About the link function in generalized linear models for categorical responses

Yinneth Leon Velasco

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Biostatistique

École doctorale I2S – Information, Structures, Systèmes

Unité de recherche UMR 5149 – IMAG – Institut Montpelliérain Alexander Grothendieck

et CIRAD - AGAP

About the link function in generalized linear models for categorical responses

Présentée par Yinneth Lorena León Velasco

Le 07/06/2022

Sous la direction de Catherine Trottier
et Jean Peyhardi

Devant le jury composé de

JACQUES Julien, Professeur, Université Lumière Lyon 2	Rapporteur
ROBIN Stéphane, Professeur, Sorbonne Université	Rapporteur
GIORDANO Sabrina, Associate Professeur, Università della Calabria	Examinatrice
BACRO Jean-Noël, Professeur, Université de Montpellier	Président du jury
TROTTIER Catherine, Maître de Conférences, Université Paul Valéry - Montpellier 3	Directrice
PEYHARDI Jean, Maître de Conférences, Université Montpellier	Encadrant
RIVALS Éric, Directeur de recherche, CNRS	Invité
JOUANNIC Stéphane, Chargé de recherche, IRD	Invité

UNIVERSITÉ DE MONTPELLIER

Remerciements

Au terme de cette thèse, l'une des parties les plus agréables est d'exprimer ma gratitude. Tout d'abord, je veux remercier l'ensemble global, l'atmosphère quotidienne, et la totalité qui a été alignée et formulée jour après jour pour que ce contrefactuel que je suis aujourd'hui puisse arriver jusqu'à ce moment. Aujourd'hui, je suis le résultat d'une série de circonstances qui font que je ne suis rien d'autre que fortunée.

Ma plus profonde gratitude et mon appréciation professionnelle vont à mes excellents superviseurs, Catherine Trottier et Jean Peyhardi, pour leur encadrement et leur soutien à chaque étape de ce parcours scientifique. Pour avoir accepté le défi de poursuivre le projet de recherche malgré la perte d'un membre essentiel de notre équipe. Pour sa grande capacité à s'adapter à mes compétences linguistiques particulières, qui étaient un mélange de français, d'anglais et d'espagnol. Pour leur patience et leur gentillesse, pour avoir toujours été à mon écoute et pour tout le temps qu'ils ont pu me dédier, malgré toutes les responsabilités qu'ils avaient. Je voudrais également profiter de cette occasion pour honorer la mémoire de Yann Guedon, qui est malheureusement décédé 3 mois après le début de ma thèse, laissant la direction de ma thèse à Catherine et Jean.

Je remercie chaleureusement à messieurs Stéphane Robin et Julien Jacques pour le temps et les efforts consacrés à la lecture de ce manuscrit. Plus généralement, à tous les membres du jury, sachez que votre présence à ma soutenance est un grand honneur pour moi. Je remercie sincèrement Hellene Adam et Stéphane Jouannic pour avoir fourni le jeu de données sur la diversité du riz et pour leur soutien sur le contexte des données. Je tiens à exprimer ma gratitude à tous les membres de l'équipe PhenoMEn, dont l'accueil et la convivialité ont été essentiels pour moi. Au CIRAD, à la Région Occitanie, et à l'Université de Montpellier pour avoir sponsorisé ce travail de recherche.

A mis amigos (a quienes no me atrevo a nombrar por miedo a que se me escape algún nombre), gracias por aligerar los momentos de estrés y por el cariño y el apoyo que manifestaron de diferentes maneras.

Finalmente, es a mi familia a quien dirijo estos últimos agradecimientos. A mí papá, que me enseña a diario lo que es ser fuerte, a quien le costó trabajo imaginarme sana y salva en otro lugar diferente de casa, a quien un día dije que me iba solo un mes fuera de Colombia y hoy ya son casi 10 años, gracias por aceptarlo, aunque no entendieras mi plan, gracias por creer en mí, gracias por ser una de mis grandes fuentes de fortaleza. A mi hermano y a mi hermana por ser ejemplos de motivación, por el rol parental que tuvieron conmigo, por darme el amor más puro que existe, muchas gracias por todo el amor que me dan como su hermanita menor. A mi madre, por su fortaleza, por creer a ojo cerrado en mí, por sus sacrificios y por su amor inagotable e infinito. A Yuyu, Kami, Zoe, y Blanca, mis fuentes de inspiración diaria. A Victor, qui est venu dans la

dernière ligne droite de cette thèse pour me soutenir et me détendre en me montrant la France comme je ne l'avais jamais vue. Y a Germán, gracias por tanto.

Je sais que, malgré mes meilleures intentions, je ne peux pas reconnaître toutes les personnes qui ont rendu cette thèse possible. Pour ceux qui ont été mentionnés ci-dessus, et pour tous ceux qui auraient dû l'être, je leur adresse un sincère merci.

Summary in french

Les données catégorielles sont fréquemment analysées dans différents domaines comme les sciences sociales, la biologie, la santé ou encore l'économétrie. On parle d'une variable catégorielle lorsque ses modalités forment un ensemble de catégories. Un tel ensemble peut être constitué de réponses à des questionnaires (diagnostics médicaux, options de choix), d'observations de certaines caractéristiques, de votes... De nombreuses variables catégorielles ne comportent que deux catégories (vrai ou faux, succès ou échec, existence ou non-existence). On parle alors de variables binaires. Avec plus de deux catégories, on distingue principalement deux types de variables. Les variables dont les catégories ne sont pas ordonnées sont appelées variables nominales (par exemple : les espèces dans un genre, le mode de transport). Les variables dont les catégories sont naturellement ordonnées sont appelées variables ordinales (par exemple : le niveau d'éducation, le statut d'une maladie, les préférences ou opinions).

Dans de nombreuses situations, pour analyser ces données catégorielles, on dispose d'informations complémentaires au travers d'autres variables que l'on souhaite relier aux proportions des catégories. Cette relation est souvent analysée en statistique par des modèles de régression. Au sein de ces modèles, les variables explicatives (également appelées variables indépendantes, covariables ou prédicteurs) participent à l'explication de la variable réponse (également appelée variable d'intérêt). Cette démarche de régression est largement répandue dans le cas d'une variable réponse gaussienne. Cependant, la loi gaussienne n'étant pas adaptée aux réponses catégorielles, la méthodologie a été étendue.

Les modèles linéaires généralisés (GLMs) ont été introduits par Nelder and Wedderburn (1972) pour relâcher le postulat d'une distribution gaussienne de la variable réponse, en particulier pour prendre en compte une réponse discrète (données de comptage ou catégorielles). Dans cette thèse, nous aborderons uniquement le cas des réponses catégorielles (où J dénote le nombre de catégories) dans trois contextes distincts :

- dans le cas d'une réponse binaire (chapitre 2),
- lorsque la réponse comporte plus de deux catégories (chapitre 3),
- lorsqu'il existe une structure hiérarchique entre les catégories (chapitre 4).

Dans chaque chapitre de cette thèse, nous analysons les différents modèles possibles en étudiant particulièrement le choix de la fonction de lien. Selon Nelder and Wedderburn (1972), un modèle linéaire généralisé est caractérisé par trois composantes : la distribution de la réponse, le prédicteur linéaire (combinaison linéaire des variables explicatives) et la fonction de lien. Dans le cas particulier d'une réponse catégorielle, la distribution est forcément la loi de Bernoulli (lorsque $J = 2$) ou la loi multinomiale (lorsque $J \geq 2$). Ainsi, seuls le prédicteur linéaire et la fonction de lien caractérisent ces modèles. Comme le prédicteur linéaire ne représente que les contraintes imposées sur les paramètres associés aux covariables, la fonction de lien devient la clé pour la définition des modèles de

régression catégorielle. Par conséquent, les différences fondamentales entre les modèles des trois contextes mentionnés ci-dessus sont attribuables au choix de leur fonction de lien. Notre travail de thèse se concentre ainsi sur l'étude de la fonction de lien dans les modèles de régression catégorielle.

Le chapitre 1 replace la problématique statistique au cœur d'une problématique biologique, et présente un état de l'art sur la fonction de lien dans les trois contextes. Nous commençons par présenter les données sur la diversité du riz qui ont motivé la réalisation de cette thèse. La variable réponse dans ces données est la classification taxonomique du riz. Elle est catégorielle, et en tenant compte des différents niveaux de classification, elle présente alors une structure hiérarchique des catégories. Les variables explicatives sont des caractéristiques phénotypiques hétérogènes qui permettent de décrire la structure du riz. Il est à noter que le paradigme classique de l'étude des phénotypes en fonction de l'information génétique a ici été renversé. Pour modéliser les données décrites ci-dessus, il est nécessaire d'examiner des modèles pour les réponses binaires, nominales et ordinales, ainsi que des modèles qui admettent une structure hiérarchique des catégories.

Dans ce chapitre, on rappelle tout d'abord la fonction de lien et l'estimation des modèles logit, probit et cauchit (entre autres) pour les réponses binaires. On s'intéresse à la question de la robustesse aux valeurs aberrantes, notamment à travers la fonction d'influence. Pregibon (1982); Copas (1988) ont alors mis en évidence une limite des modèles logit et probit, contrairement au modèle cauchit.

Ensuite, pour les variables ordinales, nous introduisons les modèles logit cumulatif (McCullagh, 1980), séquentiel (Tutz, 1991) et adjacent, tandis que pour les variables nominales, nous présentons le modèle logit multinomial (Luce, 1959). Puis nous rappelons une méthodologie qui unifie tous les modèles susmentionnés en décomposant la fonction de lien en deux parties : le ratio et la fonction de répartition (Peyhardi et al., 2015). On présente les procédures (R, SAS et STATA) couramment utilisées pour l'ajustement de ces modèles et nous soulignons l'absence d'une solution unique avec laquelle il soit possible d'ajuster la diversité des modèles pour réponse catégorielle.

Enfin, pour les réponses hiérarchiques, deux modèles logit sont présentés. Le modèle logit emboîté (McFadden et al., 1978) est défini dans le contexte des modèles de choix (basés sur la maximisation de l'utilité). Le modèle partitionné conditionnel (Zhang and Ip, 2012) est quant à lui défini comme un modèle linéaire généralisé, qui en plus de considérer les variables nominales et ordinales, convient également aux variables partiellement ordonnées. On en présente également une extension introduite par Peyhardi et al. (2016) qui permet d'utiliser un lien autre que le lien logistique à chaque niveau de la hiérarchie. Pour tous les cas précédents, il est supposé que l'arbre est connu à l'avance.

Les trois chapitres suivants présentent des propositions pour répondre aux différentes limites soulignées ci-dessus.

Dans le chapitre 2, nous considérons une variable réponse binaire, le cas le plus simple et le plus connu. Malgré l'existence d'un riche ensemble de modèles de régression binaire, dans la pratique seuls deux modèles sont largement utilisés : le modèle logit et le modèle

probit. La fonction de lien du premier est l'inverse de la fonction de répartition logistique, tandis que celle du deuxième est l'inverse de la fonction de répartition gaussienne. Comme les modèles probit et logit sont très proches, les chercheurs ont longtemps supposé qu'aucune autre fonction ne permettait d'améliorer notablement l'ajustement de ce modèle. Pourtant les modèles logit et probit sont réputés être sensibles aux perturbations des données contrairement au lien Student qui a été suggéré comme une alternative robuste. Afin d'utiliser le lien Student, nous proposons un algorithme d'estimation du degré de liberté. Ce chapitre étudie alors l'évolution de la robustesse en fonction des différents degrés de séparation (ou inversement de chevauchement) des deux classes de la réponse.

Une situation problématique en régression binaire est en effet celle de la séparation complète des données qui se produit lorsqu'une ou plusieurs variables explicatives prédisent parfaitement la réponse binaire. La séparation complète est un obstacle puisque l'estimation par maximum de vraisemblance n'existe pas de façon unique dans ce cas. Par conséquent, le chevauchement est nécessaire pour l'existence et l'unicité de cette estimation. Toutefois, un chevauchement important n'est pas requis. Dans ce travail, nous mettons en évidence qu'un faible chevauchement (c'est-à-dire proche de la séparation complète) permet d'obtenir simultanément l'existence et l'unicité des estimations du maximum de vraisemblance ainsi qu'une qualité remarquablement élevée du modèle (en termes d'ajustement et de classification). Ces bonnes performances peuvent être fortement dégradées par l'ajout de points aberrants et/ou de variables de bruit lorsque le modèle n'est pas robuste. Ainsi nous étudions le lien entre la robustesse du modèle et le degré de séparation des données. Dans ce contexte, nous montrons que le modèle Student est robuste, contrairement au modèle logit, surtout lorsque le chevauchement est faible. De plus, on constate que plus le degré de liberté est faible, plus le modèle Student est robuste. Nous obtenons ainsi un indicateur du degré de séparation des données puisque nous trouvons une forte association entre le degré de liberté et la robustesse du modèle Student. Sur la base de ces résultats, nous visons à promouvoir le lien Student dans le cadre des modèles de régression binaire. L'article résultant de cette recherche a été soumis à la revue *Computational Statistics & Data Analysis*.

Dans le chapitre 3, nous abordons la grande variété de modèles linéaires généralisés existant pour les réponses catégorielles (nominales et ordinales). On dispose aussi de plusieurs packages pour l'ajustement de réponses catégorielles (dans R : ordinal, VGAM, nnet, polr, entre autres). Ces modèles et logiciels ont été développés dans diverses disciplines et donc avec un manque d'uniformité, tant au niveau théorique que dans leur implémentation. Il arrive même que certains modèles soient identiques mais portent des noms différents car ils sont utilisés dans des contextes distincts. Par exemple, le *proportional odds model* est également connu sous le nom de *cumulative logit model*, ou tout simplement sous le nom de *ordinal model* (puisqu'il s'agit du modèle le plus populaire pour les réponses ordinales). En termes de solutions logicielles, la plupart d'entre eux ne couvrent qu'un ou quelques-uns des modèles disponibles pour les réponses catégorielles. Il n'existe donc pas de logiciel qui englobe tous ces modèles dans un cadre unique et générique. Dans ce troisième chapitre, nous répondons à ce problème en introduisant dans R le package GLMcat. Il permet d'estimer des modèles linéaires

généralisés pour réponse catégorielle implémentés sous la spécification unifiée (r, F, Z), où r représente le ratio de probabilités (référence, cumulatif, adjacent ou séquentiel), F la fonction de répartition pour le lien, et Z la matrice de design (Peyhardi et al., 2015). Notez que pour ces réponses non binaires, la fonction de lien est une composition des deux fonctions r et F. Tous les modèles classiques (et leurs variantes) pour les données catégorielles peuvent être écrits sous la forme d'un triplet (r, F, Z), et peuvent donc être estimés avec GLMcat. Cette spécification unifiée des modèles pour données catégorielles permet de souligner les propriétés de chacun ainsi que les équivalences entre certains d'entre eux. En outre, les extensions possibles pour chaque famille de modèles deviennent évidentes et peuvent être facilement implémentées.

On propose de plus un guide méthodologique et pratique pour la sélection appropriée d'un modèle (via la fonction de lien et les contraintes sur la matrice de design) en considérant la concordance entre la nature des données et les propriétés du modèle. Pour optimiser les performances des fonctions de GLMcat, la partie calculatoire de notre code a été écrite en C++ (intégrée au travers du package Rcpp). Les algorithmes sont implémentés de manière modulaire, ce qui signifie qu'une amélioration ou un ajustement peut être facilement étendu à toutes les familles de modèles. Le travail lié au développement de ce package est en révision dans la revue *Journal of Statistical Software*. Les applications de ce travail sont nombreuses tant les données catégorielles sont répandues. On espère que cet outil permettra de populariser l'utilisation des modèles de régression catégorielle différents des modèles logit.

Le chapitre 4 propose une méthodologie pour modéliser une structure hiérarchique parmi les J catégories de réponses. La fonction de lien est dans ce cas composée de l'arbre qui représente la structure hiérarchique et des modèles classiques à ajuster à chaque nœud de l'arbre. Pour obtenir ce modèle, deux tâches principales doivent être effectuées. La première est la définition de la structure hiérarchique des catégories de réponses. La deuxième consiste à trouver et à ajuster l'ensemble des modèles qui génèreront des informations spécifiques pour chacun des nœuds (non terminaux). Pour simplifier l'exécution de ces deux tâches, nous avons décidé de réduire l'espace des arbres en ne considérant que les arbres binaires qui décrivent les groupements par paires existant dans les différentes catégories de la réponse. En fait, la spécification du modèle est simplifiée pour le cas binaire puisque la caractérisation de ce modèle est donnée uniquement par F au lieu des trois composantes (r, F, Z). La fonction de lien pour ce cas est alors composée de 1) l'ensemble des fonctions de répartition de chacun des $J-1$ modèles binaires et 2) l'arbre lui-même.

Dans la plupart des applications, l'arbre de partition n'est pas connu *a priori* et l'ensemble des arbres de partitions à tester est évidemment très vaste. Dans notre contexte, nous ne considérons que les arbres de partition binaires qui sont équivalents aux dendrogrammes étiquetés et non ordonnés. Le nombre possible de ce type de dendrogrammes est connu et augmente exponentiellement en fonction du nombre de catégories J. Étant donné que l'inférence de tous les modèles serait trop longue, nous proposons une méthode de construction du dendrogramme basée sur les distances entre classes. Pour ce faire, on s'appuie sur l'algorithme de clustering hiérarchique ascendant dans lequel, au lieu de regrouper les individus, on considère comme points de départ les clus-

ters générés par les J catégories. Ensuite, nous réalisons une série de fusions successives (basées sur les distances entre les groupes) jusqu'à ce que toutes les catégories soient membres d'un seul groupe, la racine. Une fois l'arbre trouvé (ou connu), les autres éléments de la fonction de lien doivent être spécifiés. Différents degrés de séparation peuvent être observés dans les données binaires associées à chaque nœud de l'arbre (qui normalement dépendra de la profondeur du nœud). Par conséquent, nous utilisons les résultats obtenus au chapitre 2 pour sélectionner la distribution à utiliser comme fonction de lien dans les modèles binaires.

En partant du principe que les paramètres sont différents d'un nœud à l'autre, le modèle est facilement estimable puisque la log-vraisemblance du modèle total est alors égale à la somme des log-vraisemblances de chaque nœud non terminal. Cela offre une flexibilité que l'on ne trouve pas dans d'autres modèles. Bien que l'on constate que l'arbre proposé tend à avoir l'un des scores les plus élevés, une heuristique ne garantit pas de trouver la structure optimale parmi le vaste ensemble des possibilités. Par conséquent, dans la recherche d'un arbre avec le meilleur score, nous proposons deux algorithmes qui explorent l'espace des arbres voisins. Ces derniers ayant une structure de sous-arbre commune, nous utiliserons la propriété de décomposition de la log-vraisemblance afin d'éviter de la recalculer entièrement pour chaque nouveau voisin. À l'aide d'exemples numériques et de simulations, nous évaluons les performances de ce type de modélisation en terme de log-vraisemblance et de proportion de bien classés.

On terminera par présenter, dans le chapitre 5, les travaux en cours et les perspectives sur chacun des thèmes abordés dans cette thèse.

Contents

List of Figures

List of Tables

Chapter 1

State of the Art

Over the course of this chapter, we describe the state of the art of statistical modeling methods for categorical responses. The topic of this thesis emerged when determining the most appropriate statistical approach to explain the taxonomic classification of rice based on phenotypic differences. We begin this chapter by presenting the data set on rice diversity which motivates the statistical modeling methods to be treated in this thesis. We then introduce the GLMs framework for categorical responses with a particular emphasis on the link function. We describe the estimation and some characteristics of these models for both the multivariate case (more than two categories) and the univariate case (two categories). We present some of the most popular link functions for the binary models (such as the logit, probit, and cauchit), and we briefly compare the sensitivity to outliers of these models. As for the multivariate case, we present the logit models: multinomial, cumulative, adjacent, and sequential, as alternatives with different foundations. In addition, we detail a unifying modeling framework that allows us to define all of the above models and some extensions. This framework is based on a decomposition of the link function into two parts that largely shape the properties of the model. We also discuss the computational implementations of the above models and argue that none of these alternatives provides an integrated solution to fit all the possible models for categorical responses. Finally, we present three regression models for hierarchically structured responses. All of them assume that the structure is known in advance. These models share a split conditional structure suitable for different scales: nominal, ordinal, and partially ordered scales.

Contents

1.1 Context and data

Quantitative genetics relies on random-effects models whose traditional functional view is $phenotype = f(genotype)$ (Gianola, 2008). In this context, the modeling of phenotypic traits (response variable of the regression model) is strongly constrained, often a single trait or a vector of traits using (generalized) linear mixed models. On the other hand, the categorical genotype variable is often treated as a random-effect variable due to the high number of its levels. For the analysis of plant diversity, we are interested in an approach that reverses the functional view of the regression model, which results in

$$genotype = f(phenotype).$$

In this framework, there is no restriction on the phenotype variables' nature, hence, several structures can be considered. The aim is to be able to incorporate any number of heterogeneous phenotypic traits (qualitative nominal and ordinal, quantitative discrete and continuous) while modeling various genotype families using hierarchies of categories (for instance, species subdivided into subspecies, themselves having different geographical origins).

The above proposition was motivated by the phenotypic database of rice panicles developed by the UMR DIADE (Al-Tam et al., 2013) with different partnerships, including the LMI RICE 2, a joint international IRD-CIRAD-UM-USTH-AGI laboratory in Vietnam, CIAT in Colombia, and INERA in Burkina Faso. Panicle traits are among the most representative features of rice diversity; their architecture is relevant for the biological classification of plants, as well as for the improvement of cultivated rice. The rice phenotype database at hand has 752 panicles (observations). Each panicle was classified according to its geographical origin, species, subspecies, and subpopulation. For each continent (Asia, Africa), one cultivated and one wild species of rice (***Oryza***) are considered: *Sativa* (Asia-cultivated), *Rufipogon* (Asia-wild), *Glaberrina* (Africa-cultivated), and *Barthii* (Africa-wild). The hierarchical structure of this response is partially known since it is possible to create two different structures with information available about the phylogenetic tree of rice. A first structure is obtained when the species (the tree leaves) are aggregated according to first the geographical origin and then the domestication condition. The second structure is obtained by aggregating inversely, i.e., according to first the domestication condition and then to the geographical origin. The two possible hierarchical structures are represented in Figure 1. The subspecies in this Figure and in the following are coded as follows *Obar I*: ObI, *Obar II*: ObII, *Ogla I*: OgI, *Ogla II*: OgII, *Japonica*: Ja, and *Indica*: In. As explanatory variables, the database includes: rachis length, total length, number of grains, maximum number of branching order, number of nodes, and number of nodes on the rachis.

The methodology's application to the rice panicle diversity analysis was at the origin of this thesis subject. Still, the formalism is general, and other applications were investigated throughout this work.

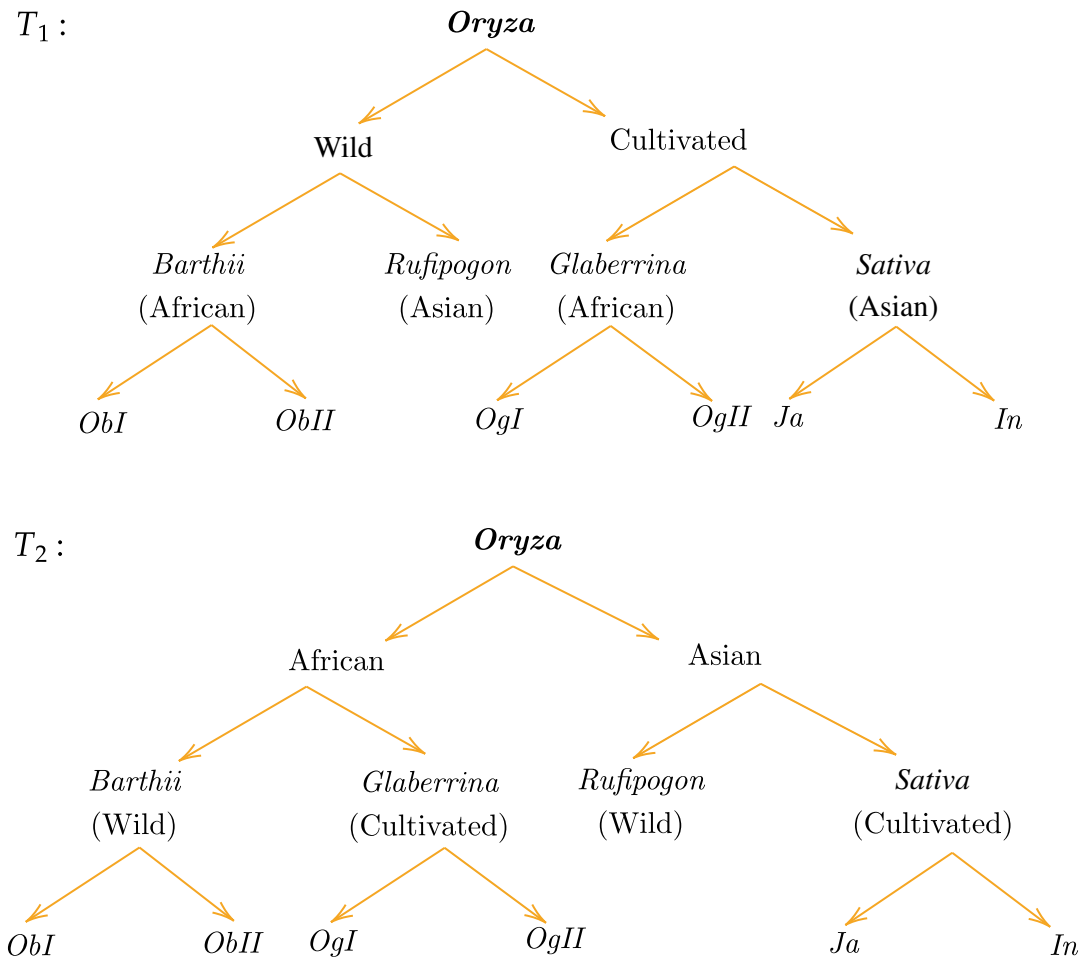

Figure 1: Hierarchical structures of the categories of the rice panicle data set.

1.1.1 Link function in GLMs for categorical responses

The GLMs class is an extension of traditional linear models that i) allows the expectation of a random variable to depend on a linear predictor through a link function and ii) considers the response probability distribution to be any member of the exponential family of distributions. This section presents the case of categorical responses in a general framework, i.e., when $J \geq 2$ from which the binary model can be deduced as a particular case. To this end, the exponential family of distributions has to be introduced in its multivariate form. Consider a random vector $\boldsymbol{y} = (y_1, \ldots, y_K)$ that lies in $\mathbb{R}^K$ whose distribution depends on a parameter $\boldsymbol{\theta} \in \mathbb{R}^K$. Its distribution belongs to the exponential family if its probability density function (pdf) can be written in the following standard form (which is a generalization of the univariate form of the exponential family)

$$f(\boldsymbol{y}; \boldsymbol{\theta}, \phi) = \exp\left\{\frac{\boldsymbol{y}^\top \boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi}\omega + c(\boldsymbol{y}, \phi)\right\}, \tag{1.1}$$

where $\boldsymbol{\theta}$ is the so-called natural parameter vector, b and c are known functions, ϕ is the

dispersion parameter and ω is a known weight.

Property 1. *Let $\boldsymbol{y}$ be a random vector whose distribution belongs to the exponential family. The function b is assumed to be twice differentiable with respect to $\boldsymbol{\theta}$. Its expected value and covariance are:*

(i) $\mathrm{E}(\boldsymbol{y}) = \nabla_b(\boldsymbol{\theta})$,

(ii) $\mathrm{Cov}(\boldsymbol{y}) = \dfrac{\phi}{\omega}\mathcal{H}_b(\boldsymbol{\theta})$,

where $\nabla_b(\boldsymbol{\theta})$ and $\mathcal{H}_b(\boldsymbol{\theta})$ denote respectively the gradient and the Hessian matrix of b with respect to $\boldsymbol{\theta}$.

Multinomial distribution as a member of the exponential family In the following, we consider the random variable Y with $J \geq 2$ categories, or equivalently its (truncated) indicator vector representation $\boldsymbol{y} = (y_1, \ldots, y_{J-1})^\top$ where $y_j = \mathbb{1}_{\{Y=j\}}$. The null vector thus corresponds to the last category. The expectation of $\boldsymbol{y}$ is denoted by the vector $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_{J-1})$ with the constraint $\sum_{j=1}^{J} \pi_j = 1$. In this framework, the discrete vector $\boldsymbol{y}$ follows the multinomial distribution

$$\boldsymbol{y} \sim \mathcal{M}(1, \boldsymbol{\pi})$$

which is a generalization of the Bernoulli distribution (obtained when $J = 2$). The probability mass function written in terms of $\boldsymbol{y}$ is then

$$f(\boldsymbol{y}; \boldsymbol{\pi}) = \left(\prod_{j=1}^{J-1} \pi_j^{y_j} \right) \left(1 - \sum_{j=1}^{J-1} \pi_j \right)^{1 - \sum_{j=1}^{J-1} y_j}.$$

Its natural parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{J-1})^\top$ is defined by

$$\boldsymbol{\theta} = \left(\log\left(\frac{\pi_1}{1 - \sum_{j=1}^{J-1} \pi_j} \right), \ldots, \log\left(\frac{\pi_{J-1}}{1 - \sum_{j=1}^{J-1} \pi_j} \right) \right)^\top,$$

and

$$b(\boldsymbol{\theta}) = \log\left(1 + \sum_{j=1}^{J-1} e^{\theta_j} \right).$$

Based on the above decomposition, the probability mass function can be simply written as

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \exp\{\boldsymbol{y}^\top \boldsymbol{\theta} - b(\boldsymbol{\theta})\}.$$

Using the weight $\omega = 1$, the dispersion parameter $\phi = 1$ and the null function $c(\boldsymbol{y}, \phi) = 0$, we see that this distribution function belongs to the exponential family of dimension $K = J - 1$ (see Equation (1.1)).

Regression model specification The aim is to model the effect on the multivariate response variable $\boldsymbol{y}$ of a given set of p covariates $\boldsymbol{x} = (x_1, \ldots, x_p)$, defined in a general

form of dimension q with $p \leq q$ (i.e., categorical variables are represented by indicator vectors). For convenience, models are presented at the individual level; thus, the subscript $i \in \{1, \ldots, n\}$ is not mentioned in the following.

According to Nelder and Wedderburn (1972), a GLM is characterized by three components: the random component, the systematic component, and the link function. The random component accounts for the conditional distribution of the response variable given the set of covariates. In the particular case of categorical outcomes, the response distribution is necessarily the multinomial $\mathcal{M}(1, \boldsymbol{\pi}(\boldsymbol{x}))$ (viewed as the multivariate extension of the Bernoulli). Hence, the only two parts that define a GLM for categorical responses are:

1. The systematic component which is determined by the linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{J-1})$. Considering the parameter vector as $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{J-1})$, the linear predictor can be written as the matrix product

$$\boldsymbol{\eta} = Z\boldsymbol{\beta},$$

   where Z denotes the design matrix composed of repetitions of 1 and $\boldsymbol{x}^\top$.

2. The link function g which relates the conditional expectation of the response variable $\boldsymbol{\pi} = \mathrm{E}[\boldsymbol{y}|\boldsymbol{x}]$ and the linear predictor $\boldsymbol{\eta}$. It connects the random and systematic components through the equality $g(\boldsymbol{\pi}) = \boldsymbol{\eta}$.

Note that the link function's domain is the simplex Δ, and the range of the linear predictor is $\mathbb{R}^{J-1}$, i.e.,

$$\begin{array}{rccc} g: & \Delta & \longrightarrow & \mathbb{R}^{J-1}, \\ & \boldsymbol{\pi} & \longmapsto & \boldsymbol{\eta}, \end{array}$$

with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{J-1}) \in \Delta$ where $\Delta = \{\boldsymbol{\pi} \in (0,1)^{J-1} : \sum_{j=1}^{J-1} \pi_j < 1\}$. The linear predictor and the mean parameter lie in different spaces. Hence, the link function should take the form according to the constraints of those spaces.

For the inference of the model (through the Fisher scoring algorithm or the Newton Raphson algorithm), the link function must be differentiable. The link function could also be invertible. Although invertibility is not a compulsory requirement, it is a desirable property. Indeed, the invertibility property allows obtaining well-defined models for any value of the linear predictor. Hence, considering together the properties of invertibility and differentiability (of the link function and its inverse), the space of possible link functions is reduced to all diffeomorphism from the simplex Δ to $\mathbb{R}^{J-1}$. The most common diffeomorphism is the *canonical link*, defined such that the natural parameter $\boldsymbol{\theta}$ equals the linear predictor $\boldsymbol{\eta}$. In this case the maximum likelihood estimation (MLE) is easily reached because the log-likelihood is strictly concave. Although most of the usual link functions are diffeomorphisms, some of them are not invertible (like the cumulative link described in section 1.3.1). As such, they bear a drawback in terms of interpretability that we will discuss later.

1.2 Generalized linear models for binary responses

When considering only two categories ($J = 2$), the response distribution is necessarily the Bernoulli. In this case, it is assumed that the probability of success π is characterized by the explanatory variables through the link $g(\pi) = \boldsymbol{z}^\top \boldsymbol{\beta}$ such that

$$\begin{array}{rccc} g: & (0,1) & \longrightarrow & \mathbb{R}, \\ & \pi & \longmapsto & \eta. \end{array}$$

A simple link function, which is not a diffeomorphism from $(0,1)$ to $\mathbb{R}$, is the identity function $\pi = \boldsymbol{z}^\top \boldsymbol{\beta}$. This linear model is used in some practical applications. Still, its fundamental problem is that whereas π is a probability, the fitted values $\boldsymbol{z}^\top \boldsymbol{\beta}$ may be less than zero or greater than one. To ensure π to be restricted to the unit interval $(0,1)$, a strictly increasing cdf F is often used, such that

$$\pi = F(\eta). \tag{1.2}$$

Note that the strict increasing assumption implies invertibility of the link function. Because of that, we can obtain a straightforward interpretation of the effect of a covariate on the response expected value. Regardless of the value of x_k, if δ_k is positive, then increasing x_k will be associated with increasing π, and if δ_k is negative, then increasing x_k will be related to decreasing π.

Binary regression models can be motivated (although it is not a strict model requirement) by the assumption of a latent continuous variable $\tilde{Y}$, for which there exists a threshold that defines whether the original observed variable Y is 0 or 1. Let the model for the latent variable be $\tilde{Y} = \alpha + \boldsymbol{x}^\top \boldsymbol{\delta} + \varepsilon$, where $\boldsymbol{x} = (x_1, \ldots, x_p)$ denotes the covariate vector, $\alpha \in \mathbb{R}$ the intercept, $\boldsymbol{\delta} \in \mathbb{R}^p$ the slope parameter vector, and ε the latent residual. Considering Y as a dichotomized version of $\tilde{Y}$ we obtain

$$\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = P(\tilde{Y} \geq 0|\boldsymbol{x}) = P(\alpha + \boldsymbol{x}^\top \boldsymbol{\delta} + \varepsilon \geq 0) = P(-\varepsilon \leq \alpha + \boldsymbol{x}^\top \boldsymbol{\delta}).$$

Allowing $-\varepsilon$ to have the distribution function F, we get the simple form (1.2). As noticed by Agresti (2018), the variance of ε is linked to the normalization of the parameters, which we will discuss later in this section.

1.2.1 Usual link functions

Logistic cdf (canonical link) The most widely adopted GLM for binary responses is the logit model which uses as inverse link the cumulative logistic distribution function

$$F(\eta) = \frac{1}{1 + \exp(-\eta)} = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

The logit regression yields a linear model for the logarithmic odds: $log(\pi/(1-\pi))$. The transformation using the exponential function results in the form

$$\frac{\pi}{1-\pi} = \exp(\alpha_0)\exp(\delta_1 x_1)\ldots\exp(\delta_p x_p),$$

which exhibits how the explanatory variables affect the odds in an exponential multiplicative form. Thus, the basic interpretation for the magnitude of δ_k is that the odds increase multiplicatively by $\exp(\delta)$ for every unit of increase in x_k. Some early uses of the logit link were in biomedical studies, and nowadays, it has been popularized in social sciences, marketing, genetics, among others. The logit link function is the canonical link in binary models, which means that the linear predictor is directly equal to the canonical parameter of the Bernoulli distribution (considering its form as a member of the exponential family). The canonical link function for binary GLMs is then defined as

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

Normal cdf A common model used in several areas of social and biological sciences (prevalent in econometrics and genetic studies) is the probit model, which is based on the standard normal distribution

$$F(\eta) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\eta}\exp(-z^2/2)dz.$$

The probit model usually yields approximately the same results as the logit link. We can see in Figure 2 that both logistic and normal cdfs are symmetric s-shaped curves, but the normal places less probability in the tails of the distribution than does the logistic. By comparing their variances, we can see that the scale (and thus the spread) of the logistic is greater than the normal. However, when standardizing the logistic curve to have variance equal to 1, the curves become virtually indistinguishable, and so does the fit. Therefore, large amounts of data are needed to obtain substantial differences between probit and logit models. A minor disadvantage of the probit model is the required numerical evaluation of $\Phi(\eta)$ in the maximum likelihood estimation of the parameter $\boldsymbol{\beta}$ (Ludwig Fahrmeir, 2013).

Cauchy cdf The cauchit link is defined as $\tan(\pi(\pi(\boldsymbol{x})-\frac{1}{2}))$, with $\pi(x)$ the probability and $\pi = 3.14159\ldots$ and the cdf of the standard Cauchy distribution is

$$F(\eta) = \frac{1}{2} + \frac{1}{\pi}\arctan(\eta).$$

The use of the Cauchy cdf in the context of GLMs is suggested if outliers (i.e., high leverage points) are suspected in the space of the linear predictor (Smithson and Verkuilen, 2006). The Cauchy distribution is very heavy-tailed and makes less extreme predictions for the expected value of the dependent variable than the normal or logistic distributions

for large values of the linear predictor. From Figure 2 we can see that the Cauchy places even less probability in the tails than the normal and thus the logistic cdfs.

Gumbel and Gompertz cdfs Another conventional model results when considering the complementary log-log link, characterized by the the Gompertz (or the Gumbel min cdf)

$$F(\eta) = 1 - \exp(-\exp(\eta)).$$

The Gumbel max (or simply Gumbel) cdf yields to the log-log model where

$$F(\eta) = \exp(-\exp(-\eta)).$$

The Gompertz and Gumbel distributions are not symmetrical. Those distributions are closely connected; if a random variable X has a Gumbel distribution, then the conditional distribution of $-X$, has a Gompertz distribution. For this reason, if $\boldsymbol{\beta}$ is the parameter vector of the Gompertz model for the response Y, then $-\boldsymbol{\beta}$ is the parameter vector for the response $1 - Y$ of the Gumbel model.

