{y},\mathbf{x})$ and its logarithm are differentiable at $\boldsymbol{\gamma} \in U_{UM}$, $\tilde{U}_{CM}$ or U_{CM}, then, by Theorem 3.3, continuity also holds for those functions at those $\boldsymbol{\gamma}$. Also Proposition 3.2 addresses differentiability.

3.4 Parameter space of the constrained POCLM

The development of asymptotic inference for the unconstrained POCLM discussed in Fahrmeir and Kaufmann (1986) assumes that the parameter set U_{UM} is in $\mathcal{R}^p$, *nonvoid*, *open*, and *convex*. U_{UM} is constrained by (2.3.2), which operates on the intercepts only. Asymptotic inference for the constrained POCLM, which also constraints the coefficients associated with the ordinal predictors, is discussed under strict monotonicity assumptions, dropping the equality sign in (3.3.1) and (3.3.2). Openness and convexity will be discussed under the context of the constrained MLE for the POCLM in the following subsections.

3.4.1 Openness of the parameter space of the constrained POCLM

In order to prove that the parameter space U_{CM} of model (2.3.4) is still an open set after imposing monotonicity constraints on its ordinal predictors, the following definitions taken from Protter et al. (2012) will be used:

Definition 3.4 (Metric space, Distance function)**.** Let S be a set and suppose d is a function with domain consisting of all pairs of points of S and with range in $\mathcal{R}^1$. That is, d is a function from $S \times S$ into $\mathcal{R}^1$. We say that S and the function d form a **metric space** when the function d satisfies the following conditions:

(i) $d(x,y) \geq 0$ for all $(x,y) \in S \times S$; and $d(x,y) = 0$ if and only if $x = y$.

(ii) $d(y,x) = d(x,y)$ for all $(x,y) \in S \times S$.

(iii) $d(x,z) \leq d(x,y) + d(y,z)$ for all x, y, z in S. (Triangle inequality.)

The function d satisfying conditions (i), (ii), and (iii) is called the **metric** or **distance function** in S. Hence a metric space consists of the pair (S,d). ■

Definition 3.5 (Open ball, closed ball)**.** Let v_0 be an element of S, a metric space, and suppose r is a positive number. $d(v,v_0)$ is assumed to be the Euclidean distance between v and v_0. The **open ball with center at** v_0 **and radius** r is the set $B(v_0,r)$ given by

$$B(v_0,r) = \{v \in S : d(v,v_0) < r\}.$$

The **closed ball with center at** v_0 **and radius** r is the set $\overline{B(v_0,r)}$ given by

$$\overline{B(v_0,v)} = \{v \in S : d(v,v_0) \leq r\}.$$

■

Definition 3.6 (Open set, closed set)**.** A set A in a metric space S is **closed** if and only if A contains all of its limit points.

A set A in a metric space S is **open** if and only if each point p_0 in A is the center of an open ball $B(p_0,r)$ which is contained in A. That is, $B(p_0,r) \subset A$. It is important to notice that the radius r may change from point to point in A. ■

The only difference between the parameter space of the unconstrained model and the one of the constrained model lies in the parameter space of the parameters associated with the ordinal predictors, which are assumed to be restricted under strict monotonicity with a pre-determined monotonicity direction for each ordinal predictor s.

Proposition 3.7. *The set*

$$\begin{aligned} U_{CM} = \{\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_t', \boldsymbol{\beta}_{(nonord)}') \in \mathcal{R}^p : -\infty < \alpha_1 < \cdots < \alpha_{k-1} < \infty, \\ \beta_{s,2} > 0, \beta_{s,h_s} > \beta_{s,h_s-1} \forall (s,h_s) \in \mathcal{I} \times \{3,\ldots,p_s\}, \\ \beta_{s,2} < 0, \beta_{s,h_s} < \beta_{s,h_s-1} \forall (s,h_s) \in \mathcal{A} \times \{3,\ldots,p_s\}\}. \end{aligned}$$

is open. ■

Proof. Take any $\boldsymbol{\gamma}_0 \in U_{CM}$ as the center of an open ball $B(\boldsymbol{\gamma}_0, r)$. If the radius r is set at some fraction (say 1/2) of the minimum distance between adjacent parameters belonging to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_s$ $\forall s \in \mathcal{I} \cup \mathcal{A}$, then the open ball $B(\boldsymbol{\gamma}_0, r)$ is contained in U_{CM}. More formally, setting $r = \min(r_\alpha, r_\beta)$ with $r_\alpha = \min_{j\in\{2,\ldots,k-1\}}([\alpha_{j,0} - \alpha_{j-1,0}]/2)$ and $r_\beta = \min_{(s,h_s)\in(\mathcal{I}\cup\mathcal{A})\times\{2,\ldots,p_s\}}(|\beta_{s,h_s,0} - \beta_{s,h_s-1,0}|/2)$, with $\beta_{s,1,0} = 0$ $\forall s \in \mathcal{I} \cup \mathcal{A}$, allows to get $B(\boldsymbol{\gamma}_0, r) \subset U_{CM}$ for any $\boldsymbol{\gamma}_0 \in U_{CM}$ as required, so U_{CM} is open. ■

3.4.2 Convexity of the parameter space of the constrained POCLM

In Section 3.3, two constrained parameter sets were defined, $\tilde{U}_{CM}$ in (3.3.3) and U_{CM} in (3.3.4). The former does not require to allocate the monotonicity directions of the effects for each ordinal predictor, and the latter refers to the case when the monotonicity directions are pre-established. It will be shown that the parameter space for the model under monotonicity constraints is not convex when the monotonicity directions are undetermined (see Remark 3.9), and it is convex when these directions are determined (see Proposition 3.10). The following definition taken from Rudin et al. (1976) will be used:

Definition 3.8 (Convex set)**.** A set $S \subset \mathcal{R}^k$ is convex if

$$\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in S \tag{3.4.1}$$

whenever $\mathbf{x} \in S$, $\mathbf{y} \in S$, and $0 < \lambda < 1$. ■

The parameter set of model (2.3.1) under monotonicity constraints on the parameters of its ordinal predictors with undetermined monotonicity directions is denoted as $\tilde{U}_{CM}$ and defined in (3.3.3).

Remark 3.9. The parameter set $\tilde{U}_{CM}$ is not convex.

Proof. Consider a model with one ordinal predictor and take $\tilde{\boldsymbol{\gamma}}_0, \tilde{\boldsymbol{\gamma}}_1 \in \tilde{U}_{CM}$ for which the parameter vector of the ordinal predictor $s = 1$ of the point $\tilde{\boldsymbol{\gamma}}_0$ is in the "isotonic" set (with constraints $(\tilde{\beta}_{1,2,0} > 0, \tilde{\beta}_{1,h_s,0} > \tilde{\beta}_{1,h_s-1,0})$ $\forall h_s \in \{3,\ldots,p_1\}$)

and the parameter vector of the ordinal predictor $s = 1$ of the point $\tilde{\boldsymbol{\gamma}}_1$ is in the "antitonic" set (with constraints $(\tilde{\beta}_{1,2,1} < 0, \tilde{\beta}_{1,h_s,1} < \tilde{\beta}_{1,h_s-1,1})\ \forall h_s \in \{3, \ldots, p_1\}$). Consider $\tilde{\boldsymbol{\gamma}}_2 = \lambda\tilde{\boldsymbol{\gamma}}_0 + (1-\lambda)\tilde{\boldsymbol{\gamma}}_1$ with $\lambda \in (0, 1)$. The third subindex (0, 1, or 2) indicates the point ($\tilde{\boldsymbol{\gamma}}_0$, $\tilde{\boldsymbol{\gamma}}_1$ or $\tilde{\boldsymbol{\gamma}}_2$) to which the parameter belongs to. To prove that $\tilde{U}_{CM}$ is not convex, it is enough to show that some component of $\tilde{\boldsymbol{\gamma}}_2$ violates some of the restrictions defined in $\tilde{U}_{CM}$ for some $\lambda \in (0, 1)$.

The first component of the parameter vector of the ordinal predictor $s = 1$ of the point $\tilde{\boldsymbol{\gamma}}_2$ (k-th component of $\tilde{\boldsymbol{\gamma}}_2$) is

$$\tilde{\beta}_{1,2,2} = \lambda\tilde{\beta}_{1,2,0} + (1-\lambda)\tilde{\beta}_{1,2,1},$$

which can take any real value, violating its corresponding restriction in (3.3.3), for example, when $\tilde{\beta}_{1,2,2} = 0$ for the case $\tilde{\beta}_{1,2,0} = -\tilde{\beta}_{1,2,1}$ and $\lambda = 0.5$. Therefore $\tilde{\boldsymbol{\gamma}}_2 \notin \tilde{U}_{CM}$ and $\tilde{U}_{CM}$ is not convex. ■

However, it will be shown in the proof of Proposition 3.10 below, that both the isotonic and antitonic subsets of the parameter space of an ordinal predictor are convex separately. Therefore, when each vector of parameters associated with the t ordinal predictors, $\boldsymbol{\beta}_s\ \forall s = 1, \ldots, t$, is constrained under a particular monotonicity direction, then the parameter set of the constrained model is convex.

Proposition 3.10. *The set*

$$\begin{aligned} U_{CM} = \{\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_t', \boldsymbol{\beta}_{(nonord)}') \in \mathcal{R}^p : -\infty < \alpha_1 < \cdots < \alpha_{k-1} < \infty, \\ \beta_{s,2} > 0, \beta_{s,h_s} > \beta_{s,h_s-1} \forall (s, h_s) \in \mathcal{I} \times \{3, \ldots, p_s\}, \\ \beta_{s,2} < 0, \beta_{s,h_s} < \beta_{s,h_s-1} \forall (s, h_s) \in \mathcal{A} \times \{3, \ldots, p_s\}\}. \end{aligned}$$

is convex. ■

Proof. Take $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1 \in U_{CM}$ and consider $\boldsymbol{\gamma}_2 = \lambda\boldsymbol{\gamma}_0 + (1-\lambda)\boldsymbol{\gamma}_1$ with $\lambda \in (0, 1)$. For U_{CM} to be convex, it must be proved that $\boldsymbol{\gamma}_2 \in U_{CM}$ for all $\lambda \in (0, 1)$, which is equivalent to show that every component of the vectors contained by $\boldsymbol{\gamma}_2' = (\boldsymbol{\alpha}_2', \boldsymbol{\beta}_{1,2}', \ldots, \boldsymbol{\beta}_{t,2}', \boldsymbol{\beta}_{(nonord),2}')$ fulfils its corresponding restriction stated in U_{CM}. The last subindex (0, 1 or 2) will indicate the point to which the vector/component belongs to ($\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ or $\boldsymbol{\gamma}_2$).

- Regarding the components of $\boldsymbol{\alpha}_2$:

  The first component of $\boldsymbol{\alpha}_2$ is $\alpha_{1,2} = \lambda\alpha_{1,0} + (1-\lambda)\alpha_{1,1}$, then it follows that $\alpha_{1,2} > -\infty \; \forall\lambda \in (0,1)$. For the subsequent components of $\boldsymbol{\alpha}_2$, i.e., $\alpha_{j,2}$ with $j = 2, \ldots, k-1$,

  $$\alpha_{j,2} = \lambda\alpha_{j,0} + (1-\lambda)\alpha_{j,1},$$

  and given the strict monotonicity (isotonic) restriction, each of these components can be rewritten as

  $$\begin{aligned}\alpha_{j,2} &= \lambda(\alpha_{j-1,0} + \delta_{j-1,0}) + (1-\lambda)(\alpha_{j-1,1} + \delta_{j-1,1}) \\ &= \lambda\alpha_{j-1,0} + (1-\lambda)\alpha_{j-1,1} + \lambda\delta_{j-1,0} + (1-\lambda)\delta_{j-1,1} \\ &= \alpha_{j-1,2} + [\lambda\delta_{j-1,0} + (1-\lambda)\delta_{j-1,1}],\end{aligned}$$

  where $\delta_{j-1,0} > 0$ and $\delta_{j-1,1} < \infty \; \forall j \in \{2, \ldots, k-1\}$. Therefore, $\alpha_{1,2} > -\infty$, $\alpha_{j,2} > \alpha_{j-1,2} \; \forall j \in \{2, \ldots, k-1\}$, and $\alpha_{k-1,2} < \infty$ as required.

- Regarding the components of $\boldsymbol{\beta}_{s,2} \; \forall s \in \mathcal{I}$:

  The first component of $\boldsymbol{\beta}_{s,2}$ is $\beta_{s,2,2} = \lambda\beta_{s,2,0} + (1-\lambda)\beta_{s,2,1}$, then it follows that $\beta_{s,2,2} > 0 \; \forall\lambda \in (0,1)$. For the subsequent components of $\boldsymbol{\beta}_{s,2}$, i.e., $\beta_{s,h_s,2}$ with $h_s = 3, \ldots, p_s$,

  $$\beta_{s,h_s,2} = \lambda\beta_{s,h_s,0} + (1-\lambda)\beta_{s,h_s,1},$$

  and given the strict monotonicity (isotonic) restriction, each of these components can be rewritten as

  $$\begin{aligned}\beta_{s,h_s,2} &= \lambda(\beta_{s,h_s-1,0} + \delta_{h_s-1,0}) + (1-\lambda)(\beta_{s,h_s-1,1} + \delta_{h_s-1,1}) \\ &= \lambda\beta_{s,h_s-1,0} + (1-\lambda)\beta_{s,h_s-1,1} + \lambda\delta_{h_s-1,0} + (1-\lambda)\delta_{h_s-1,1} \\ &= \beta_{s,h_s-1,2} + [\lambda\delta_{h_s-1,0} + (1-\lambda)\delta_{h_s-1,1}],\end{aligned}$$

  where $\delta_{h_s-1,0}, \delta_{h_s-1,1} > 0 \; \forall h_s \in \{3, \ldots, p_s\}$. Therefore, $\beta_{s,2,2} > 0$ and $\beta_{s,h_s,2} > \beta_{s,h_s-1,2} \; \forall (s, h_s) \in \mathcal{I} \times \{3, \ldots, p_s\}$ as required.

- Regarding the components of $\boldsymbol{\beta}_{s,2} \; \forall s \in \mathcal{A}$:

  By analogous arguments to those used in the previous point, it is shown that $\beta_{s,2,2} < 0$ and $\beta_{s,h_s,2} < \beta_{s,h_s-1,2} \; \forall (s, h_s) \in \mathcal{A} \times \{3, \ldots, p_s\}$ as required.

- Regarding the components of $\boldsymbol{\beta}_{(nonord),2}$:

  Every component of $\boldsymbol{\beta}_{(nonord),2}$ is $\beta_{u,2} = \lambda\beta_{u,0} + (1-\lambda)\beta_{u,1}$ with $u = 1, \ldots, v$ and $\beta_{u,0}, \beta_{u,1} \in \mathcal{R}$, then $\beta_{u,2} \in \mathcal{R}$ $\forall\lambda \in (0,1)$ as required.

As none of the components of $\boldsymbol{\gamma}_2' = (\boldsymbol{\alpha}_2', \boldsymbol{\beta}_{1,2}', \ldots, \boldsymbol{\beta}_{t,2}', \boldsymbol{\beta}_{(nonord),2}')$ violates its corresponding restriction in U_{CM}, then $\boldsymbol{\gamma}_2 \in U_{CM}$ and therefore U_{CM} is convex.

■

3.5 Asymptotic monotonicity direction and consistency

The two main objectives of this section are to show that, when the set of parameters associated with each ordinal predictor is strictly monotonic,

(O.1) the true monotonicity direction classification of the set of parameters associated with each ordinal predictor is indicated by the MLEs when $n \to \infty$ with probability one, and

(O.2) asymptotic consistency of the MLEs holds for the constrained POCLM.

These two objectives will be finally addressed in Sections 3.5.4 and 3.5.5 accordingly. Previous sections will discuss the necessary aspects to build up the connection between asymptotic theory for unconstrained generalised linear models with natural link functions and some topics of asymptotic theory for the constrained POCLM, which is a particular case of constrained generalised linear models with non-natural link function. The first objective, (O.1), is achieved by Corollary 3.18, which is associated with Theorem 3.17, both of them in Section 3.5.4. The second objective, (O.2), is achieved by Theorem 3.19 in Section 3.5.5.

Section 3.5.1 presents the general setting of GLMs that will be used in subsequent sections. Asymptotic theory for unconstrained generalised linear models with natural link functions is discussed explicitly in Fahrmeir and Kaufmann (1985). In particular, its Theorem 2 will be presented in Section 3.5.2 as Theorem 3.11. Relevant parts of its proof will be highlighted because they will be used in further sections.

An explicit extension of Theorem 3.11 from unconstrained GLMs with natural link functions to unconstrained GLMs with general link functions (including the non-natural ones) is addressed as a first step leading to the case of the constrained POCLM. This extension is included in Theorem 3.16, Section 3.5.3. The proof of Theorem 3.16 requires some arguments that are stated and proved earlier, in Propositions 3.14 and 3.15. Proposition 3.14 is of special interest because it will also provide the key arguments in order to achieve (O.1) later. All of these are contained in Section 3.5.3.

As mentioned before, (O.1) will be achieved in Section 3.5.4. There it will be shown that when $n \to \infty$, the monotonicity direction of the set of parameters for each ordinal predictor is correctly established by the constrained MLEs of the POCLM with probability one, which allows to allocate each ordinal predictor s into either $\mathcal{I}$ or $\mathcal{A}$. In order to formalise this statement, Corollary 3.18, a special case of Theorem 3.17, will be stated in that section and their corresponding proofs will be given in detail. Theorem 3.17 extends the scope of Proposition 3.14 in the sense that the monotonicity constraints of ordinal predictors and other considerations are taken into account.

Finally, (O.2) will be achieved in Section 3.5.5, where asymptotic existence and strong consistency of the MLEs for the constrained POCLM will be stated by Theorem 3.19 together with its proof. Theorem 3.19 is an extension of Theorem 3.16 because it incorporates the monotonicity constraints of the parameters associated with ordinal predictors.

3.5.1 The GLM setting

Fahrmeir and Kaufmann (1985), p.345, characterise the generalised linear models (GLMs) using the following structure:

(i) The $\{\mathbf{y}_n\}$ are k-dimensional independent random variables with densities

$$f(\mathbf{y}_n|\boldsymbol{\theta}_n) = c(\mathbf{y}_n)\exp(\boldsymbol{\theta}_n'\mathbf{y}_n - b(\boldsymbol{\theta}_n)), \quad n = 1, 2, \ldots, \tag{3.5.1}$$

of the natural exponential type, with the parameter vector $\boldsymbol{\theta}_n$ belonging to Θ^0, the interior of the *natural parameter space* Θ of all natural parameters

$\boldsymbol{\theta}$ associated with a density function belonging to the exponential family (3.5.1). $E_{\boldsymbol{\theta}_n}(\mathbf{y}_n) = \partial b(\boldsymbol{\theta}_n)/\partial \boldsymbol{\theta}_n = \boldsymbol{\mu}(\boldsymbol{\theta}_n)$.

(ii) The matrix $\mathbf{X}_n$ influences $\mathbf{y}_n$ in form of a linear combination $\boldsymbol{\eta}_n = \mathbf{X}_n'\boldsymbol{\gamma}$ where $\boldsymbol{\gamma}$ is a p-dimensional parameter.

(iii) The linear combination is related to the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ of $\mathbf{y}_n$ by the injective link function $\mathbf{g} : M \to \mathcal{R}^k$, $\boldsymbol{\eta}_n = \mathbf{g}(\boldsymbol{\mu}(\boldsymbol{\theta}_n))$, where M is the image $\boldsymbol{\mu}(\Theta^0)$ of Θ^0. These functions will be referred to as general link functions.

As a remark, Fahrmeir and Kaufmann (1985), p.345, indicates that for theoretical purposes, it is more convenient to relate $\boldsymbol{\eta}_n = \mathbf{X}_n'\boldsymbol{\gamma}$ to the natural parameter $\boldsymbol{\theta}_n$ by the injective function $\mathbf{u} = (\mathbf{g} \circ \boldsymbol{\mu})^{-1}$, i.e., $\boldsymbol{\theta}_n = \mathbf{u}(\mathbf{X}_n'\boldsymbol{\gamma})$. Natural link functions are defined as $\mathbf{g} = \boldsymbol{\mu}^{-1}$, $\mathbf{u} = \mathrm{id}$, obtaining a linear model $\boldsymbol{\theta}_n = \mathbf{X}_n'\boldsymbol{\gamma}$ for the natural parameter. Natural link functions are special cases of general link functions.

Regularity assumptions (Fahrmeir and Kaufmann (1985), p.346)

(i) The admissible parameter set is open in $\mathcal{R}^p$,

(ii) $\mathbf{X}_n'\boldsymbol{\gamma} \in \mathbf{g}(M), n = 1, 2, \ldots$, for all $\boldsymbol{\gamma}$ in the admissible parameter set,

(iii) $\mathbf{g}$ resp. $\mathbf{u}$ is twice continuously differentiable, $\det(\partial \mathbf{u}/\partial \eta) \neq 0$,

(iv) $\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ has full rank for $n \geq n_0$, say.

The score function and the Fisher information matrix of the first n observations are

$$\mathbf{s}_n(\boldsymbol{\gamma}) = \partial \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{X}_n)/\partial \boldsymbol{\gamma}, \tag{3.5.2}$$

$$\mathbf{F}_n(\boldsymbol{\gamma}) = \mathrm{cov}_{\boldsymbol{\gamma}}\mathbf{s}_n(\boldsymbol{\gamma}). \tag{3.5.3}$$

The negative derivative of the score function yields

$$\mathbf{H}_n(\boldsymbol{\gamma}) = -\partial^2 \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{X}_n)/\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'. \tag{3.5.4}$$

The matrix $\mathbf{H}_n(\boldsymbol{\gamma})$ can also be expressed as $\mathbf{H}_n(\boldsymbol{\gamma}) = \mathbf{F}_n(\boldsymbol{\gamma}) - \mathbf{R}_n(\boldsymbol{\gamma})$, with the matrix $\mathbf{R}_n(\boldsymbol{\gamma})$ given by,

$$\mathbf{R}_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{r=1}^{k} \mathbf{X}_i \mathbf{W}_{ir}(\boldsymbol{\gamma}) \mathbf{X}_i' (y_{ir} - \mu_{ir}(\boldsymbol{\gamma})) \tag{3.5.5}$$

where $\mathbf{W}_{ir}(\boldsymbol{\gamma}) = \partial^2 u_r(\mathbf{X}_i'\boldsymbol{\gamma})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'$, and $u_r(\mathbf{X}_i'\boldsymbol{\gamma})$, y_{ir}, $\mu_{ir}(\boldsymbol{\gamma})$ are the components of $\mathbf{u}(\mathbf{X}_i'\boldsymbol{\gamma})$, $\mathbf{y}_i$, $\boldsymbol{\mu}_i(\boldsymbol{\gamma})$. In general, $E_{\boldsymbol{\gamma}}[\mathbf{H}_n(\boldsymbol{\gamma})] = \mathbf{F}_n(\boldsymbol{\gamma})$, and in particular for natural link functions, $\mathbf{H}_n(\boldsymbol{\gamma}) = \mathbf{F}_n(\boldsymbol{\gamma})$.

3.5.2 Consistency of GLMs with natural link function

Theorem 2 in Fahrmeir and Kaufmann (1985) (see Theorem 3.11 below) establishes the asymptotic existence and strong consistency of the MLEs in generalised linear models with natural link functions.

The notation $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma})]$ denotes the minimum eigenvalue of the matrix $\mathbf{F}_n(\boldsymbol{\gamma})$ and $\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma})]$ denotes its maximum eigenvalue.

Define the sequence $N_{\Delta,n}(\boldsymbol{\gamma}_0)$ of neighbourhoods of the true parameter vector $\boldsymbol{\gamma}_0$, with $\Delta > 0$, as

$$N_{\Delta,n}(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\mathbf{F}_n^{T/2}(\boldsymbol{\gamma}_0)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)|| \leq \Delta\}. \tag{3.5.6}$$

Theorem 3.11 (Fahrmeir and Kaufmann (1985), p.349)**.** *If the following assumptions hold,*

(D) *Divergence:* $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$*, and*

($\mathbf{S}_\Delta$) *Boundedness of the eigenvalue ratio: there is a neighbourhood* $N \subset U_{UM}$ *of* $\boldsymbol{\gamma}_0$ *such that*

$$\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}, \quad \boldsymbol{\gamma} \in N, \quad n \geq n_1,$$

with some constants $c > 0$, $\Delta > 0$, n_1,

then there is a sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ *of random variables and a random number* n_2 *with*

(i) $P\{\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) = \mathbf{0} \quad \forall n \geq n_2\} = 1$ *(asymptotic existence),*

(ii) $\hat{\boldsymbol{\gamma}}_n \xrightarrow{a.s.} \boldsymbol{\gamma}_0$ *(strong consistency).*

The proof of Theorem 3.11 is given by Fahrmeir and Kaufmann (1985), p.351-352, and requires that, given some arbitrary $\epsilon > 0$ with $K_\epsilon(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \leq \epsilon\}$ contained in the neighbourhood N of condition (S_Δ), and with a random number n_2,

$$P_{\boldsymbol{\gamma}_0}\{\log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) > \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n) \quad \forall \boldsymbol{\gamma} \text{ with } ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon, \forall n \geq n_2\} = 1, \tag{3.5.7}$$

namely, $\boldsymbol{\gamma}_0$ is the parameter vector that maximises the log-likelihood with probability one among all those $\boldsymbol{\gamma}$ that are at a distance of ϵ from $\boldsymbol{\gamma}_0$ for all n greater than or equal to some random number n_2, with $\log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n)$ being the value of the log-likelihood function for the first n observations.

Theorem 3.11 holds for generalised linear models with natural link functions. However, the link function of the model of interest, the POCLM, is non-natural. This requires to modify Theorem 3.11, which is discussed in the next section.

3.5.3 Consistency of GLMs with general link function

To see that the link function of the POCLM is non-natural, consider the multinomial model for an unordered response to compare its link function against the one of the POCLM. Both models use the multinomial distribution, which is part of the exponential family. $b(\boldsymbol{\theta})$ in (3.5.1) is sometimes referred to as the natural parameter. When the link function $\mathbf{g}$ is chosen to be of the form of the natural parameter, then $\mathbf{g}$ is called the natural (or canonical) link function (see Agresti (2007)). For a categorical response variable with k unordered categories it is common to use the link function

$$\log[\pi_j(\mathbf{x}_i)/\pi_k(\mathbf{x}_i)] = \boldsymbol{\alpha}_j + \mathbf{x}_i'\boldsymbol{\beta}_j, \quad \text{with } j = 1, \ldots, k, \quad \pi_j(\mathbf{x}_i) = P_{\boldsymbol{\gamma}}\{y_i = j|\mathbf{x}_i\}, \tag{3.5.8}$$

which turns out to be the natural link function (see Fahrmeir and Kaufmann (1986), p.182).

When the response variable is ordinal, the link function is different to (3.5.8) because it has to take into account the order of its categories. For a categorical response variable with k ordered categories the proportional odds cumulative logit

model defined in (2.3.6) uses the link function (2.3.1), which is non-natural (see Kaufmann (1988), p.296, and Fahrmeir and Kaufmann (1986), p.182).

For generalised linear models with non-natural link functions $\mathbf{H}_n(\boldsymbol{\gamma}) \neq \mathbf{F}_n(\boldsymbol{\gamma})$, making consistency and normality more difficult to be established since the uniqueness of the MLEs cannot be guaranteed for every non-natural link function (Fahrmeir and Kaufmann (1985), p.360). Wedderburn (1976) considers different link functions for four models, including some non-natural ones. These are the normal, Poisson, binomial, and gamma models. The existence and uniqueness of the MLEs is established for them. However, the multinomial model is not considered.

In Fahrmeir and Kaufmann (1985) p.360, a side remark indicates that Theorem 3.11 remains true for non-natural link functions under (S^*_Δ), a modified version of its condition (S_Δ), which involves $\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$ rather than $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$, without stating a new theorem explicitly. As this version is of interest for the case of the POCLM, it is made explicit in Theorem 3.16 below for general link functions. The proof of Theorem 3.16 will require an argument about the value of the log-likelihood function for the true parameter vector and its comparison against the value of the log-likelihood function for some other parameter vectors fulfilling some conditions. This will be formally stated and proved in Proposition 3.14. Therefore, the latter will be analysed before stating Theorem 3.16. The proof of Proposition 3.14 is based on part of the concise arguments used in the proof of Theorem 2 in Fahrmeir and Kaufmann (1985), so it extends them in detail and incorporates the corresponding modifications to make it hold for general link functions. It also makes reference to Wu (1981)'s Lemma 2 (see Wu (1981) p.504), some results of the Rayleigh quotient (see Meyer (2000), p.550-551), and the Cauchy-Bunyakovsky-Schwarz inequality. Therefore, they are presented below as a list of resources to be used in the proof of Proposition 3.14.

Lemma 3.12 (Wu (1981), p.504)**.** *Let $\{X_i\}$ be a sequence of independent random variables with $E[X_i] = 0$ and $Var[X_i] = \sigma_i^2$ and*

$$A_n \to \infty, \quad \limsup_{n\to\infty} \frac{(\sum_{i=1}^n \sigma_i^2)^{1/2+\Delta}}{A_n} < \infty \quad \text{for some } \Delta > 0. \tag{3.5.9}$$

Then,

$$\frac{\sum_{i=1}^{n} X_i}{A_n} \to 0 \quad a.s. \tag{3.5.10}$$

Meyer (2000), p.550-551, presents some properties of Hermitian matrices in terms of their smallest and largest eigenvalues, which are based on the Rayleigh quotient. The Rayleigh quotient also holds for real symmetric matrices since real symmetric matrices are special cases of Hermitian matrices. The eigenvalues $\lambda_i[\mathbf{A}]$ $(i = 1, 2, \ldots, n)$ of a real symmetric matrix $\mathbf{A}_{n\times n}$ are real, so they can be ordered as $\lambda_1[\mathbf{A}] \leq \lambda_2[\mathbf{A}] \leq \cdots \leq \lambda_n[\mathbf{A}]$. The largest and smallest eigenvalues can be described as

$$\lambda_1[\mathbf{A}] = \min_{||\boldsymbol{\lambda}||_2=1} \boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda} \text{ and } \lambda_n[\mathbf{A}] = \max_{||\boldsymbol{\lambda}||_2=1} \boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda}. \tag{3.5.11}$$

This characterisations often appear in the equivalent forms

$$\lambda_1[\mathbf{A}] = \min_{||\boldsymbol{\lambda}||_2\neq\mathbf{0}} \frac{\boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\lambda}} \text{ and } \lambda_n[\mathbf{A}] = \max_{||\boldsymbol{\lambda}||_2\neq\mathbf{0}} \frac{\boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\lambda}}. \tag{3.5.12}$$

Consequently, $\lambda_1[\mathbf{A}] \leq (\boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda}) \leq \lambda_n[\mathbf{A}]$ for all $\boldsymbol{\lambda} \neq \mathbf{0}$. The term $\boldsymbol{\lambda}'\mathbf{A}\boldsymbol{\lambda}/\boldsymbol{\lambda}'\boldsymbol{\lambda}$ is referred to as the Rayleigh quotient. In this case, the class of squared complex Hermitian matrices is a generalisation of the class of real symmetric matrices, such as the case of $\mathbf{F}_n(\boldsymbol{\gamma})$ and $\mathbf{H}_n(\boldsymbol{\gamma})$.

Theorem 3.13 (Cauchy-Bunyakovsky-Schwarz inequality)**.**

$$|\mathbf{x}'\mathbf{y}| \leq ||\mathbf{x}||\ ||\mathbf{y}|| \quad \textit{for all } \mathbf{x}, \mathbf{y} \in \mathcal{R}^n. \tag{3.5.13}$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ *for* $\alpha = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x}$.

Lemma 3.12, equations (3.5.11) and (3.5.12), and Theorem 3.13 will be used in the proof of Proposition 3.14.

Proposition 3.14. *If the following assumptions hold,*

(D) *Divergence:* $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$*, and*

($\mathbf{S}^*_\Delta$) *Boundedness of the eigenvalue ratio: there are some constants* $c > 0$, $\Delta > 0$, n_1, *and there is a neighbourhood* $N \subset U_{UM}$ *of* $\boldsymbol{\gamma}_0$, *such that* $\forall \boldsymbol{\gamma} \in N$, *and* $\forall n \geq n_1$,

$$\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$$

holds almost surely,

and considering for some arbitrary $\epsilon > 0$ *with* $K_\epsilon(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \leq \epsilon\}$ *contained in the neighbourhood* N *of condition* (S^*_Δ) *the event*

$$Q_{n_2} = \{((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots) : \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) > \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n), \\ \forall \boldsymbol{\gamma} \text{ with } ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon, \forall n \geq n_2\}, \tag{3.5.14}$$

then, with a random number n_2 *depending on the sequence* $\{\mathbf{y}_n\}$,

$$P_{\gamma_0}\{Q_{n_2}\} = 1. \tag{3.5.15}$$

■

Proof. We now consider $\boldsymbol{\gamma}$ with $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| = \epsilon$ as holds for the elements of Q_{n_2}.

We start with the Taylor expansion of the log-likelihood, which is

$$\log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n) = \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)' \frac{\partial \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n)}{\partial \boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} \\ + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)' \frac{\partial \log^2 L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}\bigg|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\gamma}}} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0), \tag{3.5.16}$$

where $\tilde{\boldsymbol{\gamma}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$, allowing the use of the equality sign. Letting $\boldsymbol{\lambda} = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)/\epsilon$ and using the score function (3.5.2) and the negative second derivative of the log-likelihood defined by (3.5.4), then an alternative expression to (3.5.16) is

$$\log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n) - \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) = \epsilon \boldsymbol{\lambda}' \mathbf{s}_n(\boldsymbol{\gamma}_0) - \frac{1}{2}\epsilon^2 \boldsymbol{\lambda}' \mathbf{H}_n(\tilde{\boldsymbol{\gamma}}) \boldsymbol{\lambda}. \tag{3.5.17}$$

Note that $\boldsymbol{\lambda} = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)/\epsilon$ and $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon$, then

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \frac{(\gamma_1 - \gamma_{0,1})^2}{\epsilon^2} + \frac{(\gamma_2 - \gamma_{0,2})^2}{\epsilon^2} + \cdots + \frac{(\gamma_p - \gamma_{0,p})^2}{\epsilon^2} \\ = \left[\frac{||(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||}{\epsilon}\right]^2 = 1. \tag{3.5.18}$$

Based on the event Q_{n_2}, the left hand side of (3.5.17) is negative, and therefore its right hand side fulfils

$$\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0) < \frac{\epsilon}{2}\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda} \quad \text{with} \quad n \geq n_2. \tag{3.5.19}$$

For the application of Wu (1981)'s Lemma 2 it is convenient to divide (3.5.19) by $(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$ and, according to (3.5.18), to use $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, from which we get

$$\frac{\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}} < \frac{\epsilon}{2}\frac{\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}} \quad \text{with } \boldsymbol{\lambda}'\boldsymbol{\lambda} = 1, \quad n \geq n_2. \tag{3.5.20}$$

Furthermore, (3.5.20) is equivalent to the inequality of the event Q_{n_2}, which follows from (3.5.17) and (3.5.19).

The left hand side of (3.5.20) will be analysed separately from its right hand side. It will be shown that the left hand side of (3.5.20) converges almost surely and uniformly to zero, whereas the right hand side of (3.5.20) is bounded from below by $c\epsilon/2$ if $n \geq n_1$, and therefore the event Q_{n_2} has probability one.

For the left hand side of (3.5.20), define

$$\frac{\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}} = \boldsymbol{\lambda}'\mathbf{v}_n(\boldsymbol{\gamma}_0), \tag{3.5.21}$$

where $\mathbf{v}_n(\boldsymbol{\gamma}_0) = \mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$. Given that $\mathbf{F}_n(\boldsymbol{\gamma}_0)$ is the var-cov matrix of $\mathbf{s}_n(\boldsymbol{\gamma}_0)$, each component of $\mathbf{s}_n(\boldsymbol{\gamma}_0)$ has a variance less than or equal to $\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]$. This is true because using (3.5.11) we get $\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] = \max_{||\boldsymbol{\lambda}||_2=1} \boldsymbol{\lambda}'\mathbf{F}_n(\boldsymbol{\gamma}_0)\boldsymbol{\lambda}$.

A component-wise application of Wu (1981)'s Lemma 2 (see Lemma 3.12) will be used to show that $\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \xrightarrow{a.s.} \mathbf{0}$ on the left hand side of (3.5.20). Lemma 3.12 states that

$$\frac{\sum_{i=1}^n X_i}{A_n} \to 0 \quad \text{a.s.} \tag{3.5.22}$$

for a sequence $\{X_i\}$ of independent random variables with $E[X_i] = 0$ and $Var[X_i] = \sigma_i^2$ and

$$A_n \to \infty, \quad \limsup_{n\to\infty} \frac{(\sum_{i=1}^n \sigma_i^2)^{1/2+\Delta}}{A_n} < \infty \quad \text{for some } \Delta > 0. \tag{3.5.23}$$

Now, given that the score function $\mathbf{s}_n(\boldsymbol{\gamma}_0)$ is the first derivative of the log-likelihood function (2.3.7), then the score function at $\boldsymbol{\gamma}_0$ can be written as the sum of n contributions defining

$$\mathbf{s}_n(\boldsymbol{\gamma}_0) = \sum_{i=1}^{n} \frac{\partial \sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i)}{\partial \boldsymbol{\gamma}_0} = \sum_{i=1}^{n} \mathbf{s}_{n,i}(\boldsymbol{\gamma}_0),$$

with $\partial[\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i)]/\partial\boldsymbol{\gamma}_0 = \mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)$. Taking expectations we obtain $E[\mathbf{s}_n(\boldsymbol{\gamma}_0)] = \sum_{i=1}^{n} E[\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)]$, where $E[\mathbf{s}_n(\boldsymbol{\gamma}_0)] = \mathbf{0}$. Each $\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)$ can be understood as the score function at $\boldsymbol{\gamma}_0$ for the ith observation, therefore $E[\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)] = \mathbf{0}$.

To see this, consider the likelihood function for the i-th observation given by

$$\begin{aligned} L_{n=1}(\boldsymbol{\gamma}|\mathbf{y}_i, \mathbf{x}_i) &= \prod_{j=1}^{k} \pi_j(\mathbf{x}_i)^{y_{ij}} \\ &= \prod_{j=1}^{k} P(y_i = j|\mathbf{x}_i)^{y_{ij}} \end{aligned} \tag{3.5.24}$$

and the corresponding log-likelihood function for the i-th observation

$$\ell_{n=1}(\boldsymbol{\gamma}) = \sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i), \tag{3.5.25}$$

where y_i and j denote the number of the ordinal category of the response variable $y_i, j \in \{1, 2, \ldots, k\}$; $y_{i1}, \ldots, y_{ik}$ are the binary indicators of the response for the i-th observation with $y_{ij} = 1$ if the response falls in category j and 0 otherwise; and $\mathbf{y}_i$ is the response vector with k binary components for the i-th observation. Hence, the expectation of the score function for the i-th observation is

$$\begin{aligned} E[\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)] &= \sum_{j=1}^{k} \frac{\partial \log L_{n=1}(\boldsymbol{\gamma}_0|\mathbf{y}_i, \mathbf{x}_i)}{\partial \boldsymbol{\gamma}_0} P(y_i = j|\mathbf{x}_i) \\ &= \sum_{j=1}^{k} \frac{\partial L_{n=1}(\boldsymbol{\gamma}_0|\mathbf{y}_i, \mathbf{x}_i)/\partial \boldsymbol{\gamma}_0}{L_{n=1}(\boldsymbol{\gamma}_0|\mathbf{y}_i, \mathbf{x}_i)} P(y_i = j|\mathbf{x}_i) \\ &= \sum_{j=1}^{k} \frac{\partial L_{n=1}(\boldsymbol{\gamma}_0|\mathbf{y}_i, \mathbf{x}_i)}{\partial \boldsymbol{\gamma}_0} \\ &= \frac{\partial}{\partial \boldsymbol{\gamma}_0} \sum_{j=1}^{k} L_{n=1}(\boldsymbol{\gamma}_0|\mathbf{y}_i, \mathbf{x}_i) \\ &= \mathbf{0}, \end{aligned} \tag{3.5.26}$$

as previously stated.