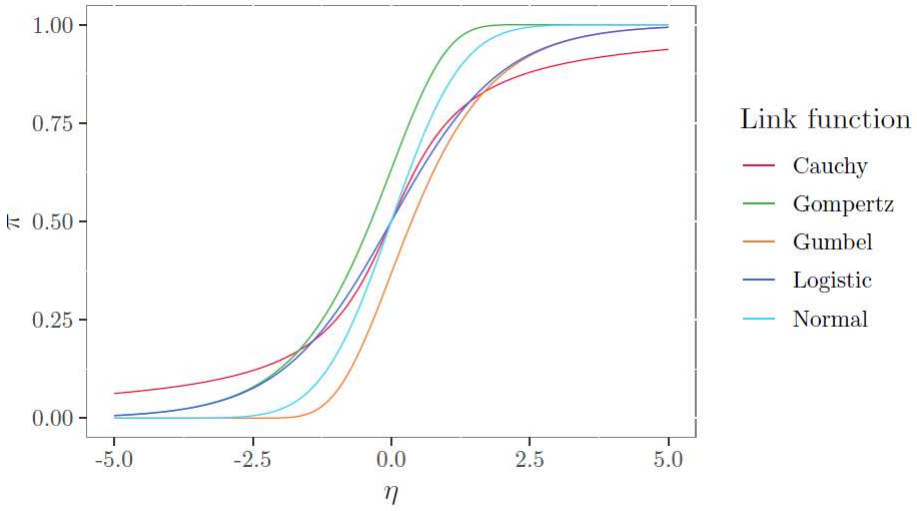

Figure 2: Shapes of the most common link functions for binary data $F(\eta) = \pi$.

1.2.2 Likelihood inference

The score for the binary regression model, described in (1.2), has the form

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = f(\eta)\frac{y - F(\eta)}{F(\eta)[1 - F(\eta)]}\boldsymbol{z}, \tag{1.3}$$

while the Fisher information matrix is given by

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}}\right) = -\frac{f^2(\eta)}{F(\eta)[1-F(\eta)]}\boldsymbol{z}\boldsymbol{z}^\top. \tag{1.4}$$

The log-likelihood is strictly concave for the logit canonical link, for which the score and the Fisher information functions are respectively reduced to $(y - F(\eta))\boldsymbol{z}$ and $-f(\eta)\boldsymbol{z}\boldsymbol{z}^\top$. Strict concavity implies that there is a unique global optimum, so there is no risk of convergence to a local optimum. But this is not the case for all the rest of link functions. The condition that both $\log(F)$ and $\log(1-F)$ be concave is sufficient for concavity of the log-likelihood (Pratt, 1981). Bagnoli and Bergstrom (2005) showed that strict log-concavity holds for distributions including normal, Gumbel, Gompertz, and Laplace, but not for others such as the Cauchy or the Student distributions. Hence, convergence problems on the estimation algorithm may occur.

Normalization of the parameters Models defined with different cdfs (i.e., different link functions) are not comparable since they refer to specific means and variances (Tutz, 2011). Often, the parameter estimates will turn out to be different even if there is no apparent discrepancy in the goodness-of-fit indicators of the models. A common approach to standardize the parameters of a binary regression model is based on the expected value and variance of ε

$$\tilde{\alpha} = \frac{\alpha - \mathrm{E}(\varepsilon)}{\sqrt{\mathrm{Var}(\varepsilon)}}, \qquad \tilde{\delta} = \frac{\delta}{\sqrt{\mathrm{Var}(\varepsilon)}}, \qquad \text{where } \varepsilon \sim F.$$

Remark that this approach is not suitable when using a cdf with undefined mean or variance as it is the case for the Cauchy cdf, or more generally, the Student (with ν degrees of freedom) distribution whose mean and variance are not defined when $\nu \leq 1$, and $\nu \leq 2$, respectively.

A propagated approach in econometrics that solves this problem is to consider the average partial effect of the variable x_k on π , i.e., $\partial\pi/\partial x_k$ as the scale factor, If x_k is a continuous variable, its partial effect on π is obtained from the partial derivative:

$$\frac{\partial \pi}{\partial x_k} = \frac{\partial F}{\partial \eta}\hat{\delta}_k \,. \tag{1.5}$$

Considering a sample of observations $(i = 1, \ldots, n)$, the average partial effect of x_k on π is then given by $n^{-1}\sum_{i=1}^{n} f(\eta_i)\hat{\delta}_k$, where f is the pdf of the associated inverse link function. The downsides of this method are that the scale factor depends on the model's input data and that it is only valid for continuous explanatory variables. Note that if f is a symmetric pdf around zero, the largest effect occurs when $\eta = 0$. For instance, for the normal pdf, this will be at $f(0) \approx 0.4$ and for the logistic pdf at $f(0) = 0.25$. A simple approach to make the magnitudes of those two cdfs roughly comparable is to multiply the probit estimates by $0.4/0.25 = 1.6$ or to multiply the logit estimates by $0.25/0.4 = 0.625$.

In chapter 3, we present a methodology that do not share the limitations of the two

previous methods. It is thus defined for Cauchy (and Student) cdf and is independent of the dataset. We detail its computational implementation, and an example of a proposal presented by Peyhardi (2020) for the normalization of parameter estimates through the location parameter and the scale parameter of the cdf F.

1.2.3 Sensitivity to perturbations

By definition, the term outlier implies the comparison of an observation with the rest of the sample (Salsas et al., 1999). From a geometric interpretation, an outlier corresponds to an observation that is far from the bulk of the data; however, there is limited scope for binary data to move in the Y direction since there are only two options. As stated by Copas (1988), an outlier for binary data is only defined from the probabilistic perspective as follows: an outlier occurs in binary data when $Y = 1$ and the corresponding fitted probability approaches zero, or when $Y = 0$ and the fitted probability approaches one. Since fitted probabilities close to zero or one occur when the linear predictor has a large absolute value, outliers are only found when the observations exhibit extreme values of the explanatory variables (see Figure 3).

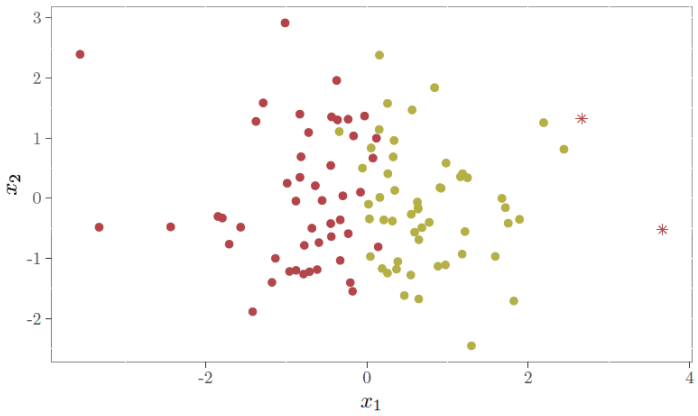

Figure 3: Scatter plot of the data set with a binary response (green and red colors correspond to the two response levels) and two covariates (specified on the axes). The observations represented with an asterisk are possible outliers on the direction of x_1.

It has been shown that the two most popular links for modeling binary data, logit and probit, are sensitive to the presence of outliers (see Pregibon, 1982). Authors including Lange et al. (1989); Liu (2005) have shown that the influence function (IF) of a binary model is unbounded for the logit and the probit link functions and bounded for other functions such as the Student and the cauchit link functions. The IF formalizes the bias caused by one outlier (Hampel, 1974). In the context of GLM, the IF of a new observation $(y^*, \boldsymbol{x}^*)$ on the MLE, is given by

$$IF[(y^*, \boldsymbol{x}^*), \hat{\beta}] = \left\{ \mathsf{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \right)_{\beta=\hat{\beta}} \right\}^{-1} \left(\frac{\partial l^*}{\partial \boldsymbol{\beta}} \right)_{\beta=\hat{\beta}},$$

where the log-likelihood computed for the original dataset $\{(y, \boldsymbol{x})\}_{i=1,\dots,n}$ and for the new

observation $(y^*, \boldsymbol{x}^*)$ are denoted by l and l^*, respectively (see Künsch et al. (1989)).

The presence of outliers is not the only disturbance that can occur in the data of a regression model. Another common one is the presence of noisy variables. In this work, we aim to assess the robustness of link functions for binary data comprehensively. Hence, in chapter 2, we study the sensitivity of different links when outliers and/or noisy variables are present in the data set. We will analyze this robustness property from different separation settings of the response levels in the covariate space.

1.3 Generalized linear models for categorical responses

The response distribution is necessarily the multinomial when considering more than two categories ($J \geq 2$). In this case, we assumed that the vector of probabilities $\boldsymbol{\pi}$ is characterized by the explanatory variables through the link $g(\boldsymbol{\pi}) = Z\boldsymbol{\beta}$ such that

$$\begin{array}{rccc} g: & \Delta & \longrightarrow & \mathbb{R}^{J-1}, \\ & \boldsymbol{\pi} & \longmapsto & \boldsymbol{\eta}, \end{array}$$

with $\boldsymbol{\pi} = (\pi_1(\boldsymbol{x}), \ldots, \pi_{J-1}(\boldsymbol{x})) \in \Delta$ where $\Delta = \{\boldsymbol{\pi} \in (0,1)^{J-1} : \sum_{j=1}^{J-1} \pi_j < 1\}$.

1.3.1 Logit models for categorical responses

Multinomial logit model (canonical link) The multinomial logit (MNL) model, also referred to as the baseline-category logit model or as the polytomous model (Luce, 1959; Engel, 1988), is designed to analyse nominal scales where there are several categories. It is the most commonly used regression model for nominal response variables. It is actually a generalization of the logistic regression for dichotomous responses where the probability of category j is given by

$$P(Y = j|\boldsymbol{x}) = \frac{\exp\left(\alpha_j + \boldsymbol{x}^\top \delta_j\right)}{\sum_{k=1}^{J} \exp(\alpha_k + \boldsymbol{x}^\top \delta_k)},$$

for $j = 1, \ldots, J$. In this model, the logits are formed by comparing each response category j to an arbitrarily chosen baseline response category. The common approach sets the last category J as the baseline (reference) category and the corresponding parameters α_J and δ_J are assumed to be zero in order to avoid identifiability problems. We thus obtain

$$P(Y = j|\boldsymbol{x}) = \frac{\exp\left(\alpha_j + \boldsymbol{x}^\top \delta_j\right)}{\sum_{k=1}^{J-1} \exp(\alpha_k + \boldsymbol{x}^\top \delta_k)},$$

for $j = 1, \ldots, J-1$. This model has been introduced in biology, sociology, and econometrics with different definitions. It can be viewed as a GLM for multivariate responses, as $J-1$ logit models with the same reference category or as a random utility model. The

associated design matrix of this model is of the form

$$Z = \begin{pmatrix} 1 & & \boldsymbol{x}^\top & & \\ & \ddots & & \ddots & \\ & & 1 & & \boldsymbol{x}^\top \end{pmatrix}. \tag{1.6}$$

with $J-1$ rows and $(J-1)(1+p)$ columns. The link function of the MNL is the canonical link in GLMs for categorical responses since the linear predictor is directly equal to the canonical parameter of the multinomial distribution. The canonical link function for categorical GLMs is defined as

$$g_j(\boldsymbol{\pi}) = \log\left(\frac{\pi_j}{1 - \sum_{k=1}^{J-1} \pi_k}\right), \tag{1.7}$$

for $j = 1, \ldots, J-1$ where $\Delta = \{\boldsymbol{\pi} \in (0,1)^{J-1} : \sum_{j=1}^{J-1} \pi_j < 1\}$.
The *baseline-categories logits* are then

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \boldsymbol{x}\top\delta_j,$$

for $j = 1, \ldots, J-1$. Those models lack of parsimony as each covariate x_k has $J-1$ parameters. In this framework, all the $\binom{J}{2}$ pairs of categories are described. The effects change according to the response paired with the baseline. The $J-1$ equations also determine parameters for logits with other pairs since

$$\log\left(\frac{\pi_j}{\pi_k}\right) = \log\left(\frac{\pi_j}{\pi_J}\right) - \log\left(\frac{\pi_k}{\pi_J}\right),$$

for $j, k \in \{1, \ldots, J-1\}$.

For the following, let the vector $\boldsymbol{\omega} = \{\boldsymbol{\omega}_j\}_{j=1,\ldots,J} \in \mathbb{R}^{qJ}$ represent the set of q alternative specific attributes. Depending on the form of the linear predictors η_j, we obtain different logit models:

- *MNL model*: $\eta_j = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\delta}_j$; for which the attributes are the same for all alternatives and the parameters depend on each alternative.
- *Conditional logit model*: $\eta_j = \alpha + \boldsymbol{x}^\top\boldsymbol{\delta} + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_J)^\top\boldsymbol{\gamma}$; for which the attributes dependent on each alternative and the parameters are the same for all alternatives (see McFadden, 1973).
- *Universal logit model*: $\eta_j = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\delta}_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_J)^\top\boldsymbol{\gamma}$; for which the attributes and the parameters are dependent on each alternative.

Cumulative logit model The cumulative logit model is designed for ordinale scales and is based on the accumulated response probabilities that denote the probability that a randomly selected observation falls in the j^{th} category of a variable (Agresti, 1981).

As considered by McCullagh (1980), this model can be seen as if the observed y was originated from the categorization of a latent continuous variable $\tilde{Y}$. This latent variable follows a linear regression model

$$\tilde{Y} = \tilde{\alpha} + \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}} + \varepsilon, \tag{1.8}$$

where $-\infty = \alpha'_0 < \alpha'_1 < \ldots \alpha'_{J-1} < \alpha'_J = \infty$ are the strictly-ordered cut-points, and ε is a noisy variable with logistic cdf. To model this categorization process, the cumulative ratio assumes that the $J-1$ cut-points partition $\tilde{Y}$ into J observable ordered categories of Y, i.e.

$$\{Y = j\} \Leftrightarrow \alpha'_{j-1} < \tilde{Y} \leq \alpha'_j,$$

for $j = 1, \ldots, J$. The cumulative probabilities are

$$\begin{aligned} P(Y \leq j|\boldsymbol{x}) &= P(\tilde{Y} \leq \alpha'_j) \\ &= P(\varepsilon \leq \alpha'_j - \tilde{\alpha} - \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}}) \\ P(Y \leq j|\boldsymbol{x}) &= \frac{\exp(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta})}{1 + \exp(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta})} \end{aligned}$$

with $\alpha_j = \alpha'_j - \tilde{\alpha}$, and $\boldsymbol{\delta} = -\tilde{\boldsymbol{\delta}}$. It is then evident that

$$\pi_j = P(\alpha'_{j-1} < \tilde{Y} < \alpha'_j).$$

The order structure is more easily interpretable using the notion of the latent continuous variable where the categories are considered as successive intervals $(\alpha'_{j-1}, \alpha'_j]$.

The cumulative logit model is usually presented as

$$\text{logit}\{P(Y \leq j|\boldsymbol{x})\} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}, \tag{1.9}$$

for $j = 1, \ldots, J-1$. Remark that the logit difference has the simple form

$$\begin{aligned} \text{logit}\{P(Y \leq j|\boldsymbol{x}_1)\} - \text{logit}\{P(Y \leq j|\boldsymbol{x}_2)\} &= \log\left\{\frac{P(Y \leq j|\boldsymbol{x}_1)/P(Y > j|\boldsymbol{x}_1)}{P(Y \leq j|\boldsymbol{x}_2)/P(Y > j|\boldsymbol{x}_2)}\right\} \\ &= (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{\delta}. \end{aligned}$$

We can notice that the log odds ratio does not depend on category j and is proportional to the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Because of this proportional odds property, the cumulative logit model is also called the proportional odds logit model (McCullagh, 1980).

Sequential logit model The sequential logit model can be derived from the assumption that the response categories $1, \ldots, J$ are reached successively. They reflect the successive transition to higher categories in a stepwise fashion. This model assumes that the successive choices between category j and the categories over j is determined

by the latent variables

$$\tilde{Y}_j = \tilde{\alpha} + \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}} + \varepsilon_j, \tag{1.10}$$

for $j = 1, \ldots, J-1$, where the residuals ε_j are independent and identically distributed according to the logistic cdf. This sequential mechanism can be viewed as a binary process at each transition, thus, it is appropriate when the assumption of a single underlying latent variable does not hold. We can write then

$$\{Y = j\} = \bigcap_{k=1}^{j-1} \{\tilde{Y}_k > \alpha'_k\} \bigcap \{\tilde{Y}_j \leq \alpha'_j\},$$

so the conditional probabilities $P(Y = j|Y \geq j)$ for $j = 1, \ldots, J$ can also be written as $P(Y = j|Y \geq j) = P(\tilde{Y}_j \leq \alpha'_j)$, then, we have

$$P(Y = j|Y \geq j; \boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta})}{1 + \exp(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta})},$$

where $\alpha_j = \alpha'_j - \tilde{\alpha}$, and $\boldsymbol{\delta} = -\tilde{\boldsymbol{\delta}}$. In this context, we can represent the probabilities of each category as

$$\pi_j = P(\tilde{Y}_j < \tilde{\alpha}_j) \prod_{k=1}^{j-1} P(\tilde{Y}_k > \tilde{\alpha}_k).$$

Remark that the sequential logit model is usually presented as

$$\text{logit}\{P(Y = j|Y \geq j; \boldsymbol{x})\} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta},$$

for $j = 1, \ldots, J-1$. The transition can be interpreted in terms of the difficulty of reaching the next category. Upper levels can only be achieved only if previous levels were visited earlier and not kept. Therefore the model is built around the conditionality principle.

Adjacent logit model The adjacent logits for $j = 1, \ldots, J-1$, have the basic form

$$\begin{aligned} \text{logit}\{P(Y = j|Y \in \{j, j+1\})\} &= \log\left\{\frac{P(Y = j)}{P(Y = j+1)}\right\} \\ &= \log\left\{\frac{\pi_j}{\pi_{j+1}}\right\}, \end{aligned}$$

for all pairs of adjacent categories. The adjacent proportional logit model is then defined by relating adjacent logits to proportional linear predictors

$$\text{logit}\{P(Y = j|Y \in \{j, j+1\}; \boldsymbol{x})\} = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}$$

for $j = 1, \ldots, J-1$. As for sequential models, the conditional form of adjacent models implies independence between all linear predictors η_j. Therefore, no constraints are required on $\boldsymbol{\eta}$ to obtain non-negative probabilities.

1.3.2 Likelihood inference

For GLMs, the parameter vector $\boldsymbol{\beta}$ is estimated by maximum likelihood. In the following, we present Fisher's scoring algorithm at iteration $t+1$

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left\{ \mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}} \right\}^{-1} \left(\frac{\partial l}{\partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}}.$$

Using the chain rule for differentiation, the score is given by

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\pi}} \frac{\partial l}{\partial \boldsymbol{\theta}},$$

where $l(\boldsymbol{\theta}) = \boldsymbol{y}^\top \boldsymbol{\theta} - b(\boldsymbol{\theta})$. Using Property 1, the expression becomes

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = Z^\top \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{x})^{-1} (\boldsymbol{y} - \boldsymbol{\pi}), \tag{1.11}$$

where the Jacobian matrix $\dfrac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}}$ depends on the link function. The Fisher information matrix is then given by

$$\begin{aligned} \mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \right) &= -\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\beta}} \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{x})^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\beta}^\top} \\ &= -Z^\top \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{x})^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}^\top} Z. \end{aligned} \tag{1.12}$$

Remark that the score and the Fisher information functions for the canonical link are simplified to the expressions $Z^\top(\boldsymbol{y} - \boldsymbol{\pi})$, and $-Z^\top \dfrac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}^\top} Z$.

1.3.3 Unified specification of GLMs for categorical data: (r, F, Z) models

Peyhardi et al. (2015) showed that all the classical link functions can be decomposed through the unified specification

$$g_j = F^{-1} \circ r_j \quad \Leftrightarrow \quad r_j(\boldsymbol{\pi}) = F(\eta_j) \qquad j = 1, \ldots, J-1, \tag{1.13}$$

where F is a continuous and strictly increasing cdf, and $\boldsymbol{r} = (r_1, ..., r_{J-1})$ is a map from the simplex Δ to the open hypercube $(0,1)^{J-1}$. The authors introduced the notation (r, F, Z) with which any classical GLM for categorical responses can be fully described.

- The first component, r, is called the ratio. This part of the linking function addresses the nature of the categorical response. The authors specified three ratios that rely on an ordering assumption among categories: the adjacent, cumulative and sequential, respectively defined as $r_j(\boldsymbol{\pi}) = \dfrac{\pi_j}{\pi_j + \pi_{j+1}}$, $r_j(\boldsymbol{\pi}) = \pi_1 + ... + \pi_j$, and

$r_j(\boldsymbol{\pi}) = \dfrac{\pi_j}{\pi_j + \ldots + \pi_J}$ for $j = 1, \ldots, J-1$. For the nominal responses, they defined the reference ratio as $r_j(\boldsymbol{\pi}) = \dfrac{\pi_j}{\pi_j + \pi_J}$ for $j = 1, \ldots, J-1$.

- The second component of the triplet is the cdf F. Several distributions exist and are appropriate for these models. Adjusting the tail weight and skewness of this distribution can markedly improve the model fit.
- The third component embodies the specification (allowing for constraints) of the linear predictor. For instance, the design matrices without constraint or with the common slope constraint have respectively the following forms

$$Z_c = \begin{pmatrix} 1 & & \boldsymbol{x}^\top & & \\ & \ddots & & \ddots & \\ & & 1 & & \boldsymbol{x}^\top \end{pmatrix} \quad \text{and} \quad Z_p = \begin{pmatrix} 1 & & \boldsymbol{x}^\top \\ & \ddots & \vdots \\ & & 1 \quad \boldsymbol{x}^\top \end{pmatrix}.$$

All the classical GLMs for categorical responses can be decomposed into the three components r, F, and Z. As an illustration, consider the cumulative logit model whose original formulation is given in Equation (1.9). This same model is rewritten as the (r, F, Z) triplet: (*cumulative*, *logistic*, *proportional*). Hence, we can express the cumulative model through the following equation

$$\pi_1 + \cdots + \pi_j = \frac{\exp(\alpha_j + \boldsymbol{x}^t\boldsymbol{\delta})}{1 + \exp(\alpha_j + \boldsymbol{x}^t\boldsymbol{\delta})}.$$

Note that the link function $g : \Delta \to \mathbb{R}^{J-1}$ is differentiable if the ratio $\boldsymbol{r} : \Delta \to (0,1)^{J-1}$ and the cdf $F : \mathbb{R} \to (0,1)$ are both differentiable. The four ratios, as well as the cdfs, are differentiable. Thus, the Fisher scoring algorithm is appropriate to estimate these categorical models. Using the decomposition of the link function presented in Equation (3.1), we can decompose further the score (1.11) as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = Z^\top \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}} \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{x})^{-1} (\boldsymbol{y} - \boldsymbol{\pi}) \,,$$

and the Fisher information matrix as

$$\mathsf{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}}\right) = -Z^\top \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}} \mathrm{Cov}(\boldsymbol{y}|\boldsymbol{x})^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}^\top} \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}^\top} Z.$$

Remark that the Jacobian matrix $\partial \boldsymbol{F}/\partial \boldsymbol{\eta}$ is the diagonal matrix of densities, i.e., $\{f(\eta_j)\}_{j=1,\ldots,J-1}$; the Jacobian matrices associated to each ratio $\partial \boldsymbol{\pi}/\partial \boldsymbol{r}$ are detailed in Appendix B.2. Note that the above calculations of the score and the Fisher information matrices concern only one observation. To obtain the total expected result, the contributions of the n observations have to be added.

Software availability In R (R Core Team, 2021) there is a variety of packages to fit categorical responses; however, most of them only cover one or a few of the types of

models. For instance, the function multinom() of the package nnet (Ripley and Venables, 2021; Venables and Ripley, 2002) fits the MNL via neural networks. For ordinal responses, the functions polr() of the package MASS (Ripley et al., 2021; Venables and Ripley, 2002) and omr() from the rms (Harrell Jr, 2021) package are often used to fit the odds proportional model. Few packages are aim to fit a whole family of models for categorical responses, one of them is the tram (Hothorn et al., 2021; Hothorn, 2020) package, which by means of the Polr() function allows for stratification, censoring and truncation in the response of cumulative models. The ordinal package (Christensen, 2019) is another option to fit the family of cumulative models. It includes a comprehensive implementation of this class of models offering great flexibility, notably in the specification of the linear predictor. To our knowledge, only the VGAM (Yee, 2021) and the ordinalNet (Wurm et al., 2020) packages consider the three families of ordinal models: cumulative, sequential, and adjacent. Nevertheless, the ratio of probabilities of the adjacent models in VGAM seems to be valid only for the logistic distribution since they consider the ratio to be $\pi_j/\pi_{j+1} = F(\eta_j)$ instead of $\pi_j/(\pi_j + \pi_{j+1}) = F(\eta_j)$. None of the above-mentioned packages encloses the four model families for categorical responses, and most of them have some limitations in terms of adding constraints to the design or in the availability of the cdfs that one can use as part of the link function. These gaps also exist in commercial statistical software like SAS (SAS Institute Inc., 2020), Stata (Stata Corp., 2015), and SPSS (IBM Corporation, 2017). An additional problem of commercial packages is that they use different techniques (which are not strictly equivalent) to fit the models. As a consequence, different estimations might be obtained when using different software, even though the same theoretical model is specified. For instance, Liu (2009) reported some differences in the estimation of an odds proportional model using the functions PROC LOGISTIC in SAS, OLOGIT in Stata, and PLUM in SPSS.

As part of the work of this thesis, we created an R package called GLMcat that we introduce and describe in chapter 3. This software solution is designed under the unified specification (r, F, Z) that allows the user to fit not only any classical generalized linear model for either nominal or ordinal responses but also to fit the models that emerge when changing specifications in the link function or the constraints of the linear predictor. In addition, we provide a practical guide on how to choose an appropriate model among a vast set of possibilities. We base our recommendations on the models' theoretical properties, which we explore in detail throughout the chapter.

1.4 Generalized linear models for categorical responses with hierarchical structure

In the following, we present alternatives for modeling a hierarchical structure of the response categories when such structure is known beforehand.

Nested logit model The nested logit model can be depicted by a tree structure that represents all the categories. The MNL model treats all alternatives equally, whereas

the nested logit model includes intermediate branches that group the categories. This model was introduced by McFadden et al. (1978) to avoid the inconsistency of the independence of irrelevant alternatives (IIA) property (also known as Luce's choice axiom) in specific scenarios. Let us illustrate these ideas through the well-known example of the blue and red buses (Debreu, 1960). Suppose that an individual has no preference between the two alternatives $A = \{\text{blue bus, car}\}$; so $P_A(\text{blue bus}) = P_A(\text{car}) = 1/2$. Imagine now that the travel company introduces red buses, so the options are now $B = \{\text{blue bus, red bus, car}\}$. The individual again has no preference between blue and red buses; meaning that $P_B(\text{blue bus}) = P_B(\text{red bus})$. The IIA property states that the ratio of any two-outcome probabilities is independent of the set containing the different alternatives, using this notion we have

$$\frac{P_A(\text{blue bus})}{P_A(\text{car})} = \frac{P_B(\text{blue bus})}{P_B(\text{car})} = 1.$$

If the color of the bus does not affect the mode choice the expected probabilities are $P_B(\text{blue bus}) = P_B(\text{red bus}) = 1/4$ and $P_B(\text{car}) = 1/2$. However, due to the IIA property and the non-preference of the bus color, we obtain that $P_B(\text{blue bus}) = P_B(\text{red bus}) = P_B(\text{car}) = 1/3$. This is a counterintuitive result because the additional *irrelevant* alternative (red bus) has decreased the choice probability of driving substantially.

In this example, the IIA property is not appropriate because two alternatives share many characteristics. Therefore, the J alternatives can be divided into L *nests* (sets) such that the choice process starts by choosing among the L choice sets and then making the specific choice within the chosen set. The nested logit model captures the similarities between close alternatives. In the presented example, the individual chooses first between bus and car according to specific factors and then between the two buses according to preferred color (see Figure 4). More generally, suppose that alternatives can

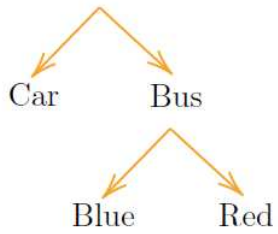

Figure 4: Illustration of the transport vehicle selection process.

be aggregated according to their similarities; this means that all alternatives of the same nest N_l share attributes x^l, whereas other alternatives do not. The nested logit model is presented with only two levels in the following. Let L be the number of nests obtained by partitioning the set of J alternatives and

$$\{1, \ldots, J\} = \bigcup_{l=1}^{L} N_l.$$

If j denotes an alternative belonging to the nest N_l, then the probability of alternative

j is decomposed as follows

$$P(Y = j|\boldsymbol{x}) = P(Y = j|Y \in N_l; \boldsymbol{x}^l)P(Y \in N_l|\boldsymbol{x}^0, IV), \tag{1.14}$$

where $IV = (IV_1, \ldots, IV_L)$ denotes the vector of *inclusive values* (described above), $\boldsymbol{x}^0$ are the attributes which influence only the first choice level between nests and $\boldsymbol{x} = (\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^L)$. Each probability of (1.14) is determined by a MNL model as follows

$$P(Y = j|Y \in N_l; \boldsymbol{x}^l) = \frac{\exp(\eta_j^l)}{\sum_{k \in N_l} \exp(\eta_k^l)},$$

and

$$P(Y \in N_l|\boldsymbol{x}^0, IV) = \frac{\exp(\eta_l^0 + \lambda_l IV_l)}{\sum_{k=1}^{L} \exp(\eta_k^0 + \lambda_k IV_k)},$$

where the inclusive value is

$$IV_l = \ln\left\{\sum_{k \in N_l} \exp(\eta_k^l)\right\}.$$

The predictors η_j^l and η_l^0 depend respectively on $\boldsymbol{x}^l$ and $\boldsymbol{x}^0$. In practice, they are linear with respect to $\boldsymbol{x}$. Because of the inclusive values, the nested logit model must be estimated in two steps. In the first step, the L models of the second level can be estimated separately because the parameters $\boldsymbol{\beta}^l$ are different in each nest. In a second step, the inclusive values IV_l of each nest can then be computed and used to estimate the first level model. The nested logit model has been extended to three and higher levels. However, the complexity of the model increases geometrically with the number of levels (Greene, 2003). This model has been found to be extremely flexible, and it is widely used to model consumer choices since it follows the utility-maximization principle. The nested model has some limitations that result from complying with the random utility model assumption. On one side, the model must include the inclusive values whose artificial nature makes interpretation difficult. On the other side, there is the constraint where $0 < \lambda_l \leq 1$ for $l = 1, \ldots, L$ (McFadden et al., 1978). Hence, the nested model would be more flexible if the random utility maximization assumption is relaxed. Remark that the particular case of $\lambda_l = 1$ leads to the simple multinomial logit model.

Partitioned conditional model for partially-ordered data Among the extensive research devoted to GLM, few efforts have been focused on the analysis of partially ordered responses. Zhang and Ip (2012) proposed the partitioned conditional model, a new class of GLM intended primarily for partially ordered responses, but which also includes nominal and ordinal responses as special cases. The main idea was to recursively partition the J categories to transform the partial order into subsets of cases with a total order or no order whatsoever. Thus, the authors proposed to use the odds proportional logit model for the total order case and the MNL model for the nominal case. The basis for the construction of the formal partitioned conditional model for partially-ordered data is the proposition 1, for which we need first to introduce the following notions:

- A partial ordered set (poset) is defined as the pair $(P;\preceq)$, where $P = 1,\ldots,J$ is the set of categories, and $\preceq$ indicates the partial order of P. Remark that a poset can be summarized by a Hasse diagram as in Figure 5.
- The order relation $j \preceq k$ is represented by an edge between the two nodes (categories) where node k is above node j.
- A chain in a poset $(P;\preceq)$ is a totally ordered subset C of P, whereas an antichain is a set A of pairwise incomparable elements.

Proposition 1. *A finite partially-ordered set can always be partitioned into antichains that are totally weakly ordered (Zhang and Ip, 2012).*

Let $(P;\preceq)$ be the poset represented by the Hasse diagram in Figure 5. The partition is defined by the antichains $N_1 = \{1\}$, $N_2 = \{2,3,4\}$ and $N_3 = \{5\}$ corresponding to each level of the Hasse diagram. As these antichains are totally (weakly) ordered, the odds proportional logit model is used to describe the cumulative probabilities $P(Y \in \bigcup_{k=1}^{l} N_k|\boldsymbol{x})$ for $l = 1,2,3$. Within each antichain N_l, the elements are not comparable, thus the MNL model is used to describe the conditional probabilities. The probability of category 3 in Figure 5 can be fully specified as

$$\begin{aligned} P(Y = 3 \mid \boldsymbol{x}) &= P(Y = 3 \mid Y \in \{2,3,4\};\boldsymbol{x}) \times P(Y \in \{2,3,4\};\boldsymbol{x}) \\ &= P(Y = 3 \mid Y \in N_2;\boldsymbol{x}) \times P(Y \in N_2;\boldsymbol{x}) \end{aligned}$$

where

$$P(Y \in N_2;\boldsymbol{x}) = \frac{\exp(\alpha_2 + \boldsymbol{x}^\top\boldsymbol{\delta})}{1+\exp(\alpha_2 + \boldsymbol{x}^\top\boldsymbol{\delta})} - \frac{\exp(\alpha_1 + \boldsymbol{x}^\top\boldsymbol{\delta})}{1+\exp(\alpha_1 + \boldsymbol{x}^\top\boldsymbol{\delta})},$$

and

$$P(Y = 3 \mid Y \in N_2;\boldsymbol{x}) = \frac{\exp(\alpha_{2,2} + \boldsymbol{x}^\top\boldsymbol{\delta}_{2,2})}{1+\exp(\alpha_{2,1} + \boldsymbol{x}^\top\boldsymbol{\delta}_{2,1}) + \exp(\alpha_{2,2} + \boldsymbol{x}^\top\boldsymbol{\delta}_{2,2})}.$$

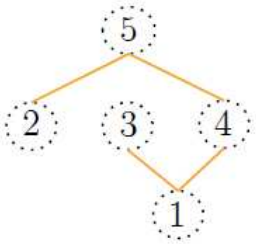

Figure 5: Hasse diagram among 5 categories.

1.4.1 Partitioned conditional generalized linear models for categorical data

Recently, Peyhardi et al. (2016) introduced the PCGLMs to analyze the hierarchical structure of a response with any number of categories. Their methodology is based on

the (r, F, Z) specification of the categorical models and on a partition tree of categories. Using the genericity of the (r, F, Z) GLMs for categorical responses, it is possible to use different link functions and explanatory variables for each partitioning step. Thus, it becomes possible to model hierarchically structured categorical responses, including nominal, ordinal, and partially ordered responses. Formally, a k–PCGLM with J response categories is specified by

- a partition tree $\mathcal{T}$ of $\{1, ..., J\}$ with $\text{card}(\mathcal{V}^*) = k$, where $\mathcal{V}^*$ denote the set of non-terminal vertices of $\mathcal{T}$.
- a collection of binary regression models $\mathcal{C} = \{(r^v, F^v, Z^v) : v \in \mathcal{V}^*\}$ for each conditional probability vector.

The probability for each category j is obtained as

$$P(Y = j \mid \boldsymbol{x}) = P(Y = j \mid Y \in Pa(j), \boldsymbol{x}^{Pa(j)}) \prod_{v \in An(\{j\})} P(Y \in v | Y \in Pa(v), \boldsymbol{x}^{Pa(v)}),$$

where $Pa(v)$ is the parent of v and An are the ancestors of v excluding the root. Figure 25 illustrates the path of the conditional probabilities used to obtain $P(Y = 2 \mid \boldsymbol{x})$ which is expressed as

$$P(Y = 2 \mid \boldsymbol{x}) = P(Y = 2 \mid Y \in \{1, 2, 3\}, (x_1, x_6, x_8)) \times P(Y \in \{1, 2, 3\} \mid (x_4, x_7, x_8))$$

where

$$P(Y \in \{1, 2, 3\} \mid (x_4, x_7, x_8)) = \frac{\exp(\alpha_1 + \boldsymbol{x}^\top \boldsymbol{\delta}_1)}{1 + \exp(\alpha_1 + \boldsymbol{x}^\top \boldsymbol{\delta}_1)}$$

and

$$P(Y = 2 \mid Y \in \{1, 2, 3\}, (x_1, x_6, x_8)) = \left(\frac{1}{2} + \frac{1}{\pi} \arctan(\alpha_{1,2} + \boldsymbol{x}^\top \boldsymbol{\delta}) - \frac{1}{2} + \frac{1}{\pi} \arctan(\alpha_{1,1} + \boldsymbol{x}^\top \boldsymbol{\delta})\right).$$

Due to the hierarchical structure of a PCGLM, the log-likelihood of the model is

$$l = \sum_{v \in \mathcal{V}^*} l^v,$$

where l^v represents the log-likelihood of the GLM for node v. Remark that at each node, a model can be separately estimated since parameters are assumed to be different between nodes.

The sequential (logit) model is a particular case of the PCGLMs. Indeed, it can be represented as a $(J-1)$–PCGLM (see Figure 7) if all of the vertices $v \in \mathcal{V}^*$ share the set of explanatory variables x and the cdf is common (for this case the logistic distribution) for all the non-terminal vertices.

Unkown partition tree The hierarchical structure is partially or even totally unknown in many real data analysis situations. Finding and modeling such a structure

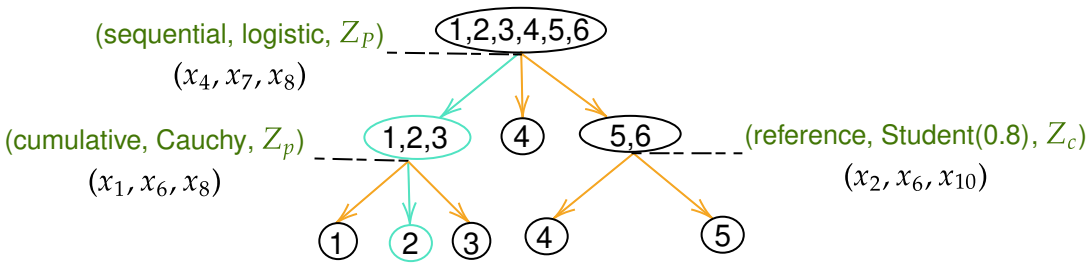

Figure 6: Illustration of a PCGLM with 5 categories where the blue color arrows highlight the path to obtain $y = 2$.