Now, the term X_i of Lemma 3.12 is defined in this proof as some component of $\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)$, i.e., $X_i = s_{n,r,i}(\boldsymbol{\gamma}_0)\ \forall r \in \{1,\ldots,p\}$. In addition, let $Var[s_{n,r,i}(\boldsymbol{\gamma}_0)] = \sigma^2_{r,i}$. As $E[\mathbf{s}_{n,i}(\boldsymbol{\gamma}_0)] = \mathbf{0}$, then $E[s_{n,r,i}(\boldsymbol{\gamma}_0)] = 0\ \forall r \in \{1,\ldots,p\}$ as required by Lemma 3.12.

Next it will be shown that the remaining conditions in (3.5.23) are fulfilled. Let $A_n = (\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$ and consider $\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$ by assumption (D) and $\Delta > 0$ by assumption (S^*_Δ). Then it follows that $(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \to \infty$. The second part of the conditions in (3.5.23) follows because it has been seen that the variance of each component of $\mathbf{s}_n(\boldsymbol{\gamma}_0)$ is less than or equal to $\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]$. Therefore, given independent observations,

$$\limsup_{n\to\infty}\left(\frac{\sum_{i=1}^n \sigma^2_{r,i}}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]}\right)^{\frac{1}{2}+\Delta} = \limsup_{n\to\infty}\left(\frac{\mathrm{var}[s_{n,r}(\boldsymbol{\gamma}_0)]}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]}\right)^{\frac{1}{2}+\Delta} < \infty, \quad \forall r \in \{1,\ldots,p\}.$$

Then, by Lemma 3.12 it follows that $\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \xrightarrow{a.s.} \mathbf{0}$ holds. As this result addresses $\mathbf{v}_n(\boldsymbol{\gamma}_0)$ only, it still remains to analyse the full term of (3.5.21), $\boldsymbol{\lambda}'\mathbf{v}_n(\boldsymbol{\gamma}_0)$, which is discussed next.

By the Cauchy-Bunyakovsky-Schwarz (CBS) inequality (also known as Cauchy-Schwarz inequality), it will be shown that the left hand side of (3.5.20) converges to zero a.s. and uniformly for all $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$.

By the CBS inequality, we can write

$$|\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}| \leq ||\boldsymbol{\lambda}||\ ||\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}||. \quad (3.5.27)$$

Using Wu's Lemma it has been seen that $\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \xrightarrow{a.s.} \mathbf{0}$. Its Euclidean norm also converges almost surely to zero, and given that $||\boldsymbol{\lambda}|| = 1$, then the right hand side of (3.5.27) converges almost surely to zero and therefore its left hand side too. Furthermore, this is true uniformly over all $\boldsymbol{\lambda}$ with $||\boldsymbol{\lambda}|| = 1$ because $\boldsymbol{\lambda}$ is constant and there is a number $n \geq n_2$ so that for every arbitrarily small $\epsilon > 0$, the left hand side of (3.5.27) is smaller than ϵ, namely it converges almost surely and uniformly for all $\boldsymbol{\lambda}$ with $||\boldsymbol{\lambda}|| = 1$, which is used later in order to show (3.5.15).

Next, the right hand side of (3.5.20) is discussed.

From condition (S^*_Δ), the right hand side of (3.5.20) is bounded from below by $c\epsilon/2$ if $n \geq n_1$ for the following reason. From (S^*_Δ) we can write

$$c \leq \frac{\lambda_{\min}\mathbf{H}_n(\boldsymbol{\gamma})}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}}. \tag{3.5.28}$$

By (3.5.11) it follows that $\lambda_{\min}\mathbf{H}_n(\boldsymbol{\gamma}) \leq \boldsymbol{\lambda}'\mathbf{H}_n(\boldsymbol{\gamma})\boldsymbol{\lambda}$. In addition, $\tilde{\boldsymbol{\gamma}}$ is between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$. Therefore, from condition (S^*_Δ), (3.5.28) also holds for $\tilde{\boldsymbol{\gamma}}$ in place of $\boldsymbol{\gamma}$, and multiplying both sides by $\epsilon/2$ we get

$$\frac{\epsilon}{2}c \leq \frac{\epsilon}{2}\frac{\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}}. \tag{3.5.29}$$

Thus, it has been shown that the right hand side of (3.5.20) is bounded from below by $c\epsilon/2$ if $n \geq n_1$.

Hence the event (3.5.20), and respectively (3.5.14), have probability one, which completes the proof. ■

It will be shown that the use of Proposition 3.14 allows to prove asymptotic existence and strong consistency of the parameter estimates as stated by Theorem 3.16.

We now analyse the case where for every n large enough, the sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot leave a compact set $N_\Delta(\boldsymbol{\gamma}_0)$, as it will also be required to prove Theorem 3.16.

Proposition 3.15. *If the following assumption holds,*

1. *there is a number n_1, so that $\forall n > n_1$ the relative frequency of every combination of possible values for $\mathbf{x}_n$ is larger than r^*, with $r^* > 0$,*

then, there is a large enough $\Delta > 0$ so that for every $n > n_1$ the sequence of global maxima $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot leave the compact set $N_\Delta(\boldsymbol{\gamma}_0)$ defined in (3.5.6). ■

Proof. Consider the mean log-likelihood function for the model:

$$\frac{1}{n}\ell(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\log\pi_j(\mathbf{x}_i)}{n}. \tag{3.5.30}$$

It will be shown that the sequence of mean log-likelihood functions for the true parameter vector $\boldsymbol{\gamma}_0$ is bounded by some constant c^* whereas for a divergent sequence of $\hat{\boldsymbol{\gamma}}_n$ tends to $-\infty$, i.e. $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$. This means that a divergent sequence of $\hat{\boldsymbol{\gamma}}_n$ cannot correspond to one of maximum likelihoods for each given n, and therefore the sequence of global maxima $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot leave a compact set.

In order to show that the mean log-likelihood function for $\boldsymbol{\gamma}_0$ is bounded by some positive constant c^*, set $\boldsymbol{\gamma}$ in (3.5.30) at the true parameter vector $\boldsymbol{\gamma}_0$, which belongs to a compact set $N_\Delta(\boldsymbol{\gamma}_0)$. Given that the values for $\mathbf{x}_n$ are bounded and $\boldsymbol{\gamma}_0$ is fixed, then the term $\pi_j(\mathbf{x}_i)$ in (3.5.30) is also bounded for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$ because some components of the parameter vector are strictly isotonic (see (2.3.2)), which prevents $\pi_j(\mathbf{x}_i)$ to be zero, and its definition (2.3.6) makes it to be smaller than one. Define c, with $c > 0$, as the lower bound of $\pi_j(\mathbf{x}_i)$ $\forall i, j$ for a given true parameter vector $\boldsymbol{\gamma}_0$. Therefore, every $\log \pi_j(\mathbf{x}_i)$ is bounded from below $\forall i, j$ by c^*, with $c^* > -\infty$. As there is a single $y_{ij} = 1$ for every $i = 1, \ldots, n$, then the term $\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i)$ is also bounded from below by c^*. Now consider the sequence of the mean log-likelihood function for $\boldsymbol{\gamma}_0$:

$$\frac{1}{n}\ell_n(\boldsymbol{\gamma}_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i)}{n}. \tag{3.5.31}$$

As in (3.5.31) the term $\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i)$ is also bounded from below by c^* for each i,

$$\frac{1}{n}\ell_n(\boldsymbol{\gamma}_0) > c^*, \tag{3.5.32}$$

and then the mean log-likelihood function for $\boldsymbol{\gamma}_0$ is bounded from below by c^* too.

Now consider the sequence $\{\hat{\boldsymbol{\gamma}}_n\}$, for which one or more of its elements diverge, meaning that $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$. At a certain value of $\mathbf{x}_i$, later it will be shown that the $\log \hat{\pi}_j(\mathbf{x}_i) \to -\infty$ as $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$ when $n \to \infty$. By assumption 1, there are more than r^*n values of $\mathbf{x}_i$ of this type. Therefore, $\log \hat{\pi}_j(\mathbf{x}_i) \to -\infty$ for more than r^*n observations. Define I^* as the set of indexes i belonging to these type of observations, $I^* = \{i : \log \hat{\pi}_j(\mathbf{x}_i) \to -\infty$ as $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$ and $n \to \infty\}$, so that the

sequence of mean log-likelihood for $\hat{\boldsymbol{\gamma}}_n$ is

$$\begin{aligned}\frac{1}{n}\ell_n(\hat{\boldsymbol{\gamma}}_n) &= \sum_{i=1}^{n}\frac{\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)}{n}\\ &= \sum_{i\in I^*}\frac{\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)}{n} + \sum_{i\notin I^*}\frac{\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)}{n}. \end{aligned} \tag{3.5.33}$$

Given that $|I^*| > r^*n$, then we can write the limit of (3.5.33) as $n \to \infty$ as

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\ell_n(\hat{\boldsymbol{\gamma}}_n) &< r^*n\frac{-\infty}{n} + \lim_{n\to\infty}\sum_{i\notin I^*}\frac{\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)}{n}\\ &= -\infty + \lim_{n\to\infty}\sum_{i\notin I^*}\frac{\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)}{n}. \end{aligned} \tag{3.5.34}$$

As $\sum_{j=1}^{k} y_{ij}\log\hat{\pi}_j(\mathbf{x}_i)$ in (3.5.34) always take a non-positive value for each i, then

$$\frac{1}{n}\ell_n(\hat{\boldsymbol{\gamma}}_n) \to -\infty \text{ as } n \to \infty. \tag{3.5.35}$$

Therefore, for n large enough, the likelihood of the $\hat{\boldsymbol{\gamma}}_n$, with $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$, is smaller than the likelihood for a value of $\boldsymbol{\gamma}$ that is in a compact set, meaning that the sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot be of global maxima, and therefore $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot leave the compact set $N_\Delta(\boldsymbol{\gamma}_0)$.

It still remain to be shown that $\log\hat{\pi}_j(\mathbf{x}_i) \to -\infty$ as $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$ when $n \to \infty$, which is discussed next.

Using $\hat{\boldsymbol{\gamma}}_n$ in (2.3.6) defines $\hat{\pi}_j(\mathbf{x}_i)$ and $\log\hat{\pi}_j(\mathbf{x}_i) \to -\infty$ is equivalent to $\hat{\pi}_j(\mathbf{x}_i) \to 0$ as $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$. Define $\hat{\boldsymbol{\gamma}}'_n = (\hat{\boldsymbol{\alpha}}'_n, \hat{\boldsymbol{\beta}}'_{1,n}, \ldots, \hat{\boldsymbol{\beta}}'_{t,n}, \hat{\boldsymbol{\beta}}'_{(nonord),n})$. An increasing sequence of $||\hat{\boldsymbol{\gamma}}_n||$ is analysed according to the different vectors of $\{\hat{\boldsymbol{\gamma}}_n\}$ and two cases are considered. Case A is when only one of the components of $\hat{\boldsymbol{\gamma}}_n$ diverges, and Case B is when more than one component of $\hat{\boldsymbol{\gamma}}_n$ does it.

Case A: If $\hat{\alpha}_{1,n} \to -\infty$ or $\hat{\alpha}_{k-1,n} \to \infty$, then $\hat{\pi}_1(\mathbf{x}_i) \to 0$ or $\hat{\pi}_k(\mathbf{x}_i) \to 0$ respectively. The cases where $\hat{\alpha}_{1,n} \to \infty$ and $\hat{\alpha}_{k-1,n} \to -\infty$ are left to the analysis of Case B because they imply that more than one component of $\hat{\boldsymbol{\alpha}}_n$ diverge. If one of the components of $\hat{\boldsymbol{\beta}}_{s,n}$ or $\hat{\boldsymbol{\beta}}_{(nonord),n}$ diverge, then (2.3.6) tends to 0 regardless of the values that other components take as long as they do not tend to infinity or negative infinity as $n \to \infty$, which is also left to the analysis of Case B.

Case B: If $\hat{\alpha}_{j,n} \to \infty$ with $j = 1, \ldots, k-2$, then every $\hat{\alpha}_{j',n}$ with $j' = j+1, \ldots, k-1$ also tends to ∞ because of their isotonic restriction according to (3.2.2). Therefore, all those subsequent probabilities $\hat{\pi}_{j'}(\mathbf{x}_i)$ with $j' = j+1, \ldots, k$ converge to 0. Similarly, if a component $\hat{\alpha}_{j,n} \to -\infty$, $j = 2, \ldots, k-1$, then the preceding components of of $\hat{\boldsymbol{\alpha}}_n$ also tend to $-\infty$ and all those probabilities $\hat{\pi}_{j'}(\mathbf{x}_i)$ with $j' = 1, \ldots, j$ converge to 0. Based on their isotonic restriction, if some components of $\hat{\boldsymbol{\alpha}}_n$ diverge in opposite directions, these arguments remain the same for each direction accordingly. If more than one component of $\hat{\boldsymbol{\beta}}_{s,n}$ and/or $\hat{\boldsymbol{\beta}}_{(nonord),n}$ diverge, then, by assumption 1, for some of the i's the sums $(\sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\hat{\beta}_{s,h_s,n}x_{i,s,h_s} + \sum_{u=1}^{v}\hat{\beta}_{u,n}x_{i,u})$ in (2.3.6) tend to either $-\infty$ or ∞, and both the first and second term on the right hand side of (2.3.6) tend to 0 or 1 correspondingly. Based on an analogous argument but using the sums $(\hat{\alpha}_{j,n} + \sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\hat{\beta}_{s,h_s,n}x_{i,s,h_s} + \sum_{u=1}^{v}\hat{\beta}_{u,n}x_{i,u})$ in (2.3.6), if one or more components of $\hat{\boldsymbol{\alpha}}_n$ diverge together with one or more components of $\hat{\boldsymbol{\beta}}_{s,n}$ and/or $\hat{\boldsymbol{\beta}}_{(nonord),n}$, by assumption 1, $\hat{\pi}_j(\mathbf{x}_i)$ will also tend to 0. The same happens if any subset of components of $\hat{\boldsymbol{\gamma}}_n$ diverge in opposite directions ($-\infty$ or $+\infty$) among each other, then they cannot cancel each other out because every one of these parameters applies for different values of the predictors and finally, by assumption 1, $\hat{\pi}_j(\mathbf{x}_i)$ will also tend to 0 anyway. Therefore, $\log\hat{\pi}_j(\mathbf{x}_i) \to -\infty$ as $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$ when $n \to \infty$ regardless of the direction ($-\infty$ or $+\infty$) and number of diverging components of the parameter vector $\hat{\boldsymbol{\gamma}}_n$. ■

Now Propositions (3.14) and (3.15) are used to prove the following theorem about asymptotic existence and strong consistency of the MLEs for unconstrained GLMs with general link functions, from which the unconstrained POCLM is a particular case.

Theorem 3.16. *If the following assumptions hold,*

(D) *Divergence:* $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$, *and*

($\mathbf{S}^*_\Delta$) *Boundedness of the eigenvalue ratio: there are some constants* $c > 0$, $\Delta > 0$, n_1, *and there is a neighbourhood* $N \subset U_{UM}$ *of* $\boldsymbol{\gamma}_0$, *such that* $\forall\boldsymbol{\gamma} \in N$, *and*

$\forall n \geq n_1$,

$$\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \geq c[\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]]^{1/2+\Delta}$$

holds almost surely,

then there is a sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ *of random variables and a random number* n_2 *with*

(i) $P\{\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) = \mathbf{0} \quad \forall n \geq n_2\} = 1$ *(asymptotic existence), and*

(ii) $\hat{\boldsymbol{\gamma}}_n \xrightarrow{a.s.} \boldsymbol{\gamma}_0$ *(strong consistency).*

Proof. As conditions (D) and (S^*_Δ) of this theorem are the same as the ones for Proposition 3.14, then Proposition 3.14 holds, meaning that given some arbitrary $\epsilon > 0$ with $K_\epsilon(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \leq \epsilon\}$ contained in the neighbourhood N of condition (S^*_Δ), and with a random number n_2 depending on the sequence $\{\mathbf{y}_n\}$,

$$P_{\gamma_0}\{\log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) > \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n) \quad \forall \boldsymbol{\gamma} \text{ with } ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon, \forall n \geq n_2\} = 1. \tag{3.5.36}$$

By Proposition 3.2 and Theorem 3.3, the log-likelihood function (2.3.7) is continuous and differentiable at $\boldsymbol{\gamma} \in U_{UM}$. For continuous functions in a compact set, the existence of its supremum and infimum is guaranteed (see Theorem 6.30 in Protter et al. (2012), p159). Consider a sequence of global maxima of the log-likelihood function given by the sequence of parameter vectors $\{\hat{\boldsymbol{\gamma}}_n\}$ contained in the compact set N of condition (S^*_Δ). By definition of a compact set (see Protter et al. (2012)), $\{\hat{\boldsymbol{\gamma}}_n\}$ has a subsequence denoted as $\{\boldsymbol{\gamma}'_n\}$ such that $\boldsymbol{\gamma}'_n \xrightarrow{a.s.} \tilde{\boldsymbol{\gamma}}$ with $\tilde{\boldsymbol{\gamma}} \in N$. Let ϵ be the distance of $\tilde{\boldsymbol{\gamma}}$ from $\boldsymbol{\gamma}_0$. Given that the sequence of parameter vectors $\{\hat{\boldsymbol{\gamma}}_n\}$ is the one that maximises the log-likelihood, then $\log L(\tilde{\boldsymbol{\gamma}}|\mathbf{y}_n, \mathbf{x}_n) \geq \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n)$ $\forall n \geq n_2$, which is a contradiction to (3.5.36). Therefore, there is no bounded sequence of $\hat{\boldsymbol{\gamma}}_n$ that are global optima that does not converge to $\boldsymbol{\gamma}_0$, from which *(ii)* follows.

Given that the log-likelihood function of the POCLM is differentiable at $\boldsymbol{\gamma} \in U_{UM}$ (see Proposition 3.2), $\boldsymbol{\gamma}_0$ is in the interior of the compact set N (see (3.5.6)), and the sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ is a sequence of global maxima of the log-likelihood function

of the POCLM, then $\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) = \mathbf{0}$, from which asymptotic existence stated in *(i)* follows.

So far we have considered the case where the sequence $\{\hat{\boldsymbol{\gamma}}_n\} \in N$. For the case when $||\hat{\boldsymbol{\gamma}}_n|| \to \infty$, it has been shown in the proof of Proposition 3.15 that for every n large enough, $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot leave a compact set N, and therefore the sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ cannot diverge to infinity. ■

3.5.4 MLEs and monotonicity direction of the effects of the ordinal predictor(s)

Theorem 3.16 showed that asymptotic existence and strong consistency of the MLEs hold for non-natural link functions as the one of the POCLM. It states that the log-likelihood associated with any $\boldsymbol{\gamma}$ belonging to the neighbourhood N of condition (S^*_Δ) and with $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon$ is lower than the one resulting from the true parameter vector $\boldsymbol{\gamma}_0$ with probability one as $n \to \infty$. However, we are interested in the case of the POCLM with monotonicity constraints, for which it will be shown that the log-likelihood supplied by any $\boldsymbol{\gamma}$ belonging to the wrong monotonicity direction is lower than the one provided by the true parameter vector $\boldsymbol{\gamma}_0$, i.e., this holds for all those $\boldsymbol{\gamma}$ for which $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \geq \epsilon$ when ϵ defines, in some way that will be discussed later, the boundary between the right and wrong monotonicity direction. In addition, the sequence of neighbourhoods is defined as follows to take into account the monotonicity restrictions,

$$N^*_{\Delta,n}(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\mathbf{F}_n^{T/2}(\boldsymbol{\gamma}_0)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)|| \leq \Delta, \boldsymbol{\gamma} \in \tilde{U}_{CM}\}. \tag{3.5.37}$$

The following theorem states that, asymptotically, the MLE of the constrained POCLM is in the right monotonicity direction.

Theorem 3.17. *If the following assumptions hold,*

(D) *Divergence:* $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$, *and*

($\mathbf{S}^*_\Delta$) *Boundedness of the eigenvalue ratio: there are some constants* $c > 0$, $\Delta > 0$, n_1, *and there is a neighbourhood* $N^* \subset \tilde{U}_{CM}$ *of* $\boldsymbol{\gamma}_0$, *such that* $\forall \boldsymbol{\gamma} \in N^*$, *and*

$\forall n \geq n_1$,

$$\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$$

holds almost surely,

and considering for some arbitrary $\epsilon > 0$ *with* $K_\epsilon(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \leq \epsilon, \boldsymbol{\gamma} \in N^*\}$ *the event*

$$Q_{n_2} = \{((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots) : \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) > \log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n),$$
$$\forall \boldsymbol{\gamma} \text{ with } ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \geq \epsilon, \forall n \geq n_2\}, \tag{3.5.38}$$

then, with a random number n_2 *depending on the sequence* $\{\mathbf{y}_n\}$,

$$P_{\boldsymbol{\gamma}_0}\{Q_{n_2}\} = 1. \tag{3.5.39}$$

Proof. This proof follows the same line of arguments used in the proof of Proposition 3.14 with some exceptions that will be mentioned as needed.

Instead of using $\boldsymbol{\gamma}$ with $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| = \epsilon$ as in Proposition 3.14, we now consider $\boldsymbol{\gamma}$ with $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \geq \epsilon$ as holds for the elements of Q_{n_2}.

Based on the arguments of the first part of the proof of Proposition 3.14, we get

$$\log L(\boldsymbol{\gamma}|\mathbf{y}_n, \mathbf{x}_n) - \log L(\boldsymbol{\gamma}_0|\mathbf{y}_n, \mathbf{x}_n) = \epsilon\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0) - \frac{1}{2}\epsilon^2\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}, \tag{3.5.40}$$

where $\tilde{\boldsymbol{\gamma}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$, allowing the use of the equality sign.

Given that (3.5.39) holds for all those $\boldsymbol{\gamma}$ with $||\boldsymbol{\gamma} - \boldsymbol{\gamma}_0|| \geq \epsilon$, then now

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \frac{(\gamma_1 - \gamma_{0,1})^2}{\epsilon^2} + \frac{(\gamma_2 - \gamma_{0,2})^2}{\epsilon^2} + \cdots + \frac{(\gamma_p - \gamma_{0,p})^2}{\epsilon^2}$$
$$= \left[\frac{||(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||}{\epsilon}\right]^2 \geq 1. \tag{3.5.41}$$

Based on the same arguments of the proof of Proposition 3.14, we can express the right hand side of (3.5.40) as

$$\frac{\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}} < \frac{\epsilon}{2}\frac{\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}} \quad \text{with } \boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1, \quad n \geq n_2, \tag{3.5.42}$$

where (3.5.42) is equivalent to the inequality of the event (3.5.38).

In the proof of Proposition 3.14 there is a similar expression to (3.5.42) (see (3.5.20)). The differences are based on the fact that in (3.5.42) $\boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1$ whereas in (3.5.20) $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$.

Now we analyse the left hand side of (3.5.42). Letting $\mathbf{v}_n(\boldsymbol{\gamma}_0) = \mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$ and using the same arguments as in the proof of Proposition 3.14 regarding $\mathbf{v}_n(\boldsymbol{\gamma}_0)$, we get that $\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \xrightarrow{a.s.} \mathbf{0}$ holds for the corresponding factor in the left hand side of (3.5.42). The full term, $\boldsymbol{\lambda}'\mathbf{v}_n(\boldsymbol{\gamma}_0)/\boldsymbol{\lambda}'\boldsymbol{\lambda}$, is discussed next to show that the left hand side of (3.5.42) converges to zero a.s. and uniformly for all $\boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1$ using the Cauchy-Schwarz inequality.

By the CBS inequality, we can write

$$\left|\frac{\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}}\right| \leq ||\boldsymbol{\lambda}|| \left|\left|\frac{\mathbf{s}_n(\boldsymbol{\gamma}_0)}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}}\right|\right| \tag{3.5.43}$$

or equivalently, using $||\boldsymbol{\lambda}|| = \sqrt{\boldsymbol{\lambda}'\boldsymbol{\lambda}}$,

$$\left|\frac{\boldsymbol{\lambda}'\mathbf{s}_n(\boldsymbol{\gamma}_0)}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}}\right| \leq \left|\left|\frac{\mathbf{s}_n(\boldsymbol{\gamma}_0)}{\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)]^{1/2+\Delta}}\right|\right| \frac{1}{||\boldsymbol{\lambda}||}. \tag{3.5.44}$$

The left hand side of (3.5.44) converges almost surely because, as it has been seen in the proof of Proposition 3.14, $\mathbf{s}_n(\boldsymbol{\gamma}_0)/(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta} \xrightarrow{a.s.} \mathbf{0}$, its Euclidean norm also converges almost surely to zero, and $\boldsymbol{\lambda}$ does not depends on n. Furthermore, this is true uniformly over all $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1$ because $\boldsymbol{\lambda}$ is constant and there is a number $n \geq n_2$ so that for every arbitrarily small $\epsilon > 0$, the left hand side of (3.5.44) is smaller than ϵ, namely it converges almost surely and uniformly for all $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1$, which is used later in order to show (3.5.39).

Next, the right hand side of (3.5.42) is discussed.

Based on condition (S^*_Δ) and following the same arguments as in the proof of Proposition 3.14 but using $\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \leq \frac{\boldsymbol{\lambda}'\mathbf{H}_n(\boldsymbol{\gamma})\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} \geq 1$, which follows from (3.5.12), and given that $\tilde{\boldsymbol{\gamma}}$ is between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}_0$, therefore we can write

$$c \leq \frac{\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}}. \tag{3.5.45}$$

Thus, by (3.5.45), the right hand side of (3.5.42) is bounded from below by $c\epsilon/2$ if $n \geq n_1$, i.e.,

$$\frac{\epsilon}{2}c \leq \frac{\epsilon}{2}\frac{\boldsymbol{\lambda}'\mathbf{H}_n(\tilde{\boldsymbol{\gamma}})\boldsymbol{\lambda}}{(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}\boldsymbol{\lambda}'\boldsymbol{\lambda}}. \tag{3.5.46}$$

Hence, the event (3.5.42), and respectively (3.5.39), have probability one. ■

Corollary 3.18. *Given a true parameter vector $\boldsymbol{\gamma}_0 \in U_{CM}$, where each set of parameters associated with each ordinal predictor belongs to either $\mathcal{I}$ or $\mathcal{A}$, the maximum likelihood estimators are in the right monotonicity direction with probability 1 as $n \to \infty$.*

Proof. Given that $U_{CM} \subset \tilde{U}_{CM}$ and $\boldsymbol{\gamma}_0$ is in the interior of U_{CM} because, by Proposition 3.7, U_{CM} is open, there exists a value for ϵ, denoted as ϵ^*, for which there is a set $K_{\epsilon^*}(\boldsymbol{\gamma}_0)$ where every $\boldsymbol{\gamma}$ belonging to the parameter space associated with the wrong monotonicity direction, $\boldsymbol{\gamma} \in \tilde{U}_{CM} \backslash U_{CM}$, is not in $K_{\epsilon^*}(\boldsymbol{\gamma}_0)$. Therefore, Corollary 3.18 follows from Theorem 3.17 with large enough Δ and given $\epsilon = \epsilon^*$. ■

3.5.5 Consistency of the constrained POCLM

In Section 3.5.3, Proposition 3.14 was used to prove asymptotic existence and strong consistency of the MLEs for the unconstrained version of the POCLM as shown in Theorem 3.16. In Section 3.5.4, Proposition 3.14 was extended to hold for the constrained POCLM stating Theorem 3.17 and to address the monotonicity classification of effects of ordinal predictors when $n \to \infty$ by Corollary 3.18.

This time the sequence of neighbourhoods considers the monotonicity restrictions associated with the true parameter vector $\boldsymbol{\gamma}_0$ as follows,

$$N^{**}_{\Delta,n}(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||\mathbf{F}_n^{T/2}(\boldsymbol{\gamma}_0)(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)|| \leq \Delta, \boldsymbol{\gamma} \in U_{CM}\}. \tag{3.5.47}$$

In the current section, Theorem 3.17 is also used to prove asymptotic existence and strong consistency of the MLEs for the constrained version of the POCLM as stated in the next theorem.

Theorem 3.19. *If the following assumptions hold,*

(D) *Divergence: $\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\gamma}_0)] \to \infty$, and*

($\mathbf{S}^*_\Delta$) *Boundedness of the eigenvalue ratio: there are some constants $c > 0$, $\Delta > 0$, n_1, and there is a neighbourhood $N^{**} \subset U_{CM}$ of $\boldsymbol{\gamma}_0$, with N^{**} defined by*

(3.5.47), *such that* $\forall \boldsymbol{\gamma} \in N^{**}$, *and* $\forall n \geq n_1$,

$$\lambda_{\min}[\mathbf{H}_n(\boldsymbol{\gamma})] \geq c(\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\gamma}_0)])^{1/2+\Delta}$$

holds almost surely,

then there is a sequence $\{\hat{\boldsymbol{\gamma}}_n\}$ *of random variables and a random number* n_2 *with*

(i) $P\{\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) = \mathbf{0} \quad \forall n \geq n_2\} = 1$ *(asymptotic existence),*

(ii) $\hat{\boldsymbol{\gamma}}_n \xrightarrow{a.s.} \boldsymbol{\gamma}_0$ *(strong consistency).*

Proof. The statement and assumptions of this theorem are the same as the ones for Theorem 3.16, except for only one difference in assumption (S^*_Δ), where U_{CM} is used instead of U_{UM}. The constrained space U_{CM} is a subset of U_{UM}. Differentiability and continuity of the log-likelihood function for all $\boldsymbol{\gamma} \in U_{CM}$ stated in Proposition 3.2 and Theorem 3.3 also hold, and, by Propositions 3.7 and 3.10, U_{CM} is still open and convex. This means that, for n large enough and with probability one, N^{**} contains both $\boldsymbol{\gamma}_0$ and also a small enough ball of center $\boldsymbol{\gamma}_0$ and radius ϵ, with $\epsilon > 0$, defined as $K_\epsilon(\boldsymbol{\gamma}_0)$, just as in Theorem 3.16. Given that $\boldsymbol{\gamma}_0$ is assumed to be in the constrained space, then there is a small enough ϵ so that $K_\epsilon(\boldsymbol{\gamma}_0)$ belongs to both the unconstrained and constrained space. Therefore, Theorem 3.19 holds with probability one for large enough n and small enough ϵ based on Theorem 3.16. ■

For not large enough n, $\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) = \mathbf{0}$ does not guarantee that $\hat{\boldsymbol{\gamma}}_n$ is a global optimum of the likelihood. It could be the case where the global optimum is on the border of the constrained space, which means that $\mathbf{s}_n(\hat{\boldsymbol{\gamma}}_n) \neq \mathbf{0}$.

3.6 Asymptotic normality

As seen in Section 3.5.5, when the true pattern of parameters associated with the ordinal predictors is monotonic, then asymptotic existence and strong consistency of the MLEs was proved for the constrained POCLM. In terms of asymptotic normality, Fahrmeir and Kaufmann (1985) show it for the unconstrained POCLM. The only difference between the unconstrained and constrained version

of the POCLM is related to their parameter spaces, U_{UM} versus U_{CM} correspondingly (defined in (3.2.2) and (3.3.4)). The parameter space U_{CM} is a subset of the one of the unconstrained model, U_{UM}, and, by Propositions 3.7 and 3.10, U_{CM} is still open and convex. Therefore, for large enough n and with probability one, the unconstrained MLEs are in the constrained space, and because of that the unconstrained MLEs will be equal to the constrained MLEs, which also means that their asymptotic distribution (normality) will be the same. Because of this, theorems about asymptotic normality in Fahrmeir and Kaufmann (1985) still hold for the constrained POCLM under the additional assumption that the true pattern of parameters associated with the ordinal predictors is monotonic.

3.7 Asymptotic confidence regions

Confidence regions for the parameters of the unconstrained POCLM were defined in Section 3.2.2. Based on Sections 3.5 and 3.6, asymptotic theory indicates that when the effects of ordinal predictors are monotonic, the asymptotic properties of the unconstrained and constrained MLEs are the same. This means that, asymptotically, for every arbitrarily small $\Delta > 0$ there is a large enough n so that the UMLE and the CMLE belong to a small ball around the true parameter $\boldsymbol{\gamma}_0$ defined as

$$B_\Delta(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : ||(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)|| \leq \Delta\}. \tag{3.7.1}$$

Now assume that $B_\Delta(\boldsymbol{\gamma}_0)$ is in the true monotonicity region, meaning that the monotonicity directions of the parameters associated with the ordinal predictors of every parameter vector belonging to $B_\Delta(\boldsymbol{\gamma}_0)$ are the same as those of the true monotonic parameter vector. Then it is possible to choose n large enough so that the UMLE and the CMLE belong to $B_{\Delta/2}(\boldsymbol{\gamma}_0)$ and, at the same time, that the confidence region is in $B_{\Delta/2}(\hat{\boldsymbol{\gamma}})$ (note that $B_{\Delta/2}(\hat{\boldsymbol{\gamma}})$ is a ball around the estimator with radius $\Delta/2$). The size of a confidence region normally decreases as n increases. Therefore, it is guaranteed that the confidence region belongs to the ball around the true parameter vector $\boldsymbol{\gamma}_0$, $B_\Delta(\boldsymbol{\gamma}_0)$. Therefore, asymptotically, the approximate confidence region for the constrained parameters is the same as the one for the unconstrained ones.

For finite n, the quality of the approximation of a confidence region for the constrained parameters is unclear. Four cases are distinguished depending on whether the parameter estimates from the constrained and unconstrained model are the same or not, and, for the first three cases, they also depend on the location of their confidence region:

Case 1: If the UMLE is the same as the CMLE and their confidence region is fully in the constrained parameter space U_{CM} with monotonicity directions indicated by the parameter estimates (also referred to as *monotonicity region of the parameter estimates*), then using the results of the asymptotic theory discussed in previous sections should not be problematic, because only monotonic patterns that share the same monotonicity direction for each ordinal predictor would be compatible with their confidence region. This is possible when the vector of parameter estimates is in the the monotonicity region of the parameter estimates and far away enough from the border of it, and when n is large enough to make the confidence regions small enough so that only monotonic patterns are compatible with them.

If this is not the case, then there are some situations for which using the asymptotic theory could be problematic. The reasons why they are considered as problematic are discussed in the next cases and their implications will be explored in the next section.

Case 2: If the UMLE is the same as the CMLE and there is just one combination of monotonicity directions in the confidence region (only one monotonicity direction for each ordinal predictor), but in addition there are not monotonic parameter vectors in it, then using the results of the asymptotic theory discussed in previous sections is problematic because, according to it, the confidence set should not contain parameter values belonging to a parameter set that violates some monotonicity constraints. This situation calls into question the validity of the approximation resulting from the asymptotic theory. The reason of this problem could normally be that n is not yet large enough so that the confidence set is not small enough to be fully in the

monotonicity region of the parameter estimates and/or that the vector of parameter estimates is not far away enough from the border of it (which is again associated with n).

Case 3: If the UMLE is the same as the CMLE but the confidence region for the parameters associated with some ordinal predictors is compatible with more than one combination of monotonicity directions, then using the approximations resulting from the asymptotic theory discussed in previous sections could be problematic because confidence regions compatible with more than one combination of monotonicity directions indicate that the finite n situation is different from the one on which the asymptotic theory is based. In this case the point constrained MLE scenario is the same as the one on which asymptotic theory is based, but the estimator of variation is too large for the finite n case. Therefore, there is some doubt about the quality of the approximation of the confidence region. It is also important to notice that when more than one combination of monotonicity directions are compatible for the same ordinal predictors, then normally there are also parameters that do not fulfil any monotonicity constraint in the confidence region. This scenario is possible when the vector of parameter estimates is in the constrained parameter space but close to the border of it, and when n is not large enough so that the confidence region is large enough to allow more than one combination of monotonicity directions for the parameters of the same ordinal predictors. Because of this, constrained parameter estimates can switch from one monotonicity direction to the other, potentially producing a multimodal distribution of parameter estimators, which is not well approximated by the normal distribution obtained in the asymptotic theory.

Case 4: If the UMLE is different from the CMLE, then this means that the monotonicity constraints were active when fitting the CMLE and therefore the UMLE does not even belong to the constrained parameter space. Hence, because this situation is different from what is required in the asymptotic theory, it is unclear why the approximation resulting from it should be good,

and therefore, theoretically there is no strong argument in using confidence regions for the constrained MLE based on the asymptotic theory discussed in the previous sections. However, this will be analysed by means of a simulation study in Section 3.7.1 in order to assess whether it is still reasonable to use a confidence region based on the asymptotic theory discussed in the previous sections, which will be defined in the following discussion.

In practice, for Cases 2, 3 and 4 the quality of the asymptotic approximation of a confidence region defined by the formula of the confidence region for the unconstrained parameters (3.2.8) could be under doubt. If the UMLE and CMLE are the same, then, for cases 2 and 3, those parameter values that make a parameter vector violate monotonicity must not be included in the confidence region of the constrained parameters. If the UMLE and CMLE are different, then, in addition to excluding non-monotonic values, a clear ambiguity is which estimator will be the centre of the confidence region. Therefore, some possible definitions of confidence regions are proposed:

1. One possibility is to use (3.7.2) defined below, a constrained confidence region that is based on the formula of the confidence region for the unconstrained parameters (3.2.8) but uses the results of the constrained POCLM, i.e., the CMLE. For the reasons discussed at the beginning of this section, this is fully correct for Case 1, but it could be doubtful for the other cases.

   For a vector with r parameters of interest, $\boldsymbol{\beta}_r$, the overall parameter vector $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ is partitioned as $(\boldsymbol{\beta}_r', \boldsymbol{\phi}')$, where $\boldsymbol{\phi}$ is a vector with the remaining $(p - r)$ parameters. The constrained MLE is now denoted as $(\hat{\boldsymbol{\beta}}'_{c,r}, \hat{\boldsymbol{\phi}}'_c)$ accordingly.

   A tentative confidence region for the parameter vector $\boldsymbol{\beta}_r$ in the context of the constrained POCLM can be constructed by:

$$\text{CCR} = \left\{ \boldsymbol{\beta}_{0r} : 2[\ell(\hat{\boldsymbol{\beta}}_{c,r}, \hat{\boldsymbol{\phi}}_c) - \ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}}_c)] \leq \chi^2_{(r);1-\alpha}, \boldsymbol{\beta}_{0r} \in \tilde{U}_{CM} \right\} \quad (3.7.2)$$

   where the degrees of freedom are r because it is the number of parameter values that are being tested to be $\boldsymbol{\beta}_{0r}$, and $\tilde{\boldsymbol{\phi}}_c$ is the vector of maximum likelihood estimators as a function of the value of $\boldsymbol{\beta}_{0r}$, where $\tilde{\boldsymbol{\phi}}_c$ is defined by

$\ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}}_c) = \max_{\boldsymbol{\beta}_r \in \tilde{U}_{CM}, \boldsymbol{\phi} \in U_{CM}, \boldsymbol{\beta}_r = \boldsymbol{\beta}_{0r}} \ell(\boldsymbol{\beta}_r, \boldsymbol{\phi})$, with U_{CM} being the monotonicity region of the CMLE $(\hat{\boldsymbol{\beta}}'_{c,r}, \hat{\boldsymbol{\phi}}'_c)$. Therefore, $\tilde{\boldsymbol{\phi}}_c$ can be thought of as the updated constrained MLE for each value of $\boldsymbol{\beta}_{0r}$ and it guarantees that there is no other best option for the values of the components of $\tilde{\boldsymbol{\phi}}$ for given $\boldsymbol{\beta}_{0r}$. If $r = p$, $\boldsymbol{\beta}_r$ is p-dimensional, the terms $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}}_c$ and $\tilde{\boldsymbol{\phi}}_c$ are omitted, and in terms of notation $\boldsymbol{\beta}$ is actually $\boldsymbol{\gamma}$. The tentative confidence region defined in (3.7.2) is referred to as CCR (constrained confidence region) because it uses the constrained MLEs to build the confidence region.

Because of the potential problem described in Cases 2, 3 and 4, the confidence region (3.7.2) might contain parameter values that violate the monotonicity constraints, in which case it can still be adjusted by dropping all the non-monotonic parameter values included in it. Whether this is a good option or not will be analysed in the next section by means of a simulation study together with other possibilities that will be discussed.

This possibility guarantees that the confidence region will contain constrained parameters because it is centred at the constrained MLE.

2. Another possibility is to use the confidence region (3.2.8) defined in Section 3.2.2, i.e., the confidence region centred at the UMLE resulting from the unconstrained model, and modify it by not including those parts of the region that violate the monotonicity assumption. Formally, the confidence region for the parameter vector $\boldsymbol{\beta}_r$ is:

$$\text{UCR} = \left\{ \boldsymbol{\beta}_{0r} : 2[\ell(\hat{\boldsymbol{\beta}}_r, \hat{\boldsymbol{\phi}}) - \ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}})] \leq \chi^2_{(r);1-\alpha}, \boldsymbol{\beta}_{0r} \in U_{CM} \right\} \qquad (3.7.3)$$

where the degrees of freedom are r because it is the number of parameter values that are being tested to be $\boldsymbol{\beta}_{0r}$, and $\tilde{\boldsymbol{\phi}}$ is the vector of maximum likelihood estimators as a function of the value of $\boldsymbol{\beta}_{0r}$, where $\tilde{\boldsymbol{\phi}}$ is defined by $\ell(\boldsymbol{\beta}_{0r}, \tilde{\boldsymbol{\phi}}) = \max_{(\boldsymbol{\beta}_r, \boldsymbol{\phi}) \in U_{UM}, \boldsymbol{\beta}_r = \boldsymbol{\beta}_{0r}} \ell(\boldsymbol{\beta}_r, \boldsymbol{\phi})$. Again, if $r = p$, then the implications on notation are the same as the ones for (3.7.2). The confidence region defined in (3.7.3) is referred to as UCR (unconstrained confidence region). The term unconstrained is in "UCR" because it uses the unconstrained MLEs to build the confidence region, but it is still constrained

because it excludes those parts that violate monotonicity with the condition $\boldsymbol{\beta}_{0r} \in U_{CM}$.

A disadvantage of using (3.7.3) is that if the UMLE is non-monotonic for the ordinal predictors, then UCR can be empty.

3. An additional option is to define the confidence region as the union of those resulting from the two previous approaches.

The performance of these three approaches will be analysed by means of a simulation study in the next section.

3.7.1 Confidence regions and coverage probability

The coverage probabilities (CPs) will be compared under different scenarios in order to assess the results of the three possible definitions of a confidence region presented in the previous section, as suggested in Morris et al. (2019) for the assessment of confidence intervals. In addition, given that Case 4 can be distinguished from Case 1, 2 and 3 by assessing whether the UMLE and CMLE are different or not, this comparison will also be analysed.

Consider model (2.3.4) with four ordinal predictors with 3, 4, 5, and 6 ordered categories each, one categorical (non-ordinal) predictor with 5 categories and one interval-scaled predictor. For every ith observation, each of the four ordinal predictors ($s = 1, \ldots, 4$) is represented in the model by dummy variables denoted as x_{i,s,h_s}, with $h_s = 2, \ldots, q_s$ and where $q_1 = 3$, $q_2 = 4$, $q_3 = 5$, and $q_4 = 6$; the nominal predictor is denoted as $x_{i,5,h_5}$ with $h_5 = 2, \ldots, 5$; and the interval-scaled predictor as $x_{i,1}$. The first category of the categorical variables is considered as the baseline so they are omitted. Thus, the model is

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \sum_{h_1=2}^{3} \beta_{1,h_1} x_{i,1,h_1} + \sum_{h_2=2}^{4} \beta_{2,h_2} x_{i,2,h_2} + \sum_{h_3=2}^{5} \beta_{3,h_3} x_{i,3,h_3} \\ + \sum_{h_4=2}^{6} \beta_{4,h_4} x_{i,4,h_4} + \sum_{h_5=2}^{5} \beta_{5,h_5} x_{i,5,h_5} + \beta_1 x_{i,1}, \qquad (3.7.4)\end{aligned}$$

where the number of categories of the ordinal response is $k = 4$, i.e., $j = 1, 2, 3$. This model was fitted for 500 data sets that were simulated as described in Section

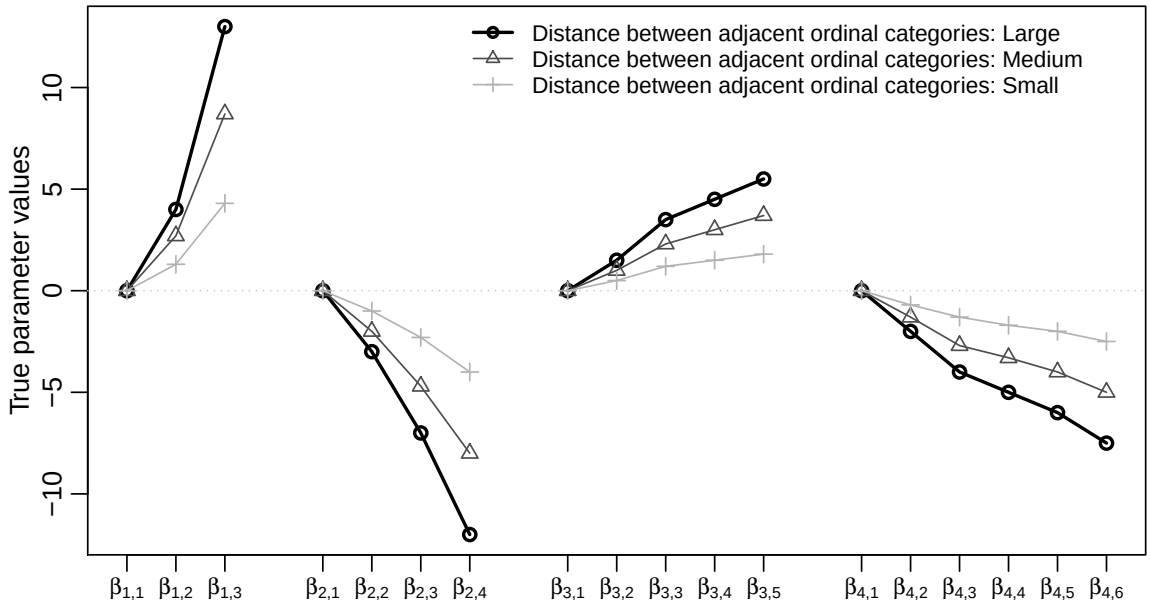

Figure 3.1: True parameter values for the simulation of coverage probabilities. Different line formats represent different distances between adjacent ordinal categories: Large, Medium, Small.

2.5 using the following true parameters: for the intercepts $\alpha_1 = -2$, $\alpha_2 = 2$, and $\alpha_3 = 5.5$; for the non-ordinal categorical predictor $\boldsymbol{\beta}_5' = (0.7, 1.4, -0.3, -1.2)$; and for the interval-scaled predictor $\beta_1 = 0.3$. The values of the ordinal predictors were drawn from the population distributions used in Section 2.5. The simulated values for the non-ordinal categorical predictor were drawn from the following population distribution: 0.2, 0.2, 0.3, 0.1, 0.2 for its corresponding categories 1, 2, 3, 4, and 5. The interval-scaled covariate x_1 was randomly generated from the normal distribution $N(1, 4)$.

Given that in simulations studies the results correspond to the design of specific scenarios (see Morris et al. (2019)), then the current simulation design offers 12 different scenarios depending on two factors: (i) distances between adjacent ordinal categories and (ii) sample sizes. The true parameter vectors of the ordinal predictors where chosen to represent three different levels of distances between their adjacent ordinal categories as shown in Figure 3.1. In addition, four different sample sizes were considered: $n = 50$, 100, 500, and 1,000.

Table 3.1 shows the results in terms of frequencies and coverage probabilities (CPs) of the three confidence regions defined in the previous section. Different scenarios were considered according to two factors: (i) the distance between the parameter values of adjacent ordinal categories representing three different *degrees of monotonicity* and (ii) four sample sizes, resulting in 12 scenarios. Factor (i) separates Table 3.1 in three sections, each one corresponding to a different level: "Small", "Medium", or "Large". For each one of the 12 scenarios, the three definitions of confidence regions explained in the previous section were considered: UCR, CCR and their union (the latter denoted as "Union").

In order to include the comparisons of cases where the unconstrained and constrained MLE are equal or not, two categories were defined: "Same MLE" and "Different MLE". The former corresponds to the group of cases 1, 2 and 3 of those discussed in Section 3.7 whereas the latter is equivalent to Case 4.

Within each one of the 12 scenarios and for each one of the three definitions of confidence regions, the frequency of cases was recorded separately depending on whether the true parameter was in or out of each confidence region. In Table 3.1, the former case is referred to as "True In" and the latter as "True Out". The confidence regions were computed using a significance level of $\alpha = 0.05$ and all the parameters ($r = p$).

The factor "distances between adjacent ordinal categories" with levels small, medium and large will be referred to as "monotonicity degree".

n	50			100			500			1000		
Conf. Region	UCR	CCR	Union	UCR	CCR	Union	UCR	CCR	Union	UCR	CCR	Union
Distances between adjacent ordinal categories: Small												
Same MLE, freq.												
True In		2			7			187			331	
True Out		0			1			4			8	
Total		2			8			191			339	
Different MLE, freq.												
True In	445	418	484	446	467	477	282	291	291	136	139	139
True Out	53	80	14	46	25	15	27	18	18	25	22	22
Total	498	498	498	492	492	492	309	309	309	161	161	161
Coverage probability, %.												
Same MLE		(100)			(87.5)			97.9			97.6	
Different MLE	89.4	83.9	97.2	90.7	94.9	97.0	91.3	94.2	94.2	84.5	86.3	86.3
Total	**89.4**	**84.0**	**97.2**	**90.6**	**94.8**	**96.8**	**93.8**	**95.6**	**95.6**	**93.4**	**94.0**	**94.0**
Distances between adjacent ordinal categories: Medium												
Same MLE, freq.												
True In		8			76			406			464	
True Out		0			1			21			29	
Total		8			77			427			493	
Different MLE, freq.												
True In	463	440	488	379	394	402	61	62	62	7	7	7
True Out	29	52	4	44	29	21	12	11	11	0	0	0
Total	492	492	492	423	423	423	73	73	73	7	7	7
Coverage probability, %.												
Same MLE		(100)			98.7			95.1			94.1	
Different MLE	94.1	89.4	99.2	89.6	93.1	95.0	83.6	84.9	84.9	(100)	(100)	(100)
Total	**94.2**	**89.6**	**99.2**	**91.0**	**94.0**	**95.6**	**93.4**	**93.6**	**93.6**	**94.2**	**94.2**	**94.2**
Distances between adjacent ordinal categories: Large												
Same MLE, freq.												
True In		11			89			460			476	
True Out		0			0			23			22	
Total		11			89			483			498	
Different MLE, freq.												
True In	471	426	485	390	388	395	13	13	13	2	2	2
True Out	18	63	4	21	23	16	4	4	4	0	0	0
Total	489	489	489	411	411	411	17	17	17	2	2	2
Coverage probability, %.												
Same MLE		(100)			100			95.2			95.6	
Different MLE	96.3	87.1	99.2	94.9	94.4	96.1	(76.5)	(76.5)	(76.5)	(100)	(100)	(100)
Total	**96.4**	**87.4**	**99.2**	**95.8**	**95.4**	**96.8**	**94.6**	**94.6**	**94.6**	**95.6**	**95.6**	**95.6**

Note: Parentheses indicate that the coverage probability was calculated on a total number under 20.

Table 3.1: Frequencies and coverage probabilities for different sample sizes, definitions of confidence regions, distances between adjacent ordinal categories, and cases according to whether the unconstrained and constrained MLE are the same or not. For the block "Same MLE, freq." the row "Total" shows the total number of cases with the same MLEs of which rows "True In" and "True Out" refer to cases where the true parameter was in or out of the corresponding confidence region accordingly. The same structure holds for the block "Different MLE, freq.".

In order to assess whether the coverage probabilities that are smaller than the confidence level of 95% are too low or not, it should be kept in mind that the 0.05 quantile of the binomial distribution with $n = 500$ (the number of simulation replicates) and $p = 0.95$ is 467. Based on a one-sided 5% test, this means that the coverage probabilities under 93.4% (467/500) are significantly smaller than the confidence level of 95%. Then 93.4% is considered as the threshold with which this assessment is done. On the other hand, if the coverage probabilities are too high, this is not a problem in itself but an indication that the confidence region is less precise than it could be, and therefore not totally desirable.

By construction, the CP resulting from Union is always higher than or equal to the one of UCR or CCR. It is shown in Table 3.1 that when the sample sizes are small (n=50, 100) the coverage probabilities for the totals are much higher for Union than for UCR or CCR, and even higher than the confidence level of 0.95 regardless of the monotonicity degree. For UCR and small sample sizes, the CPs are not significantly smaller than 95% when the monotonicity degree is medium or large, and significantly smaller (89.4% only) when the monotonicity degree is small. For CCR the CPs are all significantly smaller than 95% for n=50, ranging from 84.0% to 89.6%. In particular, when the sample size is 50, the CPs of CCR are worse than those of UCR. This is one of the implications of misclassification resulting from the MDC procedure, meaning that for some data sets the confidence region is centered around some parameter estimates that are in the wrong monotonicity direction compared to the one of the true parameter, and therefore the latter is more likely to be out of CCR than of UCR.

For the greater sample sizes, n=500, 1000, the CPs for the totals range between 93.6% and 95.6% for Union, meaning that none of them is significantly smaller than the confidence level. Only two scenarios show a CP higher than 95% though, these are the ones of (i) small monotonicity degree and n=500, and (ii) large monotonicity degree and n=1000. For Union, the CPs of the largest sample size are smaller than those for n=50 or 100, because a larger n decreases the confidence region up to the point that the CPs decrease too much to capture the true parameter value, despite the fact that asymptotically the parameter estimates get

closer to the true parameters. To see this, consider increasing n from 50 to 1000 for two extreme scenarios according to the monotonicity degrees:

Small: The initial (n=50) CP of Union is 97.2%. The proportion of cases where the UMLE and CMLE are the same ("Same MLE") increases from 0.4% to 67.8%. Therefore, the CP of 97.2% when n=50 corresponds almost completely to the case "Different MLE". When n=1000, for "Same MLE" the CP is 97.6% and for "Different MLE" is much lower, 86.3%. This means that when the MLEs are different, the confidence region is not large enough to include the true parameter value in almost 14% of the cases. Putting "Same MLE" and "Different MLE" together, the CP for n=1000 is below 95%, being 94.0%, which is still considered here as good enough taking into account the threshold of 93.4% discussed earlier and that this situation is not close to the one on which the asymptotic theory is based.

Large: The initial (n=50) CP of Union is really high, 99.2%. The proportion of cases where the UMLE and CMLE are the same increases from 2.2% to 99.6%. When n=1000, these cases show a CP of 95.6%, which is the same CP for the total ("Same MLE"+"Different MLE"). Despite the fact that there is a decrease of the CP as n increases, the final CP is still higher than 95% because of the high proportion of "Same MLE", which indicates that it is more likely to be in a situation that is close to the one on which the asymptotic theory is based.

In general, the results of the simulations are consistent with the asymptotic theory discussed in previous sections. Given the monotonicity degree, as n increases throughout the whole set of sample sizes that were considered in the simulation, the three confidence regions tend to the same coverage probability. In fact, given a monotonicity degree, they all reach the same value when $n = 1000$, except for the case when the monotonicity degree is small, because the one for UCR is smaller than the others. The latter is because under a small monotonicity degree a larger n is needed by the CMLE and UMLE to belong to the same monotonicity region.

Figure 3.2 shows, for each monotonicity degree, the CPs of the three confidence

regions without making further distinctions ("all cases", depicted by solid lines) and also those for the case when the UMLE and CMLE are different ("different MLEs", dashed lines). When comparing the CPs of "all cases" against those of "Different MLE" for n=50, they all start almost at the same point in Figure 3.2 because the proportion of "Different MLE" for any monotonicity degree is 97.8% or greater for this sample size. For larger sample sizes, the CPs for "Different MLEs" are much lower that those of "all cases". Furthermore, they increase their distance as n increases. However, the class "Different MLEs" decreases its frequency as n increases as shown in Table 3.1, reducing their proportion from 99.6% when n=50 to 32.2% when n=1000, and therefore reducing the impact of their low CPs on the total CPs.

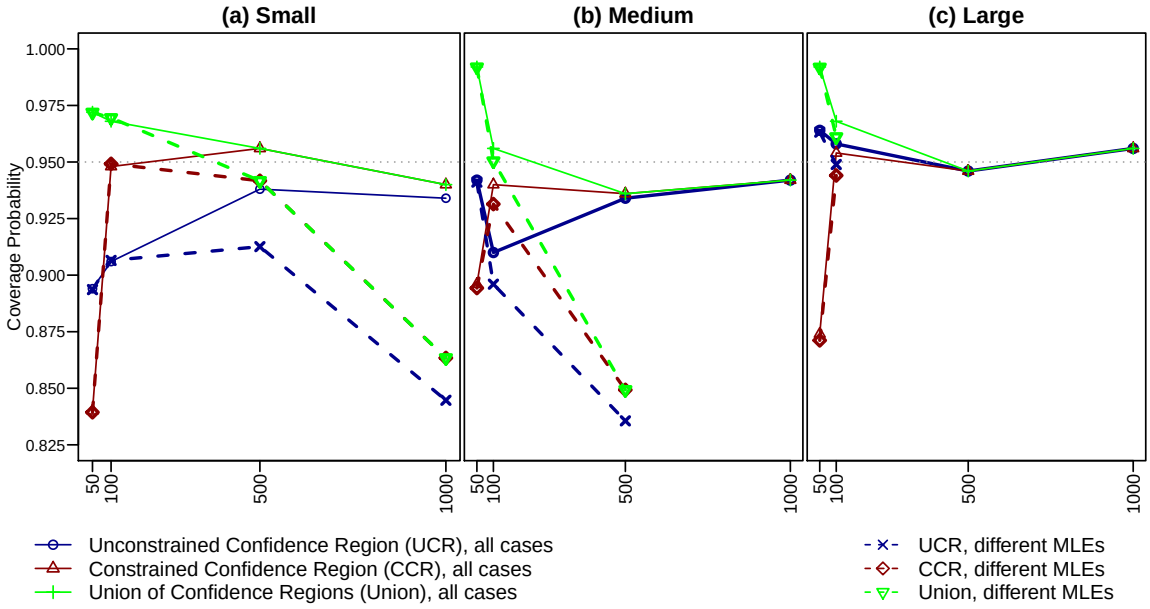

Figure 3.2: Coverage probabilities of confidence regions for different monotonicity degrees ((a) Small, (b) Medium, and (c) Large). Solid lines represent general results and dashed lines represent the results when the CMLE and UMLE were different.

Note: Points corresponding to coverage probabilities based on total numbers smaller than 20 were removed from the plot.

In terms of monotonicity degrees, the smaller the distance between adjacent

categories of the ordinal predictors, the higher CPs of the confidence regions defined as the Union compared to others. If a smaller (more precise) confidence region than Union is preferred, then the CPs of CCR are higher than those of UCR, except for n=50, despite the fact that the ones of UCR are still consistent with the asymptotic theory on which it is based. In addition, the CPs of CCR are not significantly smaller than 95% for $n \geq 100$. When $n = 50$ the approximation of any of these two confidence region shows a CP significantly smaller than 95% for the case of small monotonicity degree, whereas for higher monotonicity degrees the CPs of CCR only are significantly smaller than 95%.

Some of these results can be expected to generalise to other settings than the specific ones considered here. One of them is that as n increases, the CPs of UCR and CCR are expected to get closer, and consequently their union (Union) too. This is because as n increases, then the unconstrained and constrained MLEs get closer and, for large enough n and with probability one (see Section 3.6), the unconstrained MLEs are in the constrained space, and therefore their UCRs are expected to approximate the CCRs. Another result that could be generalised is that the larger the monotonicity degree of the parameter estimates of OPs, the faster the CPs get closer to each other as n increases. This is because when the monotonicity degree is large enough, the unconstrained MLEs require a relatively small sample size in order to belong to the constrained space, and therefore the UCR and CCR get closer.

The individual identification of Cases 1, 2 and 3 for which it is needed to diagnose whether the confidence region is either in one monotonicity region only or it also contains non-monotonic parameters or it allows more than one combination of monotonicity directions for the parameters of the ordinal predictors will be left for future work.

3.8 Asymptotic confidence intervals

In this section asymptotic confidence intervals are discussed for individual parameters, although there is still a connection with asymptotic confidence regions. As discussed in Section 3.7, asymptotic existence, strong consistency and asymp-

totic normality of the MLE for the constrained POCLM hold in the same way as they do for the unconstrained POCLM when the true parameter values associated with the ordinal predictors are monotonic. Therefore, asymptotically and as in the case of confidence regions, for every arbitrarily small $\Delta > 0$ there is a large enough n so that the UMLE and the CMLE belong to a small ball around the true parameter $\boldsymbol{\gamma}_0$ defined as in Equation (3.7.1). Following the same line of argument that was used in Section 3.7 for confidence regions but now applying it into the analysis of individual confidence intervals for individual parameters, assume that $B_\Delta(\boldsymbol{\gamma}_0)$ is in the true monotonicity region. Then, for large enough n both the UMLE and the CMLE belong to $B_{\Delta/2}(\boldsymbol{\gamma}_0)$ and, at the same time, the confidence interval belongs to the ball around the parameter estimate with radius $\Delta/2$, $B_{\Delta/2}(\hat{\boldsymbol{\gamma}})$. This is possible because the range of a confidence interval decreases as n increases, and then it will belong to an arbitrarily small $B_{\Delta/2}(\hat{\boldsymbol{\gamma}})$ if n is chosen to be large enough. Therefore, asymptotically, the confidence intervals for the parameters of the constrained POCLM are the same as the ones for the unconstrained POCLM defined in (2.4.1), Section 2.4.

In practice, for a given data set and finite n, the approximate confidence intervals could be problematic because of the same reasons discussed in Section 3.7. Values of non-monotonic parameter vectors should be removed from the confidence interval of a constrained parameter. However, each single confidence interval does not provide information to do this, because monotonicity is not a feature of a single parameter, but of a parameter vector. This is because whether a value contained in a confidence interval is monotonic or not depends on the values of parameters belonging to other confidence intervals. For instance, a particular value of a confidence interval could be part of a monotonic pattern for a given set of other parameter values, but it could also be part of a non-monotonic pattern for a given set of different parameter values. Therefore, the identification of those parts of the confidence intervals that are not compatible with monotonicity cannot be a result of analysing individual confidence intervals separately. This means that multivariate correlation must be taken into account, leading back to the analysis of confidence regions described earlier in Section 3.7.

Within a confidence interval, and even for apparently clear monotonic patterns, it is not possible to identify all those values that are part of non-monotonic parameter vectors by analysing individual confidence intervals. For example, consider an isotonic pattern of parameter estimates for one ordinal predictor with large distance between the borders of adjacent confidence intervals, but the first confidence interval allows some negative values, which violate monotonicity. These negative values belong to the first confidence interval given "other parameter values" belonging to other confidence intervals, which already converts the one-dimensional analysis into a multidimensional one. If those "other parameter values" belong to their corresponding confidence intervals given positive parameter values of the first confidence interval too, then removing the negative ones from the first confidence interval would not produce modifications on the the range of other confidence intervals. However, this again requires a multidimensional analysis rather than the analysis of the first individual confidence interval only. Furthermore, this multidimensional analysis should not rely on the analysis of confidence regions because there is not direct relationship between confidence regions and confidence intervals as the latter may not be projections of the former. This is because the limits of an individual confidence interval for a parameter at a given significance level do not take into account the distribution of other parameters, whereas confidence regions do it. This makes it possible that, for the same confidence level, a parameter value that is on the border of a confidence region might not be part of its corresponding confidence interval and vice versa. On the other hand, removing negative values from the first CI is inappropriate under the scenario that the "other parameter values" belong to their corresponding confidence intervals given that the parameter values of the first confidence interval are negative only. In this case, removing the last ones disable the "other parameter values" to be part of their CIs. However, these cannot be identified analysing their individual confidence intervals and, consequently, they cannot be removed, meaning that the range of the confidence intervals of other parameters would be overestimated. In general, the conclusion is that it is not possible to identify whether a value of a confidence interval belongs to a non-monotonic parameter vector by analysing confidence intervals separately.

The computation of confidence intervals for the constrained parameters is still of interest despite the fact that the whole set of parameter values that belong to the confidence intervals and violate monotonicity cannot be identified by analysing individual confidence intervals. If the UMLE and the CMLE are the same, there certainly is a reason to believe that the asymptotic approximations given by a confidence interval as defined in (2.4.1) (Section 2.4) are more reasonable than when they are not the same, although the quality of the approximation still depends on how close the estimates are to the border of the monotonicity region. For instance, if they are too close, it is likely that part of some confidence interval(s) will belong to a different monotonicity direction compared to the one of the parameter estimates, or even they could belong to a non-monotonic region, then it implies that there are parameter values of the confidence interval that belong to parameter vectors that violate monotonicity, bringing with it the problem related to their identification discussed earlier. If the UMLE and the CMLE are not the same, this indicates that the situation is certainly different from the one on which the asymptotic theory discussed in previous sections is based, calling into question the quality of the approximation of the confidence interval. Then, in addition to the identification problem, a clear ambiguity is which estimator will be the centre of the confidence interval. Therefore, some possible definitions of confidence intervals are proposed:

1. One possibility is to use (3.8.1) defined below, a constrained confidence interval that is based on the formula of the confidence interval for the unconstrained parameters defined in (2.4.1), Section 2.4, but now it uses the results of the constrained POCLM, i.e., it is centred at the CMLE and uses its corresponding standard errors. Thus, the approximate confidence interval of γ is defined as follows:

$$\hat{\gamma} \pm z_{\tilde{\alpha}/2}(SE_{\hat{\gamma}}), \tag{3.8.1}$$

where $z_{\tilde{\alpha}/2}$ denotes the standard normal percentile with probability $\tilde{\alpha}/2$ and $\hat{\gamma}$ can be any component belonging to either $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_{(ord)}$, or $\hat{\boldsymbol{\beta}}_{(nonord)}$. The values for all $\hat{\gamma}$ are obtained by fitting the constrained POCLM and the values

for their corresponding $SE_{\hat{\gamma}}$ are the result of computing the square root of the diagonal of the negative inverse of the Hessian matrix (see Appendix A for partial derivatives).

For the reasons discussed earlier in this section, (3.8.1) holds when the UMLE and CMLE are the same and they are in the interior of a monotonicity region, far away enough from the border, so that the situation is the one on which asymptotic theory is based. Otherwise, the quality of the approximation of the constrained confidence interval (3.8.1) could be doubtful because it could contain parameter values that are members of parameter vectors that violate monotonicity constraints.

This definition of a confidence interval for the parameters of the constrained POCLM guarantees that all the confidence intervals will contain constrained parameters.

2. Another possibility is to use the confidence intervals resulting from (2.4.1), the unconstrained model. However, some values in the confidence interval could be members of parameter vectors that are non-monotonic. Furthermore, if all the parameter values belonging to these confidence intervals are members of parameter vectors that are non-monotonic, then, if they were removed, the resulting adjusted confidence intervals would be empty.

3. An additional option is to define the confidence intervals as the union of those resulting from the two previous approaches.

3.9 Conclusions

In Section 3.3 it is shown that the likelihood function $L(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{x})$ and its logarithm are continuous and differentiable at $\boldsymbol{\gamma} \in U_{UM}, \tilde{U}_{CM}$ or U_{CM} (see proofs of Propositions 3.1 and 3.2 correspondingly). These properties together with openness and convexity of the constrained parameter space U_{CM} (see proofs of Propositions 3.7 and 3.10 in Section 3.4) are part of the arguments to show consistency in Section 3.5.5.

When $n \to \infty$ and the set of parameters associated with each ordinal predictor is strictly monotonic, then the true monotonicity direction classification of the set of parameters associated with each ordinal predictor is indicated by the one of the MLEs with probability one (see Corollary 3.18) and asymptotic consistency of the MLEs holds for the constrained POCLM (see Theorem 3.19).

In order to achieve these results, asymptotic existence and strong consistency of the unconstrained MLE in generalised linear models with natural link functions is explicitly extended to the case of non-natural link functions, as the one resulting from the POCLM, by Theorem 3.16. Next, as the interest is in the case of the POCLM with monotonicity constraints, Theorem 3.17 states that the MLE of the parameters of the constrained POCLM is in the right monotonicity direction, where Corollary 3.18 is a special case with $\boldsymbol{\gamma}_0 \in U_{CM}$. In Section 3.5.5, Theorem 3.19 states asymptotic existence and strong consistency of the MLEs for the constrained version of the POCLM based on results previously proved.

Regarding asymptotic normality for the MLE of the constrained POCLM, the reasons why theorems about asymptotic normality in Fahrmeir and Kaufmann (1985) still hold are discussed in Section 3.6 when the true parameters associated with the ordinal predictors are monotonic.

All of these results allowed to analyse confidence regions for the parameters of the constrained POCLM in Section 3.7. Asymptotically, the approximate confidence region for the constrained parameters is the same as the one for the unconstrained ones. However, for finite n the quality of the approximation of a confidence region is unclear. Four cases are distinguished, for the first three the UMLE is the same as the CMLE whereas for the fourth case they are different. The first three cases differ in terms of what is inside of their confidence region. It is either (i) fully in the constrained parameter space U_{CM}, or (ii) indicates only one combination of monotonicity directions but it also has non-monotonic parameter vectors in it, or (iii) allows multiple combinations of monotonicity directions. For (i) the use of the results of the asymptotic theory discussed in Sections 3.5 and 3.6 is not problematic, however this is unclear for cases (ii) and (iii). For the fourth case, where the UMLE is different from the CMLE, the situation is

different from what is required in the asymptotic theory and therefore there is no strong argument in using confidence regions for the constrained MLE based on the asymptotic theory discussed in Sections 3.5 and 3.6.

Three alternative definitions of a confidence region for the constrained MLE of the POCLM are proposed: the constrained confidence region (CCR) defined in Equation (3.7.2), the unconstrained confidence region (UCR) defined in Equation (3.7.3), and the third definition is the union of these two. Their performance is analysed in terms of their coverage probability in Section 3.7.1. By construction, the third proposed definition performs better than the others or at the same level because it is the union of the first two, keeping or increasing their maximum coverage probability. In general, comparisons between UCR and CCR according to Table 3.1 and Figure 3.2 show that when the sample size is small ($n = 50$) the UCR performs better than the CCR, however, for larger sample sizes ($n \geq 100$), the coverage probability of CCR is greater than or equal to the one of UCR.

Asymptotic confidence intervals are discussed in Section 3.8. Like in the case of confidence regions, asymptotically, the confidence intervals for the parameters of the constrained POCLM are the same as the ones for the unconstrained POCLM. These are defined in (2.4.1), Section 2.4. However, for finite n, the computation of an approximate confidence interval for parameters under monotonicity constraints is problematic because CIs do not allow to identify those parameter values that belong to a parameter vector that violates monotonicity. Each single confidence interval does not provide information to do this, because monotonicity is not a feature of a single parameter, but of a parameter vector. However, the computation of approximate confidence intervals for the constrained parameters is still of interest, for which three alternative definitions are proposed. Similarly to the case of the confidence region, they all hold when the UMLE and CMLE are the same and they are in the interior of a monotonicity region, far away enough from its border. However, when the situation is different from the one on which asymptotic theory is based, the quality of the approximation of the constrained confidence intervals could be doubtful.

Chapter 4

Monotonicity tests

4.1 Introduction

The discussion about the problem of testing monotonicity in the context of regression analysis is abundant in the literature. However, the solutions that have been proposed are different from what is required for the regression models discussed here because of two reasons: (i) the existing monotonicity tests hold for a single independent variable, or, when compatible with the multiple regression framework, (ii) the monotonicity test is not designed to take into account ordinal predictors. To my knowledge, there is no monotonicity test for regression models with ordinal predictors.

Several authors have proposed monotonicity tests in the framework of non-parametric regression models, see, for instance, Bowman et al. (1998), Hall and Heckman (2000), Gijbels et al. (2000), Ghosal et al. (2000), Durot (2003) and Chetverikov (2019). Many of them represent the regression model as $Y = f(X)+\epsilon$, where Y and X are scalar real valued random variables, f is an unknown smooth function, and the error term ϵ is independent of X with $E(\epsilon) = 0$. They all share the drawback of testing whether $f(\cdot)$ is non-decreasing only, restricting the analysis to one monotonicity direction only, enforcing a redefinition of $f(\cdot)$ in case testing the opposite direction is required. More importantly, these monotonicity tests work for models with a single independent variable, which is not compatible with the current context of multiple predictors. The latter was addressed by van

Beek and Daniels (2014), who proposed a non-parametric monotonicity test for what they called "partial monotonicity" in multiple regression models, but their approach is still restricted to test whether $f(\cdot)$ is non-decreasing only. On the other hand, Doveh et al. (2002) proposed a monotonicity test in the framework of parametric setting. However, it still holds for a single independent variable and it also test for a non-decreasing association only.

Despite the fact that all of these monotonicity tests work in the regression analysis context, none of them considers ordinal predictors in a multivariate context, which is addressed in this chapter.

Depending on the data set, the pattern of parameter estimates for an ordinal predictor resulting from fitting the unconstrained model (2.3.4) might indicate a clear monotonic association between an ordinal predictor and the ordinal response. This is the case when the differences between parameter estimates of adjacent ordered categories of the ordinal predictor are large enough and all positive or negative. On the other extreme, the pattern could also show a clear non-monotonic association, i.e., when the differences between adjacent parameter estimates of an ordinal predictor are all large but some of them are positive and others negative. However, the unconstrained parameter estimates could also show patterns that are not so clear in terms of their monotonic association. For instance, for an ordinal predictor of 10 categories, just one of its unconstrained parameter estimates could indicate non-monotonicity, which could be attributable to random variation of the sample. Therefore, the researcher would face the need of using a formal monotonicity test to obtain evidence about whether a pattern could be considered as monotonic or not.

The MDC procedure assists the decision on the choice of an appropriate monotonicity direction assumption for each OP when fitting model (2.3.4), but it is not a formal monotonicity test. It relies on the analysis of multiple pairwise comparisons of confidence intervals with flexibly chosen confidence levels without caring about the simultaneous error probability. Hence, two formal monotonicity test are proposed.

One of the two monotonicity test proposed in this chapter is based on the Bon-

ferroni correction, which was published already in Espinosa and Hennig (2019). The confidence intervals of the parameters of an ordinal predictor are used simultaneously to construct a monotonicity test based on the Bonferroni correction (see Miller (1981), p. 67, and Bonferroni (1936)). This test does not take into consideration the correlation between parameter estimates and its simultaneous significance level decreases as the number of categories of the OP increases. Although none of these reasons undermine the validity of the test, its results could be considered as too conservative. Hence, in order to provide a less conservative alternative test, a monotonicity test based on confidence regions is proposed in Section 4.3.

The choice of the base category for categorical variables is one of the decisions to be made by the researcher. The elementary choice for ordinal variables is the first or last categories. Then, given that there is not a unique valid alternative, the invariance under change of base category of the monotonicity test based on confidence regions is explored in Section 4.3.1.

Finally, an alternative definition of the monotonicity test based on confidence regions is analysed in Section 4.3.2. This alternative definition uses reparametrisation, with which it is possible to obtain estimates of the difference between adjacent parameters of an ordinal predictor, and it will be shown that its results are equivalent to the ones of the original proposal.

4.2 A monotonicity test based on Bonferroni correction

When analysing the monotonicity assumption on the parameters associated with an OP s, the Bonferroni correction method can be used to construct a formal monotonicity test for an OP. This monotonicity test was published already in Espinosa and Hennig (2019). The Bonferroni correction method allows to compute a set of confidence intervals achieving at least a $100(1-\alpha_s^*)\%$ confidence level simultaneously (see Miller (1981), p. 67, and Bonferroni (1936)), which is the probability that all the parameters are captured by the confidence intervals simultaneously. For a given ordinal predictor s and a pre-specified α_s^*, if each one

of the $p_s - 1$ confidence intervals is built with a $100(1 - \alpha_s^*/(p_s - 1))\%$ confidence level, then the simultaneous confidence level will be at least $100(1 - \alpha_s^*)\%$.

The null hypothesis "H_0 : The parameters $\{\beta_{s,h_s} : h_s = 1, 2, \ldots, p_s\}$ are either isotonic or antitonic" ($0 \leq \beta_{s,2} \leq \beta_{s,3} \cdots \leq \beta_{s,p_s}$ (isotonic) and $0 \geq \beta_{s,2} \geq \beta_{s,3} \cdots \geq \beta_{s,p_s}$ (antitonic)) is tested against the alternative "H_1 : The parameters $\{\beta_{s,h_s} : h_s = 1, 2, \ldots, p_s\}$ are neither fully isotonic nor fully antitonic" for a given OP s, and setting $\beta_{s,1} = 0$ as in previous sections.

For a given ordinal predictor s, and taking advantage of the ordinal information provided by its categories, it is then checked whether all the confidence intervals simultaneously are compatible with monotonicity.

In order to identify whether there are pairs of confidence intervals of β_{s,h_s} that are incompatible with monotonicity, a slight modification of equations (2.4.2) and (2.4.3) is used. Now, instead of the confidence level $\tilde{c}$, those equations use $\tilde{b} = 1 - \alpha_s^*/(p_s - 1)$. Therefore, the monotonicity test for an ordinal predictor s is

$$T_{s,\tilde{b}} = \begin{cases} \textit{reject } H_0 & \text{if } \mathcal{D}_{s,\tilde{b}} \supseteq \{-1, 1\} \\ \textit{not reject } H_0 & \text{otherwise} \end{cases} \tag{4.2.1}$$

where $\mathcal{D}_{s,\tilde{b}} = \{d_{s,h_s,h'_s,\tilde{b}}\}$ is defined as the set of distinct values resulting from using Equation (2.4.3) for the ordinal predictor s considering each confidence interval with a $100\tilde{b}\%$ confidence level (instead of $100\tilde{c}\%$) in order to achieve a simultaneous confidence level of at least $100(1 - \alpha_s^*)\%$ for the parameters associated with the OP s. The Bonferroni correction adjusts the individual confidence level of $(p_s - 1)$ confidence intervals associated with an OP s in order to obtain a simultaneous confidence level of at least $100(1 - \alpha_s^*)\%$ for the set of $(p_s - 1)$ individual CIs. Those adjusted individual confidence intervals are the ones to be used in Equation (2.4.3), which involves $p_s(p_s - 1)/2$ comparisons of the CIs' limits in order to find the $p_s(p_s - 1)/2$ indicators of relative positions of the adjusted individual confidence intervals that define $\mathcal{D}_{s,\tilde{b}}$.

If $T_{s,\tilde{b}} = \textit{reject } H_0$, then the parameters associated with the ordinal predictor s are not compatible with the monotonicity assumption with a simultaneous confidence level of at least $100(1 - \alpha_s^*)\%$.

When applying this monotonicity test to the four ordinal predictors of the illustration discussed in Section 2.5 and using a pre-specified $\alpha_s^* = 0.05$, all the ordinal predictors were found to be compatible with the monotonicity assumption.

For a given pre-determined significance level of α_s^* (say 0.1, 0.05 or 0.01), the Bonferroni correction will often be very conservative, and it will be the more conservative the higher the number of ordinal categories involved in the monotonicity test is. A higher p_s implies larger ranges of the intervals, making the test more likely to not reject H_0.

In order to show some results for the monotonicity test with ordinal predictors for which their association with the response variable is truly non-monotonic, consider a setting for model (2.3.4) with two OPs only ($t = 2$ and $v = 0$), where $p_1 = 4$, $p_2 = 5$, and $k = 4$, i.e., $j = 1, 2, 3$. The parameters for the intercepts are $\alpha_1 = -1$, $\alpha_2 = -0.5$, and $\alpha_3 = -0.1$; and the true sets of parameters of the ordinal predictors 1 and 2 represent non-monotonic associations, being $\boldsymbol{\beta}_1' = (0.4, 1.7, 0.8)$ and $\boldsymbol{\beta}_2' = (-0.25, -0.70, -0.05, 0.40)$. The distributions among categories of ordinal predictors 1 and 2 are the same as the ones shown in Figure 2.2 for OPs 2 and 3 correspondingly, and the number of observations is 2,000. This setting corresponds to the one that was published already in Espinosa and Hennig (2019).