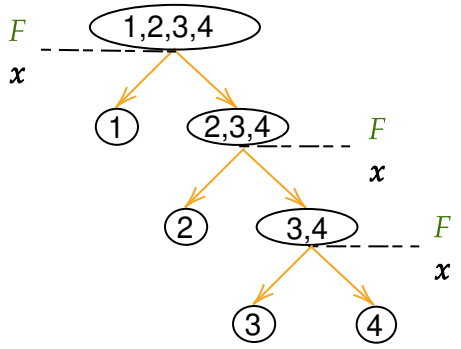

Figure 7: $(J-1)$-PCGLM.

becomes a matter of interest. However, the search for such a structure can be computationally intensive when considering the total number of possible alternatives. In fact, this number increases exponentially with the number of categories, so exploring all the alternatives in an attempt to find the best one becomes relatively ineffective. To reduce somewhat the space of possible trees, one could instead consider only binary trees. A binary tree consists of $J-1$ binary models, one at each non-terminal node of the tree. The essential advantage of this approach is that i) the link is reduced to simply (symmetric) cdfs and ii) the design matrix cannot be constrained, so it is the same (in structure) for all the nodes. The above justifies why binary models are the simplest case in the context of regression for categorical responses. Note that the estimation of this binary tree model (parallelizable) is not a primary concern. The importance then lies in the choice of the link function (equivalently, the choice of the cdf) at each non-terminal node. In chapter 2, we discuss such a choice based on the most common characteristics and/or problems of binary dependent variables.

The main objective in this modeling context is to find a proper representative structure of the presumed hierarchy. In chapter 4, we propose a methodology for finding this tree structure. The methodology is based on the exploration of the dependent variables. From the obtained initial tree, we suggest using operations defined in the space of trees in order to traverse the surrounding neighborhoods searching for a better tree alternative.

Chapter 2

GLMs for Binary Data

Abstract

The link function is the key component of regression models for binary response variables. Despite the diverse potential fits obtained from different link functions, only the logit and the probit links have been widely popularized. Models generated from these links are known to be non-robust in the presence of outliers. We demonstrate that this problem is exacerbated when the two response levels are strongly separated in the explanatory space. To address this shortcoming, we propose and encourage the use of the Student link function. We highlight its robustness to outliers and also to noisy variables, particularly when the data exhibit a strong separation setting.

Keywords: Generalized linear models, Robustness, Link function, Outliers, Noisy variables, Data separation setting

Contents

2.1 Introduction

The present work is focused on GLMs for binary responses whose levels are usually referred to as success and failure (coded respectively as $y = 1$ and $y = 0$). According to Nelder and Wedderburn (1972), a GLM is characterized by three components: the response distribution, the linear predictor, and the link function. In the particular case of binary outcomes, the response distribution is necessarily the Bernoulli. The link function thus plays a central role in the framework of binary regression models. In this paper, we are particularly interested in the link function generated by the Student cdf. In the following, we will use the term Student model to refer to a binary regression whose link function is the inverse of the Student cdf. We aim to compare the robustness of the logit and the Student models by assessing the models' quality, i.e., the goodness-of-fit (through the log-likelihood) and the classification performance (through the prediction accuracy). We will show that the Student model is more robust than the logit model when zeros and ones are well separated according to the explanatory variables.

Silvapulle (1981); Albert and Anderson (1984) showed that under (quasi-) complete separation, finite maximum likelihood estimates do not exist in binary regression models. Moreover, they showed that the maximum likelihood estimates exist and are unique for the logit and the probit models if and only if the data set has some overlap. Unfortunately, separation may go unnoticed since, after some cycles of the iterative fitting process, the log-likelihood curve becomes flat, and either the convergence criteria are met, or simply the predefined maximum number of iterations is reached. Consequently, researchers tend to ignore the problems generated by the separation pattern. The model obtained from a completely separated data set has infinite maximum likelihood estimates since its log-likelihood approaches the maximum value (i.e., zero) suggesting a perfect fit. Imagine now adding one new observation that disturbs the complete separation setting (i.e., a zero among the ones or a one among the zeros). This additional observation will create an overlap, and thus one can obtain both: a unique maximum likelihood estimate and a near-perfect model fit for the data set at hand. In this paper, we relate the number of such disturbing observations to the notion of degree of overlap (the counterpart of the degree of separation). We propose to control (in simulated data) such a degree of overlap by means of the scale parameter of the logistic cdf. We are interested in investigating the impact of different degrees of overlap on the model's quality (in terms of fit and prediction), especially at analyzing the model's sensitivity to common perturbations like the presence of outliers.

Robust regression is a class of statistical methods that has the property of reducing the sensitivity of the parameter estimates to perturbations in the data set. It has been shown that logit and probit models are sensitive to outliers; hence, these links are not considered robust alternatives (see Pregibon, 1982; Copas, 1988). Furthermore, it is believed that due to the similarities between the logit and the probit links, no other link function would produce substantial improvements to the model fit (Koenker and Yoon, 2009), so little effort has been devoted to analyzing the robustness of other links. Notable exceptions are authors like Lange et al. (1989); Liu (2005) who have recognized the links obtained with the Cauchy and the Student cdfs as robust alternatives in the presence of outliers. They have shown that the influence function (see Hampel, 1974)

of a binary model is unbounded for the logit and the probit link functions and bounded for the Student and the cauchit link functions. Indeed, the Cauchy and the Student distribution (with a low degree of freedom, i.e., $\nu \leq 1$) are very heavy-tailed, and they make less extreme predictions (further away from 0 and 1) than the normal or logistic distributions for large values of the linear predictor (Smithson and Verkuilen, 2006).

Additionally to the presence of outliers, a model perturbation can be caused by noisy variables, which would lead to over-fitting unless they are removed from the model. Using a benchmark data set (with four response categories), Peyhardi (2020) empirically showed that the Student model is less sensitive to noisy variables than the logit model. As part of the robustness assessment of the Student link function, we also evaluate and compare the sensitivity to noisy variables of the logistic and the Student regression models.

Considering together the separation and the robustness characteristics of binary regression models, we aim to investigate whether the sensitivity of the model is tied to a certain extent to the degree of overlap of the data set. This will allow us to identify the most robust model (between the logit and the Student) for particular conditions of the data set. For this purpose, we investigate the sensitivity to outliers and noisy variables on simulated data under four different degrees of overlap: one low, two intermediate (which we consider encompassing most of the real data analyses), and one high. To make use of the Student link, we propose an algorithm to estimate the unknown degree of freedom. This will enable us, as a byproduct, to get an indicator of the degree of overlap.

This chapter is structured as follows. In section 2.2 we introduce the basics of the Student regression model. In section 2.3, we present the classical notion of separation, and we extend this formulation by introducing the notion of degree of overlap. In section 2.4 and 2.5, we illustrate the robustness of the Student link through various simulations with different separation scenarios and according to the two settings 1) in the presence of outliers and 2) in the presence of noisy variables.

2.2 Student link function

Binary regression models can be motivated (although it is not a strict model requirement) by the assumption of a latent (unobserved) variable $\tilde{y}$, for which there exists a threshold that defines whether the original observed variable y is 0 or 1. Let the model for the latent variable be $\tilde{y} = \alpha + \boldsymbol{x}^t\boldsymbol{\delta} + \varepsilon$, where $\boldsymbol{x} = (x_1, \dots, x_p)$ denotes the covariate vector, $\alpha \in \mathbb{R}$ the intercept, $\boldsymbol{\delta} \in \mathbb{R}^p$ the slope parameter vector, and ε the latent residual. Considering y as a dichotomized version of $\tilde{y}$ we obtain

$$\pi(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = P(\tilde{y} \geq 0|\boldsymbol{x}) = P(\alpha + \boldsymbol{x}^t\boldsymbol{\delta} + \varepsilon \geq 0) = P(-\varepsilon \leq \alpha + \boldsymbol{x}^t\boldsymbol{\delta}).$$

Allowing $-\varepsilon$ to have the distribution function F, we get the simple form

$$\pi(\boldsymbol{x}) = F(\alpha + \boldsymbol{x}^t\boldsymbol{\delta}).$$

From this expression, we can identify the linear predictor as $\eta = \alpha + \boldsymbol{x}^t\boldsymbol{\delta}$ and the link

function as the inverse cdf F^{-1}. The most commonly cdfs used for the link function specification are the logistic and the normal cdfs, which yield the logit and probit models, respectively. In this work, our interest lies particularly in the Student link function. The cdf of the Student distribution is given by

$$F_\nu(\eta) = \frac{1}{2} + \frac{\eta\Gamma\left(\frac{\nu+1}{2}\right){}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)},$$

where ${}_2F_1$ denotes the hypergeometric function. The Student cdf is remarkably versatile, particularly in the GLM context, since it is possible to obtain the three most popular link functions by specifying different values of ν. The cauchit link is derived immediately as it corresponds to a particular case of the Student distribution. Moreover, authors such as Mudholkar and George (1978); Liu (2005) have illustrated the approximations of the Student link to the logit and the probit links. All in all, at varying ν we have:

- $\nu = 1 \Rightarrow F_\nu =$ Cauchy,
- $\nu \in (7, 9) \Rightarrow F_\nu \simeq$ logistic, and
- $\nu \to \infty \Rightarrow F_\nu =$ normal.

Theoretically, the degree of freedom of the Student distribution can be any real greater than 0. Some authors, including Lange et al. (1989); Peyhardi (2020) have discretized ν along a grid to obtain a log-likelihood profile. From those profiles, one can identify and keep the $\hat{\nu}$ with the highest log-likelihood value, $l(\hat{\nu})$. In several experiments on data sets and in the papers mentioned above, we observed that the log-likelihood curve becomes flat from $\nu = 1$ and up to the highest values of ν. Thus, one can notice almost no difference, for instance, between $l(8)$ and $l(30)$, a fact that can be related to the proximity between the logit and the probit links. On the other hand, different shapes of the log-likelihood were observed when $\nu < 1$.

We propose examining these log-likelihood profiles through a simulation based on the above information. We generated data sets with $n = 100$ observations of a binary model with the true response function having a sigmoidal form given by the Student distribution where $\nu^* \in (0.3, 1.5, 6)$. The components of the linear predictor are $\alpha = 0$, $\delta = 1$, and the covariate is drawn from a normal distribution $x \sim \mathcal{N}(0, 1)$. For this illustration, we adopt the strategy of fitting all the models within a grid from 0.25 to 8 by step = 0.05. Note that in this study, the minimum value considered for ν was 0.25 since the pdf and the cdf of the Student distribution are likely to encounter evaluation problems whenever ν is quite close to zero (see Van der Paal, 2014). In the scenarios described above and illustrated in Figure 8, we can distinguish essentially 3 types of curves: a first one where the highest log-likelihood value occurs at the smallest values of the grid ($\nu < 0.5$) and then, after $\nu > 1$ it stabilizes at a lower log-likelihood (left panel of Figure 8), a second one in which the maximum log-likelihood value falls approximately within the range $(0.2, 0.8)$ (middle panel of Figure 8), and a third one that starts with a low log-likelihood value and then increases rapidly until it plateaus when $\nu > 1$ (right panel of Figure 8). As it becomes evident, the logit or even the probit link would

unquestionably be the preferred choice in the last data setting. However, this is not the case for the other two scenarios, where visually, one should be encouraged to search for the best fit, narrowing the search range to values of ν less than 1.

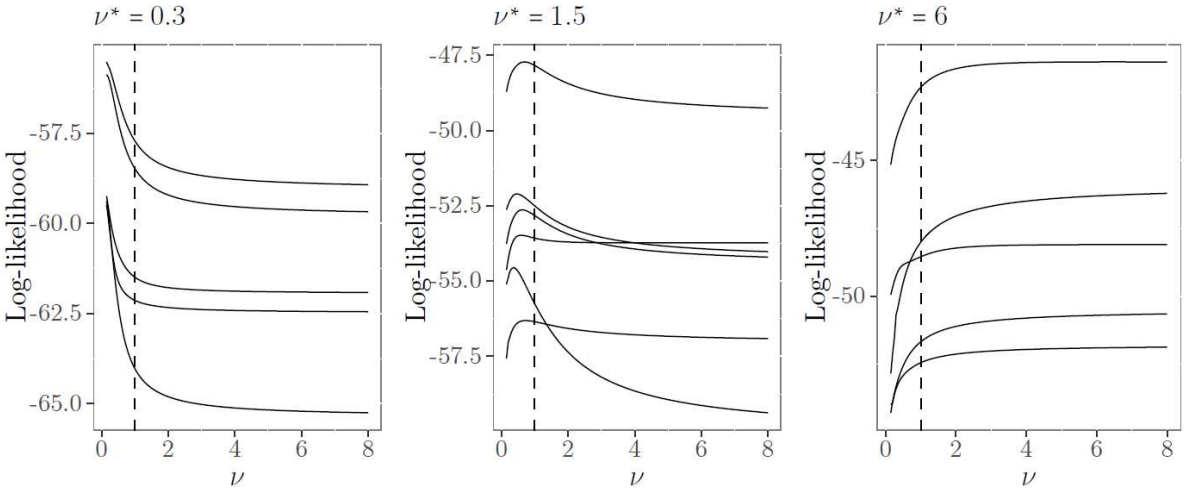

Figure 8: Log-likelihood profiles generated by different theoretical values of the degree of freedom of the Student distribution. The vertical dotted line cuts the log-likelihood profiles at the fit given by the Student(1) link function.

Considering the time-consuming computation for estimating all models within a grid and based on the above observations, we propose in Algorithm 1 a heuristic to find the $\hat{\nu}$ that best fits the data in terms of log-likelihood.

Algorithm 1: Heuristic to find the link function of a binary model.

Estimate the models with Student link where $\nu = 1$ and $\nu = 8$.
if $l_{\nu=8} > l_{\nu=1}$ **then**
  Estimate the log-likelihood l_p of a binary model with the probit link.
  if $l_p > l_{\nu=8}$ **then**
    Use the probit link.
  else
    Use the logit link.
  end
else
  Use an optimization algorithm to find the best $\nu \in (0.25, 1)$ of the Student cdf.
end

Through the built-in R function optimize(), we use as optimization algorithm (in the last step of Algorithm 1) a combination of the golden section search and the successive parabolic interpolation. This combination results in a convenient option due to the reliability of the golden section search and the fast convergence given by the parabolic interpolation (refer to Vit, 1985; Renk et al., 2009, for more details). Of course, this is not the only option. Other optimization algorithms can be implemented to search the degree of freedom in the proposed interval $(0.25, 1)$ (see Brent, 2013).

2.3 Separation of data

2.3.1 Complete separation

Formally, Albert and Anderson (1984) said that there is a complete separation of the n sample points if there exists a vector $\boldsymbol{\beta}^t = (\alpha, \boldsymbol{\delta}^t) \in \mathbb{R}^{p+1}$ that properly allocates (or predicts) all observations to their group, i.e. $\boldsymbol{z}^t\boldsymbol{\beta} > 0$ when $y = 1$, $\boldsymbol{z}^t\boldsymbol{\beta} < 0$ when $y = 0$, where $\boldsymbol{z}^t = (1, \boldsymbol{x}^t)$. If the sample points are not completely separated, they can be quasi-completely separated, if $\boldsymbol{z}^t\boldsymbol{\beta} \geq 0$ when $y = 1$, $\boldsymbol{z}^t\boldsymbol{\beta} \leq 0$ when $y = 0$, and with some points with either $y_i = 0$ or $y_i = 1$ when $\boldsymbol{z}_i^t\boldsymbol{\beta} = 0$. If the sample points exhibit neither complete separation nor quasi-complete separation, the points are said to overlap. Figure 9 shows the three classical separation settings. Note that even if only one line is plotted in the left plot, an infinity of lines can perfectly separate the data suggesting a complete separation pattern. In the middle plot, three points lie on the line that best separates the data; therefore, the data set exhibits a quasi-complete separation. The right plot shows an overlap configuration since no line separates the two response levels perfectly.

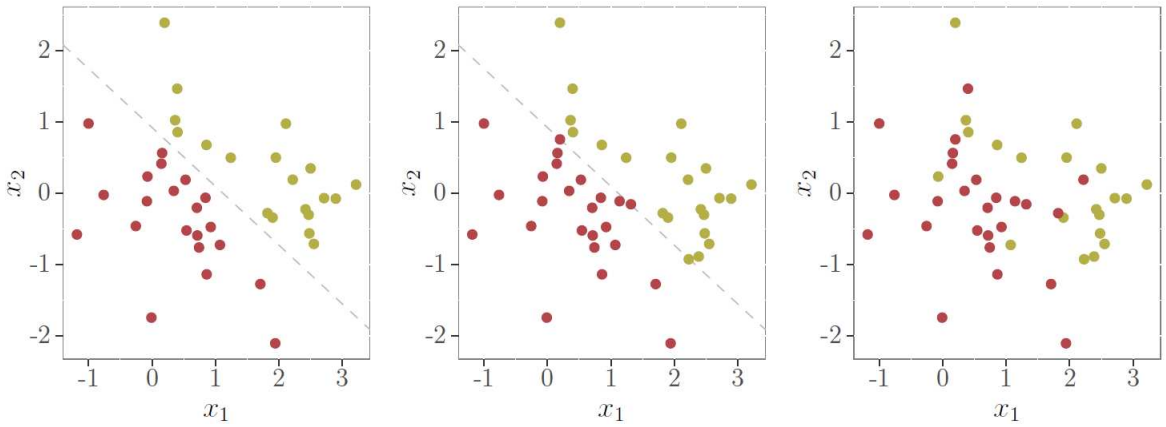

Figure 9: Possible sample points settings of a binary response (the green and red colors represent the response levels) to be explained by two explanatory variables x_1 and x_2.

Silvapulle (1981); Albert and Anderson (1984), demonstrated that overlap is necessary for the parameters of a binary regression model to be identifiable. A geometrical interpretation of this result, given by Christmann and Rousseeuw (2001), is that the maximum likelihood estimate exists if and only if no hyperplane perfectly separates the levels of the response, where the hyperplane itself may contain both ones and zeros. Indeed, let $\beta(k) = k\boldsymbol{\beta}$ for $\boldsymbol{\beta} \in A^c$ where $k > 0$, and A^c is the set of vectors satisfying the complete separation configuration. The log-likelihood as a function of $\beta(k)$ is then given by

$$l(\beta(k)) = \sum_{i=1}^{n} y_i \log(F(\boldsymbol{z_i}^t\beta(k))) + (1 - y_i)\log(1 - F(\boldsymbol{z_i}^t\beta(k)). \tag{2.1}$$

When $k \to \infty$, the cdf F returns a value close to 1 and 0 respectively for $y = 1$ and $y = 0$, therefore, Equation (2.1) tends towards a sum of n logs of one. That is how the

log-likelihood results roughly around zero (the global maximum) while the norm $\|\beta(k)\|$ approaches infinity.

2.3.2 Degree of overlap

Despite the non-convergence of the MLE when data is (quasi-) completely separated, there exists an infinite number of models with perfect prediction accuracies as the estimated probability $\hat{\pi}$ will result to be close to 1 or 0 for the n observed responses. In other words, the overlap is necessary for the existence and unicity of the MLE, but the (quasi-) complete separation of data is desirable for a good class prediction accuracy. Therefore, an ideal configuration of the data would be to overlap to a level near the complete separation. To concretize the idea, imagine a completely separated data set with 99 observations, then add a zero among the region of ones (not so far away from the zeros). In this case, the MLE exists, is unique, and if the new observation does not perturb the model too much, the proportion of correctly classified observations would be 0.99.

In the following, we aim to study the impact of adding outliers on the model's prediction accuracy and the model's log-likelihood. We will compare the performances of the Student link versus the logit link on simulated data sets with a particular interest in the overlap close to the complete separation. To perform this comparison, we need to introduce the degree of overlap (denoted in the following by d), emphasizing that the lower the degree of overlap, the closer the data configuration to complete separation. To account for different data settings, we use the logistic cdf

$$\frac{1}{1+\exp\left(-\eta/d\right)}, \tag{2.2}$$

by letting the shape parameter d to vary. In Figure 10, we illustrate four degrees of overlap considering only one explanatory variable represented in the x-axis. The parameters of the linear predictor are $\alpha = 0$ and $\delta = 1$, and x is derived from a Normal distribution. As d increases, more values of π are expected to be closer to 0.5. Thus, the overlap interval (represented by the dotted lines for respectively the minimum and maximum of each level) gets wider according to the size of d. For our following assessments on the model's robustness, we compare the Student link to the logit link according to three scenarios: a low overlap $(d = 0.05)$, standard degrees of overlap $(d \in \{0.1, 0.3\})$, and a high degree of overlap $(d = 0.8)$.

2.4 Robustness of the Student model to outliers

The most common concern of robust methods is to reduce the influence of outliers. As Copas (1988) outlined, an outlier in a binary response invokes one or both of two interpretations, a geometric one in which the point is far from the bulk of the data, and a probabilistic one in which, if the fitted model were true, the offending value of y would be most unlikely to occur, i.e., either $y = 1$ and π being close to 0, or $y = 0$ and π being close to 1. In the following, we present an illustrative example and a theoretical

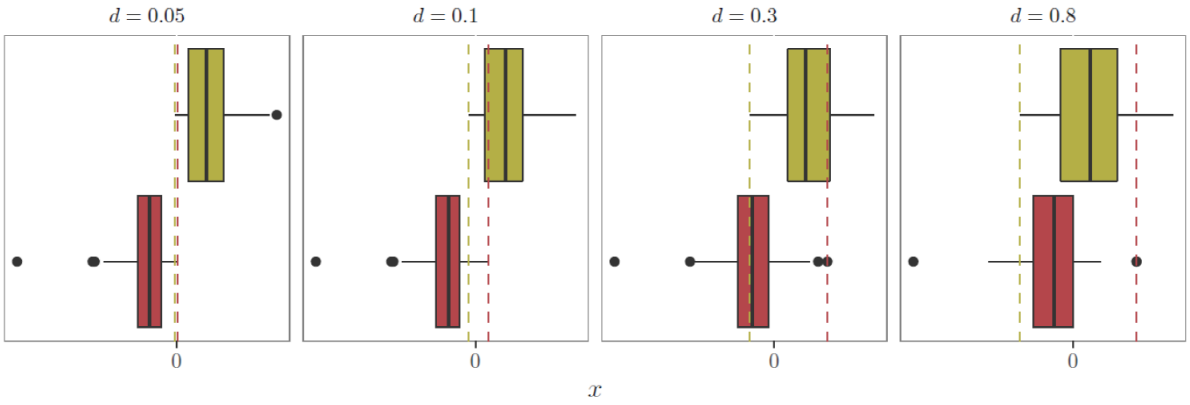

Figure 10: Simulated overlap scenarios of a binary response variable with one explanatory variable x. The colors green and red represent respectively the response levels $y = 1$ and $y = 0$. The dotted lines are drawn at the maximum value of x when $y = 0$ and at the minimum when $y = 1$.

justification of the robustness to outliers given by the Student link in contrast to the logit link. Furthermore, through simulations, we will show that this robustness is closely tied to the degree of overlap of the data.

2.4.1 Illustrative example

We illustrate the robustness of the Student model on the well-known vaso-constriction data set, already studied by authors including Finney (1947); Pregibon (1981). These authors highlighted the non-robustness of the maximum likelihood estimators in the context of the logistic regression model. We aim to contrast their results with our proposal that consists of a binary model using the Student link with a small ν. We set $\nu = 0.6$ since, in experiments with a separation setting like the current one, we have observed the appropriateness of a degree of freedom around this value. The binary outcome occurrence or non-occurrence of a reflex (*vaso-constriction*) of the skin of the digits after air inspiration is explained by two explanatory variables: x_1 the volume of air inspired and x_2 the inspiration rate (both represented in logarithms). To assess the effect of one outlier on the maximum likelihood estimator, we added one point with coordinates $(x_1^*, x_2^*, y^*) = (s, s, 1)$ to the original sample. This point will be located on different positions of the non-continuous line of Figure 11. Note that the observation would not be unusual for $s > 0$ because it would fall within the apparent space of the vaso-constriction presence ($y = 1$); for high values of s, the point is considered to be a leverage point, and it should have relatively little influence on the estimated model. In the opposite direction, i.e., $s < 0$, this observation gradually becomes an outlier, also called a bad leverage point or contamination (see Copas, 1988). Based on the 40 data points, we computed the estimated coefficients $\hat{\boldsymbol{\beta}}$ and the corresponding accuracies.

We can notice from Figure 12a that there is a clear distinction between the estimated coefficients using the logistic cdf and the Student (0.6) cdf. Both, $\hat{\delta}_1$ and $\hat{\delta}_2$, move far away from 0 using the Student (0.6) distribution when $s > -7$. In contrast, the estimates

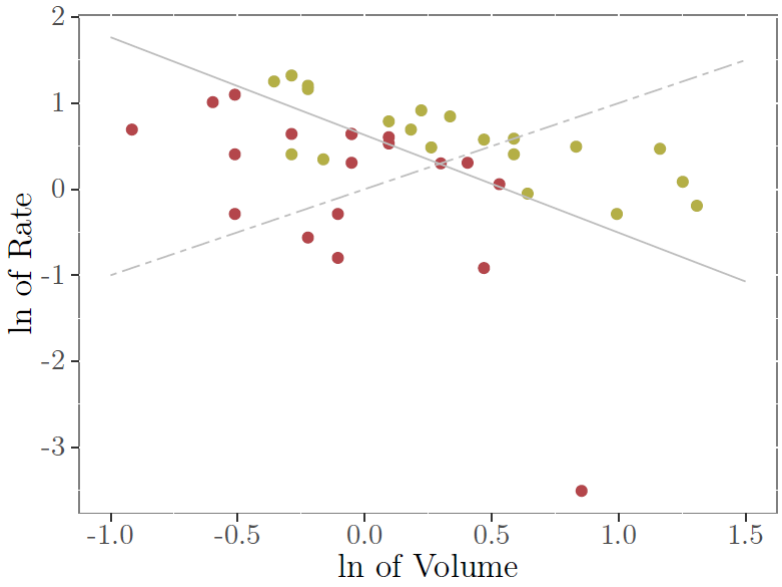

Figure 11: Scatterplot for the vaso-constriciton data set; the non-continuous line represents the position of the added observation $(s, s, 1)$. The green and red points represent the occurrence and non-occurrence of the vaso-constriction, respectively. The solid line represents the differentiation between the categories given by the logistic model.

using the logistic link are close to 0 in a broader range (approximately when $s < -3$). Hence, the obtained estimations using Student (0.6) link deteriorate less quickly as the outlier moves away. As for the number of correctly classified observations (see Figure 12b), the Student (0.6) cdf outperforms the accuracy obtained using the logistic link in almost the entire range considered for s. Another point worth noting is that the magnitude of s seems to have a constant impact on the percentage of correctly classified points for the logistic link accuracy curve. In contrast, the Student curve has a steady state after each of the few jumps along s.

Given the particular behavior of s in the above examples, let us now consider the logarithm of the volume as the sole explanatory variable. To illustrate the influence of the magnitude of x on the model fit, we added three points (one at a time) at different positions of x. Figure 13 shows the different curves obtained using the logistic and the Student(0.6) cdfs. Note that the original observations and their fitting curve are plotted in black for both links. The first point, $(x^*, y^*) = (-1, 1)$, is represented by the red color. We can observe that the red line does not diverge far from the black line for the logit link (left plot of Figure 13). However, this is not the case when considering $x^* = -3$ or $x^* = -5$ where we can see the $s-shape$ vanishing towards a straight line with decreasing slope. Note that the point at which the fit intercepts the horizontal line $\pi = 0.5$ (represented with the dashed gray line) decreases as the value of $|x^*|$ increases, implying an increase of miss-classified points. On the other hand, the curves representing the fits using the Student (0.6) cdf (right plot of Figure 13) are only slightly perturbed by the magnitude of x^*. One can notice that the three curves are on top of each other in the complete range of x and close to the fit without adding new observations. This example illustrates

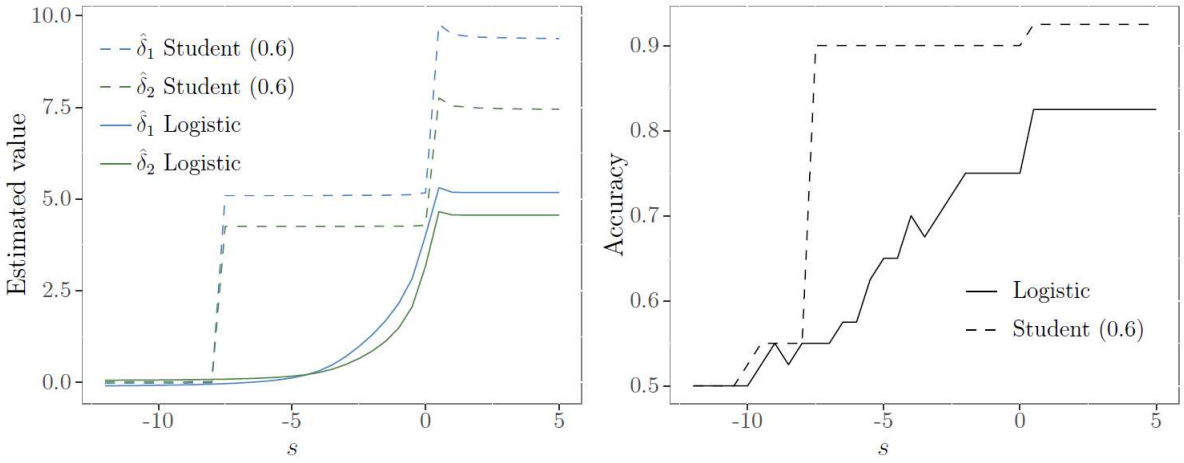

Figure 12: Figure 12a represents the estimators as a function of s (x-axis) using the logistic link (solid lines) and normalized estimators using the Student (0.6) link (dashed lines). Figure 12b represents the accuracies when adding the point $(s, s, 1)$ to the data sets

that the Student link is more stable than the logistic link to the type of perturbation generated by adding an observation near the edge of the design space with the erroneous response level.

2.4.2 Influence function

The maximum likelihood estimate of the canonical GLM is known to be sensitive to outliers because the score function is unbounded. Hence, as an alternative, one can use the Student distribution, which results in score equations with bounded influence for regression models (see Pinheiro et al., 2001; Lange et al., 1989). The IF formalizes the bias caused by one outlier (Hampel et al., 2011). In the context of GLMs, the IF of a new observation $(y^*, \boldsymbol{x}^*)$ on the MLE, is given by

$$IF[(y^*, \boldsymbol{x}^*), \hat{\beta}] = \left\{ \mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} \right)_{\beta=\hat{\beta}} \right\}^{-1} \left(\frac{\partial l^*}{\partial \boldsymbol{\beta}} \right)_{\beta=\hat{\beta}}, \tag{2.3}$$

where the log-likelihood computed for the original data set $\{(y, \boldsymbol{x})\}_{i=1,\ldots,n}$ and for the new observation $(y^*, \boldsymbol{x}^*)$ are denoted by l and l^* respectively (see Künsch et al., 1989). Given the observed data, the left factor on the right-hand side of Equation (2.3) does not depend on the new observation. Thus, it is sufficient to analyze the bounding of the new observation's score (right factor) according to $(y^*, \boldsymbol{x}^*) \in \{0, 1\} \times \mathbb{R}^p$. Without loss of generality, we focus on one coordinate of this vector, i.e., for $k \in 1, \ldots, p$ we evaluate the following upper bound

$$\sup_{y \in \{0,1\}, \boldsymbol{x} \in \mathbb{R}^p} x_k \frac{f(\eta)}{F(\eta)(1 - F(\eta))}(y - \pi).$$

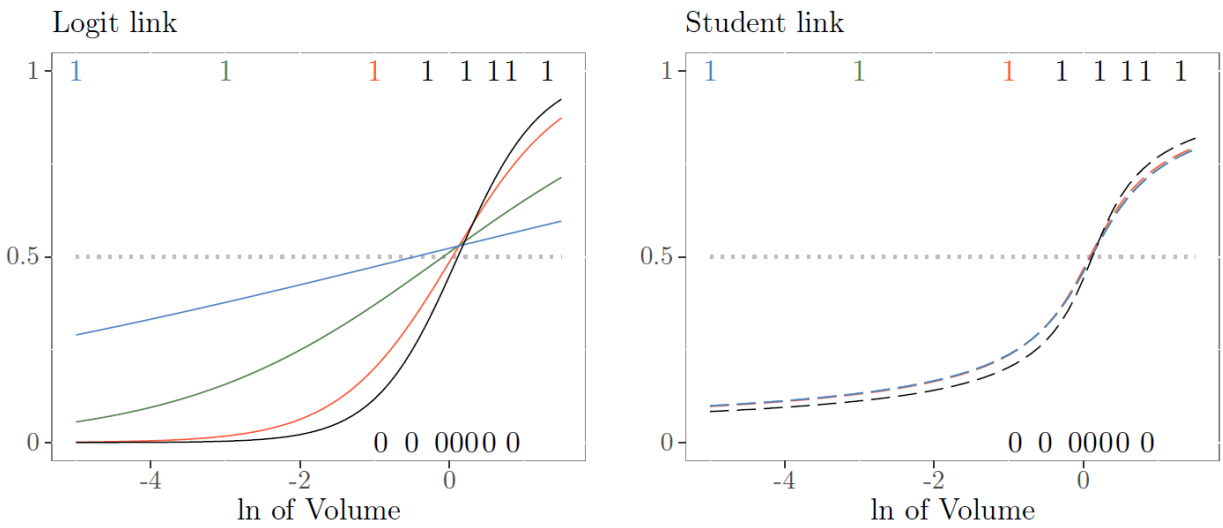

Figure 13: Fitting curves of the logistic and the Student(0.6) links. Colors correspond to the fit when adding one by one the points $(x^*, y^*) = \{(-1, 1), (-3, 1), (-5, 1)\}$ to the original observations.

Remark that $(y - \pi)$ is bounded since it lies in $(-1, 1)$. Therefore, it is sufficient to demonstrate that the following simplified function

$$\frac{\eta f(\eta)}{F(\eta)(1 - F(\eta))}, \tag{2.4}$$

is bounded. We need to study the behavior when $\eta \to +\infty$

$$\lim_{\eta \to +\infty} \frac{\eta f(\eta)}{F(\eta)(1 - F(\eta))} = \lim_{\eta \to +\infty} \frac{\eta f(\eta)}{1 - F(\eta)}. \tag{2.5}$$

By symmetry, this behavior is the same when $\eta \to -\infty$. Equation (2.5) results in the following values for each distribution:

- logistic cdf: η,
- normal cdf: η^2,
- Student(ν) cdf: ν,

(see Liu, 2005, for further details).

In Figure (14), we represent Equation 2.4 for the logistic, the normal, the Cauchy (represented as the Student(1)), and the Student distribution with different values of its degree of freedom. We can observe that while extreme values of η strongly impact the logistic and the normal distributions, the curves of the Student distributions are hardly disturbed. The cauchit link has already been recognized as an attractive option in the presence of outliers (see Koenker and Yoon, 2009). Being just a case of the Student distribution ($\nu = 1$), we intend to broaden this notion by proposing the family of links resulting from the Student (ν) (particularly, with a small ν) as a robust alternative in the framework of GLMs for binary responses.

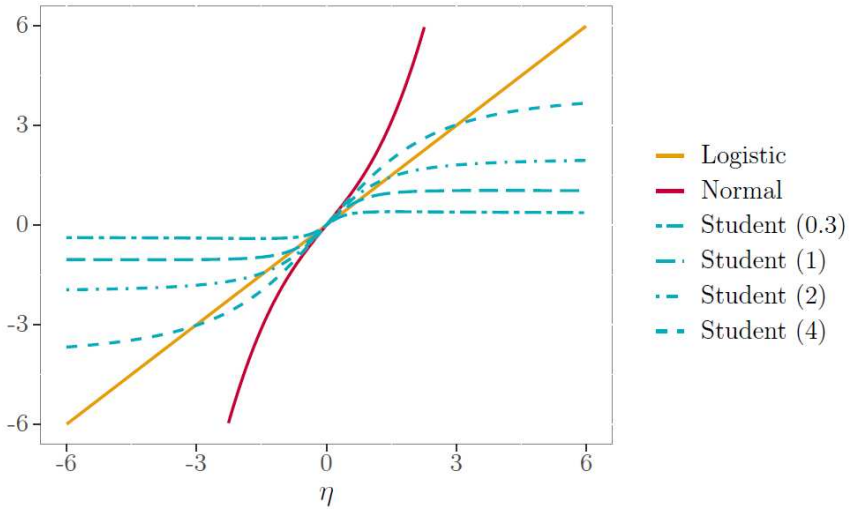

Figure 14: Representation of the boundedness of the influence function (Equation 2.4) for different link functions.

2.4.3 Impact of the overlap: a simulation study

Based on the simulation exercise designed to analyze the separation of the response levels (see section 2.3), we aim now to study the influence of outliers in the model fit at different levels of overlap. In this simulation, we consider the percentage of outliers τ to vary between 0% and 20%, and we generate 100 repetitions for each instance where $n_{training}$ = 100 and $n_{testing}$ = 30. In the previous section, we used a fixed value of ν to illustrate the robustness of the link function generated by the Student distribution. In the following, we estimate the degree of freedom ν for each model.

We present in Figure 15 the log-likelihood box-plots (of the 100 simulations) of the Student and the logit models. We can observe that the results differ notably around 6% of outliers in the sample and anywhere in the range between 1% and 15%. There are no major differences between the models outside this range, i.e., when there is no outlier, or more than 18%. These results are particularly noticeable for lower degrees of overlap, so that, the lower the overlap degree, the greater the difference between the two models. Since the response levels are highly combined as the degree of overlap increases to d = 0.8, the values intended as outliers are no longer considered as such. Therefore, the differences with respect to the canonical link fade out as the degree of overlap increases.

In real problems with binary responses, we can expect the percentage of outliers to range from 1% to 10% of the total of observations and the degree of overlap to be similar to the setting we defined when d = 0.1. And it is precisely for these situations that the Student's link largely prevails over the logistic link. Also, and not surprisingly, the percentages of correctly classified observations in the training set have similar profiles to the reported log-likelihoods; see Figure 41 in Appendix.