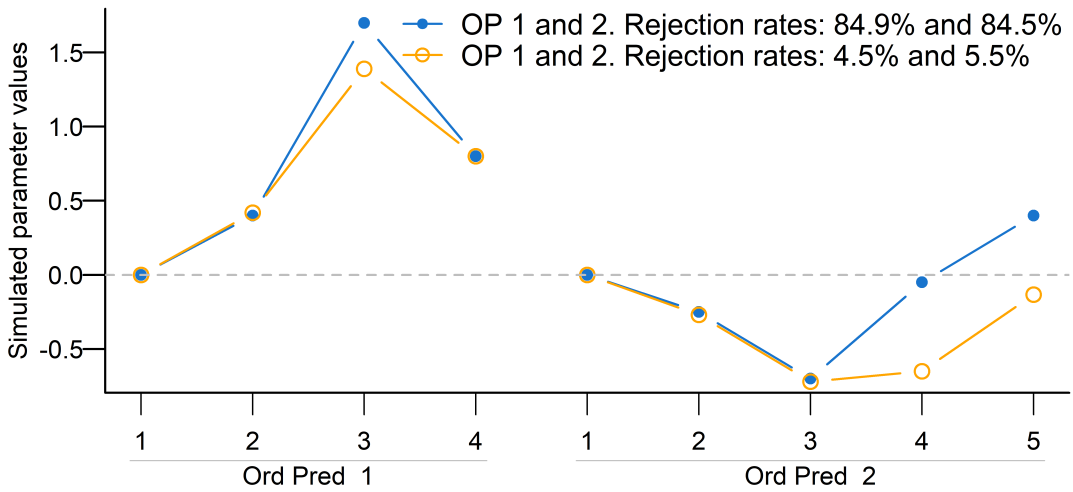

Figure 4.1: True parameter patterns simulating non-monotonicity with different rejection rates of the monotonicity test based on Bonferroni correction.

After fitting the new unconstrained model on 1,000 simulated data sets and testing for monotonicity, the rejection rate was analysed, as suggested in Morris et al. (2019) for the assessment of a hypothesis test. The null hypothesis was rejected in 84.9% of the data sets for the OP 1 and in 84.5% for the second OP, in both cases with $\alpha_s^* = 0.05$. Figure 4.1 shows the patterns of these non-monotonic ordinal predictors together with additional patterns with which rejection rates of around 5% are obtained (4.5% and 5.5% respectively).

4.3 A monotonicity test based on confidence regions

Consider the confidence region (3.2.8) for the parameters of the unconstrained POCLM model (2.3.4). The aim of the monotonicity test is to establish whether a point that is compatible with monotonicity is in this confidence region or not. Among all those points compatible with monotonicity, the one obtained through MLE is chosen, i.e., the MLE under monotonicity constraints.

Assume the base category b is chosen to be the first category of an ordinal predictor s, and consequently $\beta_{s,1} = 0$, and that we want to test whether the parameters of the ordinal predictor are monotonic in some particular direction, e.g., isotonic. Then, from the parameters $(\{\alpha_j\}, \boldsymbol{\beta})$ of the model (2.3.4), the parameter vector associated with the set of predictors $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_s, \boldsymbol{\beta}'_{(nonOrd)})$, where $\boldsymbol{\beta}_s$ is a vector with $(p_s - 1)$ parameters of the ordinal predictor s, and $\boldsymbol{\beta}_{(nonOrd)}$ is composed of all the parameters associated with the remaining predictors. The unconstrained parameter estimates are obtained by MLE of model (2.3.4). Similarly, the constrained parameter estimates are also obtained by MLE but under monotonicity constraints, which need to be defined. To set the maximisation problem of the constrained model, and following the same reasoning as in Section 2.3.2 but now with the aim of making the choice of the base category flexible, we define an $(p_s - 1)$-dimensional square matrix depending on the value

of b, the choice of the base category, which in this case is assumed to be $b = 1$:

$$\mathbf{C}_s = \mathbf{C}^I_{s,b} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & -1 & 1 \end{bmatrix}. \tag{4.3.1}$$

The way in which this matrix is defined for any choice of b is discussed below, without affecting the maximisation problem:

$$\begin{aligned} &\text{maximise } \ell(\{\alpha_j\}, \boldsymbol{\beta}) \\ &\text{subject to } \mathbf{C}_s\boldsymbol{\beta}_s \geq \mathbf{0}, \end{aligned} \tag{4.3.2}$$

where $\mathbf{0}$ is a vector of $p_s - 1$ components. (4.3.2) can be expressed as the Lagrangian

$$\mathcal{L}(\{\alpha_j\}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \ell(\{\alpha_j\}, \boldsymbol{\beta}) - \boldsymbol{\lambda}'\mathbf{C}_s\boldsymbol{\beta}_s, \tag{4.3.3}$$

where $\boldsymbol{\lambda}$ is the vector of $p_s - 1$ Lagrange multipliers.

Under the current scenario, the null hypothesis is "H_0: the parameters of ordinal predictor s are isotonic" ($0 \leq \beta_{s,2} \leq \beta_{s,3} \leq \cdots \leq \beta_{s,p_s}$) and we set a significance level α. Then, the decision rule is:

$$\text{reject } H_0 \text{ if } 2[\ell(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}}) - \ell(\{\tilde{\alpha}_j\}, \tilde{\boldsymbol{\beta}})] > \chi^2_{(p_s-1);1-\alpha}, \tag{4.3.4}$$

where $\{\hat{\alpha}_j\}$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimators of the unconstrained model (2.3.4), and $\{\tilde{\alpha}_j\}$ and $\tilde{\boldsymbol{\beta}}$ are obtained by solving (4.3.3), the constrained MLE version of the model (2.3.4) assuming an isotonic pattern for ordinal predictor s. Thus, $\ell(\{\tilde{\alpha}_j\}, \tilde{\boldsymbol{\beta}})$ will be the closest to $\ell(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})$ under H_0, and therefore any other choice of $\{\tilde{\alpha}_j\}$ and $\tilde{\boldsymbol{\beta}}$ will make the left hand side of the inequality in (4.3.4) to be even greater than the boundary of the confidence region, $\chi^2_{(p_s-1);1-\alpha}$.

In order to make the choice of the base category more flexible, for instance choosing the last category as the baseline, it could be of interest to define the base category of an ordinal predictor s as any of its categories rather than the first one only, in which case the matrix $\mathbf{C}^I_{s,b}$ defined in Equation (4.3.1) for testing an isotonic pattern is replaced by an $(p_s - 1)$-dimensional square matrix with elements

for each $\tilde{i}$-th row and $\tilde{j}$-th column defined as

$$c^I_{s,b,\tilde{i},\tilde{j}} = \begin{cases} 1 & \text{if } \tilde{i} = \tilde{j} \geq b \text{ or } \tilde{i} + 1 = \tilde{j} < b \\ -1 & \text{if } \tilde{i} = \tilde{j} < b \text{ or } \tilde{i} = \tilde{j} + 1 > b \\ 0 & \text{otherwise,} \end{cases} \tag{4.3.5}$$

for some base category b with $1 \leq b \leq m_s$ and $\tilde{i}, \tilde{j} = 1, \ldots, m_s - 1$. For example, for an ordinal predictor s with 4 categories, the corresponding matrices $\mathbf{C}^I_{s,1}$, $\mathbf{C}^I_{s,2}$, $\mathbf{C}^I_{s,3}$ and $\mathbf{C}^I_{s,4}$ for the different possible choices of the base category are

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Hence, if the researcher is testing whether the pattern of parameters is isotonic with base category b, Equation (4.3.3) uses the matrix $\mathbf{C}^I_{s,b}$ with elements defined in (4.3.5) to obtain the parameter estimates to be used in the decision rule (4.3.4). If the monotonicity direction to be tested is antitonic, then use $\mathbf{C}_s = \mathbf{C}^A_{s,b} = -\mathbf{C}^I_{s,b}$ instead of $\mathbf{C}_s = \mathbf{C}^I_{s,b}$.

Rejecting H_0 means that the pattern of parameters is not compatible with the particular monotonicity direction used in the null hypothesis with a significance level α. However, the other monotonicity direction could still be compatible. Therefore, the test should be used twice in order to establish either a specific monotonicity direction, both or none.

4.3.1 Invariance under change of base category

The monotonicity test discussed in Section 4.3 is based on comparing the unconstrained model (2.3.4) against the same model under monotonicity constraints, from which the corresponding log-likelihoods $\ell(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}})$ and $\ell(\{\tilde{\alpha}_j\}, \tilde{\boldsymbol{\beta}})$ are used in the decision rule (4.3.4). To show that the results of the monotonicity test are invariant against changes of the reference category, we show that each of these log-likelihoods is the same regardless of the base category choice.

Consider the model (2.3.4) with $v = 1$ non-ordinal predictor and $t = 1$ ordinal predictor composed of $p_1 = 4$ categories. Assuming the choice of base category as

$b = 1$, the model is

$$\text{logit}[P(y_i = j|\mathbf{x}_i)] = \alpha_j + \beta_{1,2}x_{1,2} + \beta_{1,3}x_{1,3} + \beta_{1,4}x_{1,4} + \beta_1 x_1, \tag{4.3.6}$$

where $\beta_{1,1}x_{1,1}$ is usually omitted because $\beta_{1,1} = 0$. A different choice of the base category can be understood as a reparametrisation of model (4.3.6).

Assume a the change of base category from $b = 1$ to $b = 3$. Given that

$$x_{i,s,b} = 1 - \sum_{\forall l_s \in \{1,\ldots,p_s\}\setminus\{b\}} x_{i,s,l_s}, \tag{4.3.7}$$

the reparametrised model (4.3.8) below can be re-written in the form of model (4.3.6) as follows

$$\text{logit}[P(y_i = j|\mathbf{x}_i)] = \alpha_j^* + \beta_{1,1}^* x_{i,1,1} + \beta_{1,2}^* x_{i,1,2} + \beta_{1,4}^* x_{i,1,4} + \beta_1^* x_{i,1}, \tag{4.3.8}$$

$$\begin{aligned} &= \alpha_j^* + \beta_{1,1}^*(1 - x_{i,1,2} - x_{i,1,3} - x_{i,1,4}) + \beta_{1,2}^* x_{i,1,2} + \beta_{1,4}^* x_{i,1,4} + \beta_1^* x_{i,1}, \\ &= \alpha_j^* + \beta_{1,1}^* + (\beta_{1,2}^* - \beta_{1,1}^*)x_{i,1,2} - \beta_{1,1}^* x_{i,1,3} + (\beta_{1,4}^* - \beta_{1,1}^*)x_{i,1,4} + \beta_1^* x_{i,1}. \end{aligned} \tag{4.3.9}$$

Therefore, the change of base category from $b = 1$ in model (4.3.6) to $b = 3$ in model (4.3.8) is just a reparametrisation of (4.3.6) because

$$\alpha_j = \alpha_j^* + \beta_{1,1}^*, \quad \beta_{1,2} = \beta_{1,2}^* - \beta_{1,1}^*, \quad \beta_{1,3} = -\beta_{1,1}^*, \quad \beta_{1,4} = \beta_{1,4}^* - \beta_{1,1}^* \text{ and } \beta_1 = \beta_1^*. \tag{4.3.10}$$

The parameters affected by the reparametrisation are those associated with the intercepts and the ordinal predictor of interest. The remaining parameters are not affected. This is the case for any change of base category. Thus, for an unconstrained model, $\ell(\{\hat{\alpha}_j\}, \hat{\boldsymbol{\beta}}) = \ell(\{\hat{\alpha}_j^*\}, \hat{\boldsymbol{\beta}}^*)$.

When monotonicity constraints are imposed on both model (4.3.6) and its reparametrised version (4.3.8), any change of base category does not affect the log-likelihood either. Assume that the monotonicity constraints on the ordinal predictor of model (4.3.6) are isotonic. Then, the constraints are $0 \leq \beta_{1,2} \leq \beta_{1,3} \leq \beta_{1,4}$, or

$$\begin{aligned} \beta_{1,2} &\geq 0 \\ \beta_{1,3} - \beta_{1,2} &\geq 0 \\ \beta_{1,4} - \beta_{1,3} &\geq 0, \end{aligned} \tag{4.3.11}$$

whereas for model (4.3.8), the constraints are $\beta^*_{1,1} \leq \beta^*_{1,2} \leq 0 \leq \beta^*_{1,4}$, or

$$\begin{aligned} \beta^*_{1,2} - \beta^*_{1,1} &\geq 0 \\ -\beta^*_{1,2} &\geq 0 \\ \beta^*_{1,4} &\geq 0, \end{aligned} \tag{4.3.12}$$

which, using (4.3.10), can be re-written as

$$\begin{aligned} \beta_{1,2} &\geq 0 \\ \beta_{1,3} - \beta_{1,2} &\geq 0 \\ \beta_{1,4} - \beta_{1,3} &\geq 0, \end{aligned} \tag{4.3.13}$$

which are exactly the same as (4.3.11), i.e., a change of the base category not only leads to a reparametrisation of the model but also the isotonic constraints on models (4.3.6) and (4.3.8) are equivalent. Therefore, $\ell(\{\tilde{\alpha}_j\}, \tilde{\boldsymbol{\beta}}) = \ell(\{\tilde{\alpha}^*_j\}, \tilde{\boldsymbol{\beta}}^*)$, and consequently the log-likelihood ratio used in decision rule (4.3.4) is invariant against changes in the choice of the base category.

4.3.2 A note on reparametrisation and the monotonicity test

In Section 4.3 we use monotonicity constraints on the parameters of ordinal predictor s when fitting model (2.3.4) to estimate the p-dimensional vector $(\{\tilde{\alpha}_j\}, \tilde{\boldsymbol{\beta}})$ and test whether this point is in the p-dimensional confidence region of the parameters resulting from the unconstrained model (2.3.4), i.e. it is compatible with the monotonicity direction established in the null hypothesis, or not.

Rather than estimating the effects of the parameters associated with the categories of an ordinal predictor, a researcher could be interested in estimating the differences between adjacent parameters associated with the categories of an ordinal predictor. This can be addressed by taking an alternative approach, which is to use a reparametrised version of model (2.3.4) to obtain estimates of the difference between adjacent parameters of an ordinal predictor s ($\beta_{s,l_s} - \beta_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$). For the isotonic case, all these differences must be non-negative, regardless of the choice of the base category. Therefore, to test the hypothesis of an

isotonic pattern, the reparametrised model is constrained to obtain non-negative differences and H_0 is rejected if the constrained result is out of the confidence region of the parameters of the unconstrained reparametrised model.

To see how the reparametrisation works, consider the following model for an ordinal response variable with k categories and one ordinal predictor with $p_1 = 4$ categories for which different choices of the base category are examined:

Base category $b = 1$: in this case $\beta_{1,1} = 0$ and the general model is

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \beta_{1,2}x_{i,1,2} + \beta_{1,3}x_{i,1,3} + \beta_{1,4}x_{i,1,4}, \tag{4.3.14}$$

with $j = 1, \ldots, k-1$ and where $\mathbf{x}_i$ is a vector of $p_1 - 1$ components for each ith observation representing $p_1 - 1$ dummy variables associated with the categories of the OP, excluding the one corresponding to the base category.

Model (4.3.14) is reparametrised through the following:

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j^* + \beta_{1,2}^*(x_{i,1,2} + x_{i,1,3} + x_{i,1,4}) + \beta_{1,3}^*(x_{i,1,3} + x_{i,1,4}) + \beta_{1,4}^* x_{i,1,4}. \tag{4.3.15}$$

Therefore, $\beta_{1,2} = \beta_{1,2}^*$, $\beta_{1,3} = \beta_{1,2}^* + \beta_{1,3}^*$ and $\beta_{1,4} = \beta_{1,2}^* + \beta_{1,3}^* + \beta_{1,4}^*$, which is easy to see if we re-write model (4.3.15) as

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] &= \alpha_j^* + \beta_{1,2}^*(x_{i,1,2} + x_{i,1,3} + x_{i,1,4}) + \beta_{1,3}^*(x_{i,1,3} + x_{i,1,4}) + \beta_{1,4}^* x_{i,1,4} \\ &= \alpha_j^* + \beta_{1,2}^* x_{i,1,2} + (\beta_{1,2}^* + \beta_{1,3}^*)x_{i,1,3} + (\beta_{1,2}^* + \beta_{1,3}^* + \beta_{1,4}^*)x_{i,1,4}\end{aligned} \tag{4.3.16}$$

and, more importantly, $\beta_{1,2}^* = \beta_{1,2}$, $\beta_{1,3}^* = \beta_{1,3} - \beta_{1,2}$ and $\beta_{1,4}^* = \beta_{1,4} - \beta_{1,3}$, i.e., the parameters β_{s,l_s}^* represent the difference between adjacent parameters of the original model (4.3.14), $\beta_{s,l_s}^* = \beta_{s,l_s} - \beta_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$.

Base category $b = 2$: if the second ordinal category is assumed to be the baseline, i.e., $\beta_{1,2} = 0$, the original model is

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \beta_{1,1}x_{i,1,1} + \beta_{1,3}x_{i,1,3} + \beta_{1,4}x_{i,1,4}, \tag{4.3.17}$$

and the reparametrised one is

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] &= \alpha_j^* - \beta_{1,2}^* x_{i,1,1} + \beta_{1,3}^*(x_{i,1,3} + x_{i,1,4}) + \beta_{1,4}^* x_{i,1,4} \\ &= \alpha_j^* - \beta_{1,2}^* x_{i,1,1} + \beta_{1,3}^* x_{i,1,3} + (\beta_{1,3}^* + \beta_{1,4}^*) x_{i,1,4} \quad (4.3.18)\end{aligned}$$

from which we find that $\beta_{1,2}^* = -\beta_{1,1}$, $\beta_{1,3}^* = \beta_{1,3}$ and $\beta_{1,4}^* = \beta_{1,4} - \beta_{1,3}$. Again, $\beta_{s,l_s}^* = \beta_{s,l_s} - \beta_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$.

Base category $b = 3$**:** in this case $\beta_{1,3} = 0$ and the original model is

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \beta_{1,1} x_{i,1,1} + \beta_{1,2} x_{i,1,2} + \beta_{1,4} x_{i,1,4}, \quad (4.3.19)$$

and the reparametrised one is

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] &= \alpha_j^* - \beta_{1,2}^* x_{i,1,1} - \beta_{1,3}^*(x_{i,1,1} + x_{i,1,2}) + \beta_{1,4}^* x_{i,1,4} \\ &= \alpha_j^* - (\beta_{1,2}^* + \beta_{1,3}^*) x_{i,1,1} - \beta_{1,3}^* x_{i,1,2} + \beta_{1,4}^* x_{i,1,4} \quad (4.3.20)\end{aligned}$$

from which we find that $\beta_{1,2}^* = \beta_{1,2} - \beta_{1,1}$, $\beta_{1,3}^* = -\beta_{1,2}$ and $\beta_{1,4}^* = \beta_{1,4}$. Again, $\beta_{s,l_s}^* = \beta_{s,l_s} - \beta_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$.

Base category $b = 4$**:** finally, for $\beta_{1,4} = 0$ the original model is

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \beta_{1,1} x_{i,1,1} + \beta_{1,2} x_{i,1,2} + \beta_{1,3} x_{i,1,3}, \quad (4.3.21)$$

and the reparametrised one is

$$\begin{aligned}\text{logit}[P(y_i \leq j|\mathbf{x}_i)] &= \alpha_j^* - \beta_{1,2}^* x_{i,1,1} - \beta_{1,3}^*(x_{i,1,1} + x_{i,1,2}) - \beta_{1,4}^*(x_{i,1,1} + x_{i,1,2} + x_{i,1,3}) \\ &= \alpha_j^* - (\beta_{1,2}^* + \beta_{1,3}^* + \beta_{1,4}^*) x_{i,1,1} - (\beta_{1,3}^* + \beta_{1,4}^*) x_{i,1,2} - \beta_{1,4}^* x_{i,1,3} \quad (4.3.22)\end{aligned}$$

from which we find that $\beta_{1,2}^* = \beta_{1,2} - \beta_{1,1}$, $\beta_{1,3}^* = \beta_{1,3} - \beta_{1,2}$ and $\beta_{1,4}^* = -\beta_{1,3}$. Again, $\beta_{s,l_s}^* = \beta_{s,l_s} - \beta_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$.

In general, the reparametrised version of model (2.3.4) for the differences between adjacent parameters associated with an ordinal predictor s uses a partition of both parameters and predictors. We define the parameters $(\{\alpha_j^*\}, \boldsymbol{\beta}^*)$, from which the parameter vector $\boldsymbol{\beta}^*$ is partitioned as $\boldsymbol{\beta}^{*\prime} = (\boldsymbol{\beta}_s^{*\prime}, \boldsymbol{\beta}_{(nonOrd)}^{*\prime})$, where

$\boldsymbol{\beta}_s^{*\prime} = (\beta_{s,2}^*, \ldots, \beta_{s,p_s}^*)$ is associated with the ordinal predictor s, and $\boldsymbol{\beta}_{(nonOrd)}^*$ is composed of all the parameters associated with the remaining predictors, which could contain parameters of other ordinal predictors different from OP s in which case they are treated as of nominal scale type. Correspondingly, the vector of predictors for the i-th observation $\mathbf{x}_i$ is partitioned into $\mathbf{x}_{i,s}$ and $\mathbf{x}_{i,(nonOrd)}$, where $\mathbf{x}_{i,s}$ contains $(p_s - 1)$ components, excluding the one associated with the base category b. For instance, $\mathbf{x}_{i,s}' = (x_{i,s,1}, x_{i,s,2}, x_{i,s,4})$ for an ordinal predictor with 4 categories and base category $b = 3$. Then, the reparametrised model is defined as

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j^* + \mathbf{x}_{i,s}'\mathbf{R}_{s,b}\boldsymbol{\beta}_s^* + \mathbf{x}_{i,(nonOrd)}'\boldsymbol{\beta}_{(nonOrd)}^*, \tag{4.3.23}$$

where $\mathbf{R}_{s,b}$ is a square (p_s-1)-dimensional matrix that allows the reparametrisation of the parameters of ordinal predictor s when the b-th category is chosen as the base category with $1 \leq b \leq p_s$, whose elements for each $\tilde{i}$-th row and $\tilde{j}$-th column are defined as

$$r_{s,b,\tilde{i},\tilde{j}} = \begin{cases} 1 & \text{if } \tilde{i} \geq \tilde{j} \text{ and } \tilde{j} \geq b \\ -1 & \text{if } \tilde{i} \leq \tilde{j} \text{ and } \tilde{j} < b \\ 0 & \text{otherwise,} \end{cases} \tag{4.3.24}$$

with $\tilde{i}, \tilde{j} = 1, \ldots, p_s - 1$. For example, for an ordinal predictor with 4 categories, the corresponding matrices $\mathbf{R}_{s,1}$, $\mathbf{R}_{s,2}$, $\mathbf{R}_{s,3}$ and $\mathbf{R}_{s,4}$ for the different possible choices of the base category are

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} -1 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & -1 & -1 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{bmatrix},$$

respectively.

Model (4.3.23) can be re-written as

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i^*)] = \alpha_j^* + \mathbf{x}_{i,s}^{*\prime}\boldsymbol{\beta}_s^* + \mathbf{x}_{i,(nonOrd)}^{*\prime}\boldsymbol{\beta}_{(nonOrd)}^*, \tag{4.3.25}$$

where $\mathbf{x}_i^{*\prime} = (\mathbf{x}_{i,s}^{*\prime}, \mathbf{x}_{i,(nonOrd)}^{*\prime})$ with $\mathbf{x}_{i,s}^{*\prime} = \mathbf{x}_{i,s}'\mathbf{R}_{s,b}$ and $\mathbf{x}_{i,(nonOrd)}^{*\prime} = \mathbf{x}_{i,(nonOrd)}'$, which is in the same form of model (2.3.4), and therefore, based on (3.2.1), the likelihood function is

$$\ell(\{\alpha_j^*\}, \boldsymbol{\beta}^*) = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \pi_j(\mathbf{x}_i^*). \tag{4.3.26}$$

The unconstrained MLEs of the reparametrised model (4.3.25), $(\{\hat{\alpha}_j^*\}, \hat{\boldsymbol{\beta}}^*)$, are obtained by fitting the POCLM. Each resulting parameter estimate in $\hat{\boldsymbol{\beta}}_s^*$ corresponds to the difference between adjacent parameter estimates associated with the ordinal predictor resulting from fitting the original (not reparametrised) unconstrained model (2.3.4), i.e., $\hat{\beta}_{s,l_s}^* = \hat{\beta}_{s,l_s} - \hat{\beta}_{s,l_s-1}$ with $l_s = 2, \ldots, p_s$.

Given that in the reparametrised model (4.3.25) $\boldsymbol{\beta}_s^*$ represents the differences of adjacent parameters associated with ordinal categories, the monotonicity constraints to be imposed on $\boldsymbol{\beta}_s^*$ enforces its components to be non-negative or non-positive if the ordinal predictor s is assumed to be isotonic or antitonic respectively.

Assuming an isotonic pattern for the parameters of OP s, the maximisation problem for the constrained reparametrised model is

$$\begin{aligned} &\text{maximise } \ell(\{\alpha_j^*\}, \boldsymbol{\beta}^*) \\ &\text{subject to } \boldsymbol{\beta}_s^* \geq \mathbf{0}, \end{aligned} \tag{4.3.27}$$

where $\mathbf{0}$ is a vector of $p_s - 1$ components. (4.3.27) can be expressed as the Lagrangian

$$\mathcal{L}(\{\alpha_j^*\}, \boldsymbol{\beta}^*, \boldsymbol{\lambda}) = \ell(\{\alpha_j^*\}, \boldsymbol{\beta}^*) - \boldsymbol{\lambda}' \boldsymbol{\beta}_s^*, \tag{4.3.28}$$

where $\boldsymbol{\lambda}$ is the vector of $p_s - 1$ Lagrange multipliers.

For the isotonic case, and according to the monotonicity test based on confidence regions proposed in Section 4.3, the null hypothesis is "H_0: the parameters of ordinal predictor s are isotonic" at some significance level α. Then, the decision rule is:

$$\text{reject } H_0 \text{ if } 2[\ell(\{\hat{\alpha}_j^*\}, \hat{\boldsymbol{\beta}}^*) - \ell(\{\tilde{\alpha}_j^*\}, \tilde{\boldsymbol{\beta}}^*)] > \chi^2_{(p_s-1);1-\alpha}, \tag{4.3.29}$$

where $\{\hat{\alpha}_j^*\}$ and $\hat{\boldsymbol{\beta}}^*$ are the maximum likelihood estimators of the unconstrained reparametrised model (4.3.23), and $\{\tilde{\alpha}_j^*\}$ and $\tilde{\boldsymbol{\beta}}^*$ are obtained by solving (4.3.28), the constrained MLE version of the reparametrised model (4.3.25) assuming an isotonic pattern for ordinal predictor s.

If the monotonicity direction to be tested is antitonic, then reverse the direction of the inequality in the constraints of the optimisation problem (4.3.27) and modify (4.3.28) accordingly. The decision rule (4.3.29) remains the same.

Rejecting H_0 means that the pattern of parameters is not compatible with the particular monotonicity direction used in the null hypothesis with a significance level α. However, the other monotonicity direction could still be compatible. Therefore, the test should be used twice in order to establish either a specific monotonicity direction, both or none.

4.4 Conclusion

Two monotonicity tests are proposed, one based on the Bonferroni correction and another on the analysis of confidence regions.

The former allows to make inference about the monotonicity of a pattern of parameters associated with an ordinal predictor with a simultaneous confidence level of at least $100(1-\alpha_s^*)\%$, which results from the analysis of individual $100(1-\alpha_s^*/(p_s-1))\%$ confidence intervals. Therefore, the higher the number of ordinal categories of an ordinal predictor s, the more conservative the result of the monotonicity test. The null hypothesis states that the pattern of parameters associated with an ordinal predictor s is monotonic.

The latter uses the analysis of the confidence region associated with the parameter estimates of a given ordinal predictor in order to establish whether a point that is compatible with the monotonicity direction used in the null hypothesis is in the confidence region or not with a $100(1-\alpha)\%$ confidence level. The null hypothesis states that the pattern of parameters associated with an ordinal predictor s follows a specific monotonicity direction.

The invariance under change of base category is discussed for the monotonicity test based on confidence regions, and it is shown that the results of the monotonicity test are invariant against changes of the reference category as the log-likelihoods are the same regardless of the base category choice.

In addition, when differences between adjacent parameters associated with an ordinal predictor s are of interest, then a reparametrised version of model (2.3.4) is proposed together with a version of the monotonicity test based on confidence regions discussed in Section 4.3.

The performance of these tests is investigated further under different settings

(simulated and real data sets) later in Chapter 6.

Chapter 5

Further estimation methods and variable selection

5.1 Introduction

In addition to the constrained method described in Section 2.4, which uses the three steps of the MDC procedure, another method that imposes monotonicity constraints on all of the OPs is described in Section 5.2, which chooses the model with the highest maximum likelihood over a set of models that are built according to all possible combinations of monotonicity directions. However, dropping monotonicity constraints could also be of interest. Therefore, five estimations methods will be proposed to take into account the possibility of not imposing the monotonicity constraint on some of the ordinal predictors. The five methods differ in the way they make the decision about the ordinal predictors for which their parameter estimates will not be constrained to be monotonic. Two of them consider the results of the monotonicity tests proposed in Chapter 4, one based on the Bonferroni correction and the other based on the analysis of confidence regions. They are discussed in Section 5.3.1 and Section 5.3.2 correspondingly. The remaining three proposed estimation methods discussed in Section 5.3.3 rely on the results of the steps of the MDC procedure previously proposed (see Section 2.4). Not imposing the monotonicity constraint on any of the ordinal predictors is the same as fitting the usual unconstrained POCLM proposed by McCullagh (1980).

Section 5.4 discusses the fact that the MDC procedure can also be used for variable selection. The first two steps of the MDC procedure are based on the analysis of confidence intervals. Therefore, when they classify the pattern of parameter estimates of an OP as 'both', it does not only mean that both monotonicity directions are compatible with the estimated pattern, but it also means that all of the parameter estimates are not statistically significant at the corresponding confidence level of step one or two. The use of this information as a reference for variable selection will also be discussed.

Those estimation methods based on the MDC procedure and the one that uses the monotonicity test based on the Bonferroni correction were published already in Espinosa and Hennig (2019).

5.2 Monotonicity direction classification by Maximum Likelihood over all possible combinations

When a monotonicity constraint is imposed on the effects of an OP, there are two options from which only one has to be chosen according to the monotonicity directions, 'isotonic' or 'antitonic'. Given t OPs, then there are 2^t combinations of monotonicity directions to choose from. The method of monotonicity direction classification by Maximum Likelihood over all possible combinations makes this choice by fitting 2^t constrained models, where each model differs from the others just because of the monotonicity directions that are used to impose the monotonicity constraints on the effects of the t ordinal predictors. Once the 2^t constrained models have been fitted, then the one that delivers the highest likelihood is selected and its combination of monotonicity directions is found, resulting the best option in terms of likelihood.

For scenarios where the number of OPs is high, this approach could be computationally demanding, for instance, for $t = 6$ there are 64 models to be fitted. The number of combinations of monotonicity directions rapidly increases as the number of OPs t gets greater, making the method of monotonicity classification

by Maximum Likelihood over all possibilities slower.

Another important factor to be considered is that this method does not take into account the possibility of not imposing monotonicity constraints on some or all of the OPs. In some situations, the unconstrained parameter estimates of an ordinal predictor indicate that its association with the response variable is not monotonic. If those kind of effects are constrained to be monotonic anyway, the parameter estimates of the remaining OPs will also be affected, which could even lead to misclassification of monotonicity direction as will be seen in Section 6.2.1. This is the main reason why methods that assist the researcher in making the decision of dropping monotonicity constraints for some or all of the OPs are proposed in the following section.

5.3 Dropping monotonicity constraints

The method of monotonicity direction classification by Maximum Likelihood over all possible combinations and the monotonicity direction classification procedure impose monotonicity constraints on all of the sets of parameters of the OPs. These methods do not incorporate the option of dropping monotonicity constraints. This means that, when using these methods, the researcher is forced to impose monotonicity constraints on all of the OPs, which is not a problem when the effects of all the OPs are monotonic. However, these methods are not a good approach to deal with OPs whose unconstrained effects indicate that there is a non-monotonic association with the response variable. This shows the need of more flexible methods. Therefore, five methods are proposed in this respect. Some of them are more conservative than others, requiring very strong evidence against monotonicity to determine that the monotonicity constraint should not be imposed on a set of parameter estimates of an OP. Another difference among them is that some are based on monotonicity tests and others on the MDC procedure.

5.3.1 Using the monotonicity test based on Bonferroni correction

The MDC procedure described in Section 2.4 implies that the parameter estimates of all ordinal predictors are restricted to be monotonic. However, the researcher may want to drop monotonicity constraints on ordinal predictors in case that there is clear evidence against monotonicity.

The monotonicity test proposed in Section 4.2 can be used as a complementary tool to the MDC procedure in order to assist the estimation process. If the researcher is open to the possibility of not imposing the monotonicity constraints on some ordinal predictors, then he/she could first test monotonicity on each one of them, then drop the monotonicity constraints on those ordinal predictors for which the null hypothesis was rejected, and finally perform the MDC procedure imposing monotonicity constraints on all the remaining ordinal predictors. Under this scenario, in case that monotonicity is rejected for an OP, it would be more prudent to fit unconstrained estimates on the parameters associated with it. Therefore, such an OP should not be part of $\mathcal{S}$, the set of OPs to be constrained, but rather part of the non-ordinal predictors, treating it as nominal-scaled.

5.3.2 Using the monotonicity test based on confidence regions

Similarly to the estimation method described in Section 5.3.1, an alternative complementary tool to the MDC procedure in order to assist the estimation process is the monotonicity test based on confidence regions proposed in Section 4.3 when the researcher is open to not impose monotonicity constraints on some OPs. Then, like in the previous section, the first step is to test monotonicity on each one of the ordinal predictors and drop the monotonicity constraints on those for which the null hypothesis is rejected. The parameter estimates associated with those OPs are now considered as unconstrained estimates. Therefore, those OPs are removed from the set $\mathcal{S}$, the set of OPs to be constrained, and treated as part of the non-ordinal predictors, treating them as nominal-scaled variables. Finally,

the MDC procedure is performed imposing monotonicity constraints on all the remaining OPs only.

5.3.3 Using the MDC procedure

When dropping the monotonicity constraint for some of the OPs is considered as a feasible option, then not only the approach introduced in Section 5.3.1 could be used, but also three alternative ones that are proposed in this section. As in the previous section, consider the case where the researcher might also want to explore whether the monotonicity assumption holds for all of the OPs or for a subset of them, but now using a less conservative (i.e., dropping constraints more easily) approach than the one based on the monotonicity test. We propose three additional methods. Two of them are based on the first and second steps of the MDC procedure correspondingly ('CMLE MDC S1' and 'CMLE MDC S2'), and another one is based on a slight modification of the MDC procedure ('CMLE filtered').

CMLE MDC S1

Both monotonicity constraints and monotonicity directions are established using the first step of the MDC procedure. Once it determines $\mathcal{I}_1$ and $\mathcal{A}_1$, the monotonicity constraints are dropped for the remaining ordinal predictors $\{s : s \notin (\mathcal{I}_1 \cup \mathcal{A}_1)\}$, namely $\{s : s \in (\mathcal{B}_1 \cup \mathcal{N}_1)\}$. Therefore, there is no need of executing further steps.

The model is fitted imposing monotonicity constraints on ordinal predictors $\{s : s \in (\mathcal{I}_1 \cup \mathcal{A}_1)\}$ using their corresponding monotonicity directions, and treats ordinal predictors $\{s : s \in (\mathcal{B}_1 \cup \mathcal{N}_1)\}$ as nominal-scaled variables.

This method is the least conservative one because it assumes that if a monotonicity direction is not established without adjustment of the confidence level $100\tilde{c}\%$, then the monotonicity constraint has to be dropped.

CMLE MDC S2

This method follows the same structure as the previous one but executing the MDC procedure until the end of its second step. Therefore, the third step is not executed and the model is fitted imposing monotonicity constraints on ordinal pre-

dictors $\{s : s \in (\mathcal{I}_2 \cup \mathcal{A}_2)\}$ only, using their corresponding monotonicity directions according to $\mathcal{I}_2$ and $\mathcal{A}_2$, and treating the ordinal predictors $\{s : s \notin (\mathcal{I}_2 \cup \mathcal{A}_2)\}$ as nominal-scaled variables.

CMLE filtered

An adjusted version of the MDC procedure described in Section 2.4 allows to drop the monotonicity assumption for some OPs. There are only two adjustments to be made by this approach, one in step 2.b and the other one in step 3. The first one is to set $\tilde{c}_s'^* = \tilde{c}$, i.e., the tolerance level for each OP $s \in \mathcal{N}_1$ is set to be the same as the confidence level chosen in step 1. Therefore, the second step is not performed on any ordinal predictor $s \in \mathcal{N}_1$. The second modification is to apply step 3 over the possible combinations of monotonicity directions of the ordinal predictors that were classified as 'both' by the end of step 2, i.e., the number of models to be fitted is now $2^{\#\{s:d_{s,\tilde{c}_s'}=\text{both}\}}$ instead of $2^{\#\{s:s\notin(\mathcal{I}_2\cup\mathcal{A}_2)\}}$. This implies that $\mathcal{S}$, the set of OPs to be constrained, must be updated excluding each ordinal predictor $s \in \mathcal{N}_1$ from the set of monotonicity constraints. Finally, the model should be fitted treating these OPs as nominal-scaled variables.

These adjustments are equivalent to considering the first step of the MDC procedure as a filter of OPs to be constrained, where those that are classified as 'none' by the end of this step are removed from $\mathcal{S}$ and excluded from steps 2 and 3.

5.4 MDC procedure and variable selection

The parameter estimates' patterns classified as 'both' at the end of the second step of the MDC procedure are also of interest. 'Both' refers to an ordinal predictor for which all of the parameters associated with its categories have CIs containing zero. Therefore, if this is true even for the CIs evaluated at the tolerance level, an option is to remove such an ordinal predictor from the model of interest and apply the MDC procedure again using the new model. If more than one OP is classified as 'both' and there is appetite to drop such variables, then it is advisable to do it in a stepwise fashion such as backward elimination, while checking the results of the MDC procedure in each step, because dropping an OP could affect

the monotonicity direction classification of another OP. We will not investigate this in detail here, assuming that the data is rich enough so that variable selection is not required.

5.5 Conclusions

If the researcher is open to the possibility of not imposing monotonicity constraints on some or all of the ordinal predictors and treat them as of nominal scale type, then the monotonicity tests and the steps of the MDC procedure are proposed as tools to help her/him make the decision of dropping the monotonicity constraint on the parameters of a given ordinal predictor. By construction, the method 'CMLE MDC S1' will indicate to drop the monotonicity constraint with the same or higher frequency than 'CMLE MDC S2' and 'CMLE filtered'. Similarly, 'CMLE Conf. Reg.' is less conservative than 'CMLE Bonferroni'.

The methods 'CMLE MDC S1' and 'CMLE MDC S2' do not use step 3 at all. The methods 'CMLE filtered' and the one described in Section 5.4, i.e., dropping monotonicity constraints for those ordinal predictors $s \in \mathcal{N}_1$ and dropping ordinal predictors $\{s : d_{s,\tilde{c}_s'^*} = \text{both}\}$, reduce the number of models to be fitted in step 3. If these last two methods are used simultaneously, then step 3 is avoided.

Comparisons among the results of the five estimation methods, the fully constrained estimation method of Section 2.4, and the unconstrained one will be discussed in Section 6.2. In Section 6.2.1 the differences between the two more restrictive methods will be analysed, i.e. monotonicity classification by Maximum Likelihood over all possible combinations and the monotonicity direction classification procedure. The differences among the results of using constrained methods against the use of scoring systems for the treatment of ordinal predictors will be analysed in Section 6.3. In the real data application, all of these methods will be used, however, it will be seen that one of them is the best according to the context of the analysis (see Section 6.4).

Chapter 6

Models results

6.1 Introduction

In Chapter 2 a constrained regression model for ordinal data was proposed assuming that the effects of every ordinal predictor are constrained to be monotonic in some direction (isotonic or antitonic). This assumption was relaxed in Chapter 5, where not imposing monotonicity constraints on the effects of some ordinal predictors is allowed. Five constrained methods were proposed in order to offer flexibility on the way the decision of dropping monotonicity constraints is made.