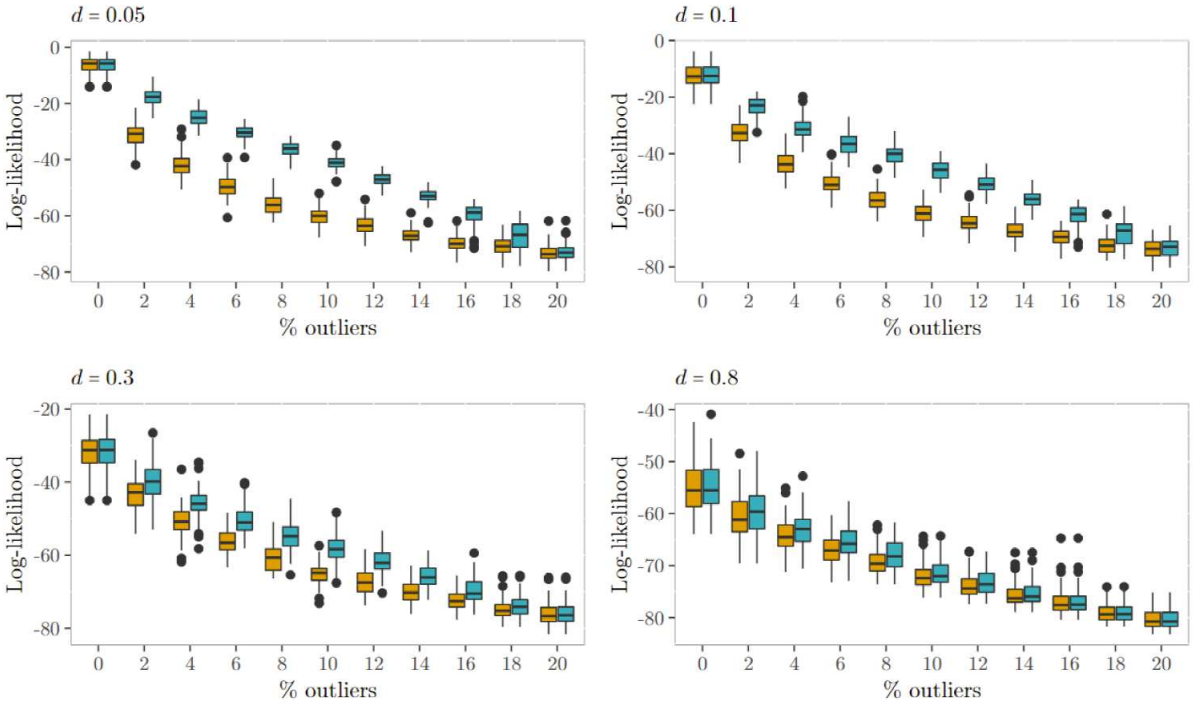

Figure 15: Log-likelihood boxplots for different percentages of outliers (x-axis) and for different overlap degrees of the data set (from $d = 0.05$ to $d = 0.8$), of the fits obtained from the logit model (yellow box-plots) and the Student model (blue box-plots).

2.5 Robustness of the Student model to noisy variables

When statistical models contain many parameters, there is a risk of overfitting the specific data set at hand. For this reason, analysts need to detect those parameters that are important and those that are not. In this section, we demonstrate how the logit model is more sensitive to noisy variables (in the presence of outliers) than the Student model. We show an illustrative example to clarify some initial notions, and then we present a simulation study to compare the robustness of the Student and logit models.

2.5.1 Illustrative example

Consider a binary response resulting from the dichomotization of $F(\eta)$ where $\eta = x_1$, i.e., $\alpha = 0$ and $\delta_1 = 1$, and F is the logistic cdf with $d = 0.1$, i.e., a low degree of overlap (refer to Equation 2.2). In the following, as in the previous section, we fix to 0.6 the degree of freedom of the Student link. We represent this scenario in the top-left graph of Figure 16 where the vertical line indicates the threshold $-\hat{\alpha}/\hat{\delta}_1$ (for which the predicted probability is equal to 1/2) of the Student and the logit models. Due to the absence of outliers, the fitted model resulting from the logit link is roughly the same as the one with the Student(0.6) link. In the top-right plot of Figure 16, we added a noisy variable (whose associated slope parameter is equal to zero) x_2, represented on the y-axis. As

expected, the estimated slope $\hat{\delta}_2$ is close to zero; hence, the lines are quasi-vertical, i.e., the noisy variable slightly perturbs the fit in the same way for both links. Based on the previous section, if we add a few outliers (see the bottom-left graph of Figure 16), we expect the logistic link to be more affected than the Student link. Precisely, this anticipated result is visible in the graph as the separation line for the logit model moves towards the outliers. In contrast, the Student(0,6) separation line remains at almost the same position as without outliers. Note that in this case, outliers only concern the discriminant variable with extreme values on x_1 and a value near zero on x_2. In the bottom-right plot of Figure 16, we add outliers in both directions, the discriminant and the noisy variable. We observe that the line of the logistic fit bends in the direction of the outliers, which indicates that the logistic link gives greater weight to these outliers. The consequence is that the logit model allows the noisy variable to influence the global fit unlike the Student model.

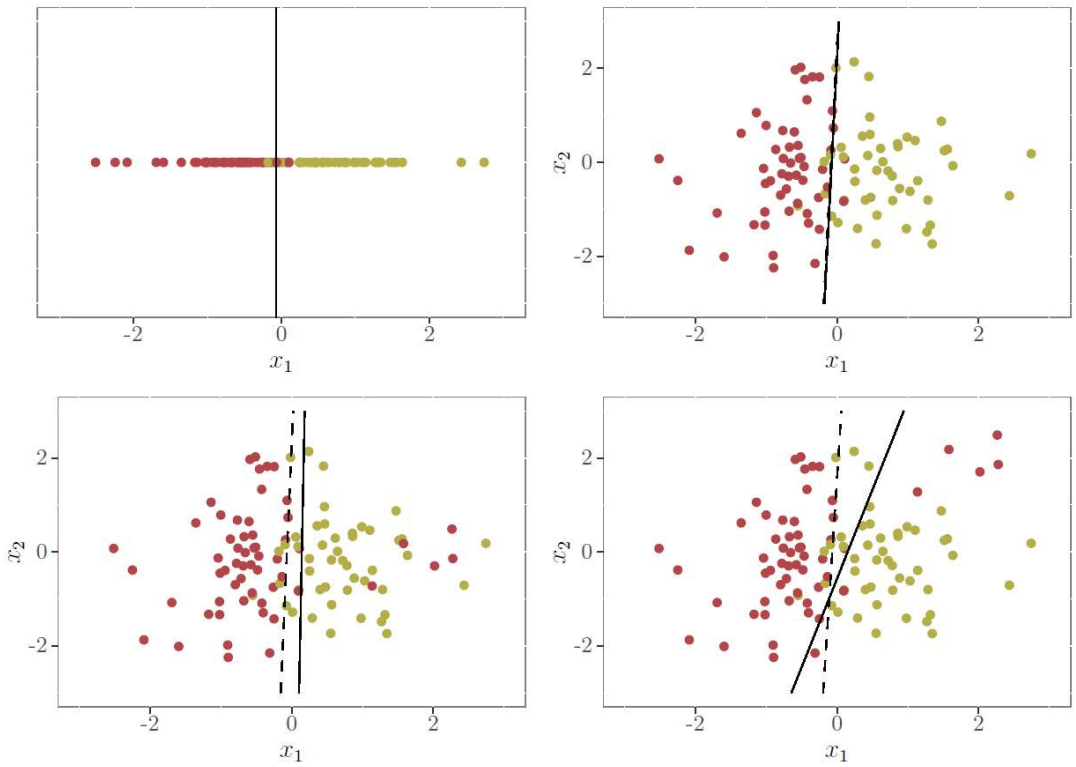

Figure 16: The dotted and solid lines refer respectively to the classification induced by the Student(0.6) and the logit models. In the top-left plot there is one explanatory variable (represented on the x-axis); the vertical line indicates the point given by $-\hat{\alpha}/\hat{\delta}_1$. In the top-right plot, the y-axis represents an added noisy variable x_2, and the lines are given by the discriminating hyperplane: $x_2 = \dfrac{-\hat{\alpha} - x_1\hat{\delta}_1}{\hat{\delta}_2}$. In the bottom-left plot, we added outliers only in the x_1 direction, while on the bottom-right, we added outliers in both directions of x_1 and x_2.

2.5.2 Simulation study

In the following, we aim to show that, compared to the logistic link, the Student link is less perturbed by noisy variables (according to the number of outliers), especially when the degree of overlap is low. For this, we evaluate the frequencies of selected noisy variables. For this study, we define p_1 discriminant variables, $x_1, \ldots, x_{p_1}$, and p_2 noisy variables $x_{p_1+1}, \ldots, x_{p_1+p_2}$, as follows: a candidate variable is a discriminant variable if its coefficient in the regression equation $\pi = F(\eta)$ is nonzero, a noisy variable otherwise. We define the parameters vector for different number of predictors as follows:

- $p_1 = 1$, $p_2 = 1$: $\boldsymbol{\delta}^t = (1, 0)$
- $p_1 = 2$, $p_2 = 2$: $\boldsymbol{\delta}^t = (0.8, -0.6, 0, 0)$
- $p_1 = 4$, $p_2 = 2$: $\boldsymbol{\delta}^t = (0.8, 0.4, -0.4, 0.2, 0, 0)$
- $p_1 = 4$, $p_2 = 4$: $\boldsymbol{\delta}^t = (0.8, 0.4, -0.4, 0.2, 0, 0, 0, 0)$

where the covariates matrix was generated as a $x \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$ with $\Sigma = \{\sigma_{ij}\}_{i,j\in\{1,\ldots,p\}}$, where $i \neq j$, $\sigma_{ij} = 0$ and $\sigma_{ii} = 1$. We conducted 100 simulation runs for each of the settings with a sample size $n = 100$ and we consider a varying overlap of the data set from $d = 0.05$, 0.1, 0.3 to 0.8. The outliers were placed in both directions of the discriminant variables (with $x_k \sim \mathcal{N}(2\delta_k, 0.5)$ for $k = 1, \ldots, p_1$) and the noisy variables (with $x_k \sim \mathcal{N}(2, 0.5)$ for $k = p_1 + 1, \ldots, p_1 + p_2$). For example, when $\boldsymbol{\delta}^t = (0.8, -0.6, 0, 0)$, n_{out} points were added where $x_1 \sim \mathcal{N}(1.6, 0.5)$, $x_2 \sim \mathcal{N}(-1.2, 0.5)$, $x_3 \sim \mathcal{N}(2, 0.5)$, and $x_4 \sim \mathcal{N}(2, 0.5)$. The summary measure of the method's performance was the average times the variable selection procedure (in this case, the stepwise-backward algorithm) retained the discriminant and the noisy variables.

In Figure 17, we represent the average number of times the Student and the logit model selected the discriminant or the noisy variables when $\boldsymbol{\delta}^t = (1, 0)$. It can be observed that the number of times the discriminant variable is retained is similar for small degrees of overlap but slightly different for $d = 0.8$, where the Student link outperforms the logit link after 5% of outliers. For the noisy variable, the discrepancies are pronounced at small degrees of overlap, notably between 2% and 15% of outliers. For the low overlap cases, $d = 0.05$ and $d = 0.1$, the Student link selects less than 10% of the times the noisy variable when there are less than 10% of outliers, while the logistic selection curve grows rapidly and selects up to 100% of the times the noisy variable much earlier than Student does.

We report in Appendix A.2 the corresponding plots of the previously defined scenarios for $p = 4$, 6, and 8. We note that the more covariates fewer differences exist between the two links. But even in those cases, the Student link excels in retaining the discriminant variables and rejecting the noisy variables, especially at lower degrees of overlap.

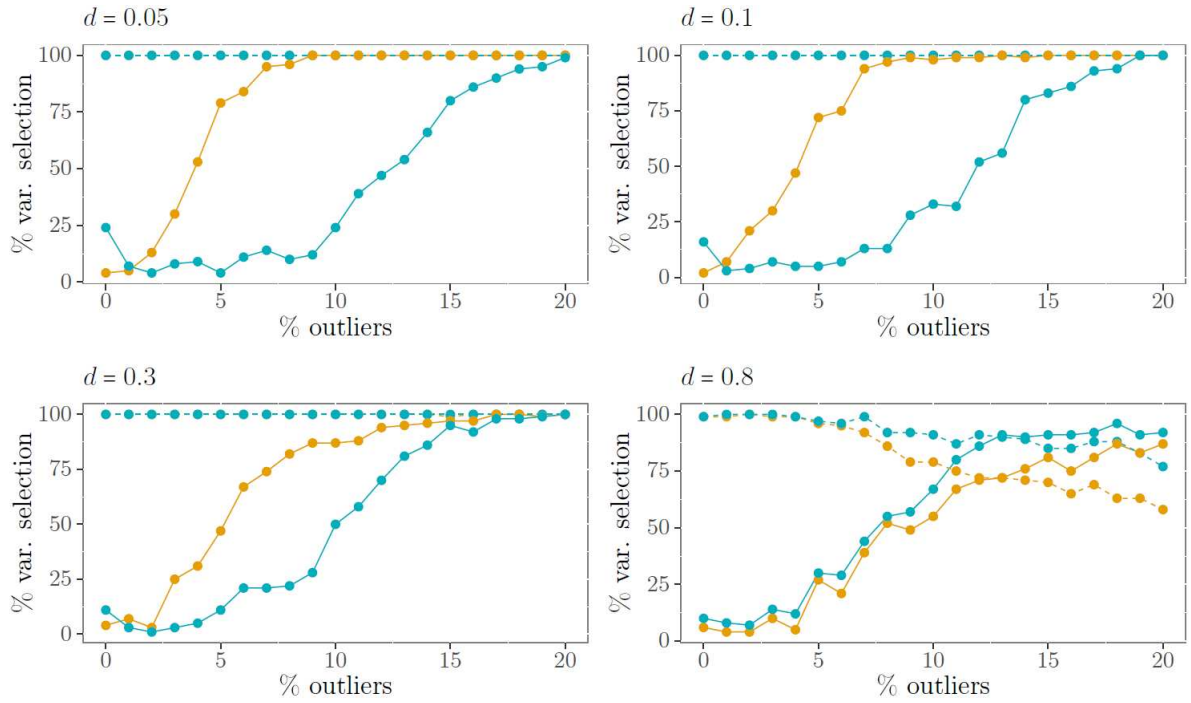

Figure 17: Average number of times the stepwise (backward) algorithm selects discriminant variables (dashed lines) and noisy variables (solid lines) where $\boldsymbol{\delta}^t = (1, 0)$. The blue and yellow colors respectively represent the Student and logistic links.

2.6 Conclusion and perspectives

In this paper, we studied the key issue of the link function in the context of GLMs for binary responses, especially when the data is subjected to any of the most characteristic perturbations: outliers and noisy variables. We have focused our research on analyzing the robustness of the Student link in reference to the classical logit link.

We examined one low, two intermediate, and one high degree of overlap to account for the different data configurations. We also considered different perturbations scenarios when adding different numbers of outliers and noisy variables. Through numerical examples and simulations, we found that the degree of overlap differently affects the quality of the model for the logit and Student models. These differences are minor when there is a high degree of overlap; in this case, it would be more convenient to choose either the logistic link (taking advantage of its benefits as the canonical link) or the probit link. We observed that the closer the data is to the complete separation configuration, the more the Student link differs from the logit link, excelling in the log-likelihood and the prediction accuracy values.

We argued that the degree of overlap should help to determine the choice of the link function. However, we remarked that the degree of overlap is closely related to the degree of freedom ν of the Student cdf. Therefore, the inference of ν serves as an indicator of the overlap configuration. If a large value of ν is obtained, indicating a high degree of overlap, the model will result to be roughly the same as the logit or the probit model. On the other hand, if it turns out to be small, then it is likely that a strong

separation setting prevails in the data set, for which the best alternative will be to use the Student link function with the estimated ν.

Furthermore, the inference of binary models might also be significantly affected if a symmetric link function is incorrectly used in place of a non-symmetric link. The Student distribution is symmetric and thus sensitive to skewed data, as the logistic and normal distributions. To overcome this problem, and as an extension of this work, we propose to use the non-central Student distribution. By setting its non-centrality parameter to a value different from zero, one will be able to control the skewness of the link function. An additional perspective would be to study the robustness of the Student cdf used as part of the link function in the GLMs for more than two categories since their basic structure is based on binary models.

Chapter 3

GLMcat: An R package for GLMs for categorical responses

Abstract

In statistical modeling, there is a wide variety of generalized linear models for categorical response variables (nominal or ordinal responses); yet, there is no software embracing all these models together in a unique and generic framework. We propose and present GLMcat, an R package to estimate generalized linear models implemented under the unified specification (r, F, Z) where r represents the ratio of probabilities (reference, cumulative, adjacent, or sequential), F the cumulative distribution function for the linkage, and Z the design matrix. All classical models (and their variations) for categorical data can be written as an (r, F, Z) triplet, thus, they can be fitted with GLMcat. The functions in the package are intuitive and user-friendly. For each of the three components, there are multiple alternatives from which the user should thoroughly select those that best address the objectives of the analysis. The main strengths of the GLMcat package are the possibility of choosing from a large number of link functions (defined by the composition of F and r) and the simplicity for setting constraints in the linear prediction, either on the intercepts or on the slopes. This paper proposes a methodological and practical guide for the appropriate selection of a model considering the concordance between the nature of the data and the properties of the model.

Keywords: GLM, categorical response, link function, cumulative models, sequential models, adjacent models, reference models

Contents

3.1 Introduction

Regression models for categorical responses have emerged in various disciplines and under different names. The underlying structures of such models may be closely related (or even the same) but they are often perceived as fundamentally different. The intrinsic differences among the classical models concern the assumed link function and the specification of the linear predictor. Once defined the linear predictor, the question to address concerns the most appropriate link function. The selected link function should reflect the nature of the response variable; for categorical responses, a broad distinction is made on the basis of the scale itself, being either nominal or ordinal.

For ordinal responses, there are three families of GLMs: the cumulative, the sequential, and the adjacent models. The family of cumulative models (simply known as ordinal regression models) is the most popular. This family includes the odds proportional logit model (McCullagh, 1980) which has been the most widely used model for ordinal data. Sequential models have initially been discussed by authors including Fienberg (1980), Armstrong and Sloan (1989) and Tutz (1991). The most iconic model within this family is the proportional hazard model which was originally developed by Cox (1972) for continuous responses. More recently, Fahrmeir and Tutz (2001) briefly proposed an extension of the adjacent logit model (Goodman, 1983; Agresti, 1989) that allows substituting the logistic distribution function by any other cdf. In this light, Peyhardi et al. (2015) detailed the estimation of such models and named them the family of adjacent models. Adjacent models have been widely adopted in Item Response Theory, a widespread model within this framework is the polytomous Rasch model (Rasch, 1961; Andersen, 1995) which is an adjacent logit model with a specific form for the linear predictor. In contrast to ordinal responses, for which there exists a variety of GLMs, until recently, the only option for nominal responses was the MNL model, introduced by Luce (1959). To fill this gap, Peyhardi et al. (2015) generalized the structure of the MNL allowing the use of several cdfs as alternatives to the logistic cdf, the resulting set of models is referred to as the family of reference models.

Notwithstanding the wide set of model options, the use of appropriate models for categorical responses seems to be rather limited in the literature (Ananth and Kleinbaum, 1997; Liddell and Kruschke, 2018). Besides, on the rare occasions when they are employed, there is often no consistency between the response variable characteristics and the model's assumptions. For instance, ordinal responses have been frequently treated by researchers as standard nominal or metric problems. Another usual and inaccurate approach is to dichotomize categorical responses with the aim of using the standard logistic or probit regression models (Sankey and Weissfeld, 1998). These pitfalls can lead to non-optimal solutions and thus to erroneous statistical inferences (see Liddell and Kruschke, 2018; Scott et al., 1997; Gutiérrez et al., 2016, for further details). Despite the current availability of statistical software to fit models which take full advantage of the ordinal nature of the response, the described poor practices are still common. As noted by Mellenbergh (1995), one possible cause may be linked to the fact that the literature for ordinal responses does not provide much support for preferring one family of models over another. In addition, we suspect that the lack of homogeneity that characterizes the literature and the software solutions for this subject may result

confusing and overwhelming. Hence, it is not surprising that, risking a loss of accuracy and interpretability, the user might opt for the most popularly used models.

In R (R Core Team, 2021) there is a variety of packages to fit categorical responses, however, most of them only cover one or a few of the types of models. For instance, the function `multinom()` of the package nnet (Ripley and Venables, 2021; Venables and Ripley, 2002) fits the MNL via neural networks. For ordinal responses, the functions `polr()` of the package MASS (Ripley et al., 2021; Venables and Ripley, 2002) and `omr()` from the rms (Harrell Jr, 2021) package are often used to fit the odds proportional model. Few packages are aimed to fit a whole family of models for categorical responses, one of them is the tram (Hothorn et al., 2021; Hothorn, 2020) package, which by means of the `Polr()` function allows for stratification, censoring and truncation in the response of cumulative models. The ordinal package (Christensen, 2019) is another option to fit the family of cumulative models, it includes a comprehensive implementation of this class of models offering great flexibility, notably in the specification of the linear predictor. To our knowledge, only the VGAM (Yee, 2021) and the ordinalNet (Wurm et al., 2020) packages consider the three families of ordinal models: cumulative, sequential, and adjacent. Nevertheless, the ratio of probabilities of the adjacent models in VGAM seems to be valid only for the logistic distribution. None of the aforementioned packages encloses the four model families for categorical responses and most of them have some limitations in terms of adding constraints to the design, or in the availability of the cdfs that one can use as part of the link function. These gaps also exist in commercial statistical software like SAS (SAS Institute Inc., 2020), Stata (Stata Corp., 2015), and SPSS (IBM Corporation, 2017). An additional problem of the commercial packages is that those use different techniques (which are not strictly equivalent) to fit the models. As a consequence, different estimations might be obtained when using different software, even though the same theoretical model is specified. For instance, Liu (2009) reported some differences in the estimation of an odds proportional model using the functions `PROC LOGISTIC` in SAS, `OLOGIT` in Stata, and `PLUM` in SPSS.

Despite the diverse origins, names, applications, and implementations of the above-mentioned models, they all share a common structure that was fully described by Peyhardi et al. (2015). The authors introduced a unified specification of GLMs for categorical responses that encompass the four families of models based on a decomposition of the link function. They introduced the notation (r, F, Z) for this decomposition where: r is the ratio that characterizes the ordering type of the response variable, F is the cdf of the link function, and Z is the design matrix where the form of the linear predictor is specified. The comprehensive description of the taxonomy of the GLMs for categorical data given by the (r, F, Z) decomposition exposes the fundamentals, relations, and equivalences of the families of models. Furthermore, the possible extensions for each model family become evident and can be easily implemented. These extensions are obtained by structuring the design matrix (for intercepts and slopes), as well as broadening the spectrum of cdfs.

We implemented the (r, F, Z) methodology in the GLMcat (León et al., 2021) package developed for R (available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/web/packages/GLMcat`). Our purpose is to provide an alternative that covers all the classical models for categorical responses and which

gives room to extend them through different components. The package supports a wide range of cdfs and allows to adapt the linear predictor at any desired extent. In consideration of all these possibilities, we intend to guide the user in the identification of the most pertinent combination of the ratio, the cdf, and the design matrix, highlighting the limitations or advantages of the resulting (r, F, Z) model. In the GLMcat package there are two main functions: `glmcat()`, which covers the four families of models for categorical responses, and `discrete_cm()`, which extends the family of reference models to take into account explanatory variables that depend on response categories (useful for discrete choice model).

The content of the paper is presented in three main sections. In section 3.2, we recall the unified specification of GLMs through the (r, F, Z) triplet and we illustrate its implementation in GLMcat. We also describe in detail each of the three components as well as the possible extensions for them. In section 3.4, we aim to characterize the different families of models for ordinal responses by outlying a series of properties inherent to each of them. We emphasize the importance of identifying the model that makes the appropriate assumptions in light of the nature of the response variable and the goals of the analysis. In section 3.5, we revisit the family of reference models in the framework of discrete choice models (Bouscasse et al., 2019; Peyhardi, 2020). We motivate the use of this family of models presenting its strengths in contrast to existing alternatives. The model fitting by means of the GLMcat package is illustrated in all the sections using different datasets and its computational implementation is described in section 3.3.

3.2 Unified specification of GLMs for categorical data

Consider the regression context where the response Y is a categorical variable (with $J \geq 2$ categories). The aim is to model the effect on Y of a given set of q explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_q)$ defined in a general form of dimension p with $p \geq q$ (i.e., categorical variables are represented by indicator vectors). In the following, we will sometimes use the univariate notation $\{Y = j\}$ or, equivalently, the indicator vector notation $\boldsymbol{Y} = (Y_1, \ldots, Y_{J-1})$ with 1 in position j and 0 elsewhere. Note that $\{Y = J\}$ would correspond to $\boldsymbol{Y} = (0, \ldots, 0)$. For convenience, models are presented at the individual level, thus, the subscript $i \in \{1, \ldots, n\}$ is not mentioned. A GLM for categorical response can be decomposed into three parts:

1. the random component which accounts for the conditional distribution of the response variable given the set of the explanatory variables. In the framework of categorical response variables, Y follows the multinomial distribution

$$\boldsymbol{Y}|\boldsymbol{x} \sim \mathcal{M}(1, \boldsymbol{\pi}(\boldsymbol{x}))$$

   with $\boldsymbol{\pi} = (\pi_1(\boldsymbol{x}), \ldots, \pi_{J-1}(\boldsymbol{x})) \in \Delta$ where $\Delta = \{\boldsymbol{\pi} \in (0,1)^{J-1} : \sum_{j=1}^{J-1} \pi_j < 1\}$.

2. The systematic component which is determined by the linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{J-1})$. For each category j, the linear predictor has the form $\eta_j = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}_j$,

where $\alpha_j \in \mathbb{R}$ is the intercept and $\boldsymbol{\delta}_j \in \mathbb{R}^p$ is the vector of slopes. Considering the parameter vector as $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{J-1})$, the linear predictor can be written as the product

$$\boldsymbol{\eta} = Z\boldsymbol{\beta},$$

where Z denotes the design matrix composed of repetitions of 1 and $\boldsymbol{x}^\top$ (see section 3.2.3 for some examples of design matrices).

3. The link function g which relates the conditional expectation of the response variable $\boldsymbol{\pi} = \mathsf{E}[\boldsymbol{Y}|\boldsymbol{x}]$ and the linear predictor $\boldsymbol{\eta}$. The equality $g(\boldsymbol{\pi}) = \boldsymbol{\eta}$ corresponds to the $J-1$ equations $g_j(\pi) = \eta_j$.

Peyhardi et al. (2015) showed that all the classical link functions can be decomposed through the unified specification

$$g_j = F^{-1} \circ r_j \quad \Leftrightarrow \quad r_j(\boldsymbol{\pi}) = F(\eta_j) \qquad j = 1, \ldots, J-1, \tag{3.1}$$

where F is a continuous and strictly increasing cdf, and $\boldsymbol{r} = (r_1, ..., r_{J-1})$ is a map from the simplex Δ to the open hypercube $(0,1)^{J-1}$. The authors introduced the notation (r, F, Z) with which any classical GLM for categorical responses can be fully described. Throughout this paper and in this framework, we interchangeably use the terms (r, F, Z) and GLM.

The GLMcat package is designed based on the (r, F, Z) decomposition. To facilitate the user experience, instead of calling a specific function for each family of models (determined by the ratio), we implemented a single function: `glmcat()`, with which any (r, F, Z) model can be fitted. In the following, we will describe more closely the components r, F, and Z and their modalities.

3.2.1 Ratio of probabilities r

In models for categorical responses, the linear predictor $\boldsymbol{\eta}$ is not directly related to the expectation $\boldsymbol{\pi}$ but to a particular transformation $\boldsymbol{r}$ of the vector $\boldsymbol{\pi}$ called the ratio. The ratios for categorical responses are defined in Table 3.1. The cumulative ratio of category j is the result of the cumulated probabilities of the precedent categories. In the sequential and adjacent ratios, each category j is compared to its following categories $j+1, \ldots, J$, and its adjacent category $j+1$, respectively. The adjacent, cumulative, and sequential ratios rely on an ordering assumption among categories. On the contrary, the reference ratio relates each category j only to a reference category (J by convention), therefore, this ratio is devoted to nominal responses. The ratios are the essential units from which a family of models is defined. For this reason, we named the model families according to their corresponding ratio.

In GLMcat we cover the four ratios described above, whereby all known models for categorical responses are within reach by handling one single package. The ratio should be specified in the `glmcat()` function as a string in the argument `ratio`. If no ratio is specified, the reference ratio will be used by default.

$r_j(\boldsymbol{\pi})$	**Cumulative** $\pi_1 + \ldots + \pi_j$	**Sequential** $\frac{\pi_j}{\pi_j + \ldots + \pi_J}$	**Adjacent** $\frac{\pi_j}{\pi_j + \pi_{j+1}}$	**Reference** $\frac{\pi_j}{\pi_j + \pi_J}$
Y	ordinal			nominal

Table 3.1: Four ratios $r_j(\boldsymbol{\pi})$ of GLMs for categorical responses ($j = 1, \ldots, J-1$).

3.2.2 Cumulative distribution function F

The link function in the binary regression framework accounts only for the cdfs that links the expected value of the response to the linear predictor of the model. Based on the decomposition presented in Equation 3.1, it is evident that for the (r, F, Z) models the cdf is just one part of the link function, which along with the ratio, characterizes the relation between $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$. Remark that the cdf is assumed to be differentiable and strictly increasing. The differentiability is necessary for the Fisher's scoring algorithm (or Newton Raphson's algorithm) computation. The condition of strict increase is necessary for parameter interpretation since the coefficients $\delta_{j,k}$ give the signs of the partial effects of the corresponding explanatory variable x_k on the probability $r_j(\boldsymbol{\pi})$.

The distinction between symmetric and asymmetric cdfs has an impact on the properties of the models as it will be demonstrated later. Moreover, the choice of the distribution might markedly improve the model fit. In literature, there are different recommendations to choose the cdf of a GLM, although the logistic distribution is the most widely used. The choice is often related to disciplines or fields. For instance, economists tend to favor the normal distribution due to its association with the utility notion; the Gumbel distribution is popular in survival and hazard analysis, since it can appropriately model the occurrence of events. The aforementioned cdfs are available in many packages. GLMcat proposes, in addition, some less popular alternatives such as the Cauchy, the Gompertz, the Laplace, the Student, and the non-central Student cdfs. In particular, the Student cdf has proven to be a robust alternative for regression models (see Lange et al., 1989; Peyhardi, 2020) and can be considered as a family of functions given that the shape of the cdf varies according to ν, the degrees of freedom. All of the cdfs presented in Table 3.2 are available in GLMcat and should be specified by its name as a string in the argument `cdf`, or, if there are some parameters to specify, the user should input a list, for instance, `list(cdf = "student", df = 7)`. If the cdf is not declared, the logistic distribution is used by default. In the following, the set of cumulative distribution functions will be denoted by $\mathfrak{F}$.

3.2.2.1 Normalization of parameter estimates

Models with different cdfs used as part of the link function are not comparable since they refer to specific means and variances (Tutz, 2011). Often, the parameter estimates will turn out to be different even if there is not an apparent discrepancy in terms of goodness-of-fit. Tutz (2011) illustrated an alternative to standardize the parameters of

Distribution	$F(\eta)$	Shape
Logistic	$\dfrac{1}{1+\exp(-\eta)}$	symmetric
Normal	$\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\eta}\exp\Big(-\frac{x^2}{2}\Big)\,dx$	symmetric
Laplace	$\begin{cases}\frac{1}{2}\exp(\eta) & \text{if } \eta<0,\\ 1-\frac{1}{2}\exp(-\eta) & \text{if } \eta\geq 0\end{cases}$	symmetric
Cauchy	$\dfrac{1}{2}+\dfrac{1}{\pi}\arctan(\eta)$	symmetric
Student(ν)	$\dfrac{1}{2}+\eta\Gamma\Big(\dfrac{\nu+1}{2}\Big)$	symmetric
Non-central t (ν, μ)	$\begin{cases}F_{\nu,\mu}(\eta) & \text{if } \eta\geq 0,^{(1)}\\ 1-F_{\nu,-\mu}(\eta) & \text{if } \eta<0\end{cases}$	- left skewed if $\mu<0$, - symmetric if $\mu=0$, - right skewed if $\mu>0$.
Gompertz	$1-\exp(-\exp(\eta))$	left skewed
Gumbel	$\exp(-\exp(-\eta))$	right skewed

Table 3.2: List of the cdfs available in GLMcat to use as part of the link function for GLMs.
(1) Refer to Appendix B.1 for the complete form of $F_{\nu,\mu}$.

a binary regression model:

$$\tilde{\alpha}=\frac{\alpha-\mathsf{E}(\varepsilon)}{\sqrt{\mathsf{VAR}(\varepsilon)}},\qquad \tilde{\delta}=\frac{\delta}{\sqrt{\mathsf{VAR}(\varepsilon)}},\qquad \text{where } \varepsilon\sim F.$$

Note that this approach is not suitable when using a cdf with undefined mean or variance as it is the case of the Student distribution (whose mean and variance are not defined when $\nu\leq 1$, and $\nu\leq 2$, respectively). A propagated approach in econometrics that solves this problem is to consider the average partial effect of the variable x_k on π_j as the scale factor, i.e., $\partial\pi_j(\boldsymbol{x})/\partial x_k$ (Wooldridge, 2012). If x_k is a continuous variable, its partial effect on $\pi_j(\boldsymbol{x})$ is obtained from the partial derivative:

$$\frac{\partial\pi_j(\boldsymbol{x})}{\partial x_k}=\frac{\partial F}{\partial\eta_j}\frac{\partial\pi_j}{\partial r_j}\hat{\delta}_{j,k}\,. \tag{3.2}$$

The average partial effect of x_k on π_j is then given by the mean value of the individuals. The downsides of this method are that the scale factor depends on the input data and that it is only valid for continuous explanatory variables. Note that Equation 3.2 results in $f(\eta_j)\hat{\delta}_k$ for the binomial regression. In this case, if f is a symmetric pdf

around zero, the largest effect occurs when $\eta = 0$. For instance, for the normal pdf, this will be at $f(0) \approx 0.4$ and for the logistic pdf at $f(0) = 0.25$. A simple approach to make the magnitudes of those two cdfs roughly comparable is to multiply the probit estimates by $0.4/0.25 = 1.6$ or to multiply the logit estimates by $0.25/0.4 = 0.625$.

Bouscasse et al. (2019) proposed a normalization of parameter estimates via the location parameter m and the scale parameter s of the cdf F. Two real points a and b are chosen such that all cdfs in $\mathfrak{F}$ have the same values for $F(a)$ and $F(b)$. It is imposed that $F(0) = 1/2$ to preserve the interpretability of the intercepts. Note that the reference, the adjacent and the sequential ratios satisfy this condition. To illustrate this, assume $\delta_j = 0$ so that the linear predictor only depends on the intercept, i.e., $\eta_j = \alpha_j$. Consider the reference family where the ratio of probabilities is given by $\pi_j/\pi_J = F(\alpha_j)/(1 - F(\alpha_j))$, hence, if the intercept is null, and setting $F(0) = 1/2$, it is evident that the probabilities π_j and π_J are equal and so are all the elements in $\boldsymbol{\pi}$. This equality is also valid for the adjacent ratio but neither for sequential nor cumulative. For the sequential family, we can find that the probabilities will correspond to $\pi_j = (1/2)^j$ for $j = 1, \ldots, J-1$, and $\pi_J = (1/2)^{J-1}$. Conversely, for the cumulative ratio, the intercepts must be strictly ordered and cannot be all equal to zero. Therefore for this ratio, the constraint $F(0) = 1/2$ is not necessary. Remark that the condition $F(0) = 1/2$ is already satisfied for the symmetric distributions and has to be imposed for asymmetric cdfs. The logistic distribution is proposed as the reference cdf since it is part of the canonical link function. Thus, the second point is given by $F(b) = e^b/(1 + e^b)$. The authors suggested to use the quantile of the logistic distribution such that $b = q_p$ for some $p > 1/2$. The normalized space is then $\mathfrak{F}_{q_p} = \{F \in \mathfrak{F} : F(0) = 1/2,\ F(q_p) = p\}$. We have that $F_{m_0,s_0} \in \mathfrak{F}_{q_p}$ if

$$\begin{cases} m_0 = \dfrac{F^{-1}(1/2) \cdot q_p}{F^{-1}(p) - F^{-1}(1/2)} \\ s_0 = \dfrac{q_p}{F^{-1}(p) - F^{-1}(1/2)}. \end{cases}$$

The normalized parameters using the above approach are: $\alpha'_j = m_0 + \alpha_j s_0$ and $\delta'_j = \delta_j s_0$ for $j = 1, \ldots, J-1$. We implemented this normalization since it works for any number of categories, for any type of explanatory variables, and because it does not depend on the dataset. In the functions of GLMcat, the normalization using the quantile $q_{0.95}$ (which can be considered as the standard case, $q_{0.95} \approx 2.94$) is obtained with the argument `normalization = 0.95`. The `summary()` function returns the transformed parameters when specifying the argument `normalized = TRUE`. An example of the normalization of parameters is illustrated in section 3.5.

3.2.3 Design matrix Z

In a linear predictor, one can define constraints to model the effect of the explanatory variables on the categorical response. Commonly, these constraints have been imposed only on the slopes and not much attention has been given to the intercepts. In GLMcat, we divide the design matrix into two blocks: I to control the intercepts and S to control

the slopes, i.e.,

$$Z = \left(\begin{array}{c|c} I & S \end{array} \right).$$

The design matrix can be fully customized using this decomposition. By default for the reference family of models, the GLMcat package proposes a complete design without any constraint, i.e., $Z_c = (I_c|S_c)$. This matrix is of dimension $(J-1) \times (J-1)(1+p)$, and has the form:

$$Z_c = \left(\begin{array}{ccc|ccc} 1 & & & \boldsymbol{x}^\top & & \\ & \ddots & & & \ddots & \\ & & 1 & & & \boldsymbol{x}^\top \end{array} \right).$$

Slope design matrix S: the most common constraint is to impose the effects of the explanatory variables to be constant across the response categories, thus, it is assumed the existence of a single global effect for each explanatory variable. This constraint is known as the parallelism or proportional assumption, and the user should verify its validity before using it (Harrell, 2015). The slope matrix associated with the parallel design is of dimension $(J-1) \times p$ and has the form:

$$S_p = \left(\begin{array}{c} \boldsymbol{x}^\top \\ \vdots \\ \boldsymbol{x}^\top \end{array} \right).$$

A more flexible framework is to consider both kinds of effects, complete and parallel, the resulting design is known as partial parallel. The following slope matrix of dimension $(J-1)\times((J-1)p_1+p_2)$ represents the design for p_1 explanatory variables $\boldsymbol{x_c} = (x_1, \ldots, x_{p_1})$ with complete design effects, and p_2 explanatory variables $\boldsymbol{x_p} = (x_1, \ldots, x_{p_2})$ with parallel effects:

$$S_{cp} = \left(\begin{array}{ccc} \boldsymbol{x_c}^\top & & \boldsymbol{x_p}^\top \\ & \ddots & \vdots \\ & \boldsymbol{x_c}^\top & \boldsymbol{x_p}^\top \end{array} \right).$$

The `glmcat()` function assumes by default a parallel design for the cumulative, sequential, and adjacent ratios. If all predictors are to be set with the complete design, one should simply specify `parallel = FALSE`. If the user opts for the partial parallel design, the variables with a parallel effect must be specified in a string vector in the argument `parallel`.