The results of the constrained methods proposed in Chapter 2 and Chapter 5 may differ in the classification of the parameter estimates of an ordinal predictor among the possible outcomes of monotonicity directions ('isotonic', 'antitonic', 'both' or 'none') and, consequently, in the value of their parameter estimates. A simulation study is conducted in Section 6.2 to analyse those differences. According to Morris et al. (2019), "simulation studies will often involve more than one data-generating mechanism to ensure coverage of different scenarios". This is why the performance of several methods is compared considering different factors such as sample sizes, number of predictors, monotonicity directions of the ordinal predictors' effects, and correlation among predictors. The first simulations setting uses two uncorrelated ordinal predictors with monotonic effects. For each combination of type of model, ordinal predictor, and sample size, an empirical distribu-

tion of the classes of monotonicity direction classification is computed ('isotonic', 'antitonic', 'both' and 'none'). These are shown in Table 6.1. The proportions of 'isotonic' and 'antitonic' are compared among constrained models using a set of McNemar hypothesis tests for differences between two dependent proportions (for an example of the process see Table 6.2). For the smallest sample size of the simulation setting ($n = 50$), the monotonicity direction classification shows a high rate of misclassification, regardless of the constrained method. However, for n=100 or larger, the classification improves significantly for the constrained methods, showing proportions of correct classification of 88.9% or higher, even when the true monotonic pattern is not that clear (OP 1), except for 'CMLE MDC S1' and 'CMLE MDC S2' for which this conclusion still holds for $n \geq 500$. All of these results also provide insights about the degree of conservativeness of each method regarding the decision of dropping monotonicity constraints, being 'CMLE Bonferroni' the most conservative method and 'CMLE MDC S1' the least conservative one.

The results of the constrained methods are compared between each other and against the ones of the unconstrained MLE in terms of their mean-squared errors. Again, pairwise comparisons are of interest, which leads to an elevated number of hypothesis test (840) because of the multiple factors to be considered (number of models, ordinal predictors, categories, and sample sizes), and therefore their analysis is addressed as an exploratory exercise (see Table 6.3 for an example of the process). Table 6.4 shows the results in terms of MSE (averaging the results associated with the categories of each OP). In general, given monotonicity of effects of ordinal predictors, the constrained methods perform better than the unconstrained one in terms of MSE.

All the analyses mentioned above were replicated after introducing correlation between the ordinal predictors, generating a new set of simulation settings. It will be seen that the results are affected but the general conclusions remain approximately the same.

The previous two simulation settings (two OPs with and without correlation) were built using monotonic patterns of effects for each OP. In order to analyse the

performance of the proposed constrained methods in a more complex context, a new pair of simulation settings was conducted and analysed. Both of them use four ordinal predictors instead of two, each one of them representing a different class of monotonicity direction ('isotonic', 'antitonic', 'both' and 'none'). The difference between these two new simulation settings is that one of them does not impose correlation among its OPs whereas the other does. Further analysis of these new scenarios are left to be discussed in the body of Section 6.2 and part of Section 6.6.

In Section 6.3 the constrained methods were compared against models using the POCLM for the treatment of an ordinal response and different scoring systems to transform ordinal predictors into interval-scaled variables. This analysis was made based on simulated data sets. The scoring systems are some of those presented in Section 1.4. In addition, a researcher could be interested in converting a set of ordinal predictors into a single interval-scaled variable. For this purpose, a latent variable model for ordinal data is used as a dimensionality reduction technique an is incorporated in the analysis (see Section 1.7.2). As in Tutz and Hechenbichler (2005), where different methods for an ordinal response were compared, the performance of these methods is assessed based on three measures of accuracy, despite the fact that they are not specially designed for ordinal data: the misclassification rate (MR), the mean absolute prediction error (MAPE), and the mean-squared prediction error (MSPE). In general, given that the true pattern of an OP is non-monotonic, the less conservative constrained methods show a better performance than the ones using scoring systems for the treatment of ordinal predictors, even when the sample size is small.

Finally, in Section 6.4 a real data application is used to illustrate in practice how the proposed methodologies work, analysing the association between a quality of life self-assessment variable (10-point Likert scale) and ordinal and other predictors from a Chilean survey, the National Socio-Economic Characterisation 2013 (CASEN). This application shows that the constrained methods are superior in terms of interpretability. In addition, all of the constrained methods were used in the context of the real data application and also some of the approaches using

scoring systems were used to compare their results against the constrained ones (see Section 6.4.1 and Section 6.4.2 correspondingly).

6.2 Constrained versus unconstrained POCLM

The model (2.3.4) with two ordinal and two interval-scaled predictors,

$$\text{logit}[P(y_i \leq j|\mathbf{x}_i)] = \alpha_j + \sum_{h_1=2}^{4} \beta_{1,h_1} x_{i,1,h_1} + \sum_{h_2=2}^{6} \beta_{2,h_2} x_{i,2,h_2} + \beta_1 x_{i,1} + \beta_2 x_{i,2}, \tag{6.2.1}$$

where $k = 5$, i.e., $j = 1, 2, 3, 4$, was fitted for 1,000 data sets simulated as described in Section 2.5 using the following parameters: for the intercepts $\alpha_1 = -1.4$, $\alpha_2 = -0.4$, $\alpha_3 = 0.3$, and $\alpha_4 = 1.1$; for the ordinal predictors' categories $\boldsymbol{\beta}_1' = (0.3, 1.0, 1.005)$, and $\boldsymbol{\beta}_2' = (-0.2, -1.5, -1.55, -2.4, -2.41)$; and for the interval-scaled predictors $\beta_1 = -0.15$ and $\beta_2 = 0.25$. The parameter vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were chosen to represent isotonic and antitonic patterns respectively. Several sample sizes were considered: n =50, 100, 500, 1000, 5000. The ordinal predictors were drawn from the population distributions used in Section 2.5 of those covariates with the same number of ordinal categories, 4 and 6. The interval-scaled covariates x_1 and x_2 were randomly generated from normal distributions, $N(0, 1)$ and $N(5, 4)$ correspondingly.

For each one of the 1,000 data sets and for every sample size, model (6.2.1) was fitted following one unconstrained estimation method and six different constrained methods:

1. UMLE (unconstrained MLE).

2. CMLE: constrained MLE based on the MDC procedure with $\tilde{c} = 0.90$ in step 1, $\tilde{c}_s'^* = 0.85$ and $\tilde{c}_s''^* = 0.999$ for $s = 1, 2$ in step 2, with versions using some or all of the steps of the MDC procedure:

   a) MDC S1 as described in Section 5.3.3,

   b) MDC S2 as described in Section 5.3.3,

c) MDC S3 as described in Section 2.4, imposing monotonicity constraints on all OPs.

3. CMLE Bonferroni: dropping monotonicity constraints on those ordinal predictors for which the null hypothesis of monotonicity was rejected as described in Section 5.3.1 and using the monotonicity test based on Bonferroni correction proposed in Section 4.2, with $\alpha_s^* = 0.05$, for $s = 1, 2$.

4. CMLE Conf. Reg.: dropping monotonicity constraints on those ordinal predictors for which the null hypothesis of monotonicity was rejected as described in Section 5.3.2 and using the monotonicity test based on confidence regions proposed in Section 4.3, with $\alpha = 0.05$.

5. CMLE filtered as described in Section 5.3.3, $\tilde{c} = 0.90$.

		True pattern	OP 1: Isotonic					OP 2: Antitonic				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000
CMLE	MDC S1	Isotonic	39.5	57.9	98.4	100	100	3.1	2.9	0.0	0.0	0.0
		Antitonic	5.6	2.0	0.1	0.0	0.0	39.1	82.1	98.1	98.5	99.7
		Both	54.8	39.8	1.4	0.0	0.0	56.9	13.2	0.0	0.0	0.0
		None	0.1	0.3	0.1	0.0	0.0	0.9	1.8	1.9	1.5	0.3
	MDC S2	Isotonic	47.5	65.5	99.2	100	100	5.3	5.6	0.0	0.0	0.0
		Antitonic	7.6	3.8	0.2	0.0	0.0	44.8	87.2	100	100	100
		Both	44.9	30.7	0.6	0.0	0.0	49.9	7.2	0.0	0.0	0.0
		None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MDC S3	Isotonic	69.2	92.9	99.8	100	100	9.1	5.7	0.0	0.0	0.0
		Antitonic	30.8	7.1	0.2	0.0	0.0	90.9	94.3	100	100	100
CMLE Bonferroni	Unconstrained		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MDC S3	Isotonic	69.2	92.9	99.8	100	100	9.1	5.7	0.0	0.0	0.0
		Antitonic	30.8	7.1	0.2	0.0	0.0	90.9	94.3	100	100	100
CMLE Conf. Reg.	Unconstrained		8.2	4.9	1.5	0.7	0.0	12.1	10.2	2.2	1.7	0.4
	MDC S3	Isotonic	68.1	89.9	98.5	99.3	100	3.2	0.9	0.0	0.0	0.0
		Antitonic	23.7	5.2	0.0	0.0	0.0	84.7	88.9	97.8	98.3	99.6
CMLE filtered	Unconstrained		0.1	0.3	0.1	0.0	0.0	0.9	1.8	1.9	1.5	0.3
	MDC S3	Isotonic	69.1	92.6	99.7	100	100	8.5	4.6	0.0	0.0	0.0
		Antitonic	30.8	7.1	0.2	0.0	0.0	90.6	93.6	98.1	98.5	99.7

Table 6.1: Classification of monotonicity direction of two OPs based on six methods with 1,000 simulated data sets, different sample sizes and independent covariates (%).

Table 6.1 shows the resulting proportions of monotonicity directions and unconstrained cases, for each sample size and OP within each one of the six constrained estimation methods discussed here. It is reasonable to consider the analysis of whether the differences between those proportions can be explained by random variation. This implied to conduct 300 hypothesis tests in total (15 comparisons for a given monotonicity direction, ordinal predictor and sample size). The analysis of the whole set of tests will be addressed as an exploratory exercise and comments about it will be focused on some selected subsets of comparisons. The type of hypothesis test, the number of tests and the way they will be interpreted are discussed next.

Given that the methods are fitted on the same data sets, the two-sided McNemar test was used, which is appropriate for differences between two dependent proportions (see Agresti (2007) Section 8.1.1), where the null hypothesis ($H_0 : \pi_{m_i} = \pi_{m_j}$) is that the probability of classifying the pattern of parameters in one of the monotonicity directions (isotonic or antitonic), for a given ordinal predictor and sample size using one of the six constrained methods (denoted with the subindex m_i), is the same as the probability of classifying the pattern of parameters in the same monotonicity direction for the same setting (ordinal predictor and sample size) using one of the remaining constrained methods (denoted with the subindex m_j).

A high number of tests were ran on paired comparisons of proportions. For each combination of ordinal predictor, sample size, and monotonicity direction (isotonic and antitonic), there are six constrained methods to be compared, leading to 15 paired differences of proportions for each setting. Table 6.2 is an example of the p-values that were obtained for a single setting (OP 1, n =500, and 'isotonic').

As shown in Table 6.2, the p-value when testing $H_0 : \pi_{MDC\ \ S3} = \pi_{CMLE\ \ Bonferroni}$ is 1 because 'CMLE Bonferroni' did not drop the monotonicity constraint for OP 1 in any of the data sets, and therefore there is no difference between 'MDC S3' and 'CMLE Bonferroni'. Those p-values 1 of other comparisons in Table 6.2 are based on the fact that the corresponding methods differ in their monotonicity classification in only one case among the 1,000 simulations.

m_i \ m_j	MDC S2	MDC S3	CMLE Bonferroni	CMLE Conf. Reg.	CMLE filtered
MDC S1	0.0133	0.0005	0.0005	1.0000	0.0009
MDC S2	-	0.0412	0.0412	0.1456	0.1306
MDC S3	-	-	1.0000	0.0009	1.0000
CMLE Bonferroni	-	-	-	0.0009	1.0000
CMLE Conf. Reg.	-	-	-	-	0.0015

Table 6.2: Example of the p-values of the McNemar tests with null hypothesis $H_0 : \pi_{m_i} = \pi_{m_j}$ for OP 1, sample size $n = 500$ and 'isotonic'.

As there are two ordinal predictors, five sample sizes, and two monotonicity directions, the total number of tests is 300. Strict interpretations of the p-values would need to take into consideration the whole set of tests using some multiple testing procedure, such as Bonferroni correction. Therefore, the analysis of the whole set of 300 two-sided paired proportions tests will be addressed as an exploratory exercise, focusing the attention on some selected subsets of comparisons, and using either direct reference to the p-value or some of the two following significance levels in order to call the results of the tests *significant*: (i) the 'Bonferroni level' of $\alpha = 0.05/300 = 1.67\text{E-}04$, or (ii) the '0.01 level' $\alpha = 0.01$.

This method of comparison between two dependent proportions will also be used for the analysis of other tables about proportions in the current section.

The MDC procedure was performed as part of the constrained approaches. Its first, second, and third steps ('MDC S1', 'MDC S2' and 'MDC S3' in Table 6.1) correctly classified OPs 1 and 2 in nearly 100% of the cases when the sample size was at least 500. For smaller sample sizes, 'CMLE MDC S2' showed significantly better results than 'CMLE MDC S1' as expected, and the third step allowed to finally classify OP 1 as 'isotonic' in 69.2% of the cases when $n = 50$, which rapidly increased to 92.9% when $n = 100$ and improved even more for larger sample sizes. Regarding OP 2, better results were obtained even with small sample sizes.

'CMLE Bonferroni' performed in exactly the same way as 'CMLE MDC S3' because the null hypothesis of monotonicity was not rejected in 100% of the data sets

for both OPs with $\alpha_s^* = 0.05$ and for any sample size. Therefore, the monotonicity constraints were not dropped.

Regarding the 'CMLE Conf. Reg.' approach, the null hypothesis of monotonicity was rejected with a confidence level of 95% ($\alpha = 0.05$) in at least 4.9% of the cases for OP 1 and 10.2% for OP 2 when $n \leq 100$, indicating that this approach is much less conservative than 'CMLE Bonferroni' for small sample sizes. For these sample sizes and except for one setting (OP 1, n=50, and 'isotonic'), all the pairwise comparisons of proportions between 'CMLE Conf. Reg.' and 'CMLE Bonferroni' for 'isotonic' or 'antitonic' show significant differences, being 'CMLE Bonferroni' more accurate. Within those cases where the monotonicity constraints where imposed, the monotonicity direction classification was more accurate compared to the one of other constrained methods. For sample sizes larger than 100, the null hypothesis of monotonicity was rejected in at most 2.2% of the cases, making the results of 'CMLE Conf. Reg.' similar to those of 'CMLE Bonferroni' or 'CMLE filtered' in terms of misclassification. However, when comparing 'CMLE Conf. Reg.' against 'CMLE Bonferroni' or 'CMLE filtered' (excluding $n = 5000$ where no significant difference exists), the proportions of 'isotonic' for OP 1 are significantly smaller for 'CMLE Conf. Reg.' (although with a minimum accurate rate of 98.5%), whereas for OP 2 there is no significant difference between the proportions of 'CMLE Conf. Reg.' and 'CMLE filtered'.

The results of 'CMLE filtered' are similar to the ones of both 'CMLE MDC S3' and 'CMLE Bonferroni' for OP 1. In fact, none of the pairwise comparisons of proportions shows significant differences. This is because the monotonicity constraints were dropped in at most 0.3% of the cases, which hardly affected the final monotonicity direction classification of this ordinal predictor compared to 'CMLE MDC S3' or 'CMLE Bonferroni'. However, for OP 2, 'antitonic', and n between 100 and 1000 the differences between 'CMLE filtered' and either 'CMLE MDC S3' or 'CMLE Bonferroni' are statistically significant, which is produced by the rejection of monotonicity in at least 1.5% of the cases for these sample sizes.

In general, smaller sample sizes provide less information to any method, increasing the misclassification rate of the monotonicity direction, or, for 'CMLE

Conf. Reg.', indicating non-monotonic patterns. However, given a monotonic association, when the value of the parameter estimate associated with the last category is further away from zero, there is less probability of misclassification irrespective of the sample size. This is the case for OP 2 (see Figure 6.1 as an example when $n = 500$), for which its last category is further away from zero than the corresponding one of OP 1. In fact, it was correctly classified in more than 90% of the cases by all of the methods but 'CMLE Conf. Reg.', even when the sample size was as small as 50.

Consider one of the 1,000 data sets as an example to illustrate the case of imposing monotonicity constraints.

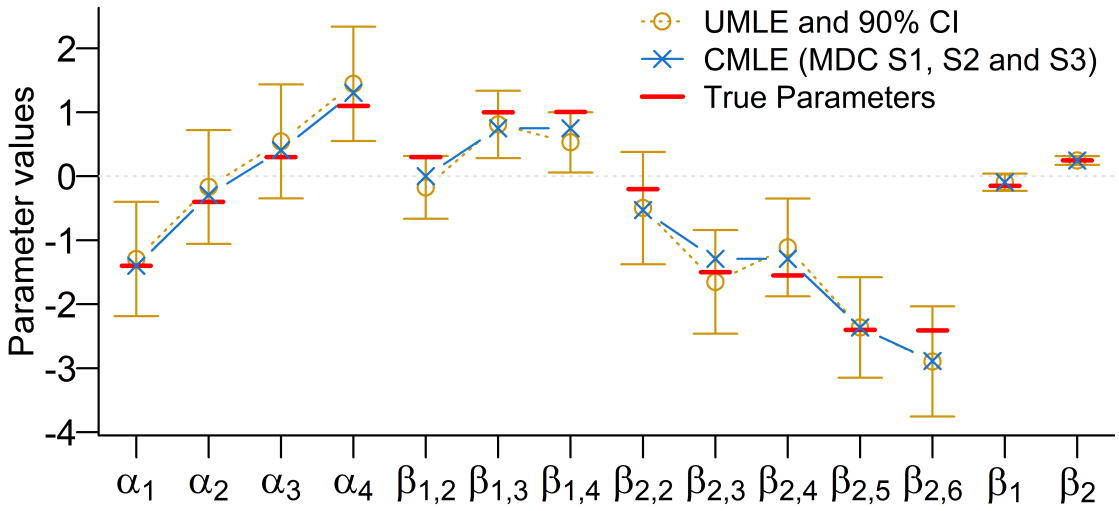

Figure 6.1: An example of unconstrained MLE and constrained MLE for a particular data set from simulations with 2 independent OPs and $n = 500$.

As shown in Figure 6.1, some unconstrained parameter estimates are incompatible with the monotonicity assumptions. Despite the fact that the OP 1 is assumed to be isotonic, the UMLE yields $\hat{\beta}_{1,2} < 0$ and $\hat{\beta}_{1,3} > \hat{\beta}_{1,4}$. Similar violations occur with the second ordinal predictor (antitonic), with $\hat{\beta}_{2,3} < \hat{\beta}_{2,4}$. By contrast, the results of the CMLEs imposed monotonicity constraints, with the estimate for $\beta_{1,2}$ being greater than zero, the estimate for $\beta_{1,4}$ being slightly greater than the one for $\beta_{1,3}$, and where the estimate for $\beta_{2,4}$ was slightly lesser than the one for $\beta_{2,3}$. The monotonicity directions were established in the first step of the MDC procedure,

therefore the methods 'CMLE MDC S1', 'CMLE MDC S2' and 'CMLE MDC S3' provided the same result. Similarly, the first step of the MDC procedure did not classify OPs 1 or 2 as 'none', and both monotonicity tests (based on Bonferroni and confidence regions) did not reject the null hypothesis of monotonicity for any of these two OPs, therefore 'CMLE Bonferroni', 'CMLE Conf. Reg.' and 'CMLE filtered' are not shown.

In this particular example, the CMLEs for the parameter estimates associated with both intercepts and interval-scaled covariates were hardly affected by the monotonicity assumption when comparing the CMLE to the UMLE.

Regardless of the sample size, imposing monotonicity constraints reduces the parameter space, which affects the distribution of the parameter estimates when they are active. To visualise this, Figure 6.2 uses boxplots to show the distribution of each parameter estimate resulting from several methods together with the true parameters used in the data generation process for the 1,000 simulation iterations with $n = 100$.

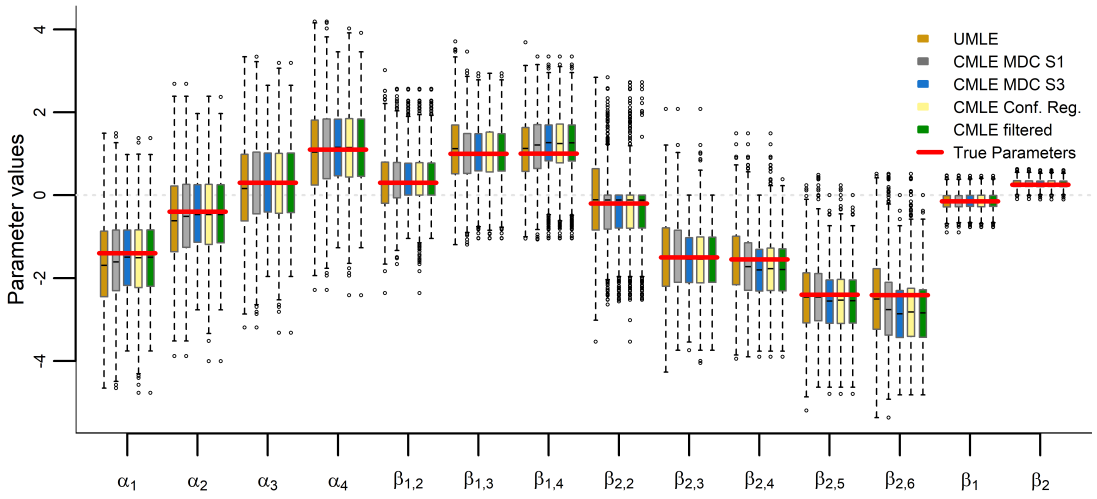

Figure 6.2: Unconstrained MLE, different methods with constrained MLE and true parameters used for 1,000 simulated data sets with 2 independent OPs, example for $n = 100$.

The effect of the monotonicity constraints is depicted by the range of values that the parameter estimates take for an OP in some of the constrained approaches,

which differs from the one of the UMLEs in two aspects. First, when the parameter estimates are correctly constrained, they are compatible with their monotonicity direction, i.e., they take positive values for the isotonic case and negative for the antitonic one. This is why the boxes of some constrained approaches seem to be truncated at zero for $\beta_{1,2}$ and $\beta_{2,2}$. The second difference is a generalisation of the first one as any constrained parameter estimate is greater/lower than the one of the preceding category rather than greater/lower than zero only. Hence, the lower extremes of their boxplots show shorter whiskers than the ones of the UMLE when there is an isotonic relationship, and the same effect occurs for the upper whiskers when the relationship is antitonic.

The results of 'CMLE MDC S1' are the closest to the ones of the unconstrained method. This is due to the fact that 'CMLE MDC S1' drops the monotonicity constraints more frequently than any other constrained method. Conversely, 'CMLE MDC S3' is the furthest because it does not drop constraints. Other constrained methods are in between these two. The approaches 'CMLE MDC S3' and 'CMLE Bonferroni' delivered the same results because the monotonicity tests did not reject monotonicity for any OP. Compared to 'CMLE MDC S3' and 'CMLE Bonferroni', the results of 'CMLE filtered' are slightly different because there are only 18 cases where the OP 2 was considered as non-monotonic and only 3 for OP 1, for which the monotonicity constraints were not imposed. The results of 'CMLE Conf. Reg.' are different to the ones of 'CMLE MDC S3', 'CMLE Bonferroni' and 'CMLE filtered' because there are 49 cases where OP 1 was classified as non-monotonic and 102 for OP 2, which makes extreme values of the parameter estimates more frequent than in other constrained methods. In general, unconstrained cases together with misclassification of the monotonicity direction are the reason why there are some negative values for the estimates of OP 1 and positive values for the ones of OP 2 in the constrained approaches.

Based on the results of the 1,000 simulation iterations, the mean-squared error (MSE) was computed for each parameter. The MSE decomposition allows to compute the variance of the estimates, and therefore their standard error (SE). The MSE and SE of the parameter estimates are shown in Table 6.4, averaged

over all parameters belonging to an OP. The values for the constrained methods are given relative to the values for UMLE.

Measures of performance are estimations, and therefore they are subject to error as a result of variability (see Morris et al. (2019)). Hence, comments involving pairwise comparisons of MSEs are based on a two-sided paired t-test for $H_0 : E(MSE_{m_i}) = E(MSE_{m_j})$ with m_i and m_j being a subindex identifying the model to be considered given a certain setting based on a combination of OP, ordinal category, and sample size. Similarly to the analysis of the comparisons of proportions, a high number of comparisons of mean-squared errors was required in order to assess whether the MSE of one method is statistically different to the MSE of another method. However, in this case the number of comparisons is much higher because one extra model is being considered (UMLE), making a total of 7 models, and the fact that there is one MSE for each parameter associated with the categories of the ordinal predictors. Therefore, for each OP, ordinal category, and sample size, there are 21 pairs of MSEs to be compared ($7 \times 6/2$). Now, given that there are three parameters to be estimated for OP 1, five for OP 2, and five sample sizes, the number of tests for OP 1 is 315 and 525 for OP 2.

As an exploratory exercise, in order to conclude about a specific comparison of models for a given OP and sample size, the Bonferroni correction was used as a multiple testing approach within each one of these settings, where the number of tests is equal to the number of parameters belonging to the corresponding OP. This will allow comments for an OP at a certain level for a given sample size. Therefore, the Bonferroni correction set a significance level of $\alpha = 0.01/3 = 3.33\text{E-}03$ for the comparisons of MSEs of parameters belonging to OP 1, and $\alpha = 0.01/5 = 0.002$ for those of OP 2. These will be referred to as the 'Bonferroni level' for OP 1 or 2. Thus, the exploratory exercise summarises the results for each OP and sample size by counting the number of tests where the null hypothesis was rejected at the corresponding 'Bonferroni level', resulting values between 0 and the number of parameter estimates associated with the categories of the corresponding OP (3 for OP 1 and 5 for OP 2) for each comparison of models. Table 6.3 shows an illustration of this for OP 1 and $n = 100$.

m_i \ m_j	MDC S1	MDC S2	MDC S3	CMLE Bonferroni	CMLE Conf. Reg.	CMLE filtered
UMLE	3	3	2	2	2	2
MDC S1	-	1	3	3	2	3
MDC S2	-	-	3	3	1	3
MDC S3	-	-	-	0	1	0
CMLE Bonferroni	-	-	-	-	1	0
CMLE Conf. Reg.	-	-	-	-	-	1

Table 6.3: Number of tests where the null hypothesis $H_0 : E(MSE_{m_i}) = E(MSE_{m_j})$ was rejected at the 'Bonferroni level' of $\alpha = 0.01/3$ for OP 1 and sample size $n = 100$. The total number of tests for each cell is three.

For example, Table 6.3 shows that the null hypothesis $H_0 : E(MSE_{MDC\ \ S2}) = E(MSE_{CMLE\ \ filtered})$ was rejected for all of the parameters associated with OP 1 when $n = 100$ at a significance level of $\alpha = 0.01/3$. This kind of results will be interpreted as that the MSE of 'CMLE filtered' is significantly different than the one of 'MDC S2' for OP 1. These kind of results are complementary to the ones of Table 6.4, which also indicates the direction in which the MSEs of these models are different. In addition, the number of tests for which the null hypothesis was rejected will usually be mentioned.

Given that there are two ordinal predictors and five sample sizes, the total number of tables like Table 6.3 is 10. Therefore, comments will be based on a selected number of comparisons only. This method of comparison between two dependent averaged MSEs will also be used for the analysis of other tables about MSEs in the current Section 6.2.

		True pattern	OP 1: Isotonic					OP 2: Antitonic				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000
UMLE		MSE_{UMLE}	1.75	0.73	0.1	0.05	0.01	64.86	1.26	0.25	0.12	0.02
		SE_{UMLE}	0.04	0.03	0.01	0.01	0.00	0.22	0.04	0.02	0.01	0.00
CMLE	MDC S1	MSE/MSE_{UMLE}	0.95	0.88	0.84	0.86	0.95	1.00	0.88	0.80	0.79	0.94
		SE/SE_{UMLE}	0.97	0.94	0.91	0.92	0.97	1.00	0.93	0.86	0.87	0.97
	MDC S2	MSE/MSE_{UMLE}	0.94	0.86	0.83	0.86	0.95	1.00	0.86	0.75	0.74	0.93
		SE/SE_{UMLE}	0.97	0.93	0.9	0.92	0.97	1.00	0.91	0.83	0.84	0.96
	MDC S3	MSE/MSE_{UMLE}	0.95	0.75	0.82	0.86	0.95	0.99	0.74	0.75	0.74	0.93
		SE/SE_{UMLE}	0.97	0.85	0.9	0.92	0.97	0.99	0.83	0.82	0.84	0.96
CMLE Bonferroni		MSE/MSE_{UMLE}	0.95	0.75	0.82	0.86	0.95	0.99	0.74	0.75	0.74	0.93
		SE/SE_{UMLE}	0.97	0.85	0.9	0.92	0.97	0.99	0.83	0.82	0.84	0.96
CMLE Conf. Reg.		MSE/MSE_{UMLE}	0.96	0.81	0.84	0.87	0.95	1.00	0.83	0.81	0.79	0.94
		SE/SE_{UMLE}	0.98	0.89	0.91	0.93	0.97	0.99	0.90	0.87	0.87	0.97
CMLE filtered		MSE/MSE_{UMLE}	0.95	0.75	0.82	0.86	0.95	1.00	0.74	0.80	0.79	0.94
		SE/SE_{UMLE}	0.97	0.85	0.9	0.92	0.97	0.99	0.84	0.86	0.87	0.97

Table 6.4: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Independent covariates.

In terms of MSE, there is a significant difference at the corresponding Bonferroni level between the unconstrained method and any constrained method in at least one of the parameters of the categories of both OPs irrespective of the sample size, where the MSE of any constrained method is lower than the one of the UMLE. For any given constrained method, the MSE ratio with respect to the one of UMLE is higher for both the smallest and largest sample sizes than for the intermediate ones. There is a reason for this to happen at each one of these extreme cases. For the largest sample size and given truly monotonic ordinal predictors as in this simulation, the constrained methods provide results close to UMLE in terms of MSE because for large enough n the UMLE reveals the true monotonic patterns, and therefore the results of the constrained and unconstrained methods get closer to each other. For the smallest sample size, the MSE results of the constrained methods are fairly close to the ones of UMLEs because the variability of their parameter estimates is affected by a considerable misclassification rate when imposing monotonicity constraints.

When comparing the MSE between pairs of constrained methods, the methods 'CMLE MDC S3', 'CMLE Bonferroni', and 'CMLE filtered' do not show a significant difference at the Bonferroni level for any of the parameters associated with the categories of OP 1 and any sample size. However, regarding OP 2, the same occurs for 'CMLE MDC S3' and 'CMLE Bonferroni' only.

For OP 1 and $n \geq 1000$, comparisons among constrained methods show not significant difference at the Bonferroni level between the MSEs of their corresponding parameters. For OP 2 and $n = 5000$ the same results were obtained. This means that for large enough n, when n increases the choice of the constrained method is less relevant in terms of MSE.

For OP 2 and $n = 50$, despite the fact that the MSEs seem to be close to each other, the one of UMLE is significantly different to the one of any other constrained method as mentioned before, with at least 3 parameters with MSE significantly smaller at the Bonferroni level of OP 2 ($\alpha = 0.01/5$). Among constrained methods, there are three that are the best in terms of MSE compared to the one of UMLE, 'CMLE MDC S3', 'CMLE Bonferroni', and 'CMLE filtered'. These do not show

a significant difference between each other at the Bonferroni level. All the other constrained methods are better than UMLE but worse than 'CMLE MDC S3', 'CMLE Bonferroni', and 'CMLE filtered', and when making comparisons among them they show at least two parameters with significant different MSE.

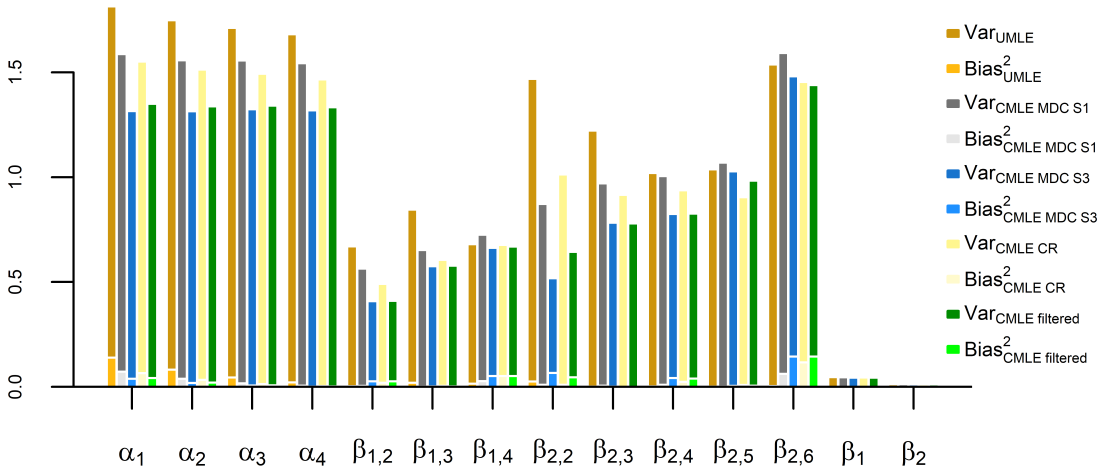

Figure 6.3: Mean-squared error for unconstrained and constrained MLEs and its decomposition, example for $n = 100$.

As an example of the analysis of the MSE, consider the results for $n = 100$ shown in Figure 6.3. The total MSE is notably smaller for the constrained approaches (depicted by the height of the bars). On average, the 'CMLE MDC S1' (grey bars) shows a 10.2% smaller MSE compared to the MSE of UMLE for the intercepts, 10.7% smaller for the first ordinal predictor, and 11.2% smaller for the second. The corresponding figures for 'CMLE MDS S3' (blue bars) are 24.2%, 24.6% and 24.9%. The figures for 'CMLE filtered' are 22.9%, 24.1% and 24.6%, being similar to the ones for 'CMLE MDS S3' because there are few cases of unconstrained parameter estimates for the categories of the ordinal predictors, whereas these percentages are 13.4%, 18.5% and 16.5% for 'CMLE Conf. Reg.', which are between the ones of 'CMLE MDS S1' and 'CMLE MDS S3', but closer to the former because there are more cases where the monotonicity constraints are dropped.

The performance of 'CMLE Bonferroni' is almost identical to 'CMLE MDC

S3'. The results of 'CMLE MDC S2' lie between those of 'CMLE MDC S1' and 'CMLE MDC S3'. These are not shown in Figure 6.3 nor later.

Despite the fact that the squared bias makes a markedly small contribution to the total MSE (lighter colours in Figure 6.3), it is clearly higher for some constrained parameter estimates, specially for those of OP 2. Its sixth category produced the highest squared bias, which represents from 3.9% of its total MSE for 'CMLE MDC S1' up to 10.0% for 'CMLE filtered'. The squared bias of the constrained approaches associated with the remaining categories of OP 2 together with the first OP and the intercepts represent, on average, between 1.4% and 3.4% of the MSE depending on the constrained method ('CMLE MDC S1' being the smallest and both 'CMLE MDC S3' and 'CMLE Bonferroni' the largest). Consequently, the MSEs are dominated by variances, which are considerably lower than the ones of the UMLE not only for the parameters associated with the ordinal predictor categories, but also for the intercepts.

The simulation was repeated with dependence among covariates. In order to simulate the predictors, a set of four variables was generated from a multivariate normal distribution with means equal to zero and unit variances for the two ordinal variables and the same means and variances that were used in the setting with independent covariates for the two interval-scaled variables. The correlation structure was set allowing different magnitudes and directions as follows:

$$\rho = \begin{bmatrix} 1 & -0.3 & 0.6 & 0.7 \\ -0.3 & 1 & -0.5 & -0.2 \\ 0.6 & -0.5 & 1 & 0.2 \\ 0.7 & -0.2 & 0.2 & 1 \end{bmatrix}.$$

The categorisation of the ordinal variables resulted from classifying each simulated value within the limits defined by the normal quantiles corresponding to the cumulative probabilities obtained from the marginal distributions that were previously set for those OPs with 4 and 6 categories (see Figure 2.2).

The monotonicity direction classification results obtained from the setting with correlated predictors are shown in Table 6.5. For sample sizes $n = 50$ and $n = 100$, there is more misclassification for OP 1 in the scenario with correlated covariates. For larger sample sizes ($n \geq 500$), the results of the setting with correlated covari-

ates are nearly as good as the ones with independent covariates for OP 1. The latter occurs for OP 2 also, but this time regardless of the sample size, including the smallest.

The proportion of cases where OP 1 was classified as 'Isotonic' is significantly smaller at the Bonferroni level ($\alpha = 0.05/300$) when comparing 'CMLE MDC S1' or 'CMLE MDC S2' versus any other method for $n \leq 500$, while for n=1000 these two methods are not significantly different from 'CMLE Conf. Reg.', whereas the latter is the only method whose proportion of 99.3% of cases classified as 'Isotonic' is significantly smaller to any other method for $n = 5000$.

Regarding the classification of OP 2 as 'Antitonic', all the proportions are not significantly different when $n = 5000$. For $n = 500$ or 1000, the proportions of 'Antitonic' for methods 'MDC S1', 'CMLE Conf. Reg.' and 'CMLE filtered', with a minimum of 97.8%, are significantly smaller to the one of any other method but not between each other at the Bonferroni level when $n = 500$ and mixed levels (Bonferroni level or 0.01) when $n = 1000$. For smaller sample sizes, $n \leq 100$, the proportions of 'Antitonic' for methods 'MDC S1', 'MDC S2' and 'CMLE Conf. Reg.' are significantly smaller to any other method, including comparisons between each other, at the Bonferroni level.

Table 6.6 shows the MSE results with correlated predictors. Compared to the scenario with independent covariates, the MSE of the version with correlated covariates is always higher, regardless of the sample size and method. In general, the MSEs decrease as n increases and the magnitude of the reduction depends on the method and the sample size. For example, for 'CMLE MDC S3' and other highly constrained methods, with correlated predictors the ratio $\text{MSE}/\text{MSE}_{\text{UMLE}}$ increases for OP 1 when n changes from 50 to 100. Despite the fact that OP 1 is often misclassified by the more restrictive methods such as 'CMLE MDC S3', their MSE ratio is still low when $n = 50$ because of the high variance of the UMLE, which is amended by the constrained methods.

All of the constrained methods show an MSE statistically smaller than the one of UMLE, for both OPs and every sample size. Among the constrained methods and for the parameters of OP 1, the MSE of 'CMLE MDC S3' or 'CMLE Bon-

ferroni' or 'CMLE Conf. Reg.' or 'CMLE filtered' is significantly smaller than the one of either 'CMLE MDC S1' or 'CMLE MDC S2' in at least one of their parameters when $n \leq 500$ only. This is still true for OP 2 but for $n \leq 100$.

		True pattern	OP 1: Isotonic					OP 2: Antitonic				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000
CMLE	MDC S1	Isotonic	25.0	35.6	87.4	98.7	100	2.9	2.3	0.0	0.0	0.0
		Antitonic	5.9	2.9	0.3	0.0	0.0	27.8	61.0	97.8	98.5	100
		Both	69.0	61.5	12.2	1.3	0.0	68.9	35.5	0.0	0.0	0.0
		None	0.1	0.0	0.1	0.0	0.0	0.4	1.2	2.2	1.5	0.0
	MDC S2	Isotonic	33.1	43.9	92.3	99.2	100	4.5	4.3	0.0	0.0	0.0
		Antitonic	7.4	4.4	0.8	0.2	0.0	35.0	67.0	100	100	100
		Both	59.5	51.7	6.9	0.6	0.0	60.5	28.7	0.0	0.0	0.0
		None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MDC S3	Isotonic	58.5	73.7	98.9	99.8	100	9.8	5.1	0.0	0.0	0.0
		Antitonic	41.5	26.3	1.1	0.2	0.0	90.2	94.9	100	100	100
CMLE Bonferroni	Unconstrained		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0
	MDC S3	Isotonic	58.5	73.7	98.9	99.8	100	9.8	5.1	0.0	0.0	0.0
		Antitonic	41.5	26.3	1.1	0.2	0.0	90.2	94.9	99.9	99.9	100
CMLE Conf. Reg.	Unconstrained		5.2	4.0	3.0	1.6	0.7	11.5	6.1	3.0	1.9	0.3
	MDC S3	Isotonic	57.9	72.3	96.9	98.4	99.3	5.5	1.6	0.0	0.0	0.0
		Antitonic	36.9	23.7	0.1	0.0	0.0	83.0	92.3	97.0	98.1	99.7
CMLE filtered	Unconstrained		0.1	0.0	0.1	0.0	0.0	0.4	1.2	2.2	1.5	0.0
	MDC S3	Isotonic	58.5	73.9	98.9	99.8	100	9.7	4.3	0.0	0.0	0.0
		Antitonic	41.4	26.1	1.0	0.2	0.0	89.9	94.5	97.8	98.5	100

Table 6.5: Classification of monotonicity direction of two OPs based on six methods with 1,000 simulated data sets, different sample sizes and correlated covariates (%).