In section 3.5, we further explore the design matrix particularly for nominal response variables for which the function `discrete_cm()` allows to specify a particular response category on which the explanatory variable(s) is expected to have an effect.

Intercept design matrix I: if a single intercept is expected in the linear predictor, the intercept matrix I_p is simply the vector $\mathbb{1}$ of size $J-1$. The design matrix I_p is obtained by specifying the string `"(Intercept)"` in the argument `parallel`. Remark that in categorical regression, the parallel design refers to the use of I_c together with S_p; I_c is used instead of I_p since it refers to the minimal model which estimates the proportions of the response categories without explanatory variables effects. The *complete* design $Z_c = (I_c|S_c)$ and the *parallel* design $Z_p = (I_c|S_p)$ are sufficient to define all the

classical models, nevertheless, the constraints on I can be further explored.

Christensen (2019) presented some constraints on the intercept for the cumulative models. For instance, if the distances between the adjacent intercepts are required to be the same for all pairs $(j, j+1)$, we can write the intercepts as $\alpha_j = \alpha_1 + (j-1)\theta$ for $j = 1, \ldots, J-1$. In that case, α_1 corresponds to the first intercept and θ to the constant distance between intercepts. This restriction implies that, regardless of the number of categories, only two parameters must be estimated. The associated design matrix is of dimension $(J-1) \times 2$ and has the form:

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & J-2 \end{pmatrix}.$$

Another form is given when the intercepts are symmetric around zero, i.e., the categories are supposed to be equally distant from the central category/categories. For an even and for an odd number of response categories, the dimension of the intercept matrix is $(J-1) \times J/2$ and $(J-1) \times (J+1)/2$, respectively, and the intercepts and their design matrices are respectively written as:

$$\alpha_j = \begin{cases} \alpha_{J/2} - \theta_j & \text{if } j < J/2, \\ \alpha_{J/2} + \theta_{J-j} & \text{otherwise,} \end{cases} \quad \text{and} \quad \alpha_j = \begin{cases} \theta_0 - \theta_j & \text{if } j < J/2, \\ \theta_0 + \theta_{J-j} & \text{otherwise,} \end{cases}$$

$$\begin{pmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ & & & -1 \\ & 0 & \cdots & 0 \\ & & & 1 \\ & & \iddots & \\ 1 & 1 & & \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -1 & & \\ \vdots & & \ddots & \\ & & & -1 \\ & & & 1 \\ & & \iddots & \\ 1 & 1 & & \end{pmatrix}.$$

The constraints on the intercepts are only available for the cumulative ratio and should be specified through the argument `threshold` by chosing the option among `"symmetric"` or `"equidistant"`. The computational instability that is frequently found in the cumulative models can be alleviated with the use of this constraint given that the number of parameters is reduced. An example of the use of the structured intercepts for cumulative models is presented by Reinhard et al. (2017). In the following, the set of design matrices will be denoted by $\mathfrak{Z}$.

3.2.4 (r, F, Z) genericity

A large number of models for categorical responses have been proposed in the literature. Depending on the scientific context, some of these models can be differently named despite having the same formulation. In consequence, the relationships among them are often unrecognized. Earlier in this paper, we mentioned that any GLM for categorical

responses can be written as the triplet (r, F, Z). In Table 3.3 we present some of the best-known models in their original formulation and decomposed into the three components r, F, and Z. For categorical responses, the (r, F, Z) specification enlarges the number of possible models to consider. Furthermore, it eases the comparison between them as we are going to demonstrate in the following.

The multinomial logit model $\mathrm{P(Y = j)} = \frac{\exp(\alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}_j)}{1 + \sum_{k=1}^{J-1}\exp(\alpha_k + \boldsymbol{x}^t\boldsymbol{\delta}_k)}$	$(reference, logistic, complete)$ $\frac{\pi_j}{\pi_j + \pi_J} = \frac{1}{1 + \exp(-\alpha_j - \boldsymbol{x}^t\boldsymbol{\delta}_j)}$
The odds proportional logit model $\ln\left\{\frac{\mathrm{P}(Y \le j)}{1 - \mathrm{P}(Y \le j)}\right\} = \alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}$	$(cumulative, logistic, parallel)$ $\pi_1 + \cdots + \pi_j = \frac{1}{1 + \exp(-\alpha_j - \boldsymbol{x}^t\boldsymbol{\delta})}$
The proportional hazard model $\ln\{-\ln\{\mathrm{P}(Y > j \mid Y \ge j)\}\} = \alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}$	$(sequential, Gompertz, parallel)$ $\frac{\pi_j}{\pi_j + \cdots + \pi_J} = 1 - \exp(-\exp(\alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}))$
The adjacent logit model $\ln\left\{\frac{\mathrm{P}(Y = j)}{\mathrm{P}(Y = j+1)}\right\} = \alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}_j$	$(adjacent, logistic, complete)$ $\frac{\pi_j}{\pi_j + \pi_{j+1}} = \frac{1}{1 + \exp(-\alpha_j - \boldsymbol{x}^t\boldsymbol{\delta}_j)}$
The continuation ratio logit model $\ln\left\{\frac{\mathrm{P}(Y = j)}{\mathrm{P}(Y > j+1)}\right\} = \alpha_j + \boldsymbol{x}^t\boldsymbol{\delta}_j$	$(sequential, logistic, complete)$ $\frac{\pi_j}{\pi_j + \cdots + \pi_J} = \frac{1}{1 + \exp(-\alpha_j - \boldsymbol{x}^t\boldsymbol{\delta}_j)}$

Table 3.3: (r, F, Z) specification of some classical GLMs for categorical responses.

3.3 Computational details and implementation

The GLMcat package can be installed within R (R Core Team, 2021) using the line of code: `install.packages("GLMcat")`. The standard arguments `formula` and `data` are already known from the `lm` and `glm` functions from the stats package. The key difference is that in the `glmcat()` function, the link of the model must be specified through the two arguments `ratio` and `cdf`. In GLMcat, the response (categorical) variable must be defined as a factor or an ordered factor. The user can specify/change the order of the factor levels by means of the `ordered()` function. Alternatively, and for ease of use, one can indicate the order as a character vector in the argument `categories_order`. An example of the syntax of the `glmcat()` function for an ordinal response is

```
R> glmcat(formula = Level ~ Age, data = DisturbedDreams,
+    categories_order = c("Not.severe", "Severe.1", "Severe.2",
+      "Very.severe"), ratio = "adjacent", cdf = "gompertz")
```

For non-ordered response variables, the user must use the reference ratio, for which by default, the reference category is set to be the last level of the response factor variable. The user can also specify manually the reference category in the argument `ref_category` as in the following

```
R> glmcat(formula = Level ~ Age, data = DisturbedDreams,
+     ref_category = "Very.severe", ratio = "reference",
+     cdf = "gompertz")
```

The object generated by the `glmcat` function is compatible with the usual generic methods: `coef()` for the parameter estimates $\hat{\boldsymbol{\beta}}$ and `confint()` for their confidence intervals, `logLik()` for the log-likelihood, `nobs()` for the number of observations n, `predict()` to obtain $\hat{\boldsymbol{\eta}}$ (if `type = "linear.predictor"`) or $\hat{\boldsymbol{\pi}}$ (if `type = "prob"`), `vcov()` to obtain the variance-covariance matrix of the parameters of the fitted object, `plot()` to represent graphically the log-likelihood profile over the iterations, and `summary()` to generate the summary of the fitted model. The `AIC()` and the `BIC()` functions are available to obtain the values of the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), respectively. As for the regressions tests (available in the function `anova()`), we implemented the Wald test to check $H_0 : \beta_j = 0$. In addition, to investigate the relevance of terms in the linear predictor, one can obtain the likelihood-ratio test that compares nested models by specifying the two models in the `anova()` function. We also adapted the `step()` function of the stats package to incorporate the classical stepwise variable selection for (r, F, Z) models. The resulting function, offers both forward and backward directions. The specification `direction = "backward"` starts with the full model and sequentially deletes predictors, it supports the definition of some or all variables with the parallel design, while the forward direction is compatible with the complete or with the parallel design but not with the partial parallel design.

Note that the link function $g : \Delta \to \mathbb{R}^{J-1}$ is differentiable if the ratio $\boldsymbol{r} : \Delta \to (0,1)^{J-1}$ and the cdf $F : \mathbb{R} \to (0,1)$ are both differentiable. All the cdfs available in GLMcat (see Table 3.2) are differentiable (i.e., there exists a density function such that $f = F'$). The four ratios are also differentiable, thus, we use the Fisher's scoring algorithm for the estimation of the model. In the following, we present the form of the algorithm in the iteration t

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left\{ \mathsf{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}} \right\}^{-1} \left(\frac{\partial l}{\partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}}.$$

Applying the chain rule to $l = \ln P(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\beta})$ we obtain the score

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\pi}} \frac{\partial l}{\partial \boldsymbol{\theta}}.$$

Since the response distribution belongs to the exponential family, it becomes

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = Z^\top \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{x})^{-1} (\boldsymbol{y} - \boldsymbol{\pi}).$$

Then, using the decomposition of the link function presented in Equation 3.1, we obtain

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = Z^\top \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}} \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{x})^{-1}(\boldsymbol{y} - \boldsymbol{\pi}) \ ,$$

and the Fisher's information matrix

$$\mathsf{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}}\right) = -Z^\top \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}} \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{x})^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{r}^\top} \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{\eta}^\top} Z.$$

Remark that the Jacobian matrix $\partial \boldsymbol{F}/\partial \boldsymbol{\eta}$ is the diagonal matrix of densities $\{f(\eta_j)\}_{j=1,\ldots,J-1}$; the Jacobian matrices associated to each ratio $\partial \boldsymbol{\pi}/\partial \boldsymbol{r}$ are detailed in Appendix B.2. Note that the above calculations of the score and the Fisher's information matrices concern only one observation. To obtain the total expected result, the contributions of the n observations have to be added.

Computational difficulties for the maximum likelihood are expected when either complete or quasi complete separation occurs in the dataset, this is due to the fact that the MLE is not unique in that case. Another situation involving such difficulties occurs for the cumulative ratio used together with a complete or partially parallel design; these models are not invertible and the algorithm might fail to converge (more details are given in section 3.4). The standard numerical optimization techniques have no way of detecting this problem and will keep iterating until the iteration's bound is reached (Albert and Anderson, 1984). The convergence criteria for the Fisher's scoring algorithm is set to be reached in GLMcat when

$$\frac{|l(\boldsymbol{\beta}^{[t+1]}) - l(\boldsymbol{\beta}^{[t]})|}{\varepsilon + |l(\boldsymbol{\beta}^{[t+1]})|} > \frac{\varepsilon}{n}, \tag{3.3}$$

where $\varepsilon = 0.0001$ by default. Thus, the algorithm will stop iterating either when the maximum number of iterations is met, or until the Expression 3.3 becomes true. In case of convergence problems, an additional strategy is to initialize the model parameters $\boldsymbol{\beta}^{[0]}$ specifying a numerical vector in the argument `control_glmcat`. For the reference, adjacent and sequential ratios, the algorithm is initiated with $\boldsymbol{\beta}^{[0]}$ as the null vector. Conversely, the intercepts of cumulative models are symmetrically and ascendingly defined around 0, thus $\alpha_1^0 < \alpha_2^0 < \ldots < \alpha_{J-1}^0$. In GLMcat, the user can also modify the number of iterations (which by default is 25), and the size of the convergence tolerance given by ε with the argument `control_glmcat`, for example: `control_glmcat(iterations = 30, epsilon = 0.0001)`.

As Wickham (2015) states, R is a high-level expressive language, and that expressivity comes at a price: speed. In order to improve the speed of the functions in GLMcat, we incorporated C++ code through the Rcpp package (Eddelbuettel et al., 2020). The algorithms are implemented in a modular manner, meaning that enhancement or adjustment can be easily extended to all the families of models.

3.4 Models for ordinal responses

Based on the common foundation exposed by the triplet (r, F, Z), it is possible to describe some properties of the models for categorical responses (Peyhardi et al., 2015). Such information empowers the practitioner to adequately choose (from a wide range of options) the model that best suits the characteristics of the data. As indicated in the past sections, the link function is composed of r and F. By changing either of them, we can obtain improvements in terms of the goodness-of-fit measures. Nevertheless, the performance of a model is not merely measured through the fit. The foremost consideration for choosing a model should be the consistency among the nature of the data, the modeling objectives, and the model's features. In the following, we introduce and illustrate on real datasets, by means of GLMcat, the properties of the GLMs for ordinal responses. We intend to guide the practitioner in the selection process of the link function.

3.4.1 Reversibility

To announce the reversibility property of the models for ordinal responses, we need to recall the following definitions introduced by Peyhardi et al. (2015):

- The models (r, F, Z) and (r', F', Z') are said to be equivalent if one is a reparameterization of the other, i.e., there exists a bijection h from Θ to Θ' such that $r^{-1} \circ \boldsymbol{F}\{Z(\boldsymbol{x})\boldsymbol{\beta}\} = r'^{-1} \circ \boldsymbol{F}'\{Z'(\boldsymbol{x})h(\boldsymbol{\beta})\}$, for all $\boldsymbol{x} \in \mathcal{X}$, and all $\boldsymbol{\beta} \in \Theta$.
- The models (r, F, Z) and (r', F', Z') are said to be equal if the corresponding distributions of $\boldsymbol{y}|\boldsymbol{x}$ are equal, i.e., if $r^{-1} \circ \boldsymbol{F}\{Z(\boldsymbol{x})\boldsymbol{\beta}\} = r'^{-1} \circ \boldsymbol{F}'\{Z'(\boldsymbol{x})\boldsymbol{\beta}\}$, for all $\boldsymbol{x} \in \mathcal{X}$, and all $\boldsymbol{\beta} \in \Theta$.

  Note that the equality between models implies that they are equivalent.
- An (r, F, Z) model is said to be invariant under a permutation σ of $\{1, \dots, J\}$, if it is equivalent to the $(r, F, Z)_\sigma$ model which is defined on the permuted vector $\boldsymbol{\pi}_\sigma = (\pi_{\sigma(1)}, \dots, \pi_{\sigma(J-1)})$.

On the basis of the above definitions, an (r, F, Z) model is said to be reversible if it is invariant under the reverse permutation $\tilde{\sigma}$ defined by $\tilde{\sigma}(j) = J - j + 1$ for all $j \in \{1, \dots, J-1\}$. The reversibility property was first studied for cumulative models with some specific distributions by McCullagh (1980). Later, Peyhardi et al. (2015) generalized it for all symmetric distributions as well as for the adjacent ratio.

Proposition 2. *The* (*adjacent*, F, Z) *and the* (*cumulative*, F, Z) *models are reversible for all symmetric cdfs* F *and all the design matrices* Z *proposed in this package.*

McCullagh (1980) suggests that depending on the application, the reversibility may be seen as an appealing property, for example, when the response is given by an ordered scale.

Moreover, for any cdf $F \in \mathfrak{F}$ and $Z \in \mathfrak{Z}$, we have that:

Proposition 3. *The* $(adjacent,\ F,\ Z)_{\tilde{\sigma}}$ *model and the* $(cumulative,\ F,\ Z)_{\tilde{\sigma}}$ *model are respectively equal to the* $(adjacent,\ \tilde{F},\ -\tilde{P}Z)$ *and the* $(cumulative,\ \tilde{F},\ -\tilde{P}Z)$*, where* $\tilde{F}(\eta) = 1 - F(-\eta)$*, and* $\tilde{P}$ *is the restricted reverse permutation matrix of dimension* $J-1$*:*

$$\tilde{P} = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix}. \tag{3.4}$$

Refer to the Appendix B.3 for the demonstration.

Note that if a cdf is symmetric then $\tilde{F} = F$; for asymmetric distributions, as the Gumbel cdf, $\tilde{F}$ corresponds to its symmetric counterpart, in this example, the Gompertz cdf. For a practical illustration of Proposition 3, consider the observations of the boys' disturbing dreams benchmark dataset presented by Maxwell (1961). This study cross-classified boys by their age $\boldsymbol{x}$ (which corresponds to the mid-point values for each stratum of 2 or 3 years, and it is treated as a continuous explanatory variable), and the severity of their disturbing dreams Y on a four-point scale of increasing severity. The data is available as the object `DisturbedDreams` in the GLMcat package. The $(adjacent, Gumbel, parallel)$ model is defined as:

```
R> adj_gumbel_p <- glmcat(formula = Level ~ Age,
+     data = DisturbedDreams, ratio = "adjacent",
+     cdf = "gumbel", categories_order = c("Not.severe",
+       "Severe.1", "Severe.2", "Very.severe"))
R> logLik(adj_gumbel_p)

'log Lik.' -279.9612 (df=4)

R> summary(adj_gumbel_p)

Level ~ Age
                ratio     cdf nobs niter     logLik
Model info: adjacent gumbel  223   (7) -279.9612
                      Estimate Std. Error z value Pr(>|z|)
(Intercept) Not.severe  0.22676    0.26157   0.867    0.386
(Intercept) Severe.1   -0.36548    0.24270  -1.506    0.132
(Intercept) Severe.2   -0.33321    0.22899  -1.455    0.146
Age                     0.07146    0.01806   3.957 7.59e-05 ***
```

Now, inverting the order of the categories in the argument `categories_order`, and using the symmetric counterpart cdf of the Gumbel, we fit the $(adjacent, Gompertz, parallel)_{\tilde{\sigma}}$ model:

```
R> adj_gompertz_rev <- glmcat(formula = Level ~ Age,
+     data = DisturbedDreams, ratio = "adjacent", cdf = "gompertz",
+     categories_order = c("Very.severe",
+       "Severe.2", "Severe.1", "Not.severe"))
R> logLik(adj_gompertz_rev)
'log Lik.' -279.9612 (df=4)

R> summary(adj_gompertz_rev)

Level ~ Age
```

```
                ratio       cdf nobs niter    logLik
Model info: adjacent gompertz  223     7 -279.9612
                        Estimate Std. Error z value Pr(>|z|)
(Intercept) Very.severe  0.33321    0.22899   1.455    0.146
(Intercept) Severe.2     0.36548    0.24270   1.506    0.132
(Intercept) Severe.1    -0.22676    0.26157  -0.867    0.386
Age                     -0.07146    0.01806  -3.957 7.59e-05 ***
```

Note that the estimations of the parameters of the last model are reversed but its log-likelihood is the same. This would also be true using any symmetric cdf. Given Property 2, the cumulative and the adjacent models are suitable for the type of responses that have an ordering scale associated with their categories. However, the reversibility property is not valid for the sequential models since these are non-invariant under the reverse permutation. The user should consider the sequential ratio if there is a time-related notion or a time-ordered process (which cannot be reversed) implicit in the response. For instance, the education level (see Figure 18) is conditioned on the completion of the previous degrees. The sequential ratio is the only one that takes into account this particularity of the response variable, for this reason, it is commonly employed in time survival analysis.

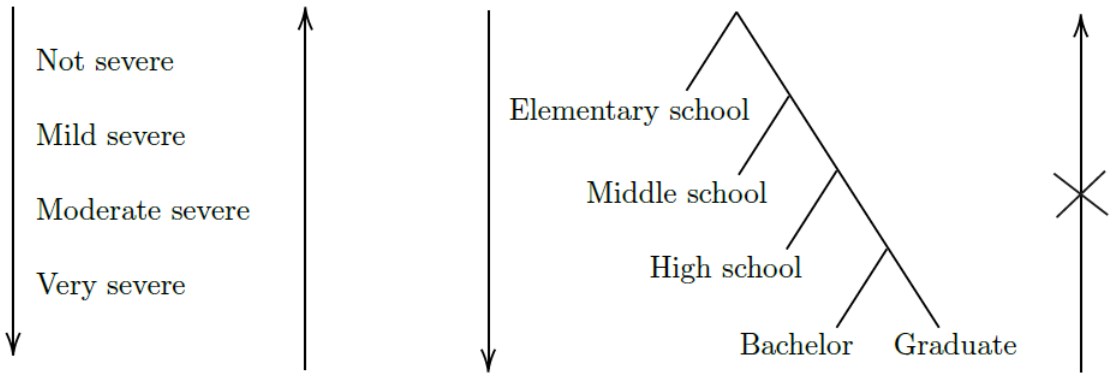

Figure 18: Scale ordering in severity of disturbed dreams versus process ordering in the educational path.

3.4.2 Latent variable interpretation

As considered by McCullagh (1980), the (*cumulative*, *logistic*, *proportional*) model can be seen as if the observed Y was originated from the categorization of a latent continuous variable $\tilde{Y}$. This latent variable follows a linear regression model $\tilde{Y} = \tilde{\alpha} + \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}} + \varepsilon$, where $-\infty = \alpha'_0 < \alpha'_1 < \ldots \alpha'_{J-1} < \alpha'_J = \infty$ are the strictly-ordered cut-points, and ε is a noise variable with cdf F. To model this categorization process, the cumulative ratio assumes that the $J-1$ cut-points partition $\tilde{Y}$ into J observable ordered categories of Y, i.e.,

$$\{Y = j\} \Leftrightarrow \alpha'_{j-1} < \tilde{Y} \leq \alpha'_j,$$

for $j = 1, \ldots, J$. The cumulative probabilities are

$$\begin{aligned} \mathsf{P}(Y \leq j|\boldsymbol{x}) &= \mathsf{P}(\tilde{Y} \leq \alpha'_j) \\ &= \mathsf{P}(\varepsilon \leq \alpha'_j - \tilde{\alpha} - \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}}) \\ &= F(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}) \end{aligned}$$

with $\alpha_j = \alpha'_j - \tilde{\alpha}$, and $\boldsymbol{\delta} = -\tilde{\boldsymbol{\delta}}$. We represent this structure (for $J = 4$) in Figure 19, where we can see that

$$\pi_j = \mathsf{P}(\alpha'_{j-1} < \tilde{Y} < \alpha'_j).$$

The order structure is more easily interpretable using the notion of the latent continuous variable where the categories are considered as successive intervals $(\alpha'_{j-1}, \alpha'_j]$. Remark that this only holds when the constraint of proportionality is assumed for all explanatory variables. In the other cases (complete or partial parallel design) the interpretation in terms of the latent variable is no longer accurate.

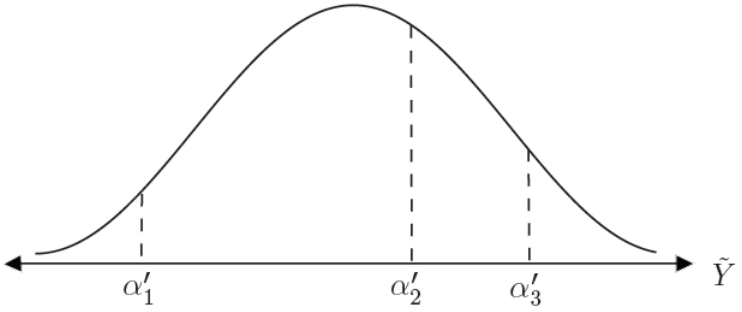

Figure 19: The cumulative model represented through a latent continuous variable.

The sequential ratio assumes that the successive choices between category j and the categories over j is determined by the latent variables $\tilde{Y}_j = \tilde{\alpha} + \boldsymbol{x}^\top \tilde{\boldsymbol{\delta}}_j + \varepsilon_j$, for $j = 1, \ldots, J-1$, where the residuals ε_j are independent and identically distributed according to the cdf F. This sequential mechanism can be viewed as a binary process at each transition, thus, it is appropriate when the assumption of a single underlying latent variable does not hold. We can write then

$$\{Y = j\} = \bigcap_{k=1}^{j-1} \{\tilde{Y}_k > \alpha'_k\} \bigcap \{\tilde{Y}_j \leq \alpha'_j\},$$

so the conditional probabilities of the event $\{Y = j|Y \geq j\}$ for $j = 1, \ldots, J$ can also be written as $\{Y = j|Y \geq j\} = \{\tilde{Y}_j \leq \alpha'_j\}$, then, we have

$$\mathsf{P}(Y = j|Y \geq j; \boldsymbol{x}) := F(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}_j),$$

where $\alpha_j = \alpha'_j - \tilde{\alpha}$, and $\boldsymbol{\delta}_j = -\tilde{\boldsymbol{\delta}}_j$. In Figure 20, we illustrated the sequential model with a process that starts from category 1. If $\tilde{Y}_1 \leq \tilde{\alpha}_1$ the process stops and we have $Y = 1$, otherwise, i.e., $\tilde{Y}_1 > \tilde{\alpha}_1$, the process continues and we know that it will at least reach category 2. The process continues in this way until the last category is reached. In this

context, we can represent the probabilities of each category as

$$\pi_j = \mathsf{P}(\tilde{Y}_j < \tilde{\alpha}_j) \prod_{k=1}^{j-1} \mathsf{P}(\tilde{Y}_k > \tilde{\alpha}_k).$$

The transition can be interpreted in terms of the difficulty of reaching the next category. Upper levels can only be achieved if previous levels were visited earlier and not kept. Therefore the model is built around the conditionality principle.

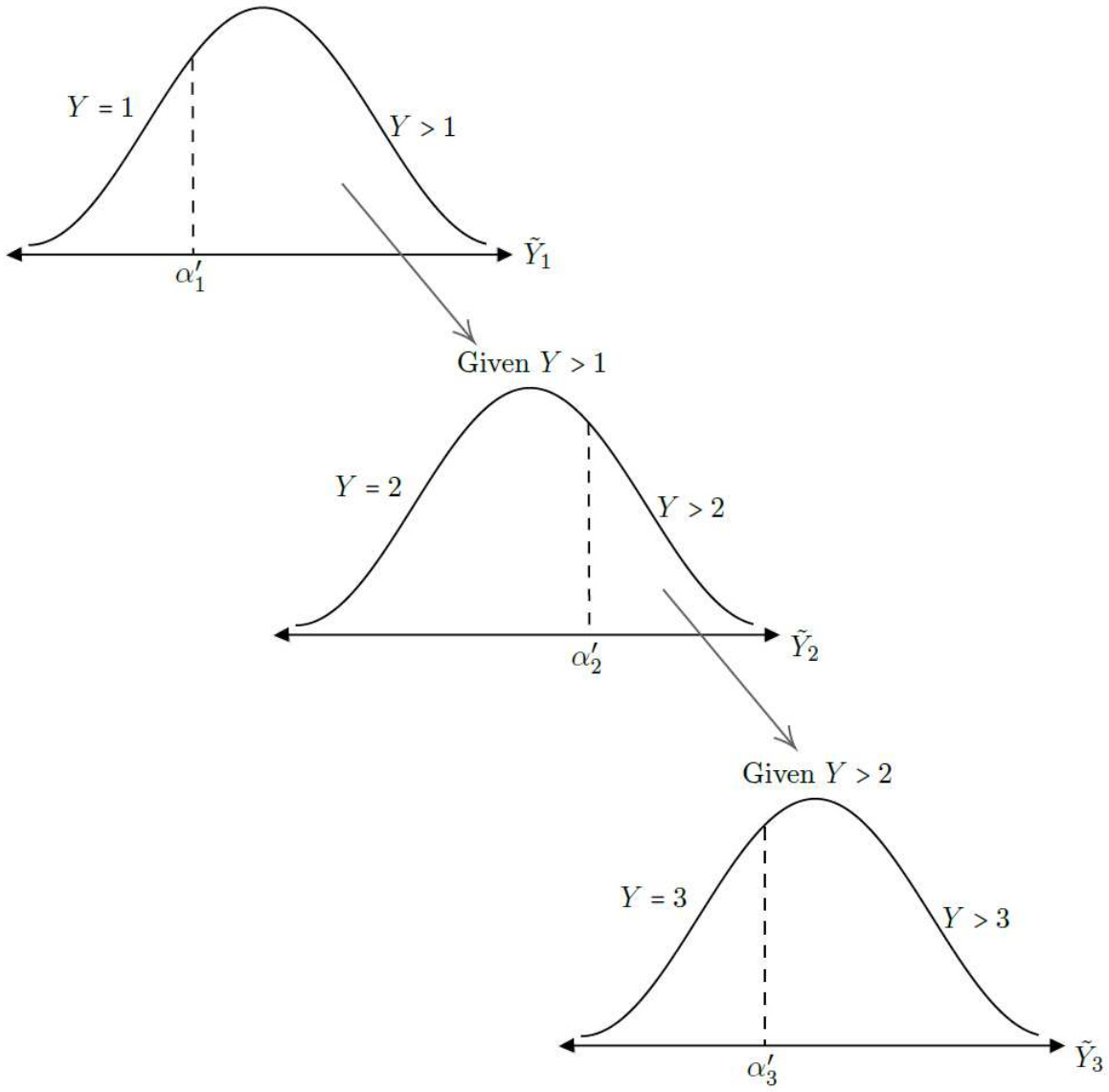

Figure 20: The sequential model represented as latent continuous variables.

The adjacent ratio describes the probability that category j rather than category $j+1$ is achieved:

$$\{Y = j | Y \in \{j, j+1\}\} = \{\tilde{Y}_j \leq \alpha'_j\}$$

for $j = 1, \ldots, J-1$. In Figure 21, we represent the adjacent ratio using latent continuous

variables. Note that each category j is present in two different latent variables $\tilde{Y}_j$ and $\tilde{Y}_{j+1}$. In contrast to the cumulative and the sequential ratio, π_j cannot be written only in terms of the latent variables:

$$\pi_j = \mathsf{P}(\tilde{Y}_j < \tilde{\alpha}_j)(\pi_j + \pi_{j+1}).$$

As a result, this ratio lacks of interpretability since there is not a natural process that leads to its formulation.

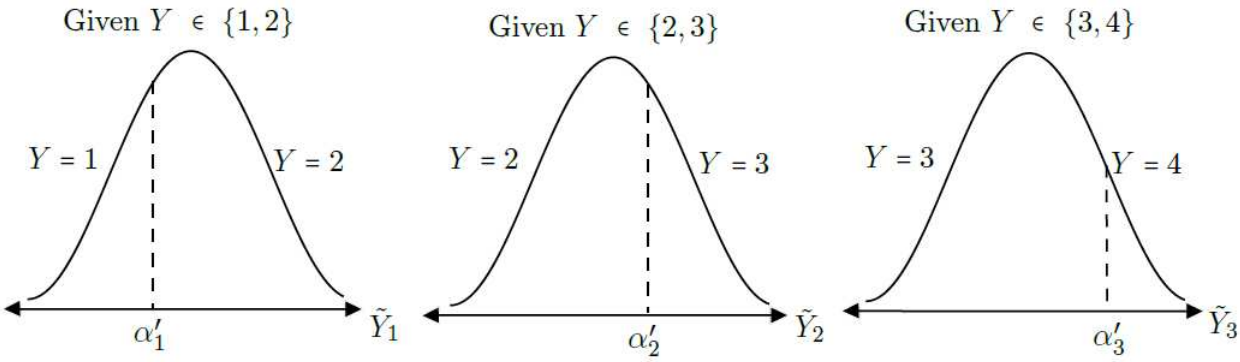

Figure 21: The adjacent model represented as a latent continuous variables.

3.4.3 Invertibility

An (r, F, Z) model is said to be invertible if its link function is invertible, i.e.,

$$\boldsymbol{\pi} = r^{-1} \circ F(\boldsymbol{\eta}) \in \Delta, \ \forall \boldsymbol{\eta} \in \mathbb{R}^{J-1}.$$

For the cumulative ratio we have

$$\pi_j = F(\eta_j) - F(\eta_{j-1}),$$

and thus, $\eta_{j-1} > \eta_j$ implies $\pi_j < 0$. Therefore, the family of cumulative models is not invertible. To illustrate the case of a non-invertible model, consider the effect on road accident severity caused by the speed limit (`speed_limit`), the road type (`road` and `urban_or_rural_area`), the light (`light_conditions`), and the weather conditions (`weather`) of the road where the accident occurred. We also considered as an explanatory variable the number of casualties (`number_of_casualties`) which is certainly an important factor in determining the severity of the accident. For this analysis, we used the data from 2019 openly available in https://data.gov.uk and accessible using the stats19 package (Lovelace et al., 2019). In the presence of the ordered response variable (accident severity with levels: slight, serious, and fatal), the ratio candidates to consider are cumulative, sequential, and adjacent. We first tried to fit the (*cumulative*, *Cauchy*, *complete*) model but due to the strong restriction of the cumulative ratio, the model failed to converge:

```
R> glmcat(accident_severity ~ road + urban_or_rural_area +
```

```
+     day_of_week + number_of_casualties + weather +
+     light_conditions + speed_limit, data = accidents,
+     parallel = F, ratio = "cumulative", cdf = "cauchy")

Warning messages:
1: In .GLMcat(formula = formula, data = data, ratio = ratio,
  cdf = cdf,  :
  Fisher matrix is not invertible. Check for convergence problems
```

One of the simplest ways of tackling this problem is to impose the constraint $\eta_{j-1} < \eta_j$ through the use of the parallel design. Evidently, the parallel constraint reduces the complexity of the Fisher's scoring algorithm since the condition to preserve the order in the successive iterations is only linked to the intercepts, i.e., $\alpha_{j-1} < \alpha_j$. However, to our knowledge, no previous research has investigated the validity of this constraint in iteration t after having imposed it in iteration 0. A widely used model with the parallel constraint is the odds proportional logit model. Its widespread popularity is due to the fact that it is ideal in terms of interpretation ease and of model parsimony (Abreu et al., 2008). However, in practice, the parallel assumption does not usually hold when considering more than one explanatory variable (Lall, 2002), thus, this restrictive assumption is often violated. Continuing with the example, the (*cumulative*, *Cauchy*, *parallel*) model is successfully fitted through:

```
R> summary(glmcat(accident_severity ~ road + urban_or_rural_area +
+     day_of_week + number_of_casualties + weather +
+     light_conditions + speed_limit, data = accidents,
+     parallel = T, ratio = "cumulative", cdf = "cauchy"))

accident_severity ~ road + urban_or_rural_area + day_of_week +
    number_of_casualties + weather + light_conditions + speed_limit
                ratio    cdf   nobs niter logLik
Model info: cumulative cauchy 109577  (10) -62593
                                Estimate Std. Error z value Pr(>|z|)
(Intercept) Slight              1.819854   0.103695   17.55  < 2e-16 ***
(Intercept) Serious            22.134261   0.544747   40.63  < 2e-16 ***
roadOne way street             -0.082418   0.096072   -0.86  0.39096
roadRoundabout                  0.247637   0.066977    3.70  0.00022 ***
roadSingle carriageway         -0.502254   0.030708  -16.36  < 2e-16 ***
roadSlip road                   0.431389   0.115313    3.74  0.00018 ***
urban_or_rural_areaUrban        0.265489   0.026713    9.94  < 2e-16 ***
day_of_weekMonday               0.033254   0.035015    0.95  0.34226
day_of_weekSaturday            -0.096729   0.033531   -2.88  0.00392 **
day_of_weekSunday              -0.196414   0.033577   -5.85  4.9e-09 ***
day_of_weekThursday            -0.036032   0.033538   -1.07  0.28266
day_of_weekTuesday              0.045625   0.034784    1.31  0.18963
day_of_weekWednesday           -0.007348   0.034093   -0.22  0.82935
number_of_casualties           -0.155693   0.009172  -16.98  < 2e-16 ***
weatherFine no high winds       0.106271   0.077432    1.37  0.16993
weatherFog or mist              0.370257   0.169997    2.18  0.02940 *
weatherRaining + high winds     0.081563   0.105066    0.78  0.43757
weatherRaining no high winds    0.226007   0.081171    2.78  0.00536 **
weatherSnowing                  0.549765   0.207627    2.65  0.00810 **
light_conditionsDaylight        0.209250   0.019735   10.60  < 2e-16 ***
speed_limit                    -0.010764   0.000921  -11.69  < 2e-16 ***
```

On the other hand, we observe that the adjacent and the sequential models are both invertible using any form of the linear predictor. We can write the probabilities of the $(adjacent, F, Z)$ models as:

$$\pi_j = \frac{\prod_{k=j}^{J-1} F(\eta_k)/(1-F(\eta_k))}{1+\sum_{k=1}^{J-1} F(\eta_k)/(1-F(\eta_k))},$$

and the probabilities of $(sequential, F, Z)$ models in the form:

$$\pi_j = F(\eta_j)\prod_{k=1}^{j-1}(1-F(\eta_k)).$$

In both cases, one can readily identify that $0 < \pi_j < 1$ for all $j \in \{1, ..., J-1\}$ such that $\sum_{j=1}^{J} \pi_j = 1$. If the slope effect is expected to be different for each category and the cumulative ratio fails to fit the model, the practitioner should consider the adjacent or sequential ratios instead. For our example, since the order of the response categories is not time-dependent, we fit the $(adjacent, Cauchy, complete)$ model obtaining:

```
R> summary(glmcat(accident_severity ~ road + urban_or_rural_area +
+     day_of_week + number_of_casualties + weather +
+     light_conditions + speed_limit, data = accidents,
+     parallel = F, ratio = "adjacent", cdf = "cauchy"))
```

```
accident_severity ~ road + urban_or_rural_area + day_of_week +
    number_of_casualties + weather + light_conditions + speed_limit
             ratio    cdf   nobs niter logLik
Model info: adjacent cauchy 109577  (14) -62060
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept) Slight                 1.97092    0.11264   17.50  < 2e-16 ***
(Intercept) Serious                9.89386    0.95812   10.33  < 2e-16 ***
roadOne way street Slight         -0.09133    0.10924   -0.84  0.40312
roadOne way street Serious         3.46606    3.39346    1.02  0.30707
roadRoundabout Slight              0.27694    0.07647    3.62  0.00029 ***
roadRoundabout Serious             6.13180    2.42201    2.53  0.01135 *
roadSingle carriageway Slight     -0.55281    0.03403  -16.25  < 2e-16 ***
roadSingle carriageway Serious    -1.06469    0.22537   -4.72  2.3e-06 ***
roadSlip road Slight               0.48098    0.13217    3.64  0.00027 ***
roadSlip road Serious             -0.36728    0.47703   -0.77  0.44134
urban_or_rural_areaUrban Slight    0.29358    0.02930   10.02  < 2e-16 ***
urban_or_rural_areaUrban Serious   0.58181    0.35379    1.64  0.10007
day_of_weekMonday Slight           0.03669    0.03844    0.95  0.33982
day_of_weekMonday Serious         -0.37198    0.31697   -1.17  0.24058
day_of_weekSaturday Slight        -0.10704    0.03657   -2.93  0.00342 **
day_of_weekSaturday Serious       -0.51803    0.29851   -1.74  0.08267 .
day_of_weekSunday Slight          -0.21198    0.03652   -5.80  6.4e-09 ***
day_of_weekSunday Serious         -0.44009    0.30439   -1.45  0.14822
day_of_weekThursday Slight        -0.03646    0.03678   -0.99  0.32151
day_of_weekThursday Serious       -0.20658    0.32653   -0.63  0.52695
day_of_weekTuesday Slight          0.05491    0.03830    1.43  0.15164
day_of_weekTuesday Serious        -0.50526    0.30869   -1.64  0.10168
day_of_weekWednesday Slight       -0.00162    0.03749   -0.04  0.96559
```

```
day_of_weekWednesday Serious          -0.58575   0.30017   -1.95  0.05101 .
number_of_casualties Slight           -0.17765   0.00967  -18.37  < 2e-16 ***
number_of_casualties Serious          -0.22305   0.03731   -5.98  2.3e-09 ***
weatherFine no high winds Slight       0.12416   0.08344    1.49  0.13675
weatherFine no high winds Serious      0.60544   0.42456    1.43  0.15385
weatherFog or mist Slight              0.38626   0.18480    2.09  0.03660 *
weatherFog or mist Serious            -0.43593   0.54573   -0.80  0.42441
weatherRaining + high winds Slight     0.10872   0.11401    0.95  0.34029
weatherRaining + high winds Serious    0.28617   0.56831    0.50  0.61458
weatherRaining no high winds Slight    0.25111   0.08761    2.87  0.00415 **
weatherRaining no high winds Serious   1.73089   0.53283    3.25  0.00116 **
weatherSnowing Slight                  0.59895   0.22914    2.61  0.00895 **
weatherSnowing Serious                 2.17619   2.17244    1.00  0.31648
light_conditionsDaylight Slight        0.23617   0.02150   10.99  < 2e-16 ***
light_conditionsDaylight Serious       0.86121   0.15687    5.49  4.0e-08 ***
speed_limit Slight                    -0.01167   0.00101  -11.60  < 2e-16 ***
speed_limit Serious                   -0.11368   0.01157   -9.82  < 2e-16 ***
```

In the analysis of the road accidents, the complete design could not be fitted using the cumulative ratio. For this reason, we considered the (*cumulative*, *Cauchy*, *parallel*) model for which the AIC results to be 125227. Then, we used the adjacent ratio aiming to investigate the complete design. From the reported output of the (*adjacent*, *Cauchy*, *complete*) model, one can observe that many of the explanatory variables are significant in their complete form. Although the number of model parameters was increased, we observed an improvement in terms of the AIC which for the adjacent model was 124199.

3.4.4 Total invariance

An (r, F, Z) model is said to be totally invariant if it is invariant under all permutations of the response categories. It is well known that the MNL or equivalently, the (*reference*, *logistic*, *complete*) model is totally invariant. Agresti (2010) demonstrated that this model is equivalent to the (*adjacent*, *logistic*, *complete*) model. To illustrate this equivalence, consider the two models: (*reference*, *logistic*, *complete*) and (*adjacent*, *logistic*, *complete*), for the `DisturbedDreams` dataset:

```
R> mod_ref_log_c <- glmcat(formula = Level ~ Age, ratio = "reference",
+     parallel = F, data = DisturbedDreams, cdf = "logistic")
R> mod_adj_log_c <- glmcat(formula = Level ~ Age, ratio = "adjacent",
+     parallel = F, data = DisturbedDreams, cdf = "logistic")
R> logLik(mod_ref_log_c); logLik(mod_adj_log_c)

'log Lik.' -277.1345 (df=6)
'log Lik.' -277.1345 (df=6)

coef(mod_ref_log_c)

(Intercept) Not.severe -2.454
(Intercept) Severe.1   -0.555
(Intercept) Severe.2   -1.125
Age Not.severe          0.310
Age Severe.1            0.060
Age Severe.2            0.112
```

```
coef(mod_adj_log_c)

(Intercept) Not.severe -1.8998
(Intercept) Severe.1    0.5700
(Intercept) Severe.2   -1.1246
Age Not.severe          0.2500
Age Severe.1           -0.0523
Age Severe.2            0.1123
```

Remark that the log-likelihoods of the last two models are equal but the estimations of the parameters are different. As demonstrated by Peyhardi et al. (2015), there exists a matrix A (see Appendix B.3 for details), such that $A\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ for the intercepts, and $A\boldsymbol{\delta} = \boldsymbol{\delta}'$ for the slopes.

```
R> A <- matrix(c(1, 0, 0, -1, 1, 0, 0, -1, 1), nrow = 3)
R> A %*% coef(mod_ref_log_c)[1:3]

[1,] -1.8998
[2,]  0.5700
[3,] -1.1246

R> A %*% coef(mod_ref_log_c)[4:6]

[1,]  0.2500
[2,] -0.0523
[3,]  0.1123
```

As well as the $(reference, logistic, complete)$, the $(adjacent, logistic, complete)$ model is totally invariant and therefore, it is inappropriate for ordinal responses. Apart from this model, any other model in the adjacent family preserves the order assumption, yet, this family has usually been ignored when dealing with ordinal responses.

3.4.5 Choice of an ordinal model

Several authors have suggested that the choice of the model to fit ordinal responses should correspond to the underlying nature of the response variable (Ananth and Kleinbaum, 1997; O'Connell, 2006; Agresti, 2010). On the basis of the above-mentioned properties, we can define some general guidelines for this choice. Firstly, it is important to differentiate whether the ordinal variable has a temporal foundation associated with the occurrence of the categories (time-ordered process); or if it was drawn from an ordered scale which would still be interpretable, even if the order of the categories is reversed (Figure 18 illustrates an example of this differentiation). In the time-ordered process scenario, it is assumed that to reach category j it was necessary to have visited the previous categories $1, \ldots, j-1$. Consider the example of the level of education attained by different people. Following a traditional academic path, it is possible to attend high school only after the completion of both elementary and middle school. In this case, the sequential ratio would be the best option to work with, since it is the one that best captures this dynamic process.

For the ordered scale response variables, either the cumulative or adjacent ratio can

be used since they are reversible. However, the adjacent ratio is invertible but there is no interpretation in terms of a latent variable. By contrast, the cumulative ratio relies on the latent variable formulation, but, it is not invertible (see Table 3.4). In practice, this means that when the practitioner wants to specify either a complete or a partial parallel design, some computation problems may occur when using the cumulative. Moreover, the interpretation via a latent variable does not hold for the cumulative ratio with a design different from the parallel. Therefore, as the adjacent ratio is invertible, the adjacent family of models should be preferred. Still, the cumulative models are the most widely used in literature. Perhaps, the unpopularity of the adjacent ratio is because, in the current software, the only possible cdf to use as part of the link function is the logistic cdf; even though, the (*adjacent*, *logistic*, *complete*) model is not appropriate for ordered responses due to its total invariance property. GLMcat is the first package that allows fitting any model from the family of models (*adjacent*, F, Z), where F and Z can be chosen from the range of options presented in the previous section.

	Reversibility	Latent variable interpretation	Invertibility	Not totally invariant
Sequential				
Cumulative		(i)		
Adjacent				(ii)

(i) true only with the parallel design Z_p,
(ii) true only if F is different from *Logistic* or if $Z \neq Z_p$.

Table 3.4: Properties of the ratios for ordinal responses; shaded cells indicate that the property is valid.

As for the matrix design, multiple alternatives can be considered. Some researchers prefer to start by using the parallel design for all the explanatory variables. If the model fits poorly they might include separate effects by considering the complete design for some or all of the explanatory variables. Other options to address the concern of an inadequate parallel assumption are using different cdfs or adding additional terms to the linear predictor. Furthermore, models with different designs can be compared using the AIC and/or the BIC as measures for parsimony. A conventional technique that aims to minimize some of these criteria is the stepwise variable selection, also available in GLMcat. Likewise, several options can be used for the cdf component. As mentioned above, the comparison of parameter estimates requires special attention if different distributions are specified. Furthermore, the assumptions of the model on the response are strongly shaped by the choice of the cdf. This is the case of the interaction between the adjacent ratio and the logistic cdf.

Figure 22 compiles the key points we have presented in this section. In summary, we recommend to use a sequential model if there is a time-ordered process among categories. If there is a scale ordering, use a cumulative or adjacent model since they are both reversible. Given that the interpretability through latent variables can be advantageous, we suggest favoring the use of a cumulative model whenever the assumption of parallelism is valid. Otherwise, opt for an adjacent model since it is invertible, and

we urge to not use the logistic cdf to avoid the total invariance. We intended to give some general recommendations, however, each analysis has its own particularities which should be addressed from each of the three angles specified by the ratio r, the cdf F, and the design matrix Z. We encourage the user to fit and compare a set of models under different criteria in order to find the (r, F, Z) triplet that best approaches their research questions.

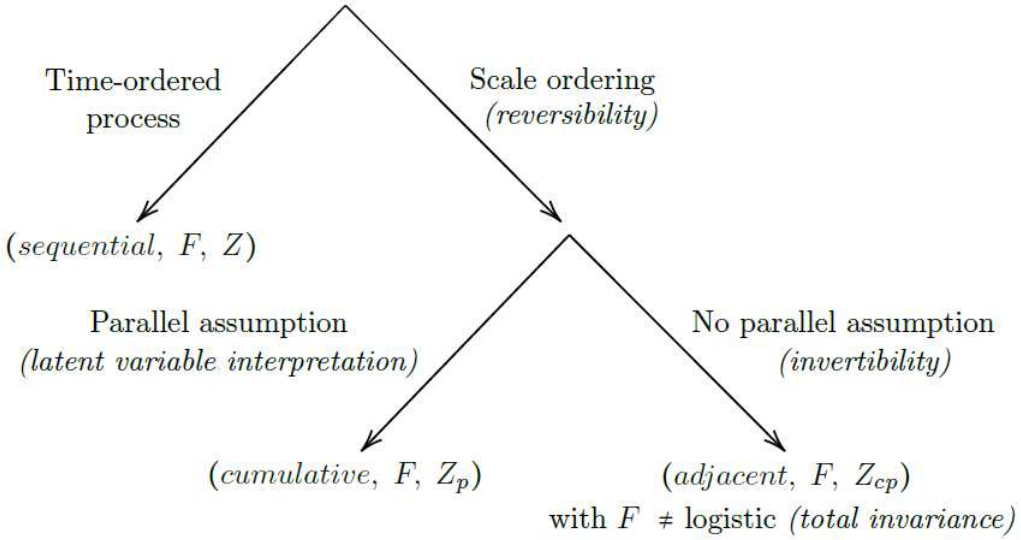

Figure 22: Schematic guide for choosing the appropriate ratio according to the characteristics of the response.

3.5 Models for nominal responses

The MNL is the most popular regression model for categorical responses. In the case of a nominal response, it is often the only model available; except in discrete choice (DC) theory where some extensions have been proposed. In this specific DC framework, the MNL can be interpreted in terms of an underlying behavioral model, the so-called random utility maximisation (RUM) model, i.e., $\mathsf{P}(Y = j) = \mathsf{P}(U_j = \max_k U_k)$, where $U_j = \eta_j + \varepsilon_j$ and ε_j's are independently Gumbel distributed. The U_j associated with each alternative j (category j) is called random utility. Two classical extensions are frequently used as RUM models: the multinomial probit (MNP) model, for which $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_J)$ follows a multivariate normal distribution, and the nested logit (NL) model, for which the residuals ε_j are independent and follow a generalized Gumbel distribution. There are some difficulties with the interpretation and the inference of these models. Since the MLE of the NL model cannot be directly obtained, the model estimation is computed either simultaneously (best alternative when there are less than four nested levels) or sequentially (which might lead to a suboptimal log-likelihood at convergence); more details about this are given by Forinash and Koppelman (1993) and Louviere et al. (2000). On the other hand, the estimation of MNP models can be complex (specially

when $J > 3$) due to the underlying multidimensional integrals of the multivariate normal density function (Geweke et al., 1994).

We propose to use an extension of the MNL in the DC framework based on the reference model's family: $(reference,\ F,\ Z)$, where Z can take into account variables specific to the alternatives $\{\boldsymbol{\omega}_j\}_{j=1,\dots,J}$. The linear predictor takes the general form: $\eta_j = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\delta}_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_J)^\top\boldsymbol{\gamma}$, for $j = 1, \dots, J-1$, thus, the design matrix, where $\tilde{\boldsymbol{\omega}}_j = \boldsymbol{\omega}_j - \boldsymbol{\omega}_J$, has the form:

$$\begin{pmatrix} 1 & & \boldsymbol{x}^\top & & & \tilde{\boldsymbol{\omega}}_1^\top \\ & \ddots & & \ddots & & \vdots \\ & & 1 & & \boldsymbol{x}^\top & \tilde{\boldsymbol{\omega}}_{J-1}^\top \end{pmatrix}.$$

Note that these models are invariant only under the permutation that fixes the reference alternative (see Peyhardi et al., 2015, for details). In other words, contrary to the MNL, changing the reference alternative leads to a different model (except if $F =$ logistic). The advantages of the reference models (versus MNP or NL) are:

- they include MNL as a special case ($F =$ logistic),
- their simple inference procedure (Fisher's scoring algorithm),
- their simple interpretation since each alternative is compared to a reference alternative $\frac{\pi_j}{\pi_j + \pi_J} = F(\eta_j)$.

Another good property is the invertibility which is evident from writing the probabilities of the model in the form:

$$\pi_j = \frac{F(\eta_j)/(1 - F(\eta_j))}{1 + \sum_{k=1}^{J-1} F(\eta_k)/(1 - F(\eta_k))}.$$

It should be remarked that the reference models are DC models but not RUM models. Moreover, the $(reference,\ normal,\ Z)$ model is different from the MNP model.

We propose to use the family of $(reference,\ Student(\nu),\ Z)$ models, which is an alternative that grants robustness and flexibility through the Student cdf. Indeed, Peyhardi (2020) showed that the influence function is bounded with the Student cdf (contrary to the logistic or normal cdfs). Consequently, these models are less sensitive to outliers than the MNL, in addition, they seem to be less sensitive to noisy explanatory variables. The Student cdf itself generates a family of models as different fits are expected when changing ν. The flexibility we previously mentioned, lies in the increase of the range of possible cdfs to consider as part of the link function. For instance, the three most popular link functions can be obtained with

- $\nu = 1 \Rightarrow F_\nu =$ Cauchy,
- $\nu = 8 \Rightarrow F_\nu \simeq$ logistic, and
- $\nu \to \infty \Rightarrow F_\nu =$ normal.

The Student distribution has been further extended by using a non-centrality parameter μ. This generalization is known as the non-central t distribution. The resulting cdf is also available in GLMcat and can be used by specifying its parameters: `cdf = list("noncentralt", df = 5, mu = 2)`. Note that the non-central t distribution is asymmetric unless $\mu = 0$ (in which case it is equivalent to the Student cdf). A detailed description of this distribution can be found in Johnson et al. (1995) and its pdf is recalled in Appendix B.1.

To estimate these models in GLMcat, we create the function `discrete_cm()` which requires data in a long format (an example is given in the following). Thus, for each individual (or decision-maker), there are multiple observations (rows), one for each of the alternatives the individual could choose. We call the group of observations for an individual a case. Each case represents a single statistical observation (although it comprises multiple observations), and the identification column of the n cases should be specified in the argument `case_id`. The user must be aware that the `discrete_cm()` function has been built for the particular case of explanatory variables specific to the alternatives. If not required, the user can call the `glmcat()` function using the reference ratio.

Application

Consider the dataset studied by Louviere et al. (2000) in which 210 passengers choose one travel mode among the $J = 4$ options: air, train, bus, and car (available in GLMcat as the `TravelChoice` object). In this analysis, the individual's attributes are the household income (hinc) and the traveling group size (psize). The alternative specific attributes for each travel mode are the generalized cost (gc) and the terminal waiting time (ttme). The dataset has a long format, i.e., the variables concerning the n individuals are detailed in $n \times J$ lines; an example for the first two individuals is:

```
R> head(TravelChoice, 8)
  indv  mode choice ttme invc invt gc hinc psize
1    1   air  FALSE   69   59  100 70   35     1
2    1 train  FALSE   34   31  372 71   35     1
3    1   bus  FALSE   35   25  417 70   35     1
4    1   car   TRUE    0   10  180 30   35     1
5    2   air  FALSE   64   58   68 68   30     2
6    2 train  FALSE   44   31  354 84   30     2
7    2   bus  FALSE   53   25  399 85   30     2
8    2   car   TRUE    0   11  255 50   30     2
```

In the following, we estimate and compare a set of models with different cdfs and with different specifications of the reference category.

Logistic cdf We first estimate the ($reference$, $logistic$, $Z^{(1)}_{\text{car}}$) model (which corresponds to the MNL) considering car as the reference category, the associated design

matrix where h represents hinc, c for gc, t for ttme, and p for psize, is:

$$Z_{\text{car}}^{(1)} = \begin{pmatrix} 1 & 0 & 0 & h & 0 & 0 & p & 0 & 0 & c_{air} - c_{car} & t_{air} - t_{car} \\ 0 & 1 & 0 & 0 & h & 0 & 0 & p & 0 & c_{bus} - c_{car} & t_{bus} - t_{car} \\ 0 & 0 & 1 & 0 & 0 & h & 0 & 0 & p & c_{train} - c_{car} & t_{train} - t_{car} \end{pmatrix}.$$

```
R> logistic_car <- discrete_cm(formula = choice ~ hinc +
+     psize + gc + ttme, case_id = "indv", alternatives = "mode",
+     reference = "car", data = TravelChoice, alternative_specific = c("gc",
+       "ttme"), cdf = "logistic")
R> summary(logistic_car)

                ratio       cdf nobs niter     logLik
Model info: reference logistic  210   (5) -177.4541
                    Estimate Std. Error z value Pr(>|z|)
X.Intercept. air    7.873608   0.986848   7.979 1.48e-15 ***
X.Intercept. bus    4.433192   0.778334   5.696 1.23e-08 ***
X.Intercept. train  5.559205   0.699139   7.952 1.84e-15 ***
hinc air            0.004071   0.012725   0.320 0.749020
hinc bus           -0.023324   0.016297  -1.431 0.152391
hinc train         -0.055185   0.014482  -3.810 0.000139 ***
psize air          -1.027423   0.265657  -3.867 0.000110 ***
psize bus          -0.030010   0.333977  -0.090 0.928402
psize train         0.302395   0.225616   1.340 0.180144
gc                 -0.019685   0.005401  -3.644 0.000268 ***
ttme               -0.101566   0.011231  -9.044  < 2e-16 ***

R> logLik(logistic_car)

'log Lik.' -177.4541 (df=11)
```

A more specific design was studied by Louviere et al. (2000, p. 157) and Greene (2003, p. 730). These analyses set the effect of the variables hinc and psize exclusively for the category air, i.e.,

$$\eta_j = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\delta}_{air} \mathbb{1}_{j=\text{air}} + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_{car})^\top \boldsymbol{\gamma} \tag{3.5}$$

for $j \in \{\text{air, bus, train}\}$. Hence, the associated design matrix is:

$$\begin{pmatrix} 1 & 0 & 0 & h & p & c_{air} - c_{car} & t_{air} - t_{car} \\ 0 & 1 & 0 & 0 & 0 & c_{bus} - c_{car} & t_{bus} - t_{car} \\ 0 & 0 & 1 & 0 & 0 & c_{train} - c_{car} & t_{train} - t_{car} \end{pmatrix}.$$

As far as we know, there is no other package in R to fit this particular design. In GLMcat, we can fit this model with the lines of code:

```
R> logistic_car_alt <- discrete_cm(formula = choice ~
+     hinc[air] + psize[air] + gc + ttme, case_id = "indv",
+     alternatives = "mode", reference = "car", data = TravelChoice,
+     alternative_specific = c("gc", "ttme"), cdf = "logistic")
R> summary(logistic_car_alt)
```

```
"choice ~ hinc[air] + psize[air] + gc + ttme+indv+mode"
                  ratio       cdf nobs niter     logLik
Model info: reference logistic  210   (5) -185.9149
                      Estimate Std. Error z value Pr(>|z|)
X.Intercept. air      7.334807   0.946436   7.750 9.19e-15 ***
X.Intercept. bus      3.591702   0.475771   7.549 4.38e-14 ***
X.Intercept. train    4.371913   0.478124   9.144  < 2e-16 ***
hinc air              0.023815   0.011189   2.128   0.0333 *
psize air            -1.173817   0.258133  -4.547 5.43e-06 ***
gc                   -0.023507   0.005084  -4.624 3.76e-06 ***
ttme                 -0.100213   0.010543  -9.505  < 2e-16 ***

R> logLik(logistic_car_alt)

'log Lik.' -185.9149 (df=7)
```

Student cdf Now, we use the Student cdf as one of the link function's components. Note that the design given by the linear predictor in Equation 3.5 depends on the reference alternative j_0. Since reference models are not invariant to a change of the reference alternative j_0, we have to select j_0. To that end, we fitted the models $(reference,\ Student(\nu),\ Z_{j_0}^{(1)})$ where the reference alternative j_0 is either air, bus, train or car, and the values of ν are taken with a 0.05-step from 0.2 to 2 and an integer-step from 2 to 20. We notice in Figure 23 that for $\nu = 8$ the log-likelihoods of the four models (one for each alternative) are close to the same value around -177. This is not surprising since the logistic cdf (which results in approximately the same fit as using $Student(8)$) is the only one to offer the invariance property under all permutations of alternatives.

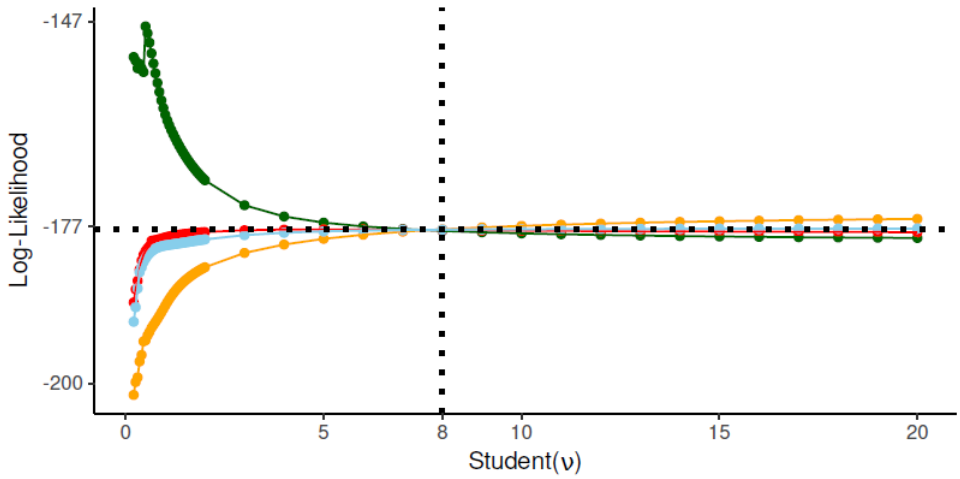

Figure 23: Log-likelihood curves for models with reference category: car (green), bus (yellow), train (red) and air (blue), and with ν with a 0.05-step from 0.2 to 2 and an integer-step from 2 to 20.

The highest log-likelihood (−147) is obtained with *car* as reference category together with $F = Student(0.496)$:

```
R> summary(discrete_cm(formula = choice ~ hinc + psize +
+     gc + ttme, case_id = "indv", alternatives = "mode",
+     reference = "car", alternative_specific = c("gc",
+         "ttme"), data = TravelChoice, normalization = 0.95,
+     cdf = list("student", df = 0.496)), normalized = T)

Normalized coefficients with s0 =  0.069308
[1] "choice ~ hinc + psize + gc + ttme+indv+mode"
                 ratio      cdf nobs niter  logLik
Model info: reference student  210    11 -147.15
                    Estimate Std. Error z value Pr(>|z|)
X.Intercept. air     5.68637    1.95182    2.91   0.0036 **
X.Intercept. bus     2.32873    1.00254    2.32   0.0202 *
X.Intercept. train   2.84782    0.93167    3.06   0.0022 **
hinc air             0.00838    0.00789    1.06   0.2878
hinc bus             0.00207    0.01145    0.18   0.8563
hinc train          -0.00698    0.00773   -0.90   0.3661
psize air           -0.22916    0.25595   -0.90   0.3706
psize bus            0.41950    0.29757    1.41   0.1586
psize train          0.29810    0.15220    1.96   0.0502 .
gc                  -0.00465    0.00262   -1.78   0.0754 .
ttme                -0.09182    0.02916   -3.15   0.0016 **
```

Based on the above model, only the terminal waiting time seems to have a significant effect on the choice of travel mode. We can notice that the model with only this explanatory variable has a log-likelihood quite close to −147:

```
R> logLik(discrete_cm(formula = choice ~ ttme, case_id = "indv",
+     alternatives = "mode", reference = "car", data = TravelChoice,
+     alternative_specific = "ttme", cdf = list("student",
+         df = 0.496)))

'log Lik.' -149.0041 (df=4)
```

This raises the question of whether the choice of transport mode can be completely determined by the explanatory variable ttme. Since the terminal waiting time for the alternative car is null, it is possible to represent the different values of this variable by points in three dimensions (air, bus, and train) with a color indicating the observed travel mode.

The reader can visualize in Figure 24 that the dataset is completely artificial. Note that there are only two triplets $(ttme_{air}, ttme_{bus}, ttme_{train})$ for which users choose car, these are: $(69, 35, 34)$ and $(64, 53, 44)$. These points are the intersection of the lines formed by the other choices. Remark that, for instance, regardless of the value $ttme_{air}$, users always choose the air option if $ttme_{bus} = 35$ and $ttme_{train} = 34$ or if $ttme_{bus} = 53$ and $ttme_{train} = 44$. Similar rules can be defined to determine when individuals choose to travel by bus or by train. Concretely, knowing the terminal waiting time, the travel mode choice becomes deterministic, all the other explanatory variables considered in this analysis are only noise. In contrast to the logistic cdf, the Student cdf allows us to discover the completely artificial nature of this classical dataset. This was possible because the Student cdf seems to be more robust to outliers and noise variables (for more details see Peyhardi (2020)).

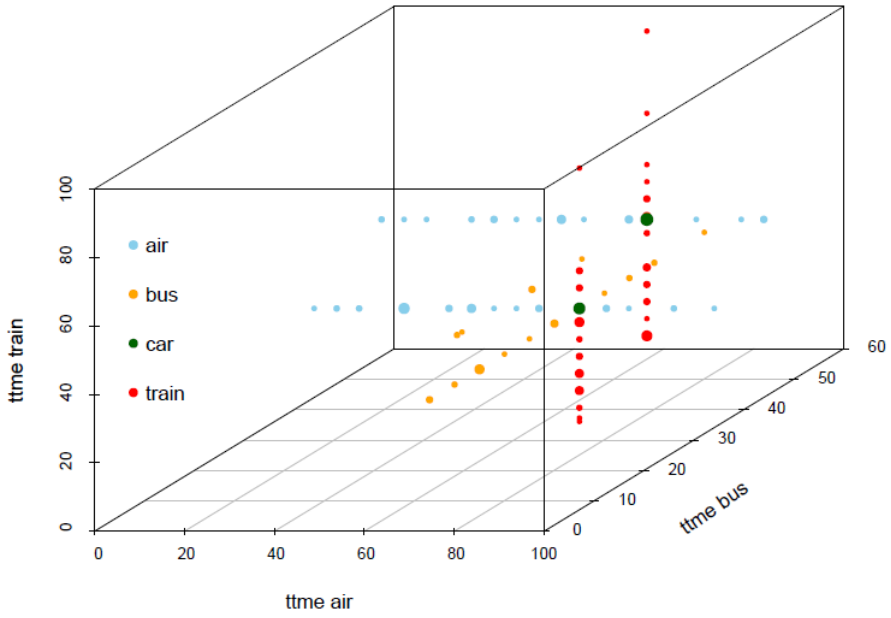

Figure 24: Three-dimensional representation of observed terminal time values. The sizes of the points are proportional to the number of individuals who chose the travel option among: air (blue), bus (yellow), car (green), train (red).

3.6 Discussion

Liu and Agresti (2005) presented an overview of developments in the analysis of ordinal responses. In their final comments, they highlighted that the current main challenge is to make these methods better known to researchers who commonly encounter this kind of data. Up to now, the models for categorical responses have been popularized in different disciplines separately. We consider that once all the models are assembled, their specific characteristics can be better understood and, thus, users can readily compare and choose a solution tailored to the objectives of their analysis. In the present article, we illustrated a generalized modeling framework for categorical responses, while introducing an R package that encompasses all these models. The contributions presented in this paper have wide applicability given that several fields of research and industry deal extensively with categorical responses. We discussed the properties of the different families of models, as well as the relevance of the choice of both the cdf and the linear predictor's form. With the GLMcat package, it is now computationally possible to test a variety of categorical regression models using one simple function. We consider that this tool allows to popularize the area of categorical data regression which has not been yet widespread on a large scale through non-logistic models.

Although the most popular cdfs often result in "similar" fits, this does not imply that all cdfs are essentially equivalent when fixing the ratio. In distributions such as the Pregibon (based on the generalized Tukey family) or the non-central t, some parameters control the symmetry, the heaviness of the tails, and/or the skewness of the distribution. Hence, one extension would be to consider an algorithm to estimate such parameters. The vast set of new possible cdfs enlarges the toolkit for modeling categorical responses,

with the use of them, subtle details might be uncovered as illustrated in the example of section 3.5. An advantage of the modularized architecture of the package is that it facilitates the inclusion of additional cdfs which will be immediately available for all four model families.

The hierarchical structure of nominal, ordinal, or partially ordinal responses has been already studied among others by Zhang and Ip (2012) and Peyhardi et al. (2016). Based on the presented methodology, we can consider the (r, F, Z) triplets as basic units of a hierarchically structured model. This general and flexible model allows taking into account possible relations among response categories. The hierarchical model is then defined by a partition tree where, for each non-terminal node, an (r, F, Z) model is specified. Remark that in this case, the link function would be composed of the tree partition, the set of ratios, and the cdfs specified for the non-terminal nodes.

The GLM presented in section 3.2 can be extended to include random effects. Some authors have already made this extension for particular models in the context of categorical responses (see Hartzel et al., 2001; Coull and Agresti, 2000; Tutz and Hennevogl, 1996). The implementation of generalized linear mixed models is envisaged for the (r, F, Z) in GLMcat.

In the regression framework, other essential tasks are the model regularization and the variable selection. These techniques aim to reduce the space of explanatory variables while improving the model estimation and the prediction accuracy. We propose within the functionalities of the GLMcat package the conventional stepwise approach. In high-dimensional problems, it is also important to consider regularization methods. For categorical variables, the elastic net penalty can be applied to categorical variables with the ordinalNet package, however, it is only designed for three ratios. As future work, we will attempt to define regularization and variable selection methods that are valid for any $(r,\ F,\ Z)$ triplet. We expect, with this extension, more detailed and accurate results, for instance, by means of the Student cdf which is less sensitive to noise variables and thus will improve the variable selection task.

Chapter 4

Hierarchically Structured GLMs for categorical responses

Abstract

The hierarchical representation of data can be meaningful within the regression framework for categorical responses. Indeed, response categories are likely to enclose others, leading to successive subdivisions. In real problems, the hierarchical structure of the categories is, in most cases, unknown beforehand. When known, the complexity is to build a model at each of the non-terminal vertices of the hierarchical structure. When unknown, an additional task is to infer the hierarchical structure of the response categories. Since the number of structures grows exponentially according to the number of categories, estimating all the structures can be difficult and time-consuming. Hence, we first propose a heuristic to build an initial structure. Then, we propose an algorithm to visit the space of structures neighboring this baseline structure, aiming to find a better fit for the data at hand.

Keywords: hierarchical GLM, categorical response, link function, binary models, tree's operations

Contents

4.1 Introduction

Within categorical variables, two types of scales are broadly differentiated: the nominal and the ordinal ones. There has been little discussion in the scientific literature about partially ordered variables for which order scales may exist in subsets of the response categories. The concept of hierarchical structure for such categories is even more general (albeit poorly known) since it does not necessarily imply an ordered relationship but refers to the grouping of categories under some similarity criterion. These structures are commonly encountered in real data scenarios, yet they are under-recognized and under-reported in the statistics literature (Zhang and Ip, 2012). For instance, when diagnosing melanoma cancer, the objective is to detect its presence and severity. Hence, a binary variable splits the levels of presence/absence, while an ordinal variable defines the severity assessed in terms of melanoma thickness (Sánchez-Monedero et al., 2018). This categorical variable is said then to be partially ordered. It can be represented graphically through a partition tree in which a first ramification splits absence and presence, and a second ramification (for the vertex presence) reveals the ordered stage of the disease. Another typical example is observed frequently in choice theory. Initially, the choice-makers are given a range of options that may be grouped according to their preferences. Hence, the first decision corresponds to choosing a group of options. Then, they evaluate the options within that chosen group. This kind of model may be defined recursively for many levels. In this example, unlike that of the diseases, there is no internal order associated with any of the response categories. However, there is still an associated hierarchical structure. In this chapter, we consider categorical variables in the broadest possible sense, which is when we expect a hierarchical structure to define the relationships (of order, partial order, or no order at all) among not only the response categories but also the possible groups of categories.

A hierarchical model for a categorical response is obtained by successively modeling the response in groups formed by homogeneous categories. To model this response, several partitioned conditional regression models have been proposed in different applied fields, including econometrics, medicine, and psychology. In econometrics, McFadden et al. (1978) introduced the nested logit model, which enables to decompose the decision mechanism in different steps but whose estimation's complexity increases when more than 2 or 3 levels are used. Tutz (1989) introduced the two-step model (also known as the cumulative compound model) to take account of subsets of categories such that the explanatory variables are informative for the between subsets choice, but maybe relatively uninformative within the subsets (see also Morawitz and Tutz, 1990). Zhang and Ip (2012) proposed the partitioned conditional model, a class of GLM intended primarily to model partially ordered responses, but which also includes nominal and ordinal responses as special cases. More recently, Peyhardi et al. (2016) introduced the partitioned conditional GLM (PCGLM) to analyze the hierarchical structure of a response with any number of categories and with the flexibility of using any model at each partition of the structure. The above-presented options are only operative if the hierarchical structure is a priori known. If the structure is unknown or partially unknown, the partition tree must first be defined based on the available data. To the best of our knowledge, no methodology exists to find such a structure automatically.

This task's complexity is directly related to the number of response categories. As the number of categories increases, the number of partition trees grows exponentially. Thus, it becomes impractical to visit the entire set of partition trees to select the best among them.

This chapter proposes a modeling approach for cases where a hierarchical structure in the response categories is expected to exist and is not known in advance. The link function in this model is specified by the partition tree, and the link functions at all non-terminal vertices. Our proposal considers only binary partition trees because of the two following advantages. The first one is that it reduces the space of possible partition trees. The second one is that the model's specification at each non-terminal vertex is simplified since neither the ratio of probabilities r nor the design matrix Z (see chapter 3) need to be tuned in binary regression models. Indeed only the cdf F is required. To find a binary partition tree of the response categories according to the explanatory variables, we make use of the notions of the hierarchical agglomerative clustering (HAC) algorithm. With the particularity that instead of grouping individuals, we consider the dissimilarities among the clusters generated by the J categories as starting points of the algorithm. The final output of this algorithm is a labeled and non-ranked dendrogram (equivalent to a binary partition tree). However, this proposal being a heuristic does not guarantee finding the most optimal structure among the large set of possibilities. Therefore, to search for a partition tree with a higher log-likelihood, we propose two greedy algorithms to explore alternative options structurally closer to the initially proposed tree. To restructure the binary partition trees in the search algorithms, we used the basic operation of trees known as *rotation* (see Lucas, 1987). We define the space of neighboring partition trees based on the possible rotations on all the non-terminal vertices of the baseline tree. As these rotated trees have a common sub-tree structure, we thus use the decomposition property of the log-likelihood to avoid recomputing the tree's log-likelihood for each new neighbor.

This chapter is structured as follows. In section 4.2, we consider the case where the partition structure is known. Here, we introduce the concepts about PCGLMs and illustrate them using the data of the rice diversity. In section 4.3, we propose a representation of a binary partition tree using a dendrogram and a matrix. In section 4.4, we present the methodology for constructing a B-PCGLM in two steps: i) the construction of the structure itself and ii) the specification of a binary model at each non-terminal vertex. We illustrate our methodology by walking step by step through the construction of a B-PCGLM for the rice diversity data set. Finally, in section 4.5, we evaluate our methodology in terms of log-likelihood and classification accuracies using a benchmark data set.

4.2 Known structure: the partitioned conditional GLM

The class of PCGLMs was introduced by Peyhardi et al. (2016) as a flexible framework for modeling nominal, ordinal, or even partially ordered responses. In the following, we introduce the necessary notations and definitions to describe this class of models.

A directed tree $\mathcal{T}$ is a connected graph $G = (V, E)$ where V is the set of vertices (also referred to as nodes) and E is the set of edges. In this framework, $\mathcal{V}^*$ denotes the set of non-terminal vertices, $Pa(v)$ is the parent vertex of v, $An^*(v)$ denotes the ancestors set of v except the root. The children must be indexed because the GLMs are not necessarily invariant under a permutation of the response categories (see section 3.4). Children $\Omega_1^v, \ldots, \Omega_{J_v}^v$ are presented from left to right and $\Omega_{J_v}^v$ is considered as the reference child by convention. The vertex with no parents is said to be the tree's root. For $u, v \in V$, a directed edge $e = (u, v) \in E$ implies that e is directed from u to v.

Definition 1. A directed tree $\mathcal{T}$ is said to be a partition tree of $\{1, \ldots, J\}$ if

- sibling vertices constitute a non-identical partition of their parent node,
- $\{1, ..., J\}$ is the root of $\mathcal{T}$,
- each singleton $\{j\}$ (the leaves) belongs to $\mathcal{T}$.