		True pattern	OP 1: Isotonic					OP 2: Antitonic				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000
UMLE		MSE_{UMLE}	8.68	0.84	0.14	0.08	0.01	87.62	43.63	0.28	0.14	0.02
		SE_{UMLE}	0.09	0.03	0.01	0.01	0.00	0.25	0.19	0.02	0.01	0.00
CMLE	MDC S1	MSE/MSE_{UMLE}	0.99	0.98	0.94	0.94	0.97	0.98	1.00	0.92	0.92	0.93
		SE/SE_{UMLE}	1.00	0.98	0.97	0.96	0.98	0.98	1.00	0.95	0.95	0.96
	MDC S2	MSE/MSE_{UMLE}	0.99	0.98	0.94	0.95	0.97	0.97	1.00	0.89	0.90	0.93
		SE/SE_{UMLE}	1.00	0.98	0.96	0.97	0.98	0.97	1.00	0.93	0.93	0.96
	MDC S3	MSE/MSE_{UMLE}	0.70	1.01	0.92	0.94	0.97	0.95	1.00	0.89	0.90	0.93
		SE/SE_{UMLE}	0.84	0.98	0.95	0.96	0.98	0.95	1.00	0.93	0.93	0.96
CMLE Bonferroni		MSE/MSE_{UMLE}	0.70	1.01	0.92	0.94	0.97	0.95	1.00	0.89	0.90	0.93
		SE/SE_{UMLE}	0.84	0.98	0.95	0.96	0.98	0.95	1.00	0.93	0.94	0.96
CMLE Conf. Reg.		MSE/MSE_{UMLE}	0.73	1.01	0.92	0.94	0.97	0.97	1.00	0.92	0.92	0.93
		SE/SE_{UMLE}	0.86	0.99	0.95	0.96	0.98	0.97	0.99	0.95	0.95	0.96
CMLE filtered		MSE/MSE_{UMLE}	0.70	1.01	0.92	0.94	0.97	0.95	1.00	0.92	0.92	0.93
		SE/SE_{UMLE}	0.84	0.98	0.95	0.96	0.98	0.95	1.00	0.95	0.95	0.96

Table 6.6: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Correlated covariates.

In the simulation presented above, no non-monotonic ordinal predictor was included and its results showed that any constrained approach performed better than the unconstrained one in almost every simulated scenario. In order to analyse their performance in presence of non-monotonic OPs, consider another simulation setting for model (2.3.4). This time an ordinal response with four categories is used, i.e., $k = 4$ and $j = 1, 2, 3$; four ordinal predictors ($t = 4$) with $p_1 = 3$, $p_2 = 4$, $p_3 = 5$, and $p_4 = 6$ categories correspondingly; and one interval-scaled predictor ($v = 1$). Again, several sample sizes were considered: $n = 50, 100, 500, 1000, 5000$. The chosen parameters for the intercepts were $\alpha_1 = -1.4$, $\alpha_2 = -0.1$, and $\alpha_3 = 1.7$; for OP 1, $\boldsymbol{\beta}_1' = (0.5, 1)$; for OP 2, $\boldsymbol{\beta}_2' = (-0.65, -0.70, -1.60)$; for OP 3, $\boldsymbol{\beta}_3' = (0, 0, 0, 0)$; for OP 4, $\boldsymbol{\beta}_4' = (-0.8, -1.6, -0.6, 0.6, 1.6)$; and for the interval-scaled predictor $\beta_1 = 0.3$. The parameters of the OPs 1 to 4 were chosen to be isotonic, antitonic, zero, and non-monotonic correspondingly. For OP 3, all the parameters were set to zero, and therefore, optimally, the monotonicity tests should not reject monotonicity and the first and second step of the MDC procedure should classify it as 'both'.

This model was fitted for 1,000 simulated data sets and for every sample size. The ordinal predictors were drawn from the population distributions showed in Figure 2.2. The interval-scaled predictor was randomly generated from a normal distribution $N(1, 4)$.

The MDC procedure was executed with a 90% confidence level in the first step ($\tilde{c} = 0.90$) and tolerance levels $\tilde{c}_s'^* = 0.85$ and $\tilde{c}_s''^* = 0.999$ for $s = 1, 2, 3, 4$ in the second step.

Table 6.7 shows the results of the MDC for the constrained estimation methods. OPs 1 and 2 follow the same trends as in the earlier simulation. OPs 3 and 4 make the constrained methods differ markedly, mainly because smaller sample sizes do not only increase the probability of misclassification of the monotonicity direction, but also decrease the probability of dropping monotonicity constraints for an OP that is truly non-monotonic, which is the case for OP 4 in this simulation. This also affects the classification of OP 3 with true pattern 'both'.

For 'CMLE MDC S1', OP 3 shows a high percentage of 'both' classifications

for any sample size, and OP 4 was correctly classified when $n \geq 500$. However, the effects of OP 4 were constrained to be either 'isotonic' or 'antitonic' in a total of 50.1% of the data sets when $n = 50$, which is relatively high considering that 'CMLE MDC S1' is the less restrictive method. The monotonicity direction classification of 'CMLE MDC S2' is hardly affected when $n \geq 1000$, whereas for smaller sample sizes it is always between 'CMLE MDC S1' and 'CMLE MDC S3', being significantly different from them with a p-value of 0.0015 or smaller for $n \leq 100$ (better than 'CMLE MDC S1' and worse than 'CMLE MDC S3' for OP 1 and 2). The classification of OPs 3 and 4 by 'CMLE MDC S3' is more evenly distributed for small sample sizes, which is not unreasonable for an OP that is set to be 'both' (OP 3) and an OP of class 'none' (OP 4). However, for larger sample sizes ($n \geq 500$), the classification of OP 3 is more concentrated in 'antitonic', whereas OP 4 is highly concentrated in 'isotonic', which is due to the fact that an isotonic monotonicity direction dominates throughout the pattern of OP 4. 'CMLE Bonferroni' does not drop monotonicity constraints of OP 4 for small sample sizes. Therefore, its performance is almost identical to the one of 'CMLE MDC S3' when $n \leq 100$, except for OP 4, $n = 100$, and 'Antitonic' classification, where the proportion resulting from these methods is not significantly different (p-value 0.0015). For larger sample sizes, the monotonicity constraints are dropped much more frequently for OP 4, and the classification of OP 3 remains consistent with its definition of 'both'. Regarding the method 'CMLE Conf. Reg.' for large sample sizes ($n \geq 500$), it drops the monotonicity constraints for OP 3 in at most 2.2% and the classification of each monotonicity direction is around 50% of the cases, being consistent with its true pattern 'both'. However, for small sample sizes ($n \leq 100$), it drops the monotonicity constraints for OP 3 more frequently than any other method, reaching 16.3% of the cases when $n = 50$, unbalancing the distribution of monotonicity directions for those cases that are still constrained. For OP 4, 'CMLE Conf. Reg.' is the method that drops the monotonicity constraints more frequently than any other even for small sample sizes being consistent with the true pattern 'none' of OP 4 (54.2% for $n = 50$ and 68.6% for $n = 100$), resulting in the best method at identifying patterns that are

truly non-monotonic. The results of 'CMLE filtered' are similar to those of 'CMLE Bonferroni' for OPs 1, 2 and 3. For OP 4, it was constrained less frequently by 'CMLE filtered' than by 'CMLE Bonferroni', regardless of the sample size, but more constrained than by 'CMLE Conf. Reg.'.

Based on the results of the average MSE (see Table 6.8) and given that there is a non-monotonic ordinal predictor, 'CMLE MDC S3' is the only method that is occasionally notably worse than UMLE, because it always imposes constraints on an OP that is non-monotonic; but for $n = 50$ the MSE of the UMLE is still so high that the one of 'CMLE MDC S3' is better, whose MSE is statistically smaller than the one of UMLE for four parameters of OP 4 at a significance level of 0.01 (and for OP 1 to 3 at least one parameter is significantly smaller too). The performance of the remaining constrained methods depends on the degree of conservativeness when establishing the set of OPs with non-monotonic effects. The less conservative the method, the closer is its MSE to the one of UMLE. The best options are 'CMLE Bonferroni' and 'CMLE filtered' because they drop constraints for OP 4 and not for other OPs, specially when $n \geq 500$, although they are still good options for smaller sample sizes. 'CMLE Conf. Reg.' is also a good option when $n \geq 500$. However, for smaller sample sizes it drops the monotonicity constraints for OP 3 more frequently than in other more conservative methods such as 'CMLE Bonferroni' and, in this case, 'CMLE filtered', making the MSEs of the parameter estimates associated with its ordinal categories higher. On the other hand, the relatively high MSE for OP 3 resulting from 'CMLE Conf. Reg.' is compensated by a smaller MSE for OP 4, which is a consequence of being the method that drops monotonicity constraints for OP 4 more frequently.

The simulation of the current model using four OPs was done again with dependence among covariates. The OPs and the interval-scaled predictor were generated from a multivariate normal distribution with the same means and variances as the

ones used in the previous simulation scenario. The correlation structure is now:

$$\rho = \begin{bmatrix} 1 & -0.5 & -0.1 & 0.3 & 0.6 \\ -0.5 & 1 & 0 & -0.4 & -0.6 \\ -0.1 & 0 & 1 & 0.2 & 0.1 \\ 0.3 & -0.4 & 0.2 & 1 & 0.7 \\ 0.6 & -0.6 & 0.1 & 0.7 & 1 \end{bmatrix}.$$

The ordinal categories of the OPs were obtained through categorisation as previously described but using the marginals of OPs according to those shown in Figure 2.2.

As an example to visualise the behaviour of the parameter estimates resulting from some selected methods under the simulation scenario with correlated covariates, Figure 6.4 shows their boxplots when $n = 500$. In general, the constrained methods perform in almost the same way as the unconstrained one for OP 1 and better for OP 2 and 3. As expected, the non-monotonic OP 4 produces more differences for 'CMLE MDC S3' than for other constrained methods, which are much closer to the unconstrained results for a non-monotonic OP.

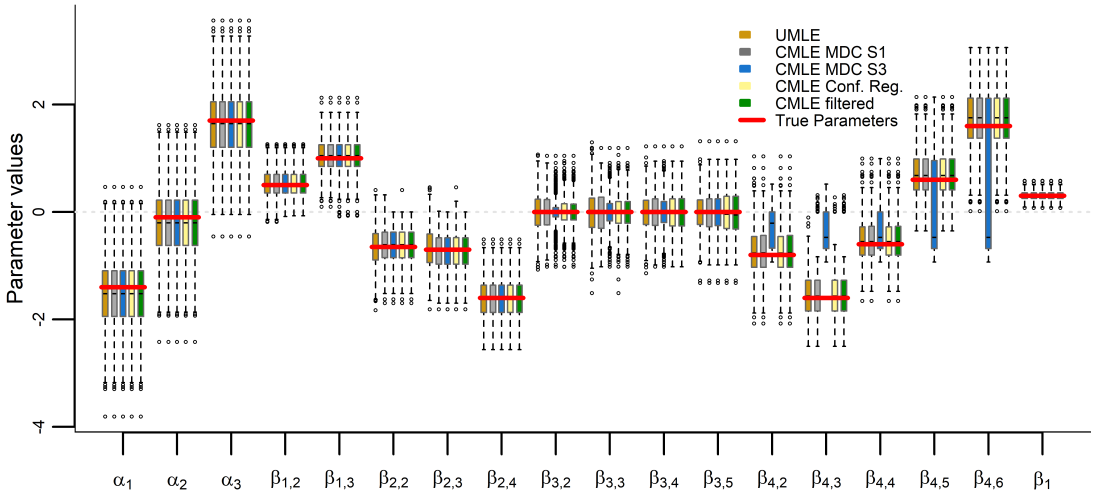

Figure 6.4: Unconstrained MLE, different methods with constrained MLE and true parameters used for 1,000 simulated data sets with 4 correlated OPs, example for $n = 500$.

Table 6.9 shows the MDC results with correlated predictors. Compared to the independent covariates scenario, the general trends remain the same. The results

of the largest sample size are hardly affected, whereas the others are somewhat worse. Regarding the MSE, the correlation among covariates increased the MSE in all the methods, specially when $n = 50$ and for OP 4 with $n \leq 100$. However, the constrained results are better or almost equal to those of the UMLE, except for 'CMLE MDC S3' when $n \geq 500$.

		True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
CMLE	MDC S1	Isotonic	44.5	63.4	99.1	100	100	3.8	1.1	0.0	0.0	0.0	17.3	15.0	13.5	14.1	11.8	26.3	33.7	2.7	0.0	0.0
		Antitonic	3.5	0.8	0.0	0.0	0.0	52.6	74.7	100	100	100	19.3	16.7	11.3	13.7	11.2	23.8	21.4	0.0	0.0	0.0
		Both	51.9	35.8	0.9	0.0	0.0	43.0	23.9	0.0	0.0	0.0	62.1	67.7	75.2	72.2	77.0	36.6	13.8	0.0	0.0	0.0
		None	0.1	0.0	0.0	0.0	0.0	0.6	0.3	0.0	0.0	0.0	1.3	0.6	0.0	0.0	0.0	13.3	31.1	97.3	100	100
	MDC S2	Isotonic	51.1	70.0	99.3	100	100	5.8	2.3	0.0	0.0	0.0	22.7	21.6	18.8	19.4	17.3	34.9	43.7	31.6	2.8	0.0
		Antitonic	4.9	1.3	0.0	0.0	0.0	60.1	80.0	100	100	100	25.2	22.8	16.7	20.3	16.3	39.3	48.4	15.4	0.0	0.0
		Both	44.0	28.7	0.7	0.0	0.0	34.1	17.7	0.0	0.0	0.0	51.9	55.6	64.5	60.3	66.4	25.6	7.7	0.0	0.0	0.0
		None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.2	53.0	97.2	100
	MDC S3	Isotonic	68.8	83.4	99.7	100	100	14.0	6.6	0.0	0.0	0.0	41.4	44.5	29.9	21.9	17.7	48.0	49.0	64.9	81.0	98.6
		Antitonic	31.2	16.6	0.3	0.0	0.0	86.0	93.4	100	100	100	58.6	55.5	70.1	78.1	82.3	52.0	51.0	35.1	19.0	1.4
CMLE Bonferroni	Unconstrained		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.4	1.3	85.2	99.7	100
	MDC S3	Isotonic	68.8	83.3	99.5	100	100	14.0	6.6	0.0	0.0	0.0	41.5	44.4	45.6	50.0	48.6	47.9	48.9	13.8	0.3	0.0
		Antitonic	31.2	16.7	0.5	0.0	0.0	86.0	93.4	100	100	100	58.4	55.6	54.4	50.0	51.4	51.7	49.8	1.0	0.0	0.0
CMLE Conf. Reg.	Unconstrained		3.5	1.1	0.0	0.0	0.0	9.3	4.6	0.5	0.1	0.0	16.3	6.1	2.2	1.5	1.7	54.2	68.6	98.4	100	100
	MDC S3	Isotonic	68.5	83.9	99.6	100	100	8.7	3.0	0.0	0.0	0.0	35.2	42.2	47.7	49.0	47.9	35.4	29.9	0.0	0.0	0.0
		Antitonic	28.0	15.0	0.4	0.0	0.0	82.0	92.4	99.5	99.9	100	48.5	51.7	50.1	49.5	50.4	10.4	1.5	1.6	0.0	0.0
CMLE filtered	Unconstrained		0.1	0.0	0.0	0.0	0.0	0.6	0.3	0.0	0.0	0.0	1.3	0.6	0.0	0.0	0.0	13.3	31.1	97.3	100	100
	MDC S3	Isotonic	68.2	83.5	99.6	100	100	13.1	5.2	0.0	0.0	0.0	40.6	42.5	48.7	50.0	48.6	45.5	41.7	2.7	0.0	0.0
		Antitonic	31.7	16.5	0.4	0.0	0.0	86.3	94.5	100	100	100	58.1	56.9	51.3	50.0	51.4	41.2	27.2	0.0	0.0	0.0

Table 6.7: Classification of monotonicity direction of four OPs based on six methods with 1,000 simulated data sets and independent covariates (%).

		True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
UMLE		MSE_{UMLE}	2.74	0.49	0.06	0.03	0.01	6.39	0.87	0.11	0.05	0.01	9.24	0.93	0.12	0.06	0.01	24.07	1.34	0.21	0.09	0.02
		SE_{UMLE}	0.05	0.02	0.01	0.01	0.00	0.08	0.03	0.01	0.01	0.00	0.09	0.03	0.01	0.01	0.00	0.15	0.04	0.01	0.01	0.00
CMLE	MDC S1	MSE/MSE_{UMLE}	0.98	0.98	0.99	1.00	1.00	0.98	0.95	0.93	0.96	0.96	0.92	0.92	0.93	0.93	0.95	0.92	1.29	1.05	1.00	1.00
		SE/SE_{UMLE}	0.99	0.99	1.00	1.00	1.00	0.99	0.97	0.96	0.97	0.98	0.97	0.96	0.96	0.97	0.97	0.97	1.11	1.02	1.00	1.00
	MDC S2	MSE/MSE_{UMLE}	0.98	0.98	0.99	1.00	1.00	0.97	0.95	0.93	0.96	0.96	0.90	0.89	0.91	0.91	0.92	0.90	1.61	2.94	1.15	1.00
		SE/SE_{UMLE}	0.98	0.98	1.00	1.00	1.00	0.98	0.97	0.96	0.97	0.98	0.96	0.95	0.95	0.95	0.96	0.96	1.07	1.56	1.07	1.00
	MDC S3	MSE/MSE_{UMLE}	0.99	1.05	1.00	1.00	1.00	0.91	0.91	0.93	0.96	0.96	0.69	0.80	0.74	0.74	0.74	0.85	1.65	5.45	9.83	39.49
		SE/SE_{UMLE}	1.00	1.03	1.00	1.00	1.00	0.94	0.94	0.96	0.97	0.98	0.84	0.89	0.85	0.84	0.84	0.92	1.07	1.52	1.73	1.15
CMLE Bonferroni		MSE/MSE_{UMLE}	0.99	1.05	1.02	1.00	1.00	0.91	0.91	0.93	0.96	0.96	0.69	0.80	0.81	0.83	0.82	0.85	1.64	1.40	1.01	1.00
		SE/SE_{UMLE}	1.00	1.03	1.01	1.00	1.00	0.94	0.94	0.96	0.97	0.98	0.84	0.89	0.89	0.91	0.90	0.92	1.08	1.16	1.01	1.00
CMLE Conf. Reg.		MSE/MSE_{UMLE}	1.00	1.05	1.01	1.00	1.00	0.87	0.91	0.94	0.96	0.96	0.80	0.86	0.84	0.84	0.84	0.96	1.08	1.03	1.00	1.00
		SE/SE_{UMLE}	1.01	1.03	1.00	1.00	1.00	0.92	0.94	0.97	0.98	0.98	0.91	0.92	0.91	0.91	0.91	0.98	1.04	1.01	1.00	1.00
CMLE filtered		MSE/MSE_{UMLE}	0.99	1.05	1.01	1.00	1.00	0.91	0.9	0.93	0.96	0.96	0.69	0.81	0.82	0.83	0.82	0.86	1.37	1.05	1.00	1.00
		SE/SE_{UMLE}	1.00	1.03	1.00	1.00	1.00	0.94	0.93	0.96	0.97	0.98	0.85	0.90	0.90	0.91	0.90	0.94	1.13	1.02	1.00	1.00

Table 6.8: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Independent covariates.

		True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
CMLE	MDC S1	Isotonic	35.9	47.6	94.6	99.9	100	5.3	2.6	0.0	0.0	0.0	16.4	14.2	12.1	12.3	11.2	12.8	24.9	5.5	0.2	0.0
		Antitonic	3.5	1.0	0.0	0.0	0.0	34.1	56.8	99.8	100	100	18.1	15.0	13.3	13.0	13.7	30.5	28.9	0.0	0.0	0.0
		Both	60.4	51.4	5.4	0.1	0.0	60.2	40.5	0.1	0.0	0.0	65.1	70.4	74.5	74.5	75.1	53.4	31.9	0.0	0.0	0.0
		None	0.2	0.0	0.0	0.0	0.0	0.4	0.1	0.1	0.0	0.0	0.4	0.4	0.1	0.2	0.0	3.3	14.3	94.5	99.8	100
	MDC S2	Isotonic	43.2	55.3	97.0	99.9	100	7.8	4.5	0.0	0.0	0.0	22.1	20.6	18.2	16.0	17.3	17.1	33.0	33.1	8.8	0.0
		Antitonic	5.3	2.0	0.1	0.0	0.0	41.9	65.1	99.9	100	100	23.5	20.4	19.4	17.4	18.6	38.1	46.3	52.4	1.4	0.0
		Both	51.5	42.7	2.9	0.1	0.0	50.3	30.4	0.1	0.0	0.0	54.4	59.0	62.4	66.6	64.1	44.8	20.7	0.0	0.0	0.0
		None	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.5	89.8	100
	MDC S3	Isotonic	60.6	75.5	99.0	99.9	100	19.6	10.2	0.0	0.0	0.0	42.1	44.2	51.9	19.7	18.3	46.4	45.3	43.5	71.0	93.8
		Antitonic	39.4	24.5	1.0	0.1	0.0	80.4	89.8	100	100	100	57.9	55.8	48.1	80.3	81.7	53.6	54.7	56.5	29.0	6.2
CMLE Bonferroni	Unconstrained		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	63.5	97.9	100
	MDC S3	Isotonic	60.6	75.5	99.2	99.9	100	19.6	10.3	0.0	0.0	0.0	42.1	44.2	47.1	46.6	50.6	46.4	45.2	22.2	2.1	0.0
		Antitonic	39.4	24.5	0.8	0.1	0.0	80.4	89.7	100	100	100	57.9	55.8	52.9	53.4	49.4	53.6	54.6	14.3	0.0	0.0
CMLE Conf. Reg.	Unconstrained		4.8	0.9	0.0	0.0	0.0	11.0	6.3	0.6	0.7	0.1	17.7	5.4	1.8	1.3	1.5	45.9	59.7	97.6	100	100
	MDC S3	Isotonic	58.3	72.5	99.3	99.9	100	13.4	6.3	0.0	0.0	0.0	32.2	39.8	48.8	46.6	49.6	38.0	34.8	2.4	0.0	0.0
		Antitonic	36.9	26.6	0.7	0.1	0.0	75.6	87.4	99.4	99.3	99.9	50.1	54.8	49.4	52.1	48.9	16.1	5.5	0.0	0.0	0.0
CMLE filtered	Unconstrained		0.2	0.0	0.0	0.0	0.0	0.4	0.1	0.1	0.0	0.0	0.4	0.4	0.1	0.2	0.0	3.3	14.3	94.5	99.8	100
	MDC S3	Isotonic	60.1	73.9	99.3	99.9	100	19.1	10.0	0.0	0.0	0.0	41.5	41.6	48.9	46.8	50.6	45.8	42.5	5.5	0.2	0.0
		Antitonic	39.7	26.1	0.7	0.1	0.0	80.5	89.9	99.9	100	100	58.1	58.0	51.0	53.0	49.4	50.9	43.2	0.0	0.0	0.0

Table 6.9: Classification of monotonicity direction of four OPs based on six methods with 1,000 simulated data sets and correlated covariates (%).

		True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
		Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
UMLE		MSE_{UMLE}	6.99	0.71	0.09	0.04	0.01	23.52	1.03	0.15	0.08	0.01	15.59	1.09	0.13	0.06	0.01	70.3	17.62	0.24	0.14	0.03
		SE_{UMLE}	0.08	0.03	0.01	0.01	0.00	0.15	0.03	0.01	0.01	0.00	0.12	0.03	0.01	0.01	0.00	0.24	0.12	0.02	0.01	0.01
CMLE	MDC S1	MSE/MSE_{UMLE}	0.99	0.98	0.99	1.00	1.00	0.99	0.97	0.93	0.96	0.97	0.93	0.94	0.94	0.94	0.94	0.94	0.93	1.09	1.00	1.00
		SE/SE_{UMLE}	0.99	0.99	1.00	1.00	1.00	1.00	0.98	0.96	0.98	0.99	0.97	0.97	0.97	0.97	0.97	1.00	0.98	1.05	1.00	1.00
	MDC S2	MSE/MSE_{UMLE}	0.99	0.98	0.99	1.00	1.00	0.99	0.97	0.92	0.96	0.97	0.91	0.91	0.91	0.92	0.92	0.92	0.91	4.91	1.49	1.00
		SE/SE_{UMLE}	0.99	0.99	1.00	1.00	1.00	1.00	0.98	0.96	0.98	0.99	0.96	0.95	0.96	0.96	0.96	0.99	0.97	1.59	1.20	1.00
	MDC S3	MSE/MSE_{UMLE}	0.79	1.04	1.03	1.00	1.00	0.94	0.91	0.92	0.96	0.97	0.74	0.80	0.76	0.74	0.73	0.82	0.90	5.45	7.64	28.01
		SE/SE_{UMLE}	0.89	1.02	1.01	1.00	1.00	0.97	0.94	0.96	0.98	0.99	0.87	0.89	0.87	0.83	0.82	0.93	0.96	1.48	1.70	1.82
CMLE Bonferroni		MSE/MSE_{UMLE}	0.79	1.04	1.02	1.00	1.00	0.94	0.91	0.92	0.96	0.97	0.74	0.80	0.81	0.82	0.83	0.82	0.90	2.29	1.07	1.00
		SE/SE_{UMLE}	0.89	1.02	1.01	1.00	1.00	0.97	0.94	0.96	0.98	0.99	0.87	0.89	0.89	0.90	0.90	0.93	0.96	1.42	1.04	1.00
CMLE Conf. Reg.		MSE/MSE_{UMLE}	0.74	1.07	1.01	1.00	1.00	0.98	0.93	0.93	0.97	0.97	0.83	0.85	0.84	0.84	0.84	0.91	1.00	1.03	1.00	1.00
		SE/SE_{UMLE}	0.87	1.03	1.01	1.00	1.00	0.99	0.96	0.96	0.98	0.99	0.92	0.92	0.91	0.91	0.91	0.96	1.00	1.02	1.00	1.00
CMLE filtered		MSE/MSE_{UMLE}	0.79	1.06	1.01	1.00	1.00	0.94	0.91	0.92	0.96	0.97	0.74	0.81	0.83	0.83	0.83	0.82	0.90	1.09	1.00	1.00
		SE/SE_{UMLE}	0.9	1.02	1.01	1.00	1.00	0.97	0.94	0.96	0.98	0.99	0.87	0.90	0.90	0.90	0.90	0.94	0.97	1.05	1.00	1.00

Table 6.10: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Correlated covariates.

6.2.1 CMLE MDC S3 versus monotonicity direction classification by Maximum Likelihood over all possible combinations

As discussed in Chapter 5, the MDC procedure described in Section 2.4 and the method of monotonicity direction classification by Maximum Likelihood over all possible combinations (see Section 5.2) are two methods that impose monotonicity constraints on all of the OPs. These two methods will be referred to as 'CMLE MDC S3' and 'MDC ML'. In this section it will be shown through the analysis of simulations that their results differ and the reasons why this is so will be discussed. Two of the previous settings used in Section 6.2 will be replicated but now with the purpose of assessing these two restrictive methods only. The first setting that will be analysed corresponds to the one with two OPs (OP 1: 'isotonic' and OP 2: 'antitonic') and independent covariates (see the first simulation setting of Section 6.2), and the second setting to be used is the one with four OPs (OP 1: 'isotonic', OP 2: 'antitonic', OP 3: 'both' and OP 4: 'none') and independent covariates (see the third simulation setting of Section 6.2). The first is called simulation setting MO (monotonic only) and the second simulation setting M&NM (monotonic and non-monotonic).

The simulation setting MO uses two OPs whose effects are truly monotonic. Given that both 'CMLE MDC S3' and 'MDC ML' impose monotonicity constraints on those two OPs, the monotonicity direction classification resulting from these methods tends to be the same as n increases according to the results shown in Table 6.11. For small sample size ($n = 50$), 'MDC ML' shows a better accuracy of the monotonicity direction classification for both, OP 1 and OP 2. This is because this method fits four models according to the four possible combinations of monotonicity directions and then chooses the one that maximises the likelihood across the four models, whereas 'CMLE MDC S3' fails more frequently because of its initial steps one and two (see Section 2.4) where the relative position among confidence intervals are compared, which are affected by the small sample size leading to misclassification.

In terms of MSE, the results shown in Table 6.12 indicate that for large enough n (see $n = 5000$) the results of these two restrictive methods are the same. However, for most of the cases of smaller n the average MSE of 'MDC ML' is greater than the one of 'CMLE MDC S3', and the case of $n = 50$ is the only exception.

Both the accuracy in terms of monotonicity direction and the average MSE (Tables 6.11 and 6.12) indicate that, when the effects associated with the OPs are monotonic, the performance of 'CMLE MDC S3' is better than or equal to the one of 'MDC ML' for $n \geq 100$, whereas for $n = 50$ the performance of 'MDC ML' is better than the one of 'CMLE MDC S3'.

The simulation setting M&NM uses four OPs, where the true monotonicity directions are the following: OP 1 is 'isotonic', OP 2 is 'antitonic', OP 3 is 'both', and OP 4 is 'none'. Given that both 'CMLE MDC S3' and 'MDC ML' impose monotonicity constraints on all of the OPs, the monotonicity direction classification resulting from these methods is heavily affected by imposing monotonicity constraints on OP 4, whose true effects are non-monotonic, as shown in Table 6.13. The main difference between these methods is in the monotonicity direction classification of OP 1. As n increases, the proportion of cases where OP 1 is classified as 'isotonic' by 'MDC ML' decreases, whereas for 'CMLE MDC S3' it rapidly increases reaching 99.7% of accuracy for $n = 500$.

Something similar occurs in terms of average MSE. The one of 'MDC ML' is much worse than the one of 'CMLE MDC S3' for OP 1 when $n \geq 500$ as shown in Table 6.14. For smaller sample sizes, the results of the two methods are similar to each other when comparing for $n = 50$ and they are mixed for $n = 100$. The largest differences are the ones for OP 1 when $n \geq 500$.

According to the results of accuracy and average MSE, see Tables 6.13 and 6.14, 'MDC ML' performs poorly compared to 'CMLE MDC S3' for OP 1 and $n \geq 100$. To see why, an exemplary simulated data set is used to illustrate one of the 911 cases where OP 1 was misclassified by 'MDC ML' for the largest sample size used in the simulations, $n = 5000$ (see Figure 6.5). Under the same setting, there is no case like this in the results of 'CMLE MDC S3'.

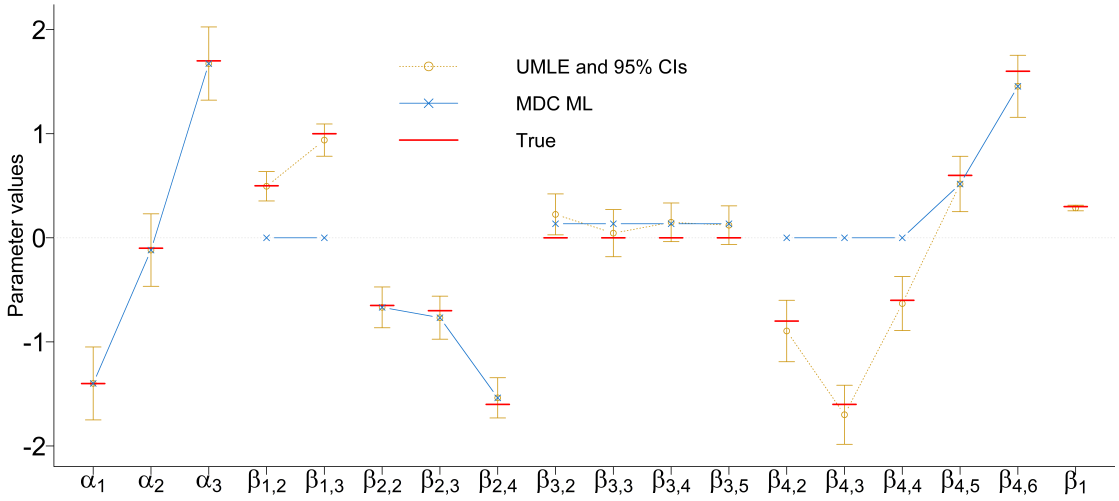

Figure 6.5: An exemplary simulated data set for which OP 1 is classified as 'antitonic' by 'MDC ML', $n = 5000$.

The misclassification of OP 1 produced by 'MDC ML', like the one shown in Figure 6.5, happens because when 'MDC ML' imposes monotonicity constraints on OP 4, it forces the parameter estimates of $\beta_{4,2}$, $\beta_{4,3}$, and $\beta_{4,4}$ to be positive despite the fact that their true parameters are actually negative, and, in most of the cases, their unconstrained MLEs too. This implies that the maximum likelihood approach modifies other constrained parameter estimates like the ones of OP 1 mainly, which in the simulation setting M&NM leads to misclassification too frequently. The same situation does not happen with 'CMLE MDC S3' because it is not based on maximum likelihood only. It compares the relative positions of the confidence intervals of the unconstrained parameter estimates for each OP in its steps 1 and 2 in order to classify their monotonicity direction and then, for those OPs that are still not classified as 'isotonic' or 'antitonic', uses the maximum likelihood approach to choose the monotonicity direction of those remaining unclassified OPs. This means that 'CMLE MDC S3' offers two instances where the monotonicity direction could be found before using constrained maximum likelihood. In this setting, OP 1 was classified as 'isotonic' in step 1 mainly, and therefore its classification was not affected by the maximum likelihood approach used in step 3.

	True pattern	OP 1: Isotonic					OP 2: Antitonic				
	Sample size	50	100	500	1000	5000	50	100	500	1000	5000
CMLE MDC S3	Isotonic	69.2	92.9	99.8	100	100	9.1	5.7	0.0	0.0	0.0
	Antitonic	30.8	7.1	0.2	0.0	0.0	90.9	94.3	100	100	100
MDC ML	Isotonic	84.1	90.8	100	100	100	5.5	2.6	0.0	0.0	0.0
	Antitonic	15.9	9.2	0.0	0.0	0.0	94.5	97.4	100	100	100

Table 6.11: Classification of monotonicity direction of two OPs based on two methods with 1,000 simulated data sets, different sample sizes and independent covariates (%).

	True pattern	OP 1: Isotonic					OP 2: Antitonic				
	Sample size	50	100	500	1000	5000	50	100	500	1000	5000
UMLE	MSE_{UMLE}	1.75	0.73	0.10	0.05	0.01	64.86	1.26	0.25	0.12	0.02
	SE_{UME}	0.04	0.03	0.01	0.01	0.00	0.22	0.04	0.02	0.01	0.00
CMLE MDC S3	MSE/MSE_{UMLE}	0.95	0.75	0.82	0.86	0.95	0.99	0.74	0.75	0.74	0.93
	SE/SE_{UMLE}	0.97	0.85	0.90	0.92	0.97	0.99	0.83	0.82	0.84	0.96
MDC ML	MSE/MSE_{UMLE}	0.86	0.90	0.89	0.92	0.95	0.99	0.82	0.89	0.88	0.93
	SE/SE_{UMLE}	0.92	0.93	0.93	0.95	0.97	0.99	0.89	0.93	0.93	0.96

Table 6.12: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Independent covariates.

	True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
	Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
CMLE MDC S3	Isotonic	68.8	83.4	99.7	100	100	14.0	6.6	0.0	0.0	0.0	41.4	44.5	29.9	21.9	17.7	48.0	49.0	64.9	81.0	98.6
	Antitonic	31.2	16.6	0.3	0.0	0.0	86.0	93.4	100	100	100	58.6	55.5	70.1	78.1	82.3	52.0	51.0	35.1	19.0	1.4
MDC ML	Isotonic	69.8	70.5	50.2	34.3	8.9	9.5	2.0	0.0	0.0	0.0	39.4	32.7	17.7	22.8	54.1	71.6	77.7	94.0	99.2	100
	Antitonic	30.2	29.5	49.8	65.7	91.1	90.5	98.0	100	100	100	60.6	67.3	82.3	77.2	45.9	28.4	22.3	6.0	0.8	0.0

Table 6.13: Classification of monotonicity direction of four OPs based on two methods with 1,000 simulated data sets and independent covariates (%).

	True pattern	OP 1: Isotonic					OP 2: Antitonic					OP 3: Both					OP 4: None				
	Sample size	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000	50	100	500	1000	5000
UMLE	MSE_{UMLE}	2.74	0.49	0.06	0.03	0.01	6.39	0.87	0.11	0.05	0.01	9.24	0.93	0.12	0.06	0.01	24.07	1.34	0.21	0.09	0.02
	SE_{UMLE}	0.05	0.02	0.01	0.01	0.00	0.08	0.03	0.01	0.01	0.00	0.09	0.03	0.01	0.01	0.00	0.15	0.04	0.01	0.01	0.00
CMLE MDC S3	MSE/MSE_{UMLE}	0.99	1.05	1.00	1.00	1.00	0.91	0.91	0.93	0.96	0.96	0.69	0.8	0.74	0.74	0.74	0.85	1.65	5.45	9.83	39.49
	SE/SE_{UMLE}	1.00	1.03	1.00	1.00	1.00	0.94	0.94	0.96	0.97	0.98	0.84	0.89	0.85	0.84	0.84	0.92	1.07	1.52	1.73	1.15
MDC ML	MSE/MSE_{UMLE}	0.94	1.14	5.64	13.76	109.72	0.92	0.83	0.93	0.96	0.96	0.68	0.71	0.45	0.44	0.53	0.90	1.38	4.16	8.02	38.83
	SE/SE_{UMLE}	0.98	1.08	1.78	2.31	3.11	0.95	0.89	0.96	0.97	0.98	0.84	0.83	0.61	0.62	0.72	0.93	0.96	0.93	0.65	0.40

Table 6.14: Average of the MSEs and average of the SEs associated with the categories of each OP when using UMLE (MSE_{UMLE} and SE_{UMLE}). Ratio of the average of the MSEs associated with the categories of each OP when using other methods to MSE_{UMLE}, and ratio of the average standard errors of a constrained method to the one of the UMLE (MSE/MSE_{UMLE} and SE/SE_{UMLE}). Independent covariates.

6.3 CMLE models versus scoring systems for the treatment of ordinal predictors

In the context of regression analysis for an ordinal dependent variable with ordinal predictors, in some literature with real data applications the chosen model is the proportional odds cumulative logit model whereas the ordinal predictors are transformed into interval-scaled variables (e.g. Alvarez-Galvez et al. (2013); Lanfranchi et al. (2014); Corathers et al. (2017)). In this section, this approach was taken in order to compare the results of using different scoring systems for the treatment of ordinal predictors against the use of the six constrained models proposed in previous sections: the most restrictive method "CMLE S3" (Sections 2.3 and 3.2.1) and five other methods that allow dropping monotonicity constraints (see Section 5.3).

The study was based on simulations. The number of simulated data sets was 1,000, from which 500 were used to train the models and 500 to test them. For each one of the 1,000 data sets the number of categories for the dependent variable, their distribution, the number of ordinal predictors together with their categories and distributions, the type of non-ordinal predictor, and all the parameters were the same as those described in Section 6.2 for the last simulation setup. In addition, the number of observations defined four different scenarios, n =100, 200, 500 and 1,000.

For each simulated data set, the six constrained methods and 10 other methods were fitted. The unconstrained methods were used according to 10 different scoring systems. As in Tutz and Hechenbichler (2005), the assessment of the results considered three measures of accuracy:

1. Misclassification rate (MR):

$$(1/n)\sum_{i=1}^{n} 1_{y_i \neq \hat{y}_i}. \tag{6.3.1}$$

2. Mean absolute prediction error (MAPE):

$$(1/n)\sum_{i=1}^{n} |y_i - \hat{y}_i|. \tag{6.3.2}$$

3. Mean-squared prediction error (MSPE):

$$(1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{6.3.3}$$

All of these measures are not specially designed for ordinal variables. On the one hand, the main drawback of the first measure, the misclassification rate, is that it does not consider the order of categories because failures in the prediction do not incorporate information about the error distance at all, assuming that prediction errors have the same value no matter whether the predictions are far away from the true category or not, which is a key distinction between nominal-scaled variables and ordinal variables. On the other hand and in contrast to the first measure of accuracy, the mean absolute predicted error and mean-squared predicted error do take into account the error distances. However, they both assume that the distance between adjacent categories is known, which is indeed a key difference between interval-scaled variables and ordinal variables. As in Tutz and Hechenbichler (2005), the mean absolute predicted error and mean-squared predicted error assume that the four ordinal categories of the dependent variable are transformed using the linear scores 1, 2, 3, 4, for the computation of the accuracy measure only. This suggests that the development of a measure of accuracy for ordinal variables is still required. One possible approach is to compute, for a given observed ordinal category, the distribution of predicted categories, from which one of the predicted categories will be the true one and all the others will be misclassification. This produces as many distributions as ordinal categories of the dependent variable, which could be compared against those obtained from models with different treatments of ordinal predictors. However, it still requires to find a way of aggregating these distributions, which is left for future work, meaning that the measures of accuracy presented above are used in the current section.