Consider the regression context where the response Y is a categorical variable with J categories. Suppose that these categories are subject to be grouped (possibly at more than one level) according to specific criteria such as similarity, order, or partial order. A partition tree $\mathcal{T}$ appropriately represents this scenario where the root is the complete set of categories $\{1, \ldots, J\}$, the leaves are the individual categories $i \in \{1, \ldots, J\}$, $\mathcal{V}^*$ is the set of groups of categories, and $\mathcal{T}$ is directed from the root to the leaves.

Definition 2. A partitioned conditional GLM of categories $\{1, \ldots, J\}$ (PCGLM) is specified by

- a partition tree $\mathcal{T}$ of $\{1, \ldots, J\}$,
- a collection of models $\mathfrak{C} = \{(r^v, F^v, Z^v(x^v)) \mid v \in \mathcal{V}^*\}$ for each conditional probability vector $\pi^v = (\pi_1^v, \ldots, \pi_{J_v-1}^v)$, where $\pi_j^v = P(Y \in \Omega_j^v | Y \in v; x^v)$ and x^v is a sub-vector of x associated with vertex v.

Based on the above definition, the probability of each category j is obtained as

$$P(Y = j|x) = P(Y = j|Y \in Pa(j), x^{Pa(j)}) \prod_{v \in An^*(\{j\})} P(Y \in v|Y \in Pa(v), x^{Pa(v)}),$$

where $P(Y \in v|Y \in Pa(v), x^{Pa(v)})$ is described by the GLM of $\mathfrak{C}$ associated with vertex $Pa(v)$. Note that the GLMs in $\mathfrak{C}$ are specified through the (r, F, Z) specification proposed by Peyhardi et al. (2015) and discussed in detail in chapter 3. An example of a PCGLM is shown in Figure 25 for a reponse with 6 categories. Remark that there exists an order of the groups of response categories $\{\{1, 2, 3\}, \{4\}, \{5, 6\}\}$ described by the sequential ratio. In addition, the individual categories of the group $\{1, 2, 3\}$ are also ordered and, in this case, modeled with the cumulative ratio. Since the response of the group $\{5, 6\}$ is binary, there is no choice to be made about the ratio in this vertex.

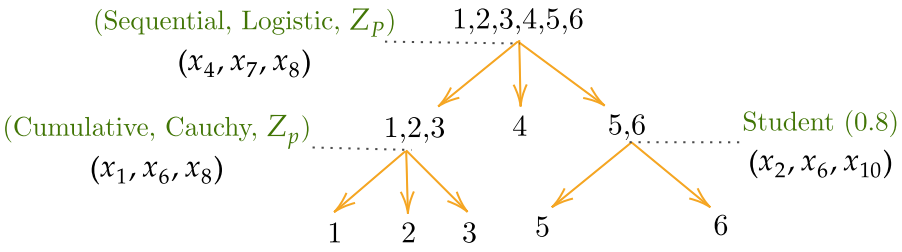

Figure 25: Illustration of a PCGLM with 6 categories.

Estimation of PCGLMs

Using the partitioned conditional structure of the model, the log-likelihood can be decomposed as follows

$$l = \sum_{v \in \mathcal{V}^*} l^v, \tag{4.1}$$

where l^v represents the log-likelihood of the model (r^v, F^v, Z^v). If all parameters are not constrained to equality from one vertex to another, each component l^v can be maximized separately on the sub-data set $\{(y, x)|y \in v\}$ using equation

$$\frac{\partial l^v}{\partial \boldsymbol{\beta}} = Z^{v\top} \frac{\partial \boldsymbol{F}^v}{\partial \boldsymbol{\eta}^v} \frac{\partial \boldsymbol{\pi}^v}{\partial \boldsymbol{r}^v} \mathsf{COV}(\boldsymbol{Y}^v|\boldsymbol{x}^v)^{-1} (\boldsymbol{y}^v - \boldsymbol{\pi}^v) \, .$$

Binary partitioned conditional GLM

A binary partition tree corresponds to a partition tree of a PCGLM with exactly $J-1$ non-terminal vertices. We represent this type of structure in Figure 26. The collection of binary regression models $\mathfrak{C}$ is simplified. Hence, we obtain $\pi^v = F^v(\alpha^v + \boldsymbol{x}^v \boldsymbol{\delta}^v)$ for all $v \in \mathcal{V}^*$. Remark that the $(sequential, F, complete)$ model can be represented as a B-PCGLM due to its successive ordered binary splits.

4.2.1 Application to the rice diversity data set

Panicle traits are among the most representative features of rice diversity; their architecture is relevant for the biological classification of plants, as well as for the improvement of cultivated rice. Groups of genotypes are often represented as a hierarchy of categories (e.g., species subdivided into groups of varieties or subspecies). For this type of data, it is usually aimed to unconstrainedly incorporate heterogeneous phenotypic traits (qualitative, quantitative, ordinal, and count variables) into a model decomposed according to such structure. Based on the PCGLMs, we intend to explain the taxonomic classification of rice given a collection of phenotypic features. The data set that we use in this application consists of 960 panicles of rice (see Al-Tam et al., 2013, for further reference on the data set). Each plant is classified according to its geographical origin, species, subspecies, and sub-population. For each continent (Asia, Africa), one cultivated and

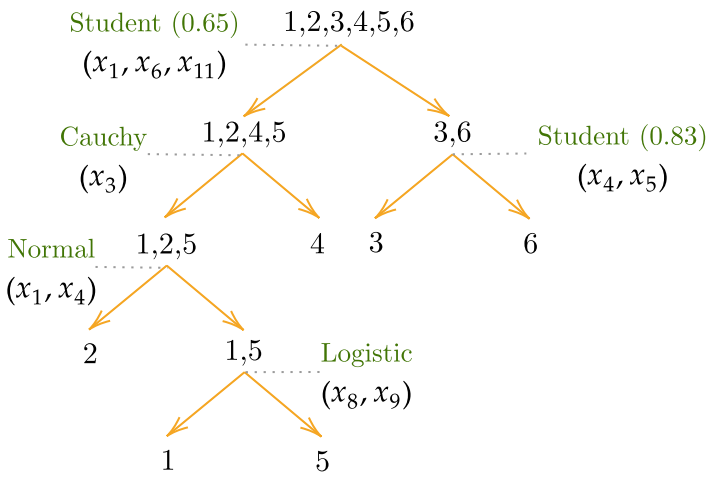

Figure 26: Illustration of a B-PCGLM with 6 categories.

one wild species are considered: *Sativa* (Asia-cultivated), *Rufipogon* (Asia-wild), *Glaberrina* (Africa-cultivated), and *Barthii* (Africa-wild). The abbreviations for the species in the following graphical representations of the rice diversity analysis are as follows: B for *Barthii*, R for *Rufipogon*, G for *Glaberrina*, S for *Sativa*; and for the sub-species: OBI for *Obar I*, $OBII$ for *Obar II*, OGI for *Ogla I*, $OGII$ for *Ogla II*, Ja for *Japonica*, Tr for *Tropical Japonica*, Te for *Temperate Japonica*, In for *Indica*, and, Ad for *Admix (Aromatic - Te)*. The explanatory variables in our context are the phenotypic traits. In the results presented in this chapter, we represent them using the following abbreviations: *lr*: length of the rachis, *tl*: total length, *ng*: number of grains, *mo*: maximum number of branching order, *nn*: number of nodes, and *nnr*: number of nodes in the rachis. The structure of this response is partially known since there are two possible ways of constructing the taxonomic hierarchy for the subspecies. In the partition tree presented in Figure 27, the data are initially divided according to the geographical origin (Africa or Asia), and secondly according to the domestication trait (wild or domestic). Whereas in the second partitioning tree presented in Figure 28, the splits are initially based on the domestication trait and subsequently on the geographic origin.

Since we knew the hierarchical structure of the categories beforehand, the task was to identify the best model (r, F, Z) for each non-terminal vertex. That is, select among the options for the ratio r: cumulative, sequential, adjacent, or reference. Select in turn the cdf that best characterizes the response variable and matches r appropriately. And select the relevant phenotypic variables for each of the non-terminal vertices of the partitioning tree. If a vertex has more than two children, the choice of the ratio is essential to reflect the notion of order among the categories correctly. For the particular case of the three subspecies of *Sativa* (with 3 children), Huang et al. (2012) presented a demographic scenario in which he claimes that *Japonica* was first domesticated from the wild species *Or-IIIa*, while *Indica* later developed from *Or-I* with the adoption of numerous domestication alleles from *Japonica*. The *Aromatic* group is considered to have been domesticated shortly after *Japonica* and *Indica* came into existence. After

fitting the models resulting from combinations of the four different ratio options and various cdfs, we observed that the best fit in terms of log-likelihood was obtained with the sequential ratio together with the Cauchy cdf. This result is in line with the chronological approach outlined previously, as the sequential ratio is the one that best captures a non-reversible temporal order (see chapter 3 for the reversibility property). Since the remaining non-terminal vertices have only two children, their corresponding models are binary; hence, all that remained to be chosen was the cdfs and the set of explanatory variables that maximized the log-likelihood. For this analysis, we estimated all the models with the logistic, normal, Cauchy, and Gompertz cdfs and used the BIC-stepwise algorithm (combining the backward and the forward direction) to find the explanatory variables in each of the non-terminal vertices. In Figures 27 and 28, we present the models that yielded the highest log-likelihood.

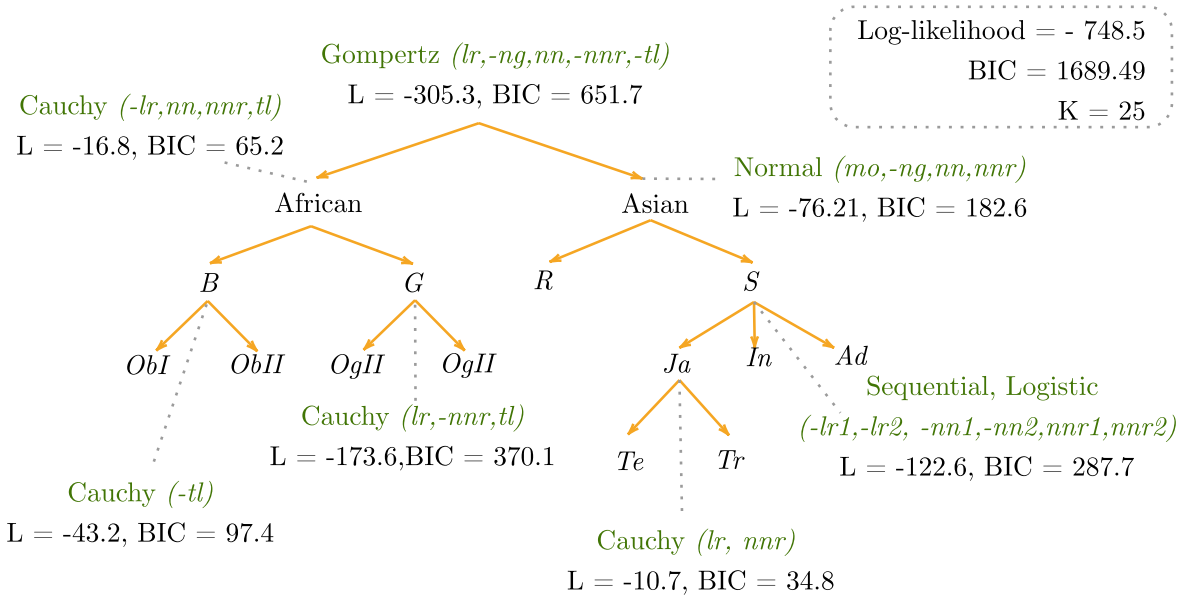

Figure 27: PCGLM for rice diversity data where the data are divided according first to the geographical origin (Africa or Asia) and second to the domestication trait (wild or domestic). The vertex at the right position is the reference category for each model estimation.

The tree presented in Figure 28 differs from the previous one only in the binary models that consider wild and cultivated subspecies and then the geographic origin as the response levels. When the two trees are compared using the BIC criterion, the PCGLM in Figure 28 (BIC = 1636) is favoured over the PCGLM in Figure 27 (BIC = 1689). Although the difference is not large, these results are consistent with published studies showing that differences in phenotypic traits are more pronounced between wild and cultivated groups than between Asian and African groups. However, population genomic analyses and data sequencing techniques also demonstrate strong genetic differentiation between Asian and African rice.

The number of nodes in the panicle is one of the most discriminating variables in

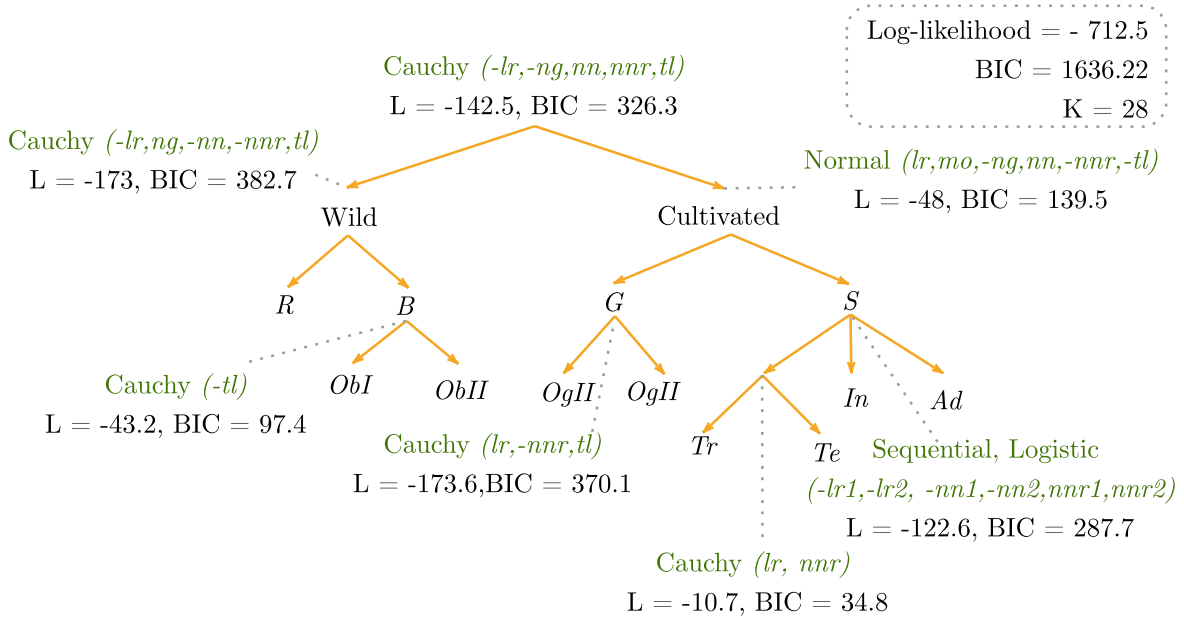

Figure 28: PCGLM for rice diversity data where the data are divided according first to the domestication trait (wild or domestic) and second to the geographical origin (Africa or Asia). The vertex at the right position is the reference category for each model estimation.

each PCGLM. It is strongly correlated with the grain quantity, i.e., the productivity, as seen in the simple interaction model (not shown here), in which this variable is found to be the most important in differentiating between cultivated and wild rice types.

We identified the effects of phenotypic traits on the description and differentiation of taxonomic categories. Given the slight difference between the BICs of each tree, it is not clear that one of the considered partitions is better than the other. One of the main difficulties in analyzing plant diversity is that, although phenotypic traits are generally specific to each taxonomic category, they also vary under different environmental conditions. This limitation is evident in other studies that only manage to describe relatively subtle differences for each taxonomic level.

4.3 Representation of a binary partition tree

The model for hierarchically structured categorical variables comprises the partition tree (representing the structure) and the collection of classical models to be fitted at each non-terminal vertex of the partition tree. This situation is common in different disciplines; however, to our knowledge, there is no methodology to find such a hierarchical structure for the groups generated by the categories. One solution would be to simply estimate all the possible models and to choose the best among them in terms of log-likelihood (or any other measure with which to compare them). However, considering the number of models to be estimated, this methodology is naive and inefficient.

To simplify the execution of this task, we decided to reduce the tree space by considering only binary trees, which describe the pairwise groupings that occur within the different response categories. By considering binary models, the model specification is furthermore simplified since it is characterized by only the cdf F instead of the full combination (r, F, Z). The binary partition trees are represented through a dendrogram and its associated matrix for the following implementations and analyses.

4.3.1 Dendrogram representation

Dendrograms are characterized by whether or not the terminal vertices are labeled and whether or not ranks are associated with the vertices. In our context, we consider only *labeled, non-ranked (L-NR)* dendrograms since those correspond to the binary partition trees. The number of L-NR dendrograms defined for J objects is given by the formula (see Murtagh, 1984):

$$b(J) = \frac{(2J-2)!}{2^{J-1}(J-1)!}.$$

As an example, let us consider $J = 5$, in Figure 29, we illustrate their three possible dendrograms' shapes. There are 60 possible different labellings for dendrogram (i), 30 for dendrogram (ii), and 15 for dendrogram (iii).

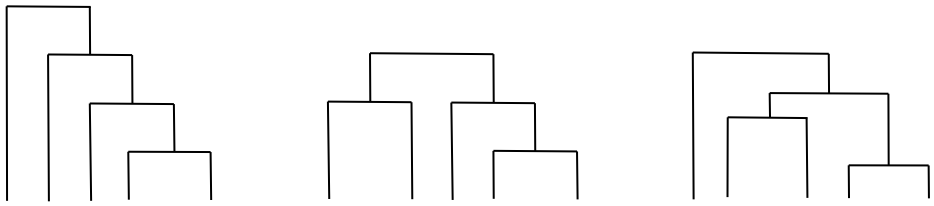

Figure 29: Structures for all possible labelled, non-ranked dendrograms for 5 categories

For more than 5 categories, the amount of dendrograms grows to a point where it becomes difficult to define and explore all possibilities (see Table 4.1).

J	3	4	5	6	7	8	9	10
b(J)	3	15	105	945	10395	135135	2027025	34459425

Table 4.1: Number of dendrograms according to the number of leaves J.

Since the binary partition tree can be represented as a labeled and non-ranked dendrogram, in the following, we will interchangeably use the terms dendrogram and binary partition tree.

4.3.2 Matrix representation

The dendrograms can be computationally represented employing a *merge matrix* which is a $J-1 \times 2$ array that shows at each step which two items are combined. In our

framework, these items are either the individual categories (represented with a negative sign) or the clusters created at a previous stage (represented with the row number of the merge).

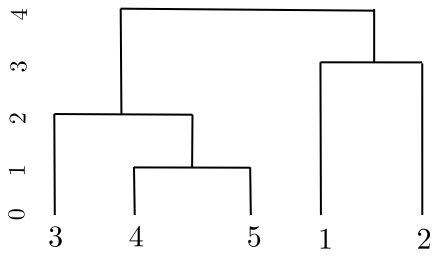

Figure 30: Representation of the hierarchical structure as a tree.

Let us consider the dendrogram with $J = 5$ leaves of Figure 30. Ignoring the heights at which the vertices have been joined, this dendrogram can be represented as

$$\begin{bmatrix} -2 & -1 \\ -5 & -4 \\ -3 & 2 \\ 1 & 3 \end{bmatrix}, \text{ or } \begin{bmatrix} -5 & -4 \\ -3 & 1 \\ -2 & -1 \\ 2 & 3 \end{bmatrix}, \text{ or } \begin{bmatrix} -5 & -4 \\ -2 & -1 \\ -3 & 1 \\ 2 & 3 \end{bmatrix}, \tag{4.2}$$

or as any of all the other matrices that are generated by switching the elements in the rows. Implementing the search algorithms requires a unique computational representation of the dendrograms to avoid re-estimating the models and reduce thus the computational burden of the algorithm. To obtain this unique representation, we propose the following rules for the merge matrix:

- for each row, the two values are ordered from left to right,
- the first rows are composed only of the negative elements, i.e., only categories merging with other categories,
- the matrix is sorted based on the second column in ascending order.

The only representation that meets the previous rules is the one on the right in Expression (4.2). When representing its corresponding heights as a vector, we obtain the unique expression:

$$\begin{bmatrix} -5 & -4 \\ -2 & -1 \\ -3 & 1 \\ 2 & 3 \end{bmatrix} \begin{Bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{Bmatrix}. \tag{4.3}$$

The graphical representation of the above merging matrix is illustrated in Figure 31. Note that for the purposes of our methodology, the tree represented in Figure 30 and the one presented in Figure 31 are the same since the order of the merges is irrelevant for the model estimation.

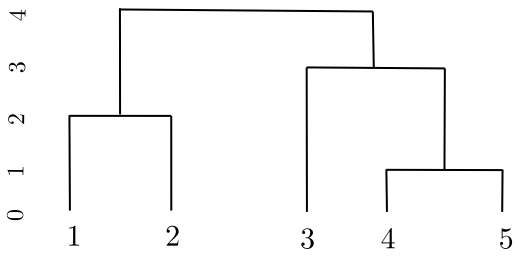

Figure 31: Representation of the tree given by Expression 4.3.

4.4 Construction of a B-PCGLM

In the following, we propose a heuristic to build a model using the explanatory variables. We then have to find the set of models $\mathfrak{C}$ to generate specific insights for each of the $J-1$ vertices. The link function for this case is then composed of the set of distribution functions of each of the $J-1$ binary models and the partition tree itself. We illustrate the proposed methodology by performing each step on the rice diversity data set.

4.4.1 Binary tree construction

The inference of all possible binary trees can be computationally expensive when the number of categories increases (as the number of dendrograms explodes when J increases; see Table 4.1). We propose a heuristic for constructing the binary tree, based on the hierarchical agglomerative clustering (HAC) algorithm but with a particular modification. Instead of grouping individuals, we consider as starting points the groups $E_j := \{1 \leq i \leq n : y_i = j\}$ for $j = 1, \ldots, J$, and then we proceed with a series of successive fusions of the groups until all of them are members of one single cluster, the root. To make this concrete, we must define the meaning of similarity or difference for two objects: individuals as well as groups. This is often a domain-specific decision that must be considered based on knowledge of the data set being studied.

Dissimilarity matrix (inter-individuals)

The essential tool for hierarchical clustering is a measure of dissimilarity or proximity between two individuals. In order to find the distances between individuals according to the p covariates, a previous common step is to transform all numerical covariates to a common scale by a standardization procedure. However, wrong weights might be assigned using inaccurate standardization approaches. For instance, the standardization that results from the division by the standard deviation may imply that the importance of a variable is assumed to decrease with an increasing variability. If all the explanatory variables have the same level of importance, the most appropriate procedure is to standardize the k^{th} covariate, dividing by its range $r_k = \max\limits_{1\leq i\leq n} x_{i,k} - \min\limits_{1\leq i\leq n} x_{i,k}$.

There exist different measures of dissimilarity according to the type of variables. For quantitative variables, the Euclidean and the Manhattan distances are the most popular,

while for ordinal and nominal variables, the best-known dissimilarities are Bray-Curtys and Sokal-Michener; for more details, see Michel Marie Deza (2016). Moreover, in most regression problems, explanatory variables are of different nature. Some are continuous, some are binary, and others are categorical on a nominal scale or an ordered scale, see (Tutz and Berger (2015)). The most appropriate approach to deal with mixed variables is to use the Gower distance (see Gower, 1971; Kaufman and Rousseeuw, 2009), defined for two individuals i and i' as:

$$D_{i,i'} = \frac{1}{p}\sum_{k=1}^{p} d_{i,i'}^{k}$$

where

- $d_{i,i'}^{k} = \frac{|x_{i,k}-x_{i',k}|}{r_k}$ if the k^{th} covariate is numerical,
- $d_{i,i'}^{k} = \mathbf{1}_{\{x_i \neq x_{i'}\}}$ (indicator function) if it is categorical on a nominal scale,
- and if it is an ordinal variable, it should be replaced by integer codes representing the order, and this new variable is then treated as a numerical covariate.

Other approaches are of interest, for example, the correlation-based distance, which considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. The correlation-based distance focuses on the shapes of observation profiles rather than on their magnitudes. The choice of the measure of dissimilarity in any given application is often a matter of subjective choice. Several authors have stressed the importance of the selection of an appropriate dissimilarity measure, arguing that it has even more impact on the results than the choice of the clustering algorithm (Friedman et al., 2001; James et al., 2013). In the rice diversity data example, we used Gower's distance (accounting for the different types of covariates) and obtained an inter-individual dissimilarity matrix (where the individuals are the rice panicles), in this case of size 960×960.

Dissimilarity matrix (inter-categories)

Based on the dissimilarity matrix between individuals D, we have to find the dissimilarities between the J groups $E_1, \ldots, E_J$ denoted by $\Delta_{J\times J}$. Several dissimilarity definitions between groups exist, and their associated methodologies are known as linkage methods. Among the most popular linkage methods are the single, complete, and average linkages. The dissimilarity in single linkage is that of the closest pair of individuals, with one individual in each group:

$$\Delta_{j,j'} := \min_{i\in E_j, i'\in E_{j'}} D_{i,i'}.$$

The single linkage is a good choice when clusters are obviously separated. On the other hand, the complete linkage calculates the maximum distance between clusters before merging. Hence, it can be sensitive to outliers. While the distance between two clusters in the average linkage is the mean distance between an observation in one cluster and an

observation in the other cluster (see Figure 32). None of these algorithms is uniformly the best for all clustering problems because they all have different properties. For instance, while the dendrograms of the single linkage and complete linkage methods are invariant under monotone transformations of the dissimilarities between pairs, this property does not hold for the average linkage method. Another difference is that average linkage depends on the size of the clusters, whereas single linkage and complete linkage do not. Single linkage usually results in long chains of clusters linked by single points close to each other so that close elements of the same cluster have small distances. However, elements at opposite ends of a cluster may be much farther away from each other than two elements of different clusters. This result is not desirable in practice (Izenman, 2008). On the other hand, complete linkage tends to produce many small, compact clusters. The average linkage method is a hybrid between the simple and complete linkage methods, avoiding extremes of large or compact clusters. And, unlike other methods, the average linkage method performs better on ball-shaped clusters in the feature space (Yang, 2017). All in all, this method is known to be a robust alternative (see Brian S. Everitt, 2011).

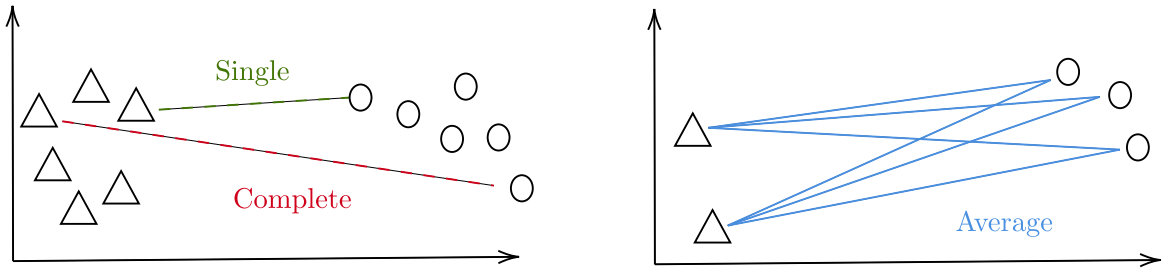

Figure 32: Single, complete, and average linkage methods.

We use the average linkage to find the dendrogram based on the dissimilarity matrix between the panicles of the rice diversity data set. The first dissimilarity matrix calculated for the response categories is presented in matrix (4.4) where we also highlighted the smallest entry which corresponds to the subspecies *Obar-I* and *Obar-II*.

$$\begin{array}{c} \\ Ad \\ In \\ ObI \\ ObII \\ OgI \\ OgII \\ R \\ Te \end{array} \begin{array}{c} \begin{array}{cccccccc} In & ObI & ObII & OgI & OgII & R & Te & Tr \end{array} \\ \begin{pmatrix} 0.152 & 0.303 & 0.368 & 0.198 & 0.187 & 0.319 & 0.226 & 0.220 \\ & 0.238 & 0.302 & 0.141 & 0.136 & 0.260 & 0.178 & 0.210 \\ & & 0.098 & 0.176 & 0.183 & 0.106 & 0.146 & 0.326 \\ & & & 0.236 & 0.243 & 0.120 & 0.196 & 0.394 \\ & & & & 0.102 & 0.205 & 0.143 & 0.206 \\ & & & & & 0.211 & 0.147 & 0.208 \\ & & & & & & 0.171 & 0.337 \\ & & & & & & & 0.249 \end{pmatrix} \end{array} \tag{4.4}$$

Building the final hierarchy using Δ

Most approaches to hierarchical clustering are agglomerative algorithms that follow a

simple methodology and proceed by repeatedly applying four steps: (i) choose the pair of vertices with the highest similarity; (ii) merge the pair of vertices into a new node/-cluster; (iii) update the dissimilarity matrix Δ after having calculated the similarities between the new vertex and the former existing vertices; and (iv) repeat the procedure until only one vertex is left. The crucial step is the update of the similarity values. The nature of the update is determined by the specification of the linkage method, which embodies the similarity of subsets of vertices. Starting from the groups $E_1, \ldots, E_j$, new groups are sequentially created as the union of the subgroups with greater similarity. Remark that this is the same principle used in HAC (see Brian S. Everitt, 2011, for more detail of the HAC), but in our framework, the initial vertices are already clusters. This process sets the stages for creating the intended dendrogram, which summarizes the hierarchical structure that reflects the response categories' differences and/or similarities.

In the rice analysis, the subspecies *Obar-I* and *Obar-II* constitute the first group V_1 from matrix (4.4). Hence, a second dissimilarity matrix was calculated and presented in (4.5). The elements with the smallest dissimilarity in this new matrix are the subspecies *Ogla-I* and *Ogla-II*, thus fused to form a second group V_2.

$$\begin{array}{c} \\ Ad \\ In \\ V_1 \\ OgI \\ OgII \\ R \\ Te \end{array} \begin{array}{c} \begin{array}{ccccccc} In & V_1 & OgI & OgII & R & Te & Tr \end{array} \\ \begin{pmatrix} 0.152 & 0.365 & 0.198 & 0.187 & 0.319 & 0.226 & 0.22 \\ & 0.299 & 0.141 & 0.136 & 0.26 & 0.178 & 0.21 \\ & & 0.233 & 0.24 & 0.119 & 0.193 & 0.391 \\ & & & 0.102 & 0.205 & 0.143 & 0.206 \\ & & & & 0.211 & 0.147 & 0.208 \\ & & & & & 0.171 & 0.337 \\ & & & & & & 0.249 \end{pmatrix} \end{array} \tag{4.5}$$

The dissimilarity matrix is updated sequentially as new groups are fused and until the final merging of the two remaining clusters into one. As a result of these successive groupings, we obtained the dendrogram presented in Figure 33 in which, through the heights (although irrelevant for the estimation of the model), we can identify the order in which groups were fused in each of the iterations of the algorithm.

In the previous analyses of the rice diversity, we used a subset of the database since there were some missing labels for the response variable (the subspecies); nevertheless, the labels corresponding to the species themselves were available. To make use of these missing observations, we transformed the species label by creating the new categories $\{B^*, G^*, S^*\}$ corresponding respectively to the missing data from *Barthii*, *Glaberrina*, and *Sativa*. We implemented the previously described methodology to the data set with the additional observations and in which there were 12 categories instead of 9. We obtained the structure presented in Appendix C. In that dendrogram, the species categories were positioned quite close to their related subspecies, so actually, this structure resulted in being essentially identical to the one presented in Figure 33. The new data did not alter the dissimilarity matrices or the linkage method. This fact confirms the reliability of the proposed methodology since we have obtained the expected tree

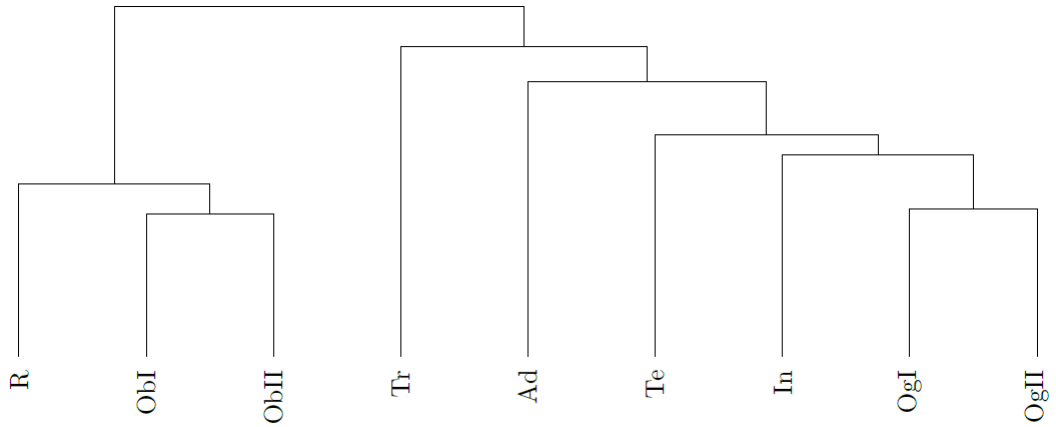

Figure 33: Partition tree structure for the rice diversity data set.

structure according to the information known beforehand.

4.4.2 Binary model for each non-terminal vertex

An issue that emerges when partitioning the data set according to the hierarchical structure is the fact that the binary responses at each vertex have specific degrees of separation (or overlap). The separation configurations are highly dependent on the linkage method employed when constructing the initial partition tree. For instance, the most similar categories (being fused in the first step of the algorithm) may have a low degree of separation in the covariate space. Conversely, the vertices closer to the partition tree's root are likely to have a high degree of separation of the response levels. This separation's characteristic has been described in scientific literature as a problem since a unique maximum likelihood estimation might not exist in this case. At the same time, as the explanatory variables in each non-terminal vertex remain to be selected, the sub-datasets may contain noisy variables which are not immediately identifiable. To overcome these issues, we suggest using the results reported in chapter 2 to find the binary model that best fits the data at each non-terminal vertex while accounting for the different degrees of separation. Through the algorithm 1 presented therein, one can adequately address the most common perturbations in binary data (outliers and noisy variables) in the frame of the separation problem.

Variable selection for each non-terminal vertex

Since $J-1$ models are to be estimated for the resulting groups of categories, the question about the explanatory variables that influence each vertex should be addressed. In our methodology, we proposed using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm that performs both regularization and variable selection. To ensure that the link function fully corresponds to the data, after selecting the covariates for each non-terminal vertex, we recommend to re-estimate the link function given the chosen set of covariates x^v.

We found the link function for each vertex $v \in \mathcal{V}^*$ of the tree structure for the rice diversity data set presented in Figure 33 by following algorithm 1. The global model with its link represented by the partition tree structure and the set of models $\mathfrak{C}$ (and their corresponding explanatory variables) is represented in Figure 34. The log-likelihood, as well as the BIC of this model, are respectively higher and lower than those obtained for the pre-defined partition trees. The found partition tree is similar to the partition tree that first splits the sub-species according to the domestication specification (see Figure 28). However, these structures differ in the sub-partition generated for the cultivated sub-species. The partition structure in Figure 28 was identified in section 4.2 as the best among the known partition trees for modeling rice diversity. Assuming the structure as unknown, and with the methodology presented above, we were able to obtain a better fit for this specific data.

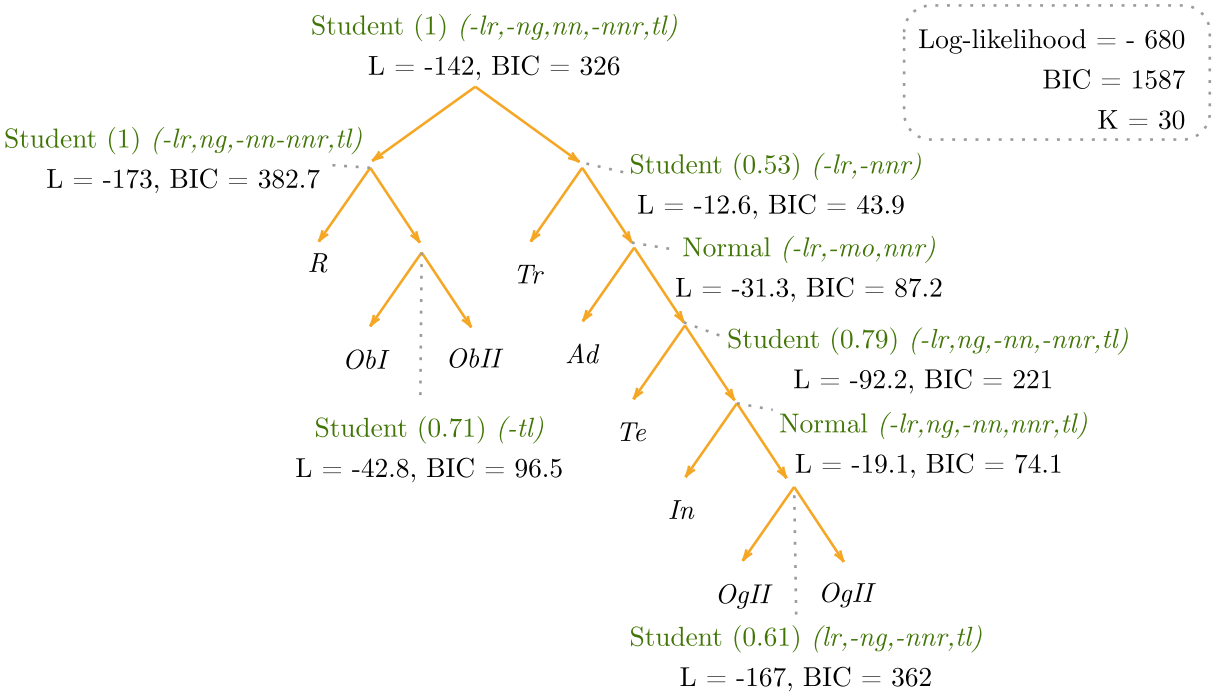

Figure 34: B-PCGLM obtained for the rice diversity data set.

4.5 Visiting neighboring trees

Whatever the quality of the B-PCGLM obtained by the proposed method, there is no certainty of having found the best one. Therefore, we now propose two search algorithms to estimate the neighboring trees in an attempt to find a better score. For this purpose and in the following, we define the operations on the binary partition trees known as rotations, and we clarify other terms related to these operations.

- The $J-2$ segments joining two vertices in a binary tree are called the internal edges (IEs). We represent an IE with the digits that indicate the order in which the groups were fused.

- Any internal edge $\alpha\beta$ connects two non-terminal vertices α and β (each specified by its height) such that β is the parent of α.

- One rotation consists of switching element a for element b, and the other results from switching a and c (see Figure 35).

- There are two possible rotations for each internal edge so that in total, there are $2(J-2)$ binary trees in the rotational neighborhood generated by one particular binary tree.