In Section 1.4 several scoring systems were presented, from which those that fit the context of the current simulation setup are Veenhoven (see (1.4.3)), Cumulative proportions (see (1.4.4)), Ridits (see Section 1.4.2), Normal median (see (1.4.7)), Midranks (see Section 1.4.3), Van Der Waerden (see (1.4.11)), Blom (see (1.4.12)), Tukey (see (1.4.14)) and Normal mean (see (1.4.15)). As seen in Section 1.4.3,

Equation (1.4.6), there is a linear relationship between Ridits and Midranks that makes them equivalent, which will be confirmed by the results of the simulations. Other approaches in Section 1.4 work in a different context, were extra variables provide information to transform each ordinal predictor into interval-scaled predictors, which is not assumed in the current simulation setup. In addition, the latent variable models for ordinal data (see Section 1.7.2) was used in the current context as a dimensionality reduction technique (referred to as LVM), where the four ordinal predictors were transformed into one latent variable and then factor scores were computed for each one of the 360 possible combination of ordinal categories (the number of categories are 3, 4, 5, and 6). Those factors were obtained using the expected a posteriori scores defined in Equation (1.7.16), which is available as the method `EAP` in the function `factor.scores` of the `R` package `ltm` (see Rizopoulos (2006)).

The latent variable approach is supported by the fact that the ordinal predictors were simulated according to what is described in the last simulation setup of Section 6.2, assuming that there is a correlation structure among them. However, the simulated ordinal predictors do not result from the data generation process of a true latent variable model. Therefore, loss of information was observed as a consequence of the dimensionality reduction, as expected.

Given that the observed distribution of each ordinal predictor changes depending on the set where it is obtained, i.e.,

(i) training set,

(ii) test set, or

(iii) both sets together,

the scores resulting from the training set might not be the same compared to either the ones resulting from (ii) or (iii). Thus, the following approach was taken. First, the scores for the ordinal predictors based on the training set were computed, with which the numeric assignments were defined for each scoring system, and then those scores were used to transform the ordinal predictors according to

the information provided by (ii), the test set. This approach was used because using (ii) or (iii) to update the scores after having learnt the scores in the training set would change the scores every time that new observations are included in the test set, changing the definition of the transformed ordinal predictors and therefore the parameter estimates resulting from the training set would not necessarily correspond to those that maximise the likelihood of the model if those new scores were to be used in the training set.

As mentioned before, the models were assessed using the accuracy measures MR, MAPE and MSPE. They were computed for each test set and every model. Given that there were 500 simulated test sets, there are 500 MR, MAPE and MSPE for each model. Hence, the analysis is based on their means and confidence intervals for all of the sample sizes. The $(1-\alpha)\%$ confidence intervals were computed according to the formula:

$$\bar{A} \pm t_{0.975,499}\frac{\hat{\sigma}_A}{\sqrt{N}},$$

for the accuracy measures A = MR, MAPE, or MSPE, and where $N = 500$, the number of simulation replicates. In addition, boxplots for the selected sample sizes n =100 and 1,000 were analysed (see Figures 6.6, 6.8, and 6.9). Comments involving significance are stated using a 95% confidence level.

Figure 6.6(a) shows for n=100 (black lines) a higher MSPE for the proposed constrained methods than for those using scoring systems as the treatment of ordinal predictors, except for 'CMLE Conf. Reg.', which is not statistically different from them at a 95% confidence level (lower mean MSPE than 'Veenoven' though). A special case is 'LVM', for which the loss of information of the dimensionality reduction implied a decrement in accuracy compared to models where the number of predictors was not reduced. However, for the smallest sample size of the analysis, this information loss did not produce a significant difference in the mean of MSPE compared to the constrained models, except for 'CMLE Conf. Reg.'.

Increasing the sample size from 100 to 200 significantly improved the mean MSPE of the constrained methods. In fact, for any given constrained model, all of the differences between the mean MSPE resulting from sample sizes 100 and

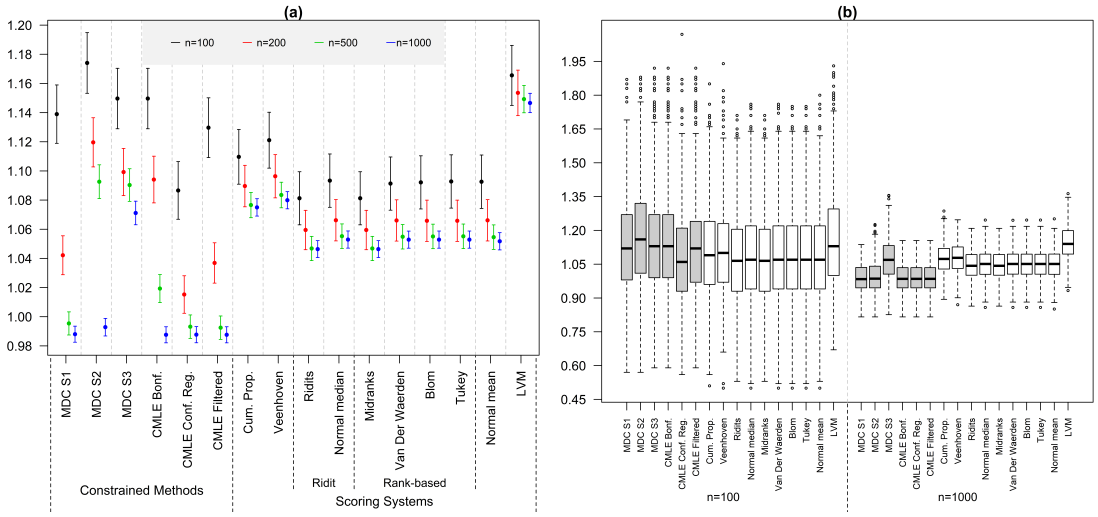

Figure 6.6: Constrained methods versus methods using some scoring systems for OPs. (a) Mean MSPE and 95% confidence intervals for different sample sizes, (b) boxplots of MSPE resulting from different methods, for n=100 and 1000, grey boxplots correspond to constrained methods.

200 are significant, which does not occur for the models using scoring systems. In general, the effect of the sample size is much greater on the constrained methods than on those using scoring systems.

For n=200 (red lines) is already possible to see that some constrained methods perform better than other models using scoring systems. 'CMLE Conf. Reg.' shows a significantly smaller mean MSPE compared to any other model using scoring systems. 'CMLE filtered' is significantly better than any other model using scoring systems except when being compared to 'Ridits' or 'Midranks'. In fact, it can be seen that these two models show the same results. This is because one is a linear function of the other, as shown by Equation (1.4.6) in Section 1.4.3 (see also Agresti (2010)). Another constrained method that rapidly becomes better than other models using scoring systems as n increases is 'MDC S1', which results to be significantly better than 'Cum. Prop.' and 'Veenhoven'.

For greater sample sizes, n=500 or 1000, 'MDC S1', 'CMLE Bonf.', 'CMLE Conf. Reg.' and 'CMLE Filtered' are all significantly better than any other model using scoring systems. The mean MSPE of 'MDC S2' is significantly higher than the one of many other models using scoring systems when n=500, but it is

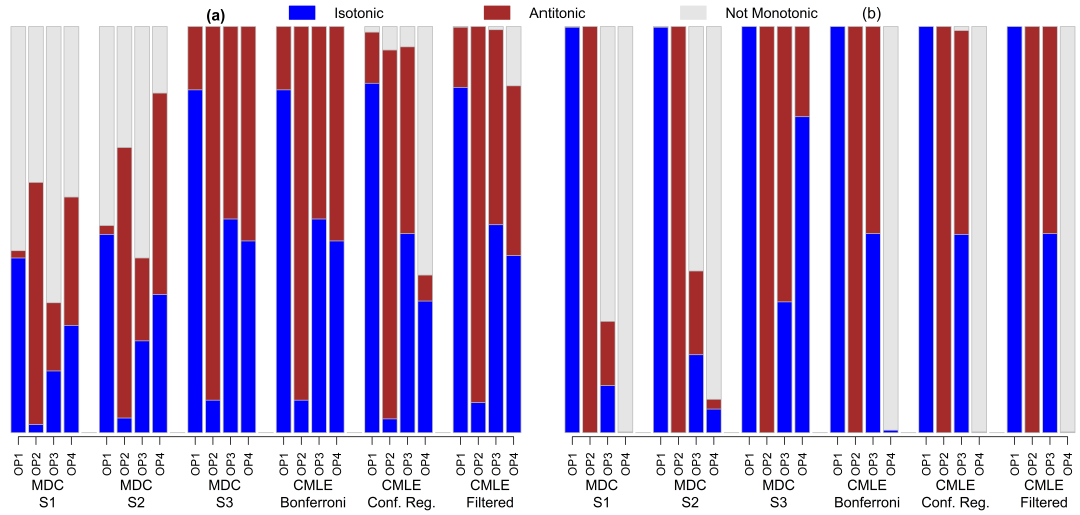

Figure 6.7: Distribution of monotonicity direction classification of different constrained methods for each one of the ordinal predictors. (a) n=100, (b) n=1,000.

significantly smaller than any of them when n increases to 1,000. 'MDC S3' is the only constrained method that is not significantly better than any other model using scoring systems for any sample size of the simulation study. This is because it is the only constrained method that imposes monotonicity constraints on every ordinal predictor, regardless of the values that their parameter estimates take. 'MDC S3' imposed monotonicity constraints to all of the OPs, including OP 3 and 4, which are assumed to represent 'both' monotonicity directions and 'none' monotonicity direction correspondingly. Figure 6.7 shows that 'MDC S3' is the only method with no 'not monotonic' classifications. In the simulation setup OP 1 was assumed to be isotonic, OP 2 antitonic, for OP 3 all the parameters were set at zero, indicating that both monotonicity directions were compatible with it, and OP 4 was assumed to be non-monotonic. In particular, 'MDC S3' classified OP 3 and OP 4 as 'Isotonic' in 52.6% and 47.2% of the 500 training sets when $n = 100$, despite the fact that their true patterns are 'both' and 'non-monotonic'. The corresponding figures when $n = 1000$ are 32.2% and 77.8%. These classifications of model 'MDC S3' force the parameters of OP 3 and OP 4 to be monotonic, leading to prediction errors.

Not only the confidence intervals for the mean of MSPE show great difference in the results for large sample sizes but also the boxplots of the MSPE shown in

Figure 6.6(b), where the interquartile ranges for the models with n=1,000 decrease to approximately a third (on average) with respect to the ones for n=100. In addition, when n=1,000 the boxes of the constrained models are almost fully under the ones of the models using scoring systems (except for 'MDC S3' because of the reasons discussed earlier). However, when n=100, the boxplots do not show great difference in their locations.

Regarding the mean absolute prediction error, Figure 6.8(a) shows that the effect of the sample size is still greater for constrained models than for models using scoring systems.

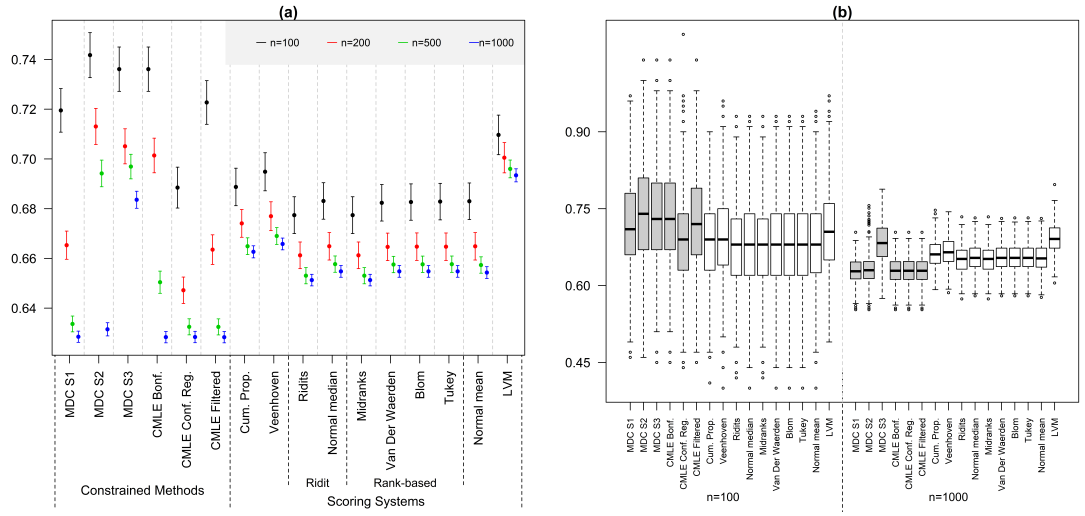

Figure 6.8: Constrained methods versus methods using some scoring systems for OPs. (a) Mean MAPE and 95% confidence intervals for different sample sizes, (b) boxplots of MAPE resulting from different methods, for n=100 and 1,000, grey boxplots correspond to constrained methods.

In general, the main results described for the analysis of the mean MSPE remain approximately the same for the analysis of the mean MAPE when n =100 or 200 (compare Figure 6.8(a) against Figure 6.6(a)). However, for greater sample sizes, there are some conclusions that are different. For n=500, the mean MAPE shows that 'CMLE Bonf.' is significantly better than some models using scoring systems, while the mean MSPE showed that it is significantly better than all of them. Also, according to the mean MAPE, 'MDC S2' and 'MDC S3' became worse than any other method except when being compared to 'LVM', while the mean

MSPE showed that their performance was not significantly different from the one of 'Cum. Prop.' and 'Veenhoven'. For n=1,000, all of the constrained models show a better mean MAPE than those models using scoring systems, except for 'MDC S3', which is indeed worse than any other.

The boxplots of the MAPE (see Figure 6.8(b)) show that the constrained methods perform better than the models using scoring systems, like the ones of the MSPE, when n=1,000. However, the boxplots of the MAPE show a slightly worse performance than the ones based on the MSPE.

The general conclusions based on the analysis of the mean MAPE are not so different compared to those from the analysis of the mean misclassification rate (mean MR) shown in Figure 6.9. However, despite the fact that many of the differences between pairs of methods remain significant and in the same direction (greater or smaller to each other), the mean MR of models using scoring systems get closer to the ones of the constrained methods when the latter perform better than the former in mean MAPE (n =500 or 1,000), and, vice versa, the mean MR of models using scoring systems get further away from the ones of the constrained methods when the latter perform worse than the former in mean MAPE (n =100 or 200), namely, the relative performance of constrained methods in mean MR is good but not as good as the one based on mean MAPE when n=500 or 1,000, and the relative performance of constrained methods in mean MR is bad and even worse than the one based on mean MAPE when n=100 or 200.

By definition, given that MSPE assigns larger values to greater distances between the predicted and observed ordinal category than MAPE, then the larger the prediction error, the greater the MSPE with respect to the MAPE. In addition, comparisons between the results of the mean MSPE and the ones of the mean MAPE, shown in Figures 6.6 and 6.8 correspondingly, indicate that the results of the constrained methods compared to models using scoring systems are even better when they are assessed using the mean MSPE. Consequently, for large sample sizes, n =500 and 1,000, the constrained methods (except for 'MDC S3' and in some settings 'MDC S2') not only show a better classification rate, but also a much better MSPE than the models using scoring systems, whereas for smaller

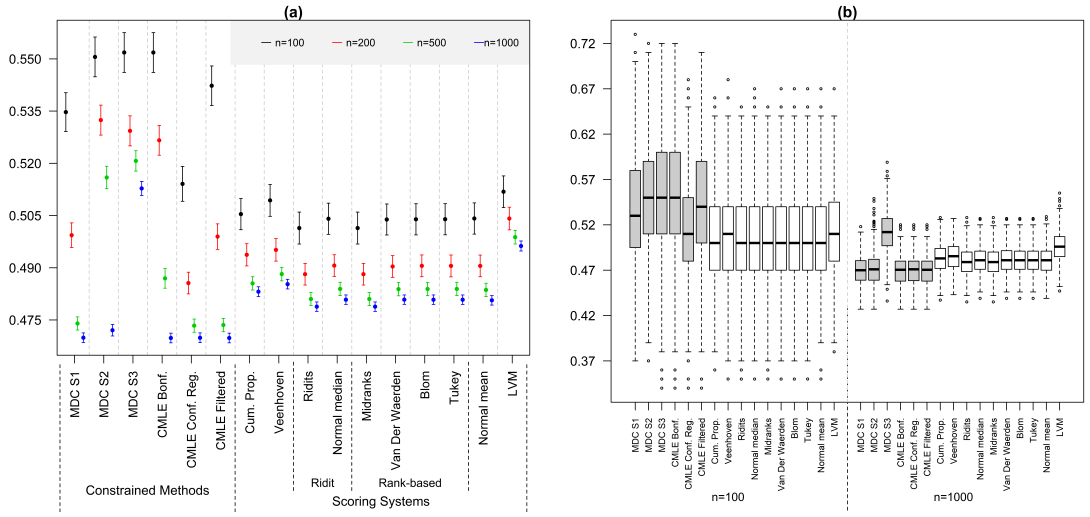

Figure 6.9: Constrained methods versus methods using some scoring systems for OPs. (a) Overall MR and 95% confidence intervals for different sample sizes, (b) boxplots of MR resulting from different methods, for n=100 and 1,000, grey boxplots correspond to constrained methods.

sample sizes, n =100 and 200, the misclassification rate is in general higher for the constrained methods, which is compensated by a relative improvement of the mean MSPE for the constrained methods, where in some settings the differences in mean MSPE with respect to the ones of the models using scoring systems are not significant or even show a better mean MSPE, even when the sample size is small.

6.4 Application to quality of life assessment in Chile

As an illustration of the proposed methodologies, we analyse the association between a quality of life self-assessment variable (10-point Likert scale) and ordinal and other predictors from a Chilean survey, the National Socio-Economic Characterisation 2013 (CASEN). More recent surveys, 2015 and 2017, do not include the question about quality of life self-assessment. This survey retrieves information with the aim of characterising the population of people and households. Our analysis is based on 7,374 householders, namely those who live in the capital and have reported the quality of life self-assessment.

The set of covariates was chosen on the basis of previous research in the field (for example, Di Tella et al. (2003); Cheung and Lucas (2014); Boes and Winkelmann (2010)). The data set was published by the Ministry of Social Development of Chile and it is available online at: `http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php`. The detailed data preprocessing is described in Appendix B.

The response variable is a self assessment of the quality of life (QoL). The question was 'Considering everything, how satisfied are you with your life at this moment?'. The possible alternatives were: '1 Completely Unsatisfied', '2',. . ., '9', '10 Completely Satisfied'.

The model was fitted with ordinal, ratio and nominal-scaled covariates. For the ordinal and nominal-scaled ones, the first category to be mentioned was considered as the baseline. The ordinal covariates are *Educational Level (Edu)* with categories 'Not Educated', 'Primary', 'Secondary', and 'Higher'; *Income Quintile (Inc)* with levels from 'Q1' to 'Q5' where 'Q5' represents the highest income; *Health Status (Hea)*, a health self-assessment reported as ordinal Likert scale from 1 to 7, with 7 being the best possible status; *Overcrowding (Ove)*, which is an index representing the number of people living in the household per bedroom, with categories 'Not Overcrowded' for less than 2.5, '[2.5,3.5)', '[3.5,5.0)', and '5.0 or more'; and *Children (Chi)*, a grouped version of the number of people under 15 years old living in the household, with categories '0','1', '2', '3', and '4 or more'. The ratio-scaled variable is *Age*. The nominal-scaled ones are *Activity (Act)*, with categories 'Economically Inactive', 'Unemployed', and 'Employed'; and *Sex* ('Male', 'Female'). Therefore, the set of ordinal predictors is $\mathcal{S} = \{Edu, Inc, Hea, Ove, Chi\}$.

An unconstrained version of the model (2.3.4) was fitted to obtain the parameter estimates and their standard errors. The unconstrained parameter estimates and their 95% confidence intervals are shown in Figure 6.10. The definition of the variables suggests a monotonic association with respect to the response and the unconstrained results seem to be consistent with the monotonicity assumption for all the OPs. We also used the monotonicity tests described in Section 4.2 and Section 4.3 as a complementary assessment of the monotonicity assumptions. Both

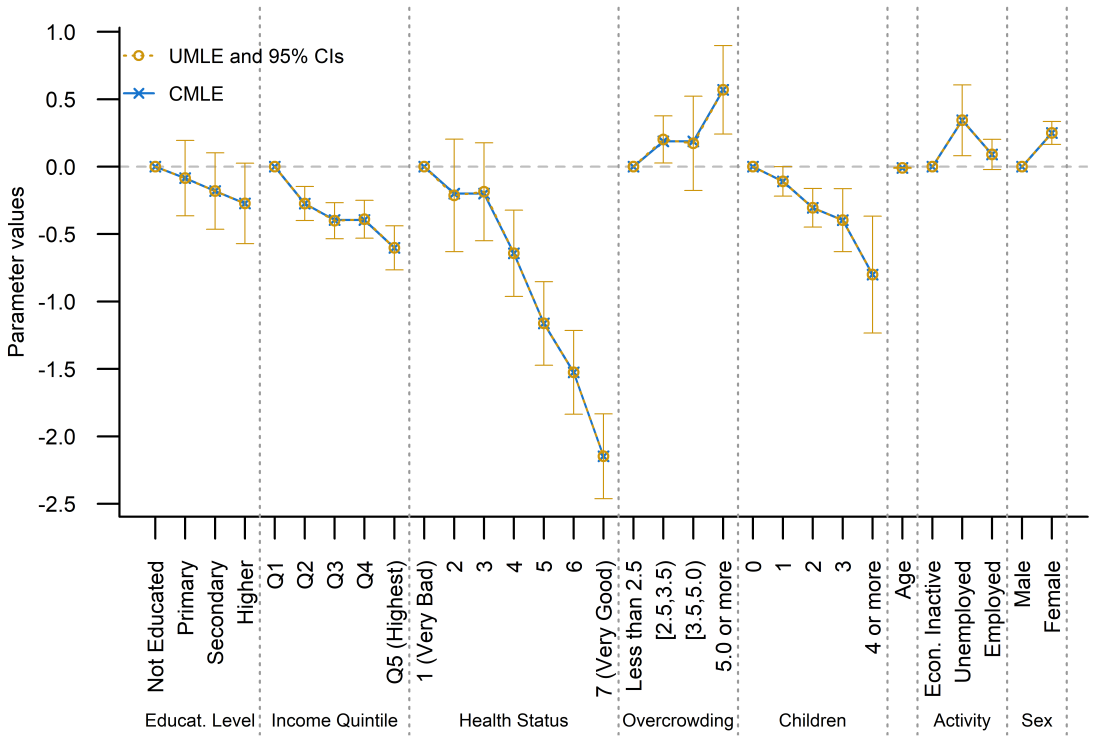

Figure 6.10: CMLEs and UMLEs for a model applied on real data with an ordinal response, ordinal predictors and others. The first category of each ordinal or nominal predictor is assumed as the reference category. Intercept parameter estimates omitted. The 95% confidence intervals correspond to the UMLEs.

of them did not reject the null hypothesis of monotonicity for any of the OPs with $\alpha^* = 0.05$, in fact, their p-values were grater than 0.998 in both tests for every OP. Therefore, the assumption of monotonicity was imposed on all of them and the approach 'CMLE MDC S3' was chosen to be the constrained method to compute the CMLEs.

Each set of parameter estimates associated with the ordinal predictors in $\mathcal{S}$ was classified as either 'antitonic' or 'isotonic'. The interpretation for the relationship between an ordinal predictor and the response variable with 'antitonic' pattern is that the further away an ordinal category is from its baseline, the smaller $P(y_i \leq j|\mathbf{x}_i)$ is, i.e., the probability of self-assessing QoL in the jth category or smaller. In other words, 'antitonic' patterns mean that higher categories of ordinal variables are associated with more probability of self-assessing QoL in a higher part of the

scale. The inverse interpretation applies for 'isotonic' patterns.

With a 95% individual confidence level ($\tilde{c} = 0.95$), the MDC procedure classified the sets of parameters associated with three ordinal variables as 'antitonic' in its first step (*Income Quintile*, *Health Status*, and *Children*), whereas *Overcrowding* was classified as 'isotonic' and *Educational Level* as 'both'. There was no ordinal predictor classified as 'none' by the end of the first step. Therefore, there was no need of making a decision on whether dropping the monotonicity constraints for variables classified as 'none'. Hence, $\mathcal{A}_1 = \{Inc, Hea, Chi\}$, $\mathcal{I}_1 = \{Ove\}$, and $\mathcal{B}_1 = \{Edu\}$.

Educational Level was the only variable in the MDC procedure's second step. To perform this step, a tolerance level of 0.9 was set together with steps of 1% when gradually decreasing the confidence level starting from the one analysed in step one, 95%. As a result of this step, *Educational Level* was classified as 'antitonic' with a 92% confidence level for each confidence interval.

There was no need to execute the third step of the MDC procedure because all of the monotonicity directions were established earlier. All the ordinal predictors were finally classified as 'antitonic' except for *Overcrowding*, which was classified as 'isotonic'. Therefore, only one constrained model was fitted.

Some of the parameter estimators resulting from UMLE are not in line with the monotonicity assumption. For example, keeping all the other variables constant, an improvement in the *Income Quintile* from 'Q3' to 'Q4', i.e., an increment in the income level, increases the probability of self-assessing *QoL* in lower categories of the scale, according to the UMLE. The same happens with *Health Status*, for which changes from '2' to '3', i.e., improving the health status, seemingly increases the probability of reporting a low self-assessment of *QoL*. These particular unconstrained results are counterintuitive. Therefore, it is reasonable to think that these may have been the result of random variation, and to impose the monotonicity assumption.

In fact, in these cases there is little difference between neighbouring UMLEs, so in terms of the parameter values constrained and unconstrained results are fairly similar, but the proposed methodology can assure the user that monotonicity is

compatible with the data.

For the OP *Educational Level*, the UMLE allows both positive and negative values in all confidence intervals, but after having classified this OP as antitonic, with the baseline parameter fixed at zero and using the CMLE, all further parameters can only be negative.

In general, the UMLEs are compatible with a monotonic association between ordinal predictors and the response variable, but the parameter estimates produce violations of monotonicity. The CMLEs avoid these, and allow for a simpler and more consistent interpretation.

Given that the sample size is relatively large, the individual confidence intervals are relatively small, which allows the first step of the MDC procedure to classify all but one OP as either isotonic or antitonic. In order to explore a situation with a smaller sample size, we ran the methodology on a random subsample of $n = 200$, i.e., 2.7% of the full sample size. All of the OPs were classified by the end of the MDC procedure in the same way as the one of the previous setting, although *Educational Level* ($s = 1$) and *Overcrowding* ($s = 4$) remained classified as 'both' until the end of step 2 even with a low tolerance level of $\tilde{c}_s'^* = 0.8$ for $s = 1$ and $s = 4$. This is an appropriate reflection of the bigger uncertainty in classification when using a smaller sample size.

6.4.1 Results based on the proposed constrained methods

In Section 5.3 five estimation methods were proposed taking into account different ways to make the decision of dropping monotonicity constraints. In the current real data application, the unconstrained results shown in Figure 6.10 indicate that imposing the monotonicity constraints on all of the OPs seems to be a sensible decision. This is supported by the results of the remaining five proposed constrained methods because of the following reasons:

- 'CMLE MDC S1' was the only method that dropped the monotonicity constraint of an ordinal predictor. Given that *Educational Level* was classified as 'both' in the first step of the MDC procedure, 'CMLE MDC S1' did not impose a monotonicity constraint on its parameter estimates. However, the

unconstrained parameter estimates for *Educational Level* were the same as the constrained ones, which can also be seen in Figure 6.10. This is because the unconstrained pattern of parameter estimates resulted to be monotonic, meaning that the monotonicity constraints were not needed. Therefore, 'CMLE MDC S3' is confirmed as a sensible option.

- 'CMLE MDC S2', 'CMLE Bonferroni', 'CMLE Conf. Reg.' and 'CMLE Filtered' did not drop any monotonicity constraint. Therefore, their results were exactly the same as the ones of 'CMLE MDC S3'.

6.4.2 Using scoring systems for the treatment of ordinal predictors

As in Section 6.3, several scoring systems were used to transform the ordinal predictors into interval-scaled variables. The scoring systems correspond to those that were described in Section 1.4 and fit the context of the real data application, that is Veenhoven (see (1.4.3)), Cumulative proportions (see (1.4.4)), Ridits (see Section 1.4.2), Normal median (see (1.4.7)), Midranks (see Section 1.4.3), Van Der Waerden (see (1.4.11)), Blom (see (1.4.12)), Tukey (see (1.4.14)) and Normal mean (see (1.4.15)). Other approaches in Section 1.4 do not fit the context of the real data application because they require extra variables that provide information to transform each ordinal predictor into interval-scaled predictors, which is not part of the current analysis. In addition, the latent variable models for ordinal data (see Section 1.7.2) was used as a dimensionality reduction technique, however its results are excluded from the analysis because the low correlation among ordinal predictors does not contribute to make this approach a good candidate to be used.

The whole sample of 7,374 observations was split in two parts randomly, 50% for the training set and 50% for the test set. As discussed in Section 6.3, the scores for the OPs resulting from the training set might not be the same compared to ones obtained based on the test set. Thus, the scores for the ordinal predictors were computed based on the training set first, and then the resulting numeric assignments were used to transform the ordinal predictors in the test set. This

avoids to make the scores a function of the set being used for each replication. The total number of replicates for the splits is $N = 500$.

Once the ordinal predictors were transformed using the scoring systems listed above, the new variables were used to fit the POCLM for the ordinal response QoL in the training set and then to make predictions in the test set. In order to compare the results among models, two measures of accuracy were used in the test set, the misclassification rate defined in Equation (6.3.1) and the mean-squared prediction error (MSPE) defined in Equation (6.3.3).

Rather than comparing the results among different scoring systems, they will be compared to the ones of the 'CMLE MDC S3', the method used in the real data application of Section 6.4. In turn, the unconstrained results (UMLE) were also included to be compared against the constrained ones too.

The differences between the accuracy measures of 'CMLE MDC S3' versus the ones of other methods are significant in most of the cases. The observed mean MR is shown in Figure 6.11(A) and the mean MSPE in Figure 6.11(B). Confidence intervals were computed following the same reasoning as in Section 6.3. Their relative positions with respect to the one of 'CMLE MDCS3' suggest that some differences could be not significant, therefore this was tested. In order to test whether their means are equal or not, a hypothesis test was conducted for each pairwise comparison between the MR or MSPE population mean of 'CMLE MDC S3' against the one of other methods. Given that each measure of accuracy was computed for each method based on the same training/test set split (500 times), then the comparisons are based on a two-sided paired t-test. According to the results of these tests for differences of MR population means, the 10 p-values of the pairwise comparisons of 'CME MDC S3' against the other methods are smaller than 1%, where the highest is 0.0022 for 'CMLE MDC S3' versus 'Ridits' or 'Midranks' (recall these two methods are mathematically equivalent as previously stated), which means that there is even stronger evidence against equality of MR population means for pairwise comparisons between the one of 'CMLE MDC S3' versus the one of any other method. Regarding the mean MSPE, the results are even more extreme except for the case of 'CMLE MDC S3' versus 'UMLE', where

the null hypothesis $H_0 : E(\text{MSPE}_{\text{CMLE MDC S3}}) = E(\text{MSPE}_{\text{UMLE}})$ is not rejected. All other pairwise tests were rejected with p-values even smaller than 0.001%.

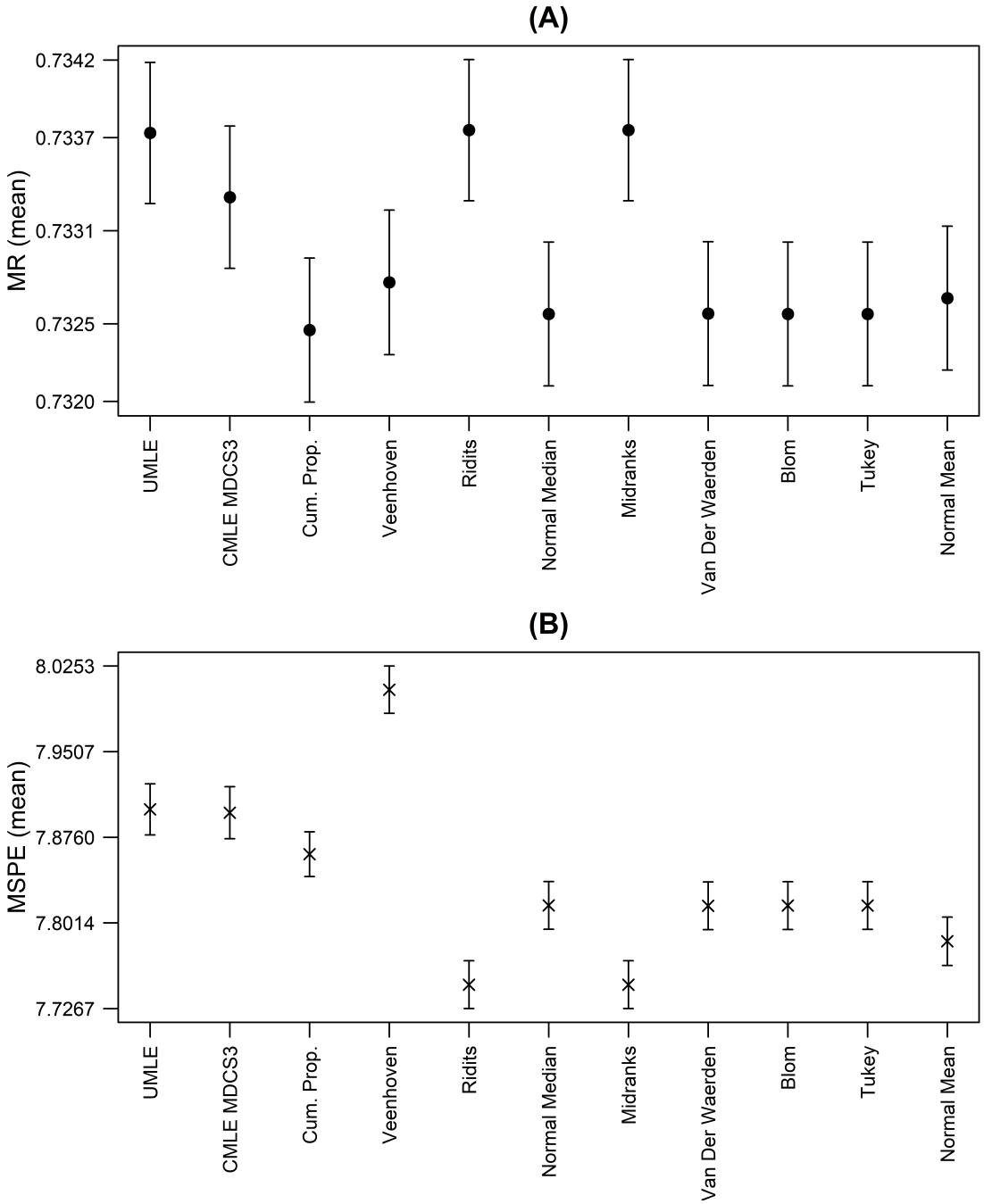

Figure 6.11: (A) mean of the misclassification rate (MR) and (B) mean of the mean-squared prediction error (MSPE). Lines correspond to 95% CIs based on 500 replicates. Both include the results of the UMLE, CMLE, and other methods using scoring systems for the treatment of ordinal predictors.

When comparing the results of the unconstrained POCLM, referred here to as UMLE, against the ones of 'CMLE MDC S3', the performance of the latter is better than the one of the former or at least the same. The misclassification rate (MR), which treats an ordinal response as categorical in terms of measure of accuracy, is much better for 'CMLE MDC S3' than for 'UMLE' as shown in Figure 6.11(A), whereas the analysis of the mean-squared prediction error (MSPE), which treats an ordinal response as of interval scale type in terms of measure of accuracy (see Figure 6.11(B)), does not provide evidence against the null hypothesis of equal population means. This means that, on average, 'CMLE MDC S3' makes the right prediction more frequently than 'UMLE', but when it fails the distance from its predicted value to the right one is, on average, higher than the one of 'UMLE'. In this case, a similar performance in at least one measure of accuracy (see Figure 6.11(B)) was expected because the monotonicity constraints are active on only two parameters of the ordinal predictors *Income Quintile*, *Health Status*, and *Overcrowding*, and the effect of the monotonicity constraints produces a not significant change in the parameter estimates, although they contribute to improve the interpretability of the results and to reduce the misclassification rate.

The 'CMLE MDC S3' showed to be a better option than 'UMLE' because of the advantages discussed above. However, this conclusion is not the same when comparing 'CMLE MDC S3' against other methods using scoring systems to transform ordinal predictors. 'CMLE MDC S3' provides information about the parts of the scale within each ordinal predictor that produce smaller or greater effects on the response variable, which is information that is not possible to obtain when using scoring systems. Also, 'CMLE MDC S3' is significantly better than 'Ridits' and 'Midranks' in MR and better than 'Veenhoven' in MSPE. However, its performance is worse than the one of other methods using scoring systems in terms of both MR and MSPE. Despite the fact that scoring systems assign to the resulting interval-scaled variable information that is not provided by the original ordinal scale, in practice their use has shown to be respectable in some instances in terms of the measures of accuracy used in this section and Section 6.3.

6.5 Implementation

An `R` package called `crov` (constrained regression for ordinal variables) was created containing the implementation of the constrained regression models for ordinal predictors discussed in Section 2.3 and Section 5.3. In addition to the constrained and unconstrained parameter estimates, it makes available the results of each step of the MDC procedure in terms of their monotonicity direction classification and their corresponding confidence levels according to what is defined in Section 2.4. The main function is called `mdcp` (monotonicity direction classification procedure) and its optional parameter `method` allows to specify the required method, for which the current alternatives are: "`MDCS1`", "`MDCS2`", "`MDCS3`", "`CMLEbonferroni`", "`CMLEconfReg`", and "`CMLEfiltered`". Furthermore, the monotonicity test based on the Bonferroni correction and the one based on confidence regions (see Section 4.2 and Section 4.3 correspondingly) are available too.

6.6 Conclusions

In the simulation study of Section 6.2 the following four simulation settings were considered:

(A) Two uncorrelated ordinal predictors with monotonic effects.

(B) Two correlated ordinal predictors with monotonic effects.

(C) Four uncorrelated ordinal predictors, each one with parameters representing a different monotonicity direction classification: 'isotonic', 'antitonic', 'both' or 'none'.

(D) Four correlated ordinal predictors, each one with parameters representing a different monotonicity direction classification.

All of them for sample sizes $n = 50$, 100, 500, 1,000, and 5,000. For each setting, the relative performance of the constrained methods proposed in Chapter 2 and Chapter 5 was analysed from two perspectives:

(i) the monotonicity direction classification ('isotonic', 'antitonic', 'both' or 'none'), and

(ii) the deviation of the parameter estimates from their corresponding true values.

The first will be referred to as the 'MDC results' and the second as the 'MSE results'.

The empirical distributions of monotonicity direction classifications are shown in Table 6.1 for setting (A). The MDC results show a high proportion of correct monotonicity direction classifications of at least 88.9% of the cases when $n \geq 100$ except for 'CMLE MDC S1' and 'CMLE MDC S2'. Given that the distances between adjacent true parameters of each OP are not large enough to prevent their confidence intervals to overlap (see Figure 6.1), these two methods classify the ordinal predictors as 'both' in upto 56.9% of the cases when $n \leq 100$. However, this figure rapidly reduces to 1.4% when $n \geq 500$. In addition, when $n \geq 500$ 'CMLE Conf. Reg.' is the one that drops monotonicity constraints more frequently, showing that for large sample sizes this method is the less conservative in terms of identifying not monotonic patterns.

In terms of the MSE results of setting (A), the constrained methods show a lower MSE compared to the one of the unconstrained model (upto 26% lower), except for particular cases, namely the less restrictive methods when $n = 50$ and for OP 2 only (see Table 6.4).

With correlated covariates (setting (B)), the MDC results require a larger sample size compared to the ones of setting (A) to show a high proportion of correct monotonicity direction classifications. When $n \geq 500$ this proportion is at least 96.9% of the cases except for 'CMLE MDC S1' and 'CMLE MDC S2', whereas for smaller sample sizes, the misclassification rate is higher than the one of setting (A) for OP 1, the ordinal predictor with less clear monotonicity of effects (see Table 6.5). In terms of MSE, the ones corresponding to the constrained methods remain significantly smaller or equal to the ones of the unconstrained model (see Tables 6.4 and 6.6).

With monotonic ordinal predictors only (settings (A) and (B)), the MDC results show that more restrictive constrained methods are highly accurate when $n \geq 100$ and the MSE results show that any constrained approach performs better than the unconstrained one in almost every simulated scenario.

For setting (C), the MDC and MSE results associated with the 'isotonic' and 'antitonic' ordinal predictors (OP 1 and 2 correspondingly) are similar to the ones of setting (A), whereas the OPs representing patterns 'both' and 'none' (OP 3 and 4) contribute to differentiate the results of the constrained methods (see Tables 6.7 and 6.8).

Given that the true pattern of OP 3 is 'both', its monotonicity constraint is expected to be not dropped and its MDC results are expected to show an evenly distributed proportion of 'isotonic' and 'antitonic' monotonicity directions. These expectations are met by the constrained methods, except for 'CMLE MDC S3' because it is too restrictive in this context. This method imposes monotonicity constraints on all of the ordinal predictors, classifying OP 4 as either 'isotonic' or 'antitonic' despite the fact that it is non-monotonic, which in turn affects the classification of OP 3. Therefore, when there is evidence of non-monotonicity for some OP, 'CMLE MDC S3' not only assigns a monotonicity direction to a non-monotonic OP but it might also affect the monotonicity direction classification of others. On the positive side of the results, the constrained method that drops monotonicity constraints more frequently for the non-monotonic ordinal predictor is 'CMLE Conf. Reg.', and it only requires a sample size of 500 observations to do it in 98.4% of the cases. When $n = 50$ it drops the constraints in only 54.2% of the cases. However, its MSE for the smallest sample size is always significantly smaller or equal than the one of the unconstrained model. For greater sample sizes this still holds for the MSE of all or the majority of the parameters associated with any OP (see Table 6.8).

For setting (D), where correlation among covariates is considered, the general trends of the conclusions of setting (C) remain the same for $n \geq 500$. For smaller sample sizes, the MDC results increase their misclassification rate, but the MSE results of the constrained methods are better (see Tables 6.9 and 6.10).

Given that multiple factors can affect the results of the models, it is not possible to generalise many of the conclusions. However, the effect of the sample size seems to be one of the common effects for the different settings under analysis. In general, the MDC results show that smaller sample sizes increase the misclassification of monotonicity directions or indicate not monotonic patterns, but it is necessary to reach $n = 100$ or 500 to observe accuracy rates higher than 90% (see Tables 6.1, 6.5, 6.7 and 6.9). In terms of the MSE results, for the smallest and largest sample sizes, the MSE of the constrained methods gets closer to the one of the unconstrained model (except for 'CMLE MDC S3' with non-monotonic OP for the reasons discussed earlier). In the first case, when $n = 50$, high misclassification rates make the constraint methods to increase the variance of their parameter estimates because they are being constrained in the wrong parameter space more frequently. In the second case, when $n = 5,000$, the estimates of all the methods, including the unconstrained one, get closer to the true model, and therefore their MSE too, meaning that for large enough n, when n increases the choice of the constrained method is less relevant.

Regarding the comparison between 'CMLE MDC S3' and 'MDC ML' discussed in Section 6.2.1, for $n \geq 100$, the former is better than the latter when there are OPs whose effects are non-monotonic, and the performance of 'CMLE MDC S3' is better than or equal to the one of 'MDC ML' when all of the OPs are monotonic. For $n = 50$ the results show a better performance of 'MDC ML', which requires to fit 2^t models though, where t is the number of OPs.

In Section 6.3 the results of using the constrained methods were compared against the ones of models using scoring systems to transform ordinal predictors into interval-scaled variables. The simulation setting includes a non-monotonic ordinal predictor, therefore the results of the most restrictive constrained approach 'CMLE MDC S3' is the only one that is not superior than the methods using scoring systems even for the largest sample size, $n = 1000$ (see Figure 6.6). The best constrained method is 'CMLE Conf. Reg.', which only requires $n \geq 200$ to show a significantly smaller MSPE than any method using scoring systems. In addition, 'CMLE MDC S1' and 'CMLE filtered', that are among the less restrictive

methods too, show a better performance than methods using scoring systems in terms of MSPE when $n \geq 200$.

Regarding the performance according to the MAPE, 'CMLE Conf. Reg.' remains as the best method for $n \geq 200$. Similarly, 'CMLE MDC S1' and 'CMLE filtered' are still statistically better than every other method using scoring systems when $n \geq 500$, but not for $n = 200$ as in MSPE (see Figure 6.8).

In terms of misclassification rate, 'CMLE MDC S1', 'CMLE Conf. Reg.' and 'CMLE filtered', which are constrained methods among the less restrictive ones, confirm their better performance when $n \geq 500$ (see Figure 6.9). However, when $n = 100$ or 200, they are not statistically different from others, and sometimes they are even worse. Given that the response variable is ordinal, the misclassification rate should not be considered as the only accuracy measure to assess the performance of a method, but rather it should be a complementary measure to MSPE, keeping in mind that the latter is not an accuracy method for ordinal responses either, but at least it takes into consideration the distance of each prediction error from the true value.

In general, for $n \geq 500$ the results of less restrictive constrained methods perform better than any other method using scoring systems. However, when $n = 100$ or 200 the performance of constrained methods is not significantly better than the one of methods using scoring systems, except for 'CMLE MDC S1', 'CMLE Conf. Reg.' and 'CMLE filtered' that are still better than methods using scoring systems in terms of MSPE when $n = 200$.

Finally, the proposed methods were used in a real data application in Section 6.4. Given that the unconstrained results showed patterns that were close to be considered as monotonic, the most restrictive method was chosen to be used, 'CMLE MDC S3', imposing monotonicity constraints on all of the OPs (see Figure 6.10). This real data application is a good example of a case when a constrained method improves interpretability compared to the unconstrained one, in particular, the interpretation of the constrained effects of *Income Quintile* and *Health Status*. The results of other constrained methods supported the choice of 'CMLE MDC S3' as the one to be used (see Section 6.4.1). In addition, although the use

of scoring systems to transform ordinal predictors is not considered here as a good practice because it overstate the information provided by the order of categories of OPs, the results of using 'CMLE MDC S3' were compared against methods using scoring systems for the treatment of ordinal predictors. Section 6.4.2 shows that the performance of 'CMLE MDC S3' is better than the one of 'UMLE' but worse than the one of some methods using scoring systems. However, 'CMLE MDC S3' is still a valid option based on the relative advantages that it offers, for instance, it does not transform the ordinal predictors, it estimates different effects for each category of the ordinal predictors providing more information about the association between the ordinal covariate and the response variable than the methods using interval-scaled transformations, and it also improves their interpretability.

Chapter 7

Concluding remarks

7.1 Contributions

A constrained regression model for an ordinal response with ordinal predictors is proposed in Chapter 2, which can involve other types of predictors. The information provided by the order of categories of the ordinal predictors is used appropriately for ordinal data, rather than ignoring it (treating categories as of nominal scale) or overstating it as interval-scaled.

Each set of parameters associated with an ordinal predictor's categories can be enforced to be monotonic. For those that are assumed to be monotonic, the monotonicity direction classification procedure is also proposed in Chapter 2. It decides automatically whether associations between ordinal predictors and the response variable are isotonic or antitonic, and it can also classify variables as compatible with both monotonicity directions or none. The researcher may sometimes prefer to leave out variables compatible with both directions and statistically not significant parameters, and to drop the monotonicity constraint for variables incompatible with either direction, which can easily be done within the framework presented here.

The MDC relies on the choice of a pre-specified range of confidence levels between $\tilde{c}_s'^*$ and $\tilde{c}_s''^*$, but the regression model itself does not require a tuning parameter and does deliver monotonic parameter estimates.

In Chapter 3, the contribution is the development of asymptotic theory for the

constrained MLE of the POCLM. Asymptotic existence and strong consistency of the unconstrained MLE of the POCLM are analysed in detail, starting from the analysis of the corresponding unconstrained results in Fahrmeir and Kaufmann (1985). Asymptotic normality is also discussed. All of these properties of the constrained MLE allow to find that, under monotonic effects of ordinal predictors, the unconstrained and constrained MLEs are asymptotically equivalent. Consequently, the approximate confidence region for the constrained parameters is asymptotically the same as the one for the unconstrained ones. However, for finite n, there are some situations in which there is some doubt about the quality of the approximation. These situations are classified in three cases and discussed. Another contribution is the definition of different confidence regions, which are compared through simulations. Similarly, asymptotic confidence intervals for the constrained and unconstrained MLE are also the same under the assumption that the parameters associated with the ordinal predictors of the POCLM are monotonic. For finite n, the problem of identifying a parameter value that belongs to a confidence interval and violates monotonicity is analysed. Given that the computation of confidence intervals is still of interest, then some possible definitions of confidence intervals are proposed, despite the fact that there is an identification problem for those parameter values that violate monotonicity.

In Chapter 4, two monotonicity tests are proposed to assess the validity of the monotonicity assumption for an ordinal predictor. One is based on the Bonferroni correction and the other on the analysis of confidence regions. The first checks whether the set of confidence intervals belonging to the parameters of an ordinal predictor is compatible with monotonicity or not. As this is based on the Bonferroni correction of confidence levels, it can be very conservative, and therefore a more powerful tests was also developed, which is the second monotonicity test. It uses the confidence region of the unconstrained parameters associated with an ordinal predictor to assess whether monotonic parameter vectors resulting from the constrained MLE belong to the confidence region or not. It is shown that the monotonicity test based on confidence regions is invariant under change of base category and that equivalent results are obtained when using a reparametrised

model based on the differences between adjacent parameters associated with an ordinal predictor.

Six different approaches for the estimation method are proposed depending on whether the researcher wishes to impose monotonicity constraints on all of the OPs or on a subset of them. In the first case, the MDC procedure proposed in Chapter 2 is fully applied ('CMLE MDC S3'). Otherwise, the five remaining approaches differ in the way they identify the subset of OPs on which the monotonicity assumption is not imposed. These methods are proposed in Chapter 5. 'CMLE MDC S1' imposes monotonicity constraints only in step 1 of the MDC procedure and gives variables the biggest chance to be classified as either 'none' or 'both'. 'CMLE MDC S2' will re-classify some of these variables as monotonic. 'CMLE MDC S3' will impose monotonicity on all OPs. 'CMLE Bonferroni' uses the monotonicity test based on the Bonferroni correction for the decision of dropping constraints, whereas 'CMLE Conf. Reg.' uses the monotonicity test based on the confidence region to make this decision. 'CMLE filtered' will enforce monotonicity except if the MDC gives a strong indication against it. This happens somewhat earlier than under 'CMLE Bonferroni'. Due to the conservativeness of the Bonferroni test, its main use is to provide a test with a guaranteed low type I error probability, whereas the other methods are probably more appropriate for classification in connection with parameter estimation. In practice, the researcher will need to decide whether monotonicity should be always enforced ('CMLE MDC S3'), whether there is a clear preference to impose monotonicity except if there is a clear indication against it ('CMLE filtered', 'CMLE Bonferroni' or 'CMLE Conf. Reg.' in case that the significance level needs to be guaranteed), or whether it is fine to drop monotonicity constraints more easily in case of doubt ('CMLE MDC S1'), possibly together with dropping variables completely that are classified as 'both'; 'CMLE MDC S2' is a compromise that will probably not play much of a role in practice but was analysed here because it adds insight in the overall procedure.

The proposed approaches offer the researcher alternatives as a response to various legitimate interests. The researcher may be in the first place interested in the precision of the resulting estimates. However, in many applications, e.g., in

social sciences, the precise numerical values can be of less interest than qualitative statements about the monotonicity of the OPs. Monotonicity may be favoured because of better interpretability in some cases in which OPs are by and large approximately monotonic even if the true parameters show a mild deviation from monotonicity. If sample sizes are small, monotonicity may be favoured because constraints can support both precision and interpretation. However, in this case the researcher cannot expect a strong power to detect non-monotonicity, and there is always the risk that non-monotonic OPs are treated as monotonic, with loss of precision. In some instances, particularly with small sample sizes and a relatively high number of categories of the OPs, the researcher may prefer making decisions about monotonicity based on the meaning of the OPs rather than in a data driven manner. In addition, a large number of categories p_s for an OP will imply that the Bonferroni test is very conservative and a large number of observations may be required to detect moderate deviations from monotonicity. It may be reasonable in such a case to pool some categories and to make statements about monotonicity at lower "granularity" with better power.

In Chapter 6, a set of simulation studies and a real data application are analysed with the purpose of comparing the performance of the constrained methods against the one of the unconstrained POCLM with ordinal predictors treated as of either nominal or interval scale types through the use of scoring systems. The constrained methods are better than the UMLE of the POCLM with ordinal predictors treated as of nominal scale type when associations between the OPs and the response are truly monotonic, in which case the more restrictive the better. On the other hand, if there is a truly non-monotonic association, the most restrictive method 'CMLE MDC S3' could be bad depending on the sample size (e.g., for $n \geq 500$), whereas the other constrained methods are good options, from which the researcher can choose according to its degree of conservativeness when establishing non-monotonic effects, with 'CMLE Bonferroni' and possibly 'CMLE filtered' being the more conservative ones. In addition, the constrained methods perform better than the UMLE when $n = 50$, despite the fact that their misclassification rate increases as n decreases and that they drop the monotonicity

constraints less frequently (or never).

Regarding the comparison between the constrained methods and the UMLE of the POCLM with ordinal predictors treated as of interval scale type, it is hard to make a comparison because there is no measure of accuracy specifically designed for ordinal responses. However, as in Tutz and Hechenbichler (2005), the misclassification rate (MR), the mean absolute prediction error (MAPE), and the mean-squared prediction error (MSPE) are used. The simulation results show that when $n \geq 500$ the constrained methods perform better than the others in terms of MR, MAPE and MSPE, except for the more restrictive ones because the simulation setting considers four ordinal predictors of which one represents both monotonicity directions and another is non-monotonic. When $n = 100$ or 200, the results are mixed and in general against the constrained methods for $n = 100$, except for 'CMLE Conf. Reg.', which is in general not significantly different from methods using scoring systems but much better when $n \geq 500$.

For the real data application, 'CMLE MDC S3' enabled a consistent interpretation for the ordinal variables' categories, which would not have been the case for the UMLE.

The approaches of imposing monotonicity constraints on ordinal predictors allowing for both isotonic and antitonic patterns described in Section 5.3 can also be used in situations in which the response variable is non-ordinal. In addition, the MDC procedure itself can be performed on an ordinal predictor in models for responses of any scale of measurement, as well as the monotonicity tests. An `R` package `crov` was made available at CRAN with all of the proposed constrained methods (see Chapter 2 and Chapter 5) and monotonicity tests (see Chapter 4).

7.2 Future work

The performance of the approximation of confidence regions for finite n based on the analysis of coverage probabilities is studied in Section 3.7.1. The coverage probabilities are computed for two types of cases according to the comparison between the constrained and unconstrained MLE. These two cases correspond to whether the MLE are the same or different. When they are different (Case 4 in

Section 3.7.1), the UMLE does not belong to the constrained parameter space, and therefore it is unclear why the approximation resulting from the asymptotic theory developed in Chapter 3.7.1 should be good, meaning that there is no strong theoretical argument in using confidence regions for the constrained MLE. This case is analysed in Section 3.7.1. However, when the constrained and unconstrained MLE are the same, the finite n situation may also correspond to the one on which the asymptotic theory of Chapter 3 was developed (Case 1 in Section 3.7.1) or not (Case 2 and 3 in Section 3.7.1). Despite the fact that the situation "Same MLE" was already analysed as a whole, the analysis of its three sub-cases, that are already defined as Case 1, 2 and 3 in Section 3.7.1, is left for future work. This requires to identify whether the confidence region is either fully in the constrained parameter space or contains not monotonic parameter vectors or allows more than one possible combination of monotonicity directions for the constrained parameters of the ordinal predictors.

Regarding monotonicity tests, given that to my knowledge there is no monotonicity test for ordinal data in regression analysis, then one extension is to use scoring systems to transform ordinal variables into interval-scaled variables and then apply an existing monotonicity tests, such as the one proposed by van Beek and Daniels (2014). This would give some insights about the association between the transformed ordinal variable and the response variable. The results could be compared against the ones without using scoring systems, i.e., using the monotonicity test proposed in Chapter 4 on the POCLM with ordinal variables treated as ordinal.

Another possible extension is about using the MDC procedure as a tool for variable selection. When the MDC procedure classifies the pattern of parameter estimates of an OP as 'both' at the end of its second step, it means that all of the corresponding CIs contain zero, and therefore the OP as a whole can be considered as not significant. This could be explored further by comparing this approach against more formal tests, such as the log-likelihood test. In addition, when more than one OP is classified as 'both', dropping those variables should be carried out in a stepwise fashion, which could also be analysed in depth.

Another future work arose when comparing models with different treatment of ordinal predictors. In Section 6.3 the POCLM using constrained MLE versus the POCLM using unconstrained MLE and transformations of ordinal predictors into interval-scaled variables are compared. Different scoring systems were used for the second set of models. The comparison of the performance among models was based on the quality of their predictions, for which measures of accuracy were required. However, to my knowledge, there is no specific definition of a measure of accuracy for ordinal responses, and therefore the development of it is left for future work. Next, some ideas about how to address this issue are presented.

Consider an ordinal response with k categories. Compute the distribution of predicted categories for every given observed ordinal category. Therefore, there are k distributions with k categories each. For each one of these distributions there is one category being the true one and the others represent misclassification. In order to obtain a single measure of accuracy, it is required to aggregate them.

Imposing monotonicity constraints as a treatment of ordinal predictors can be extended from the POCLM to any other regression model with any other type of response variable.

Appendix A

Partial derivatives

Recall the Lagrangian presented in equation 2.3.11,

$$\mathcal{L}(\{\alpha_j\}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \ell(\{\alpha_j\}, \boldsymbol{\beta}) - \boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\beta}_{(ord)}, \tag{A.0.1}$$

where the log-likelihood function for the model (2.3.7) can be expressed as

$$\begin{aligned}\ell(\{\alpha_j\}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \Bigg[& \frac{e^{\alpha_j + \sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\beta_{s,h_s}x_{i,s,h_s} + \sum_{u=1}^{v}\beta_u x_{i,u}}}{1 + e^{\alpha_j + \sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\beta_{s,h_s}x_{i,s,h_s} + \sum_{u=1}^{v}\beta_u x_{i,u}}} \\ & - \frac{e^{\alpha_{j-1} + \sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\beta_{s,h_s}x_{i,s,h_s} + \sum_{u=1}^{v}\beta_u x_{i,u}}}{1 + e^{\alpha_{j-1} + \sum_{s=1}^{t}\sum_{h_s=2}^{p_s}\beta_{s,h_s}x_{i,s,h_s} + \sum_{u=1}^{v}\beta_u x_{i,u}}} \Bigg] \\ = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log & \left[\frac{e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}}{1 + e^{\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i}} \right].\end{aligned}$$

Now, for notation purposes, consider

$$b_{i,j} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i, \quad G(b_{i,j}) = \frac{e^{b_{i,j}}}{1 + e^{b_{i,j}}}, \text{ and } g(b_{i,j}) = \frac{e^{b_{i,j}}}{(1 + e^{b_{i,j}})^2}.$$

In addition, the elements of $\mathbf{C}$ are denoted as $c_{s,row,column}$, where s specifies the submatrix and the remaining subindexes represent the number of row and column within the pre-specified $\mathbf{C}_s$ as usual.

After differentiation, the likelihood equation for an effect parameter β_l, with l being an index representing either the dummy variable corresponding to an ordinal predictor and its category in the form (s, h_s) or a non-ordinal covariate in the form

(u), is

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_l} &= \Bigg\{ \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \Bigg(\frac{x_{i,l}e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}) - x_{i,l}e^{2(\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i)}}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})^2} \\
&\quad - \frac{x_{i,l}e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}) - x_{i,l}e^{2(\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i)}}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})^2} \Bigg) \\
&\quad \Big/ \left(\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}} \right) \Bigg\} \\
&\quad - \sum_{s=1}^{t}\sum_{h_s=2}^{p_s} (\delta_{(s,h_s),l} c_{s,h_s-1,h_s-1}\lambda_{s,h_s} + \delta_{(s,h_s-1),l} c_{s,h_s-1,h_s-2}\lambda_{s,h_s}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l} \left(\frac{\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})^2} - \frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})^2}}{\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}} \right) \\
&\quad - \sum_{s=1}^{t}\sum_{h_s=2}^{p_s} (\delta_{(s,h_s),l} c_{s,h_s-1,h_s-1} + \delta_{(s,h_s-1),l} c_{s,h_s-1,h_s-2})\lambda_{s,h_s} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l} \left(\frac{g(b_{i,j}) - g(b_{i,j-1})}{G(b_{i,j-1}) - G(b_{i,j-1})} \right) \\
&\quad - \sum_{s=1}^{t}\sum_{h_s=2}^{p_s} (\delta_{(s,h_s),l} c_{s,h_s-1,h_s-1} + \delta_{(s,h_s-1),l} c_{s,h_s-1,h_s-2})\lambda_{s,h_s} = 0,
\end{aligned}$$

with $\delta_{(s,h_s),l} = 1$ if the index $l = (s, h_s)$ and 0 otherwise.

Similarly, for the intercepts α_l,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_l} &= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \Bigg\{ \Bigg(\frac{\delta_{j,l}e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}) - \delta_{j,l}e^{2(\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i)}}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})^2} \\
&\quad - \frac{\delta_{j-1,l}e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}) - \delta_{j-1,l}e^{2(\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i)}}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})^2} \Bigg) \\
&\quad \Big/ \left(\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}} \right) \Bigg\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \left(\frac{\delta_{j,l}\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})^2} - \delta_{j-1,l}\frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})^2}}{\frac{e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}} - \frac{e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}{1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}}} \right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \frac{\delta_{j,l}g(b_{i,j}) - \delta_{j-1,l}g(b_{i,j-1})}{G(b_{i,j}) - G(b_{i,j-1})} = 0.
\end{aligned}$$

The first derivatives of $\mathcal{L}$ with respect to the Lagrange multipliers λ_l are

$$\frac{\partial \mathcal{L}}{\partial \lambda_l} = -\sum_{s=1}^{t}\sum_{h_s=2}^{p_s}(c_{s,h_s-1,h_s-1}\beta_{s,h_s} + c_{s,h_s-1,h_s-2}\beta_{s,h_s-1})\delta_{(s,h_s),l} = 0.$$

Regarding the second partial derivatives,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \beta_l \partial \beta_m} &= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l}\left\{ \frac{\left[\frac{\partial}{\partial \beta_m}g(b_{i,j}) - \frac{\partial}{\partial \beta_m}g(b_{i,j-1})\right][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\left. - x_{i,m}\frac{[g(b_{i,j}) - g(b_{i,j-1})]^2}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l}\left\{ \frac{[x_{i,m}g(b_{i,j-1})\frac{(e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})} - x_{i,m}g(b_{i,j})\frac{(e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})}][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\left. - x_{i,m}\frac{[g(b_{i,j}) - g(b_{i,j-1})]^2}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l}x_{i,m}\left\{ \frac{[g(b_{i,j-1})\frac{(e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})} - g(b_{i,j})\frac{(e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})}][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\left. - \frac{[g(b_{i,j-1}) - g(b_{i,j})]^2}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\}.
\end{aligned}$$

Now, for $\frac{\partial^2 \mathcal{L}}{\partial \beta_l \partial \alpha_m}$, we have:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \beta_l \partial \alpha_m} &= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l}\left\{ \frac{\left[\frac{\partial}{\partial \alpha_m}g(b_{i,j}) - \frac{\partial}{\partial \alpha_m}g(b_{i,j-1})\right][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\left. - \frac{\left[\frac{\partial}{\partial \alpha_m}G(b_{i,j}) - \frac{\partial}{\partial \alpha_m}G(b_{i,j-1})\right][g(b_{i,j}) - g(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}x_{i,l}\left\{ \frac{[g(b_{i,j})\delta_{jm} - g(b_{i,j-1})\delta_{j-1,m}][g(b_{i,j-1}) - g(b_{i,j})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\left. + \frac{\left[g(b_{i,j})\delta_{jm}\frac{(e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})} - g(b_{i,j-1})\delta_{j-1,m}\frac{(e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})}\right][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\}
\end{aligned}$$

To continue, the second partial derivative of $\mathcal{L}$ with respect to β_l and λ_m is

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_l \partial \lambda_m} = -\sum_{h_s=1}^{p_s}(\delta_{(s,h_s),l}c_{s,h_s-1,h_s-1} + \delta_{(s,h_s-1),l}c_{s,h_s-1,h_s-2})\delta_{(s,h_s),m}.$$

The corresponding result for $\frac{\partial^2 \mathcal{L}}{\partial \alpha_l \partial \alpha_m}$ is

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \alpha_l \partial \alpha_m} &= \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\left\{ \frac{\left[\frac{\partial}{\partial \alpha_m}\left(\delta_{j,l} g(b_{i,j})\right) - \frac{\partial}{\partial \alpha_m}\left(\delta_{j-1,l} g(b_{i,j-1})\right)\right][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\quad \left. - \frac{\left[\frac{\partial}{\partial \alpha_m} G(b_{i,j}) - \frac{\partial}{\partial \alpha_m} g(b_{i,j-1})\right][\delta_{j,l} g(b_{i,j}) - \delta_{j-1,l} g(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} \\
&\quad y_{ij}\left\{ \frac{\left[\delta_{j-1,l} g(b_{i,j-1})\delta_{j-1,m} \frac{(e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_{j-1}+\boldsymbol{\beta}'\mathbf{x}_i})} - \delta_{j,l} g(b_{i,j})\delta_{jm} \frac{(e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i}-1)}{(1+e^{\alpha_j+\boldsymbol{\beta}'\mathbf{x}_i})}\right][G(b_{i,j}) - G(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right. \\
&\quad \left. - \frac{[g(b_{i,j})\delta_{jm} - g(b_{i,j-1})\delta_{j-1,m}][\delta_{j,l} g(b_{i,j}) - \delta_{j-1,l} g(b_{i,j-1})]}{[G(b_{i,j}) - G(b_{i,j-1})]^2} \right\}.
\end{aligned}$$

Finally,

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_l \partial \lambda_m} = 0,$$

and,

$$\frac{\partial^2 \mathcal{L}}{\partial \lambda_l \partial \lambda_m} = 0.$$

Appendix B

Reproducibility of real data for QoL self-assessment in Chile

In Section 6.4, a real data application to quality of life self assessment in Chile was discussed. In order to assist reproducibility, the criteria used in the data preprocessing stage to get the final data set from the raw data that is publicly available online at: `http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php` is given.

B.1 Response variable and sample definition

The response variable is a self assessment of the quality of life (QoL), the name of this variable in the original data set is `r20` and its possible values are integers from 1 to 10, representing the possible answers: '1 Completely Unsatisfied', '2',..., '9', '10 Completely Satisfied' correspondingly.

The sample is defined as those householders who live in the capital and reported the quality of life self assessment. In the original data set, householders are identified with the value 1 of the variable `pco`, whereas the capital corresponds to `region=13` and the valid responses of QoL lie between 1 and 10.

B.2 Predictors

B.2.1 Ordinal predictors

Educational Level: This variable takes into account the educational level, years of schooling, and whether the householder knows how to read and/or write. All these variables are treated in the following sequential steps:

1. Variable `educ` is grouped into four categories: values 0 and 99 in "Not educated", values 1 and 2 in "Primary", values from 3 to 6 in "Secondary", and from 7 to 12 in "Higher".
2. Those classified as "Secondary" are moved to "Primary" if their years of schooling are less than 9 (variable `ESC`>9).
3. Those classified as "Not educated" and with `educ`=99 are moved to "Primary" if their years of schooling are more than 0 (variable `ESC`>0).
4. Those classified as "Primary" and with `educ`=99 are moved to "Secondary" if their years of schooling are more than 8 (variable `ESC`>8).
5. Those classified as "Secondary" and with `educ`=99 are moved to "Higher" if their years of schooling are more than 12 (variable `ESC`>8).
6. Those classified as "Secondary" or "Higher", and declared that they do not know how to read and/or write (variable `e1` is 2, 3, or 4), are moved to "Primary".
7. Those classified as "Not educated" and with `educ`=99 (value for 'do not know/do not answer') are removed from the sample (28 cases, 0.37%).

Income Quintile: Raw variable `QAUTR_MN` is used.

Health Status: Variable `s16` is used. Values from 1 to 7 are considered only and those observations with value 99 (value for 'do not know') are removed from the sample (36 cases, 0.48%).

Overcrowding: Variable `hacinamiento` is used. Values from 1 to 4 are considered only and those observations with value 9 (value for 'NA') are removed

from the sample (21 cases, 0.28%).

Children: This is a special case, we use the whole data set and a dummy variable to identify those people under 15 years old. Then grouping by the house identifier called `folio` we get the number of children by house. We incorporate this information back in the sample of householders living in the capital and reported the quality of life self assessment. Finally, we grouped the number of children when it was greater than or equal to 4.

B.2.2 Non-Ordinal predictors

None of the non-ordinal predictors was transformed. The name of variables "Age", "Activity", and "Gender" are `edad`, `activ`, and `sexo`, correspondingly.

Bibliography

Alan Agresti. An introduction to categorical data analysis, 2nd edn. hoboken, 2007.

Alan Agresti. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.

Javier Alvarez-Galvez, Maria Rodero-Cosano, Emma Motrico, Jose Salinas-Perez, Carlos Garcia-Alonso, and Luis Salvador-Carulla. The impact of socio-economic status on self-rated health: Study of 29 countries using european social surveys (2002–2008). *International journal of environmental research and public health*, 10(3):747–761, 2013.

Silvia Bacci, Francesco Bartolucci, and Michela Gnaldi. A class of multidimensional latent class irt models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods*, 43(4):787–800, 2014.

RE Barlow and HD Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, 2011.

DJ Bartholomew. Latent variable models for ordered categorical data. *Journal of Econometrics*, 22(1-2):229–243, 1983.

Rudolf Beran and Lutz Dümbgen. Least squares and shrinkage estimation under bimonotonicity constraints. *Statistics and computing*, 20(2):177–189, 2010.

Gunnar Blom. *Statistical estimates and transformed beta-variables*. PhD thesis, Almqvist & Wiksell, 1958.

Stefan Boes and Rainer Winkelmann. The effect of income on general life satisfaction and dissatisfaction. *Social Indicators Research*, 95(1):111–128, 2010.

C Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8: 3–62, 1936.

AW Bowman, MC Jones, and Irene Gijbels. Testing monotonicity of regression. *Journal of computational and Graphical Statistics*, 7(4):489–500, 1998.

Patrick L Brockett. A note on the numerical assignment of scores to ranked categorical data. *Journal ofMathematical Sociology*, 8(1):91–110, 1981.

Irwin DJ Bross. How to use ridit analysis. *Biometrics*, pages 18–38, 1958.

Oleg Burdakov, Anders Grimvall, and Mohamed Hussian. A generalised pav algorithm for monotonic regression in several variables. In *COMPSTAT, Proceedings of the 16th Symposium in Computational Statistics*, volume 10, pages 761–767, 2004.

Sara Casacci and Adriano Pareto. Methods for quantifying ordinal variables: a comparative study. *Quality & Quantity*, 49(5):1859–1872, 2015.

Denis Chetverikov. Testing regression monotonicity in econometric models. *Econometric Theory*, 35(4):729–776, 2019.

Felix Cheung and Richard E Lucas. Assessing the validity of single-item life satisfaction measures: results from three large samples. *Quality of Life research*, 23(10):2809–2818, 2014.

Sarah D Corathers, Jessica C Kichler, Nora F Fino, Wei Lang, Jean M Lawrence, Jennifer K Raymond, Joyce P Yi-Frazier, Dana Dabelea, Angela D Liese, Sharon H Saydah, et al. High health satisfaction among emerging adults with diabetes: Factors predicting resilience. *Health Psychology*, 36(3):206, 2017.

Jan De Leeuw, Forrest W Young, and Yoshio Takane. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):471–503, 1976.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

Jan De Leeuw, Patrick Mair, et al. Gifi methods for optimal scaling in r: The package homals. *Journal of Statistical Software*, 31(4):1–20, 2009.

Rafael Di Tella, Robert J MacCulloch, and Andrew J Oswald. The macroeconomics of happiness. *The review of Economics and Statistics*, 85(4):809–827, 2003.

E Doveh, A Shapiro, and PD Feigin. Testing of monotonicity in parametric regression models. *Journal of Statistical Planning and Inference*, 107(1-2):289–306, 2002.

Cécile Durot. A kolmogorov-type test for monotonicity of regression. *Statistics & probability letters*, 63(4):425–433, 2003.

Richard L Dykstra, Tim Robertson, et al. An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 10(3):708–716, 1982.

Allen Louis Edwards and Kathryn Claire Kenney. A comparison of the thurstone and likert techniques of attitude scale construction. *Journal of Applied Psychology*, 30(1):72, 1946.

Javier Espinosa and Christian Hennig. A constrained regression model for an ordinal response with ordinal predictors. *Statistics and Computing*, 29(5):869–890, 2019.

Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of

the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985.

Ludwig Fahrmeir and Heinz Kaufmann. Asymptotic inference in discrete response models. *Statistische Hefte*, 27(1):179–205, 1986.

David A Frederick and Brooke N Jenkins. Height and body mass on the mating market: Associations with number of sex partners and extra-pair sex among heterosexual men and women aged 18–65. *Evolutionary Psychology*, 13(3): 1474704915604563, 2015.

John Gaito. Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 1980.

Subhashis Ghosal, Arusharka Sen, and Aad W Van Der Vaart. Testing monotonicity of regression. *Annals of statistics*, pages 1054–1082, 2000.

Irene Gijbels, Peter Hall, MC Jones, and Inge Koch. Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, 87(3):663–673, 2000.

Linda L Golden and Patrick L Brockett. The effect of alternative scoring methods on the analysis of rank order categorical data. *Journal of Mathematical Sociology*, 12(4):383–414, 1987.

Rod Haggarty. *Fundamentals of mathematical analysis*. Addison-Wesley New York, 4th edition, 1993.

Peter Hall and Nancy E Heckman. Testing for monotonicity of a regression mean by calibrating for linear functions. *Annals of Statistics*, pages 20–39, 2000.

H Leon Harter. Expected values of normal order statistics. *Biometrika*, 48(1/2): 151–165, 1961.

Arne Henningsen and Ott Toomet. maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011.

Carl Hensler and Brian Stipak. Estimating interval scale values for survey item response categories. *American Journal of Political Science*, 23(3):627–649, 1979.

Darko Hren, Ivan Krešimir Lukić, Ana Marušić, Ivana Vodopivec, Ana Vujaklija, Maja Hrabak, and Matko Marušić. Teaching research methodology in medical schools: students' attitudes towards and knowledge about science. *Medical education*, 38(1):81–86, 2004.

Johs Ipsen and Niels K Jerne. Graphical evaluation of the distribution of small experimental series. *APMIS*, 21(2):343–361, 1944.

William G Jacoby. opscale: A function for optimal scaling. *The R Journal*, 10, 2015.

Wim Kalmijn. From discrete 1 to 10 towards continuous 0 to 10: The continuum approach to estimating the distribution of happiness in a nation. *Social indicators research*, 110(2):549–557, 2013.

H Kaufmann. On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models. *Metrika*, 35(1):291–313, 1988.

Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

Sanford Labovitz. Some observations on measurement and statistics. *Social Forces*, 46(2):151–160, 1967.

Sanford Labovitz. The assignment of numbers to rank order categories. *American Sociological Review*, 35(3):515–524, 1970.

Sanford Labovitz. In defense of assigning numbers to ranks. *American Sociological Review*, 36(3):521–522, 1971.

M Lanfranchi, C Giannetto, and A Zirilli. Analysis of demand determinants of high quality food products through the application of the cumulative proportional odds model. *Applied mathematical sciences*, 8(65-68):3297–3305, 2014.

Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, third edition edition, 2005.

Erich Leo Lehmann and HJ D'abrera. *Nonparametrics: statistical methods based on ranks*. Holden-Day, 1975.

Mariëlle Linting and Anita van der Kooij. Nonlinear principal components analysis with catpca: a tutorial. *Journal of personality assessment*, 94(1):12–25, 2012.

Frederic M Lord. On the statistical treatment of football numbers. *Scaling*, page 402, 1953.

Patrick Mair and Jan de Leeuw. Rank and set restrictions for homogeneity analysis in r: The" homals" package. In *JSM 2008 Proceedings, Statistical Computing Section*. American Statistical Association, Alexandria, 2008.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press Inc, first edition edition, 1979.

Nick Martin and Hermine Maes. *Multivariate analysis*. Academic press, 1979.

Lawrence Mayer. Comment on "the assignment of numbers to rank order categories". *American Sociological Review*, 35(5):916–917, 1970.

Peter McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.

Peter McCullagh and James A Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.

Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

Rupert G Miller. *Simultaneous statistical inference*. Springer, 1981.

Yuichi Mori, Masahiro Kuroda, and Naomichi Makino. *Nonlinear Principal Component Analysis and Its Applications*. Springer, 2016.

Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

Irini Moustaki. A latent variable model for ordinal variables. *Applied psychological measurement*, 24(3):211–223, 2000.

Irini Moustaki. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56(2):337–357, 2003.

David J Pasta. Learning when to be discrete: continuous vs. categorical predictors. In *SAS Global Forum*, volume 248, 2009.

Murray H Protter, B Charles Jr, et al. *A first course in real analysis*. Springer Science & Business Media, 2012.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

Dimitris Rizopoulos. ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25, 2006.

Walter Rudin et al. *Principles of mathematical analysis*. McGraw-hill New York, 3rd edition, 1976.

Kaspar Rufibach. An active set algorithm to estimate parameters in generalized linear models with ordered predictors. *Computational Statistics & Data Analysis*, 54(6):1442–1456, 2010.

Manuel Santina and Jorge Perez. Health professionals' sex and attitudes of health science students to health claims. *Medical education*, 37(6):509–513, 2003.

Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103 (2684):677–680, 1946.

Quentin F Stout. Isotonic regression for multiple independent variables. *Algorithmica*, 71(2):450–470, 2015.

John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

Gerhard Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43(1):39–55, 1990.

Gerhard Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11(3):275–295, 1991.

Gerhard Tutz and Jan Gertheiss. Rating scales as predictors - the old question of scale level and some answers. *Psychometrika*, 79(3):357–376, 2014.

Gerhard Tutz and Klaus Hechenbichler. Aggregating classifiers with ordinal response structure. *Journal of Statistical Computation and Simulation*, 75(5): 391–408, 2005.

Misha van Beek and Hennie AM Daniels. A non-parametric test for partial monotonicity in multiple regression. *Computational Economics*, 44(1):87–100, 2014.

BL Van der Waerden. Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)*, volume 55, pages 453–458. Elsevier, 1952.

Louis Vargo. Comment on "the assignment of numbers to rank order categories". *American Sociological Review*, 36(3):517–518, 1971.

Vassilis GS Vasdekis, Silvia Cagnone, and Irini Moustaki. A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika*, 77(3):425–441, 2012.

Ruut Veenhoven, Joop Ehrhardt, Monica Sie Dhian Ho, and Astrid de Vries. *Happiness in nations: Subjective appreciation of life in 56 nations 1946–1992.* Erasmus University Rotterdam, 1993.

Paul F Velleman and Leland Wilkinson. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1):65–72, 1993.

RWM Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32, 1976.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Chien-Fu Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, pages 501–513, 1981.

Thomas W. Yee. *VGAM: Vector generalized linear and additive models*, 2018. URL `https://CRAN.R-project.org/package=VGAM`. R package version 1.0-5.

Forrest W Young. Quantitative analysis of qualitative data. *Psychometrika*, 46 (4):357–388, 1981.

Forrest W Young, Jan De Leeuw, and Yoshio Takane. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):505–529, 1976.

G Alastair Young and Richard L Smith. *Essentials of statistical inference*, volume 16. Cambridge University Press, 2005.