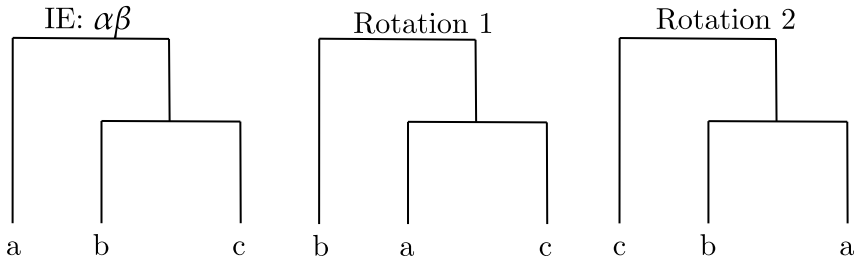

Figure 35: Rotations of the internal edge $\alpha\beta$.

In the merge matrix representation 4.3, the set of internal edges is $IE = \{13, 24, 34\}$. For illustration purposes, let us consider the rotations of $IE = 24$. The elements to obtain the two rotations are identified as $a = 3$, $b = -1$, and $c = -2$. Note that the a element is a tree itself. The obtained rotations of this dendrogram are represented in Figure 36.

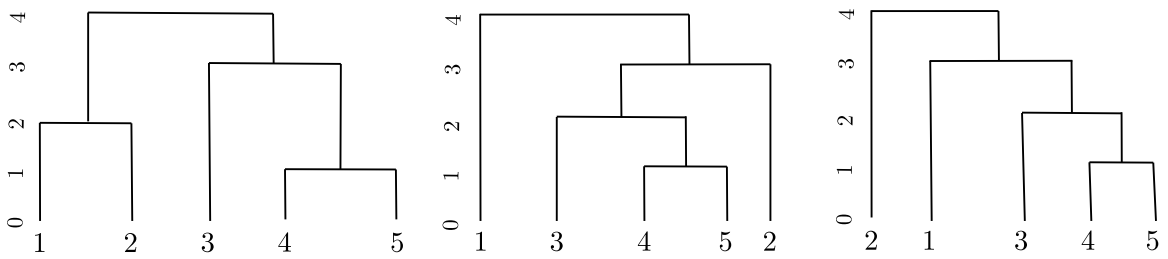

Figure 36: Trees obtained from rotation of the internal edge 24.

We aim now to find a higher score through a walk (to be defined iteratively) where the starting point is the proposed tree T_0 whose score is denoted as s_0. For the following, $R(T) = \{T_1, \ldots, T_{2(J-2)}\}$ represents the set of trees generated by all possible rotations in

T.

Algorithm 2: Estimating all partition trees in the generated neighborhoods

Input: T, $s(T)$, ϵ
$s_p = -\infty$
$s_c = s(T)$
$T_c = T$
$H = \{T\}$
while $s_c - s_p > \epsilon$ **do**
  $R = R(T_c) \setminus H$
  $s_p = s_c$
  $T_p = T_c$
  if $R = \varnothing$ **then**
    break;
  else
    $T_c = \text{argmax}\{s(T) : T \in R\}$
    $H = H \cup R$
    $s_c = s(T_c)$
return T_p, s_c

The graphical representation in Figure 37 of algorithm 2 recreates a scenario in which there are $J = 4$ categories so that in total, there are $b(J) = 15$ possible partition trees. Suppose that the starting partition tree found following the methodology proposed in section 4.4.1 is the number 6. We obtain $2 \times (J - 2) = 4$ different partition trees by executing the possible rotations on its internal edges. Two of them (2 and 12) have a lower score. Among the two with a higher score (4 and 11), 4 has the maximum score. In the next iteration of the algorithm, partition tree 4 is the baseline, and its neighbors are found. Note that it is possible to obtain partition trees that were already estimated (in this case, 6 and 11). Those are discarded by identifying the equality between their unique merge matrix. The two new partition trees have a higher score (3 and 10), and the score of 3 is the maximum. In the next iteration, two repeated partition trees (6 and 10) are obtained again as the rotations of partition tree 3. Of the two new partition trees, neither has a higher score; hence the algorithm's output is the partition tree 3.

The graphical representation in Figure 38 recreates the same scenario described above, but, in this case, we illustrate the iterations of algorithm 3. In the first iteration of this algorithm, the score of one of the possible structures (randomly selected) within the neighborhood of partition tree 6 is estimated. This score (corresponding to partition tree 2) is not higher, so another partition tree within the neighborhood is randomly chosen (partition tree 11). This time, the score is higher, thus, it becomes the baseline for the next iteration. The first randomly selected partition tree (10) scored higher than the baseline partition tree in the second iteration. A third iteration is performed with partition tree 10 as the baseline, and partition tree 3 is selected for the next iteration. In the fourth iteration, no higher score is found after estimating (in random order) all the scores of the neighboring partition trees. Therefore, the output of the algorithm is partition tree 3.

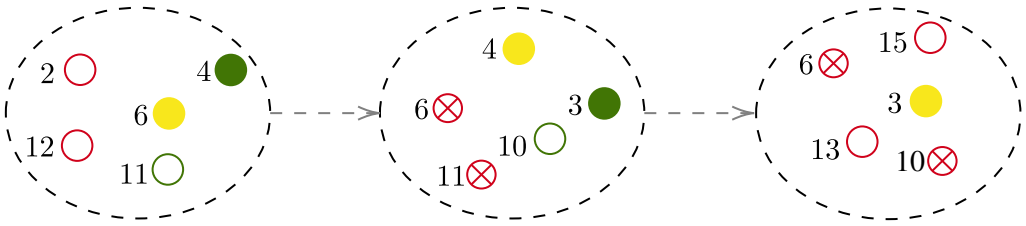

Figure 37: Graphical representation of Algorithm 2. Each dashed circle stands for an iteration. Within each dashed circle, the yellow color indicates the baseline partition tree. The green and red colors represent the partition trees with higher and lower scores, respectively. The green-filled circle depicts the partition tree with the highest score. The cross-marked circles indicate that the score of that partition tree was already estimated in a previous iteration.

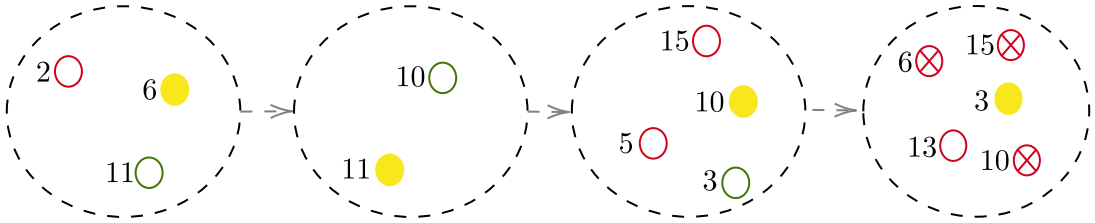

Figure 38: Graphical representation of Algorithm 3 Each dashed circle stands for an iteration. Within each dashed circle, the yellow color indicates the baseline partition tree. The red color represents a partition tree with a lower score than the one of the baseline partition tree. The green color represents the first found partition tree with a higher score than the one of the baseline partition tree. The cross-marked circles indicate that the score of that partition tree was already estimated in a previous iteration.

The log-likelihood decomposition of the PCGLM (see equation (4.1)) implies that for each new partition tree, the maximum number of new binary models to be estimated is 2. For instance, the first neighbor resulting from rotation 1 of the baseline partition tree in Figure 36 only requires estimating the binary models with response levels $\{\{3,4,5\},\{2\}\}$ and $\{\{3,4,5,2\},\{1\}\}$ since all the others have already been calculated for the baseline partition tree's estimation. Nevertheless, such two models may as well have already been estimated in previous iterations. Therefore, to avoid re-estimating models, it is essential to keep a record of the response levels and their corresponding model's summary.

4.5.1 Performance evaluation of the methodology

In the following, we aim to evaluate the quality of the obtained partition tree by contrasting it with all the other possible trees. There are 9 categories in the rice data set, using equation (4.1), we found that there are 2027025 different partition trees to estimate. The definition of all the structures, as well as the computational cost of the calculations (considering applying cross-validation) are two very demanding tasks. To simplify this evaluation, we work with the Cleveland heart disease data set, which has

Algorithm 3: Random walk on the partition trees in the generated neighborhoods

```
Input: T, s(T), ϵ
s_p = −∞
s_c = s(T)
T_c = T
H = {T_c}
while s_c − s_p > ϵ do
    R = R(T_c) ∖ H
    T_p = T_c
    s_p = s_c
    if R = ∅ then
        break;
    T_c = Uniform(R)
    s_c = s(T_c)
    H = H ∪ T_c
    while s_c − s_p < ϵ & R ≠ ∅ do
        R = R ∖ T_c
        if R = ∅ then
            break;
        T_c = Uniform(R)
        s_c = s(T_c)
        H = H ∪ T_c
return T_p, s_p
```

5 categories, so there are a total of $b(5) = 105$ different binary partition trees. This data set is originally located in the UCI machine learning repository (Dua and Graff, 2017). The partitioned data sets (obtained using 10-folds cross-validation) were extracted from the KEEL repository.

The Cleveland heart disease data set contains 303 instances with 13 attributes that were taken from patients with heart problems. The task is to distinguish the presence (values 1,2,3,4) from the absence (value 0) of the heart disease in the patient. In the context of GLM for categorical data, the multinomial logit (MNL) is the default choice for analyzing the effect of the explanatory variables on a categorical response since it corresponds to the canonical link function. We aim to compare such a model with the one obtained using the above-described methodology. To compare specifically the partition tree structure, we must use logit models with the complete set of explanatory variables in each of the $J - 1$ non-terminal vertices. In addition, to evaluate the performance of the obtained structure (beyond the multinomial model), we intend to contrast it with the other 104 B-PCGLMs.

Figure 39 presents the log-likelihoods of several models from which we formulate the following comparisons:

- The first is that of the MNL (orange line), which is the only model invariant to the permutations of the categories, against the B-PCGLM (blue line) whose

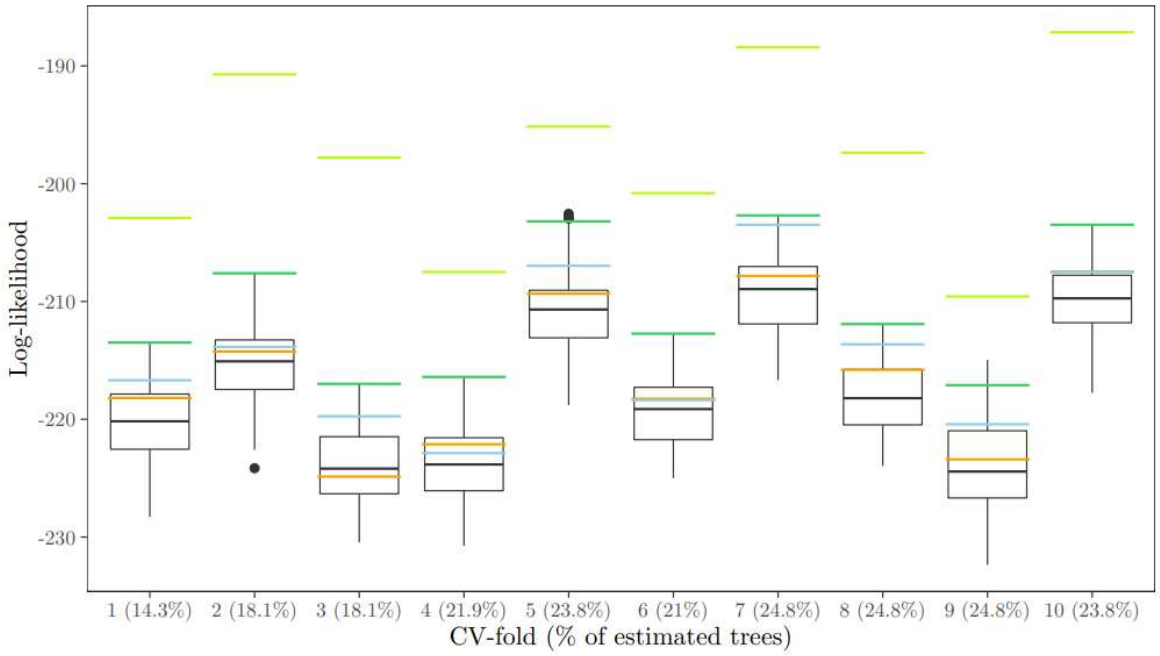

Figure 39: Box plots of the 105 B-PCGLMs log-likelihoods corresponding to the 10 samples resulting from the 10-folds cross-validation procedure (x-axis). The orange, blue, and dark green lines correspond respectively to the log-likelihood of the multinomial model, the initial partition tree, and the best partition tree found using algorithm 2. The above partition structures used the logistic link function at all non-terminal vertices. The light green line is the log-likelihood of the model with the same partition tree structure found in the neighborhood search, but the cdfs at each non-terminal vertex were selected using algorithm 1.

link functions are in all cases the logistic cdf. The log-likelihood of the found B-PCGLMs in all cases except the fourth fold turned out to be higher than the log-likelihood of the MNL. It is considerably higher in 6 of these cases, while the log-likelihoods are very close in the other 3 cases.

- The second comparison evaluates the selected structure inside the complete set of possibilities described by the box plots. As for the selected B-PCGLM, we note that for all cases, the log-likelihood is higher than the median of all possible B-PCGLMs log-likelihoods. In fact, 7 of them are within the best quartile of the likelihoods (while none of the MNLs in it).

- The third comparison is that of the B-PCGLM resulting from the search with algorithm 2 (green line) against the complete set of structures and the baseline partition tree itself. We found the best B-PCGLM in 8 folds (and in the other 2, the log-likelihood were very close to the best one) by visiting on average 21% of the total binary partitions. Note that the percentage of visited trees for each CV-fold is indicated in the labels of the $x-$ axis.

- And the fourth comparison is the B-PCGLM (light green line) whose link functions

are estimated according to algorithm 1 of chapter 2 against the best B-PCGLM that uses the logistic link in all the set of models. The average log-likelihood was reduced by approximately 7% when specific links were assigned to each non-terminal vertex.

4.6 Conclusion and perspectives

In this chapter, we studied the modeling of a hierarchically structured categorical response, and we focus on the case where the structure itself is unknown beforehand. We opted to work with binary partition trees to simplify the task of searching and the task of modeling the hierarchical structure. We first proposed a methodology for finding a candidate baseline tree of such a hierarchical structure following the grounds of the HAC. Further on, we proposed two greedy search algorithms that aim to walk through the space of partition trees neighboring the initial partition tree in order to find a partition tree with a better score. Here, we only considered the log-likelihood score, but any other characteristic measure of the model, for instance, the percentage of correct predicted classifications, can be optimized. The computational implementation of the proposed methodologies involves the estimation of several models. However, the log-likelihood decomposition of a model allows the different models to be estimated in parallel. Moreover, the search for neighboring partition trees only involves estimating two additional binary partition trees that can also be estimated in parallel. Thus, the computational burden can be highly minimized through good programming practices.

The success of this proposal is primarily determined by having found a good baseline partitioning tree, for which the methodology is based on the fundamentals of the HAC algorithm. As usual in clustering tasks, different results may be obtained when using different dissimilarity measures (to be defined for both individuals and groups). One could obtain results that correctly capture the variability of the data but that do not effectively address the problem's objective. Similarly, the choice of linkage method greatly determines the characteristics of the partition tree to be obtained. Therefore, it is crucial to carefully select the dissimilarity measure, as well as the linkage method, according to the data set at hand.

Different degrees of separation are possible for the vertices of the partition tree. Despite the simplicity of binary regression models, these are greatly affected when there is a high degree of separation of the response variable (see chapter 2). Hence, to select the link function at each non-terminal vertex, we used an algorithm to address this problem even in the presence of common perturbations for binary data.

Chapter 5

Conclusions and Perspectives

In this thesis, we addressed regression models for categorical responses from the simplest case, which corresponds to binary variables, going through ordinal and nominal variables, to the most general case, when there is a hierarchical grouping structure among the categories. Our approach and the common thread of this research were focused on the specification, description, and characterization of the link function that connects the linear predictor to the expected value of the categorical response variable. According to Nelder and Wedderburn (1972), a generalized linear model is characterized by three components: the response distribution, the linear predictor (linear combination of explanatory variables), and the link function. In the particular case of a categorical response, the distribution is necessarily the Bernoulli distribution (when $J = 2$) or the multinomial distribution (when $J \geq 2$). Thus, only the linear predictor and the link function characterize the differences among these models. However, since the linear predictor only represents the constraints on the parameters associated with the covariates, the link function becomes the key to characterize categorical regression models. The robustness to the presence of outliers, the order type, or the grouping structure of categories are three challenges rarely addressed by analysts when modeling categorical responses. This thesis focused on analyzing the composition of the link function to meet these challenges.

Chapter 2 focused on binary responses, for which, despite the existence of a rich set of models, only the logit model and the probit model are widely used in practice. Since they are very similar, researchers have long assumed that no other function could significantly improve the fit of these models. However, the logit and probit models are known to be sensitive to data perturbations, whereas the Student link has been suggested as a robust alternative. To use the Student link, we proposed to estimate its degree of freedom (ν) each time the Student model with $\nu = 1$ results in a higher log-likelihood value than the one with $\nu = 8$. This allowed us to effectively differentiate the fits of the Student model (with a small ν) from the logit and probit models. We investigated the sensitivity of the models according to the different degrees of separation (or, conversely, overlap) of the two levels of the response. We observed that the Student model is robust, unlike the logit model, especially when the degree of overlap is small. In other words, the lower the degree of overlap, the lower the estimated ν, and so the more robust the Student link is compared to the logit link. Therefore, the estimation of ν might be used as an indicator of the overlap configuration. An additional remark for binary models is about their inference, which might be significantly affected if a symmetric link function is incorrectly used instead of a non-symmetric one. The Student distribution is symmetric and thus sensitive to skewed data, as the logistic and the normal distributions. To overcome this problem, and as an extension of the work pre-

sented in chapter 2, we propose to use the non-central Student distribution. By tuning its non-centrality parameter, one will be able to control the skewness of the link function.

In chapter 3, we presented the problem of non-homogeneity of regression models for categorical responses, both in formal writing and in software solutions. To overcome this problem, we created and introduced GLMcat, an R package which enables the estimation of generalized linear models for categorical responses. The models in this package are implemented using the unified specification (r, F, Z), where r represents the ratio (reference, cumulative, adjacent, or sequential), F the distribution function for the link, and Z the design matrix (Peyhardi et al., 2015). All classical models (and their variants) for categorical data can be written as a triplet (r, F, Z), and thus can be estimated using GLMcat. Here, we implemented an alternative to the existing packages that covers all classical models and offers the possibility of new models through the different combinations of the components. The functions are user-friendly and fairly intuitive, offering the possibility to choose from an extensive range of models. Indeed, the package supports a wide range of cumulative distribution functions (cdfs) and allows the linear predictor to be tailored as desired. The unified specification of the models for categorical data enabled us to analyze their properties and highlight some existing equivalences. We also proposed a methodological and practical guide for the appropriate selection of a model (through the link function and the constraints of the design matrix), considering the correspondence between the nature of the data and the properties of the model. The contributions presented in this chapter have broad applicability since various fields of research and industry deal extensively with categorical responses. We believe that this tool makes it possible to popularize the area of categorical data regression, which has not yet been extended on a large scale for non-logistic models.

One advantage of the modularized architecture of the package is that it facilitates the inclusion of additional cdfs that will be immediately available for all four model families. Although the most popular cdfs often result in "similar" fits, this does not imply that all cdfs are essentially equivalent when describing the response. In chapter 2, we specifically investigated the properties of the Student distribution as the link function in binary models. We found strengths regarding the robustness of the model when specifying its degree of freedom as $\nu \leq 1$. As a methodological extension, we propose to evaluate the properties of other cdfs with requirements that allow improving the fit. For instance, in distributions such as Pregibon (based on the generalized Tukey family) or the non-central Student, some parameters control the symmetry, the heaviness of the tails, and/or the skewness of the distribution. An extension for the package is to consider algorithms to estimate such parameters.

The unified structure (r, F, Z) can be extended to include random effects. Some authors have proposed this extension for particular models in the context of categorical responses (see Hartzel et al., 2001; Coull and Agresti, 2000; Tutz and Hennevogl, 1996). The implementation of generalized linear mixed models is envisaged as an additional functionality in the GLMcat package.

In the regression framework, other essential tasks are model regularization and variable selection. These techniques aim to reduce the space of explanatory variables and improve model estimation and prediction accuracy. In the case of categorical variables,

the elastic net penalty can be obtained for categorical responses with the package ordinalNet. However, not all the families of models are available within this package. In future work, we will attempt to define regularization and variable selection methods that are valid for any triplet (r, F, Z).

The random utility maximisation (RUM) principle is used to model the choices of individuals. It is assumed in these models that an individual's preferences among the available alternatives can be described by a utility function $(U_j = \eta_j + \epsilon_j)$. The individual chooses the alternative with the highest utility. The utility of an alternative depends on η_j, which is determined by the characteristics of the decision-maker and the attributes of the alternatives, and on ϵ_j, which accounts for the effects on preferences of unobserved attributes. The multinomial logit (MNL) emerges when one assumes that the epsilons independently follow a Gumbel distribution. The MNL can also be derived and represented as a GLM by using the triplet $(reference,\ logistic,\ Z_c)$ (see Figure 40). There are also other alternatives as the multinomial probit (when $\epsilon_j \sim Normal$) or the nested logit model (when $\epsilon_j \sim Generalized\,Gumbel$). The multinomial probit has a straightforward interpretation in terms of the latent utilities. However, as a disadvantage, it does not have a closed-form expression, thus, in practice, this model becomes difficult to estimate for more than four alternatives. This is not the case for GLMs which are easily estimable. At using the family of reference models, there is also a great flexibility in using different distributions for the link function. An additional perspective of this work is to use the *adjacent* ratio as an alternative to the $reference$ ratio. The $(adjacent,\ logistic,\ Z_c)$ model is invariant under all the permutations of categories, thus appropriate for nominal responses (although it has been reported as a model for ordinal responses). Whenever the distribution is set differently from the logistic, or the design is defined differently from the complete form, their combination with the adjacent ratio (invariant only under the reverse permutation) could be employed as a choice model. With this proposal, we would obtain the first model designed for ordered choices, which can consider explanatory variables that might depend on the ordered choice alternatives.

Chapter 4 studied the modeling of a hierarchically structured categorical response. Whenever the partition tree structure is known, the task is simply to build the best model for each non-terminal vertex; this means choosing the link function and the pertinent covariates. Using the rice data set which motivated this work, we built two PCGLMs corresponding to the two known partition trees. One considered first the division of the subspecies according to the geographic origin and then according to the domestication factor. The other considered the opposite splits, first according to the domestication factor and then according to the geographic origin.

A particular emphasis is then given in this chapter to the case where the structure itself is unknown beforehand. We opted to work with the assumption of binary partition trees to simplify both the task of searching and the task of modeling the hierarchical structure. We proposed a methodology for finding a candidate baseline tree of such a hierarchical structure following the grounds of hierarchical agglomerative clustering (HAC). This methodology has been tested on the rice data set and has led to find a structure closely related to the (hypothesized) partition tree that first splits the data

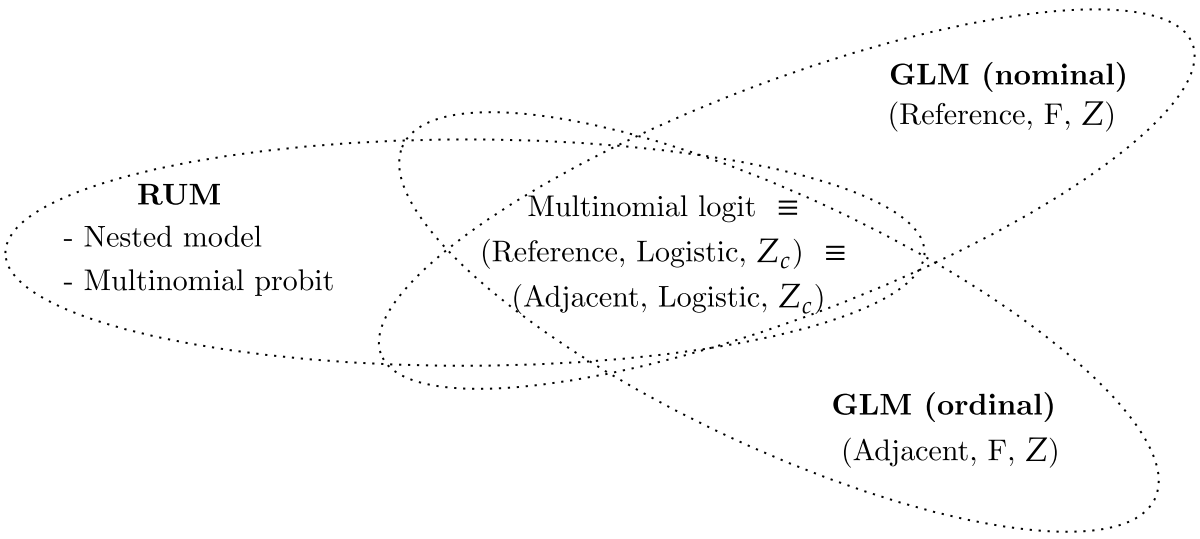

Figure 40: Types of choice models.

according to the domestication factor. This partition tree had better results in terms of the log-likelihood and the BIC. By selecting the link functions for the non-terminal vertices (based on the results of chapter 2), these measures were further improved for this data set.

Further on, we proposed two greedy search algorithms to find a partition tree with a better score. These algorithms aim to walk through the space of trees iteratively, starting from the baseline found with the HAC-based methodology. The partition trees to visit are the neighboring partition trees defined through operations on a baseline structure. We considered the log-likelihood score as the summary measure of these partition trees. However, any other characteristic measure of the model, for instance, the percentage of correct predicted classifications, can be optimized. The computational implementation of the proposed methodologies involves the estimation of several models. Nevertheless, the log-likelihood decomposition of a model allows the different sub-models to be estimated in parallel. Moreover, the search for neighboring trees only requires estimating two additional binary models that can also be estimated in parallel. Hence, the computational burden can be highly minimized through good programming practices.

This proposal's success is primarily determined by having found a good baseline partitioning tree, for which the methodology is based on the fundamentals of a clustering algorithm. As usual in clustering tasks, different results may be obtained when using different dissimilarity measures (to be defined for both individuals and groups). One could obtain results that correctly capture the variability of the data but that do not effectively address the problem's objective. Therefore, it is crucial to carefully select the dissimilarity measure to be employed. Similarly, the choice of linkage method greatly determines the characteristics of the partition tree to be obtained. For example, single linkage tends to produce unbalanced and straggly clusters. Complete linkage tends to produce compact clusters with equal diameters. Average linkage is an intermediate

between single and complete methods, or Ward's method, that tends to find clusters of equal size and is known to be sensitive to outliers. Some of these linkage methods even rely on one specific similarity measure. Taking advantage of all of these options, in the proposed methodology, one can analyze the clusters obtained using different dissimilarity measures and different linkage methods, attempting to find a common structure in all of them. Consensus algorithms synthesize different dendrograms by summarizing the concordant parts relative to the discrepant parts (Darlu and Tassy, 1993). In other words, the algorithm finds a consensus tree that optimally represents various clustering approaches resulting from the choice of the dissimilarity measure and the linkage method. Some approaches to obtain a consensus tree are given by Gordon and Vichi (2001); Mouchet et al. (2008); Lapointe and Cucumel (1997). The performance of our proposed methodology may be enhanced using these consensus trees.

As discussed by Tutz (2021), Likert-type items (defined by Likert (1932)) can be analyzed assuming a hierarchical structure of the categories. In particular, he defined a binary model (at the root level) to differentiate the neutral category, which corresponds to the ambivalence of neither agrees nor disagrees, from the ordered categories. He also proposed binary separations for the following levels (disagreement and agreement) to take full advantage of the binary partitions, which allowed him to consider the problem of the tendency of respondents to extreme categories. He proposed a particular contrast in the form of a linear predictor that allows to increase symmetrically the tendency to extreme categories. Extensions to the PCGLMs (not necessarily binary) would be to tailor the linear predictor of the models to respond to specific requirements as the one previously exposed. In this situation, the inference of the model is a topic to investigate because the log-likelihood would no longer be separable due to the generated relationship between linear predictors of two distinct partitions.

The selection of the link function as well as the selection of predictors has emerged as a common discussion in all the chapters of this thesis. Nonparametric methods have been proposed in the literature to select predictors and estimate the link function. Wang et al. (2018) proposed an algorithm that uses p-splines to estimate the link function that is guaranteed to be monotonic. More generally, Tutz and Petry (2012) employed a boosting type method to select both the predictors and the link function simultaneously. Given the great flexibility that can be achieved by using nonparametric methods, we consider of interest to use and compare those with the parametric counterpart described throughout this thesis.

In chapter 2 we showed that despite the simplicity of binary regression models, their quality is greatly affected when there is a high degree of overlap of the data. We argued that a good model fit can be obtained using the Student link function as long as there is a small degree of overlap. According to that, we would expect the best model to result from a partition tree whose non-terminal vertices sharply separate the response levels according to the covariates. Consequently, the degree of freedom for the Student link function estimated at each vertex of the partition tree is expected to be generally lower than one. Following these ideas, the best B-PCGLM would correspond to the tree with the lowest summary measure (for instance, the mean weighted by the number of

observations at each vertex) of these degrees of freedom. Future work will be devoted to the study of this.

Although the data that motivated this research comes from a biological context (as they refer to the taxonomic diversity of rice), throughout this thesis, we managed to situate the statistical problem in examples from different research areas, illustrating the popularity of categorical responses and thus the high applicability of these regression models.

Appendix of chapter 2

A.1 Accuracies outliers simulation

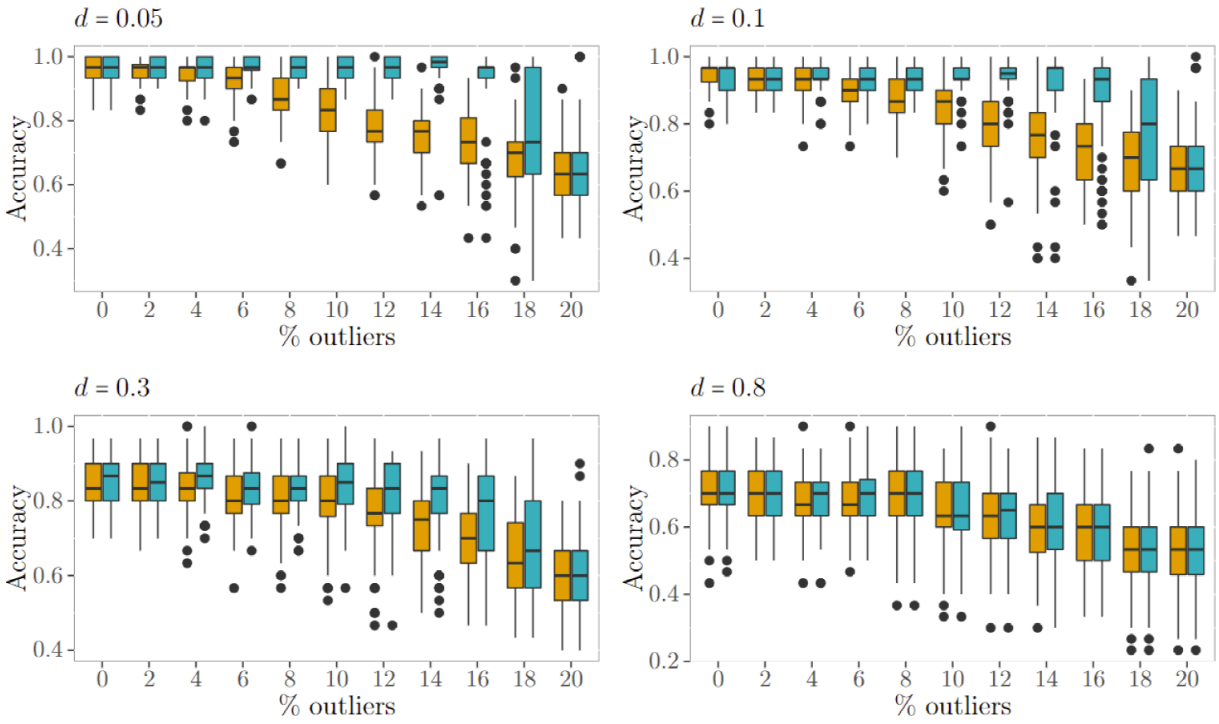

Figure 41: Classification accuracies (y−-axis) of the simulations in section 2.4.3. The x−-axis represents the percentage of outliers, and the blue and yellow colors correspond respectively to Student and the logit link.

A.2 Simulations with different numbers of discriminant and noisy variables

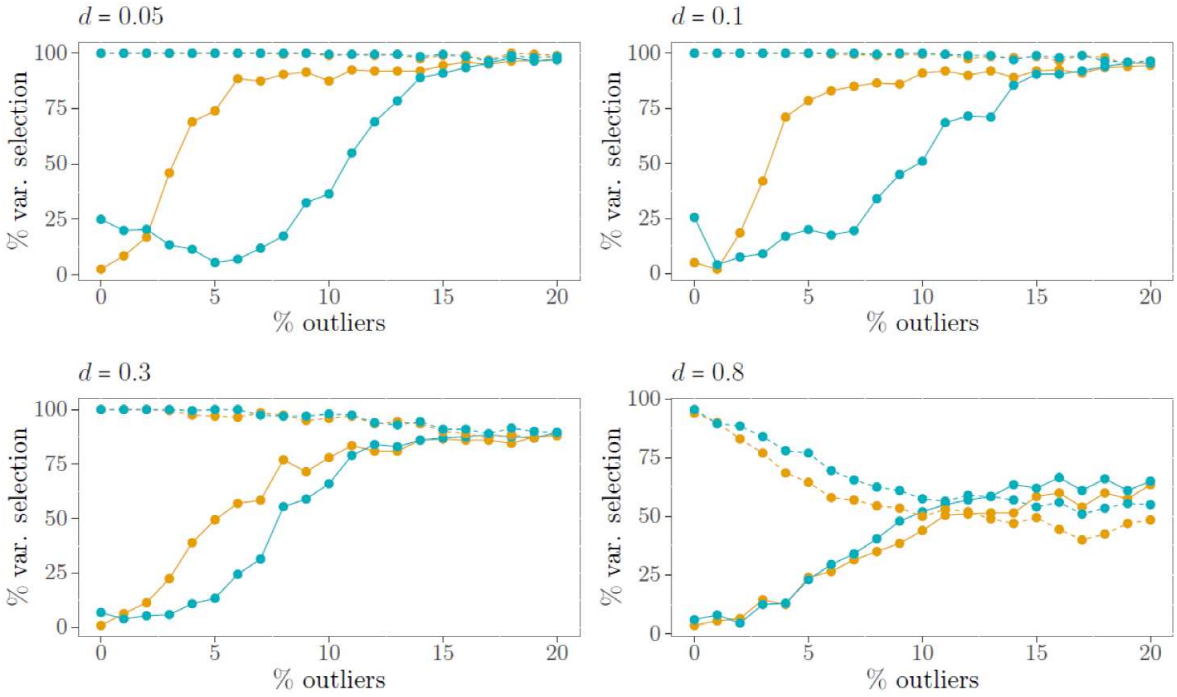

Figure 42: Average number of times the stepwise (backward) algorithm selects discriminant variables (dashed lines) and noisy variables (solid lines) where $\boldsymbol{\delta}^t = (0.8, -0.6, 0, 0)$. The blue and yellow colors respectively represent the Student and logistic links.

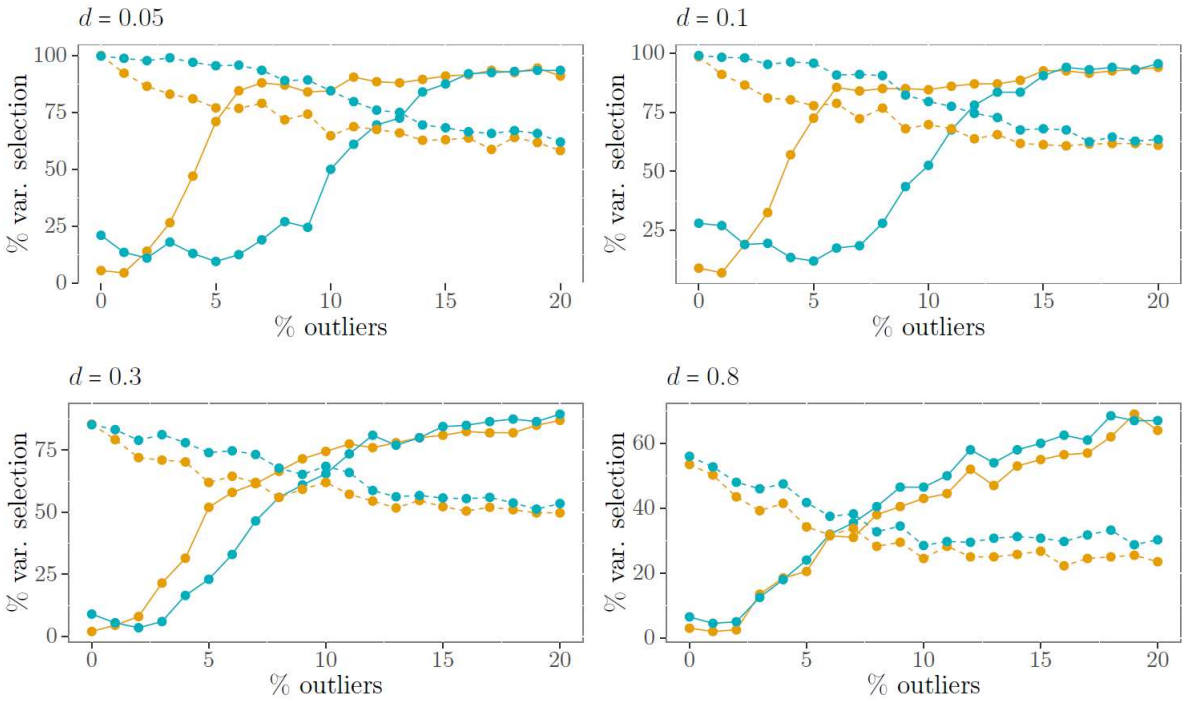

Figure 43: Average number of times the stepwise (backward) algorithm selects discriminant variables (dashed lines) and noisy variables (solid lines) where $\boldsymbol{\delta}^t = (0.8, 0.4, -0.4, 0.2, 0, 0)$. The blue and yellow colors respectively represent the Student and logistic links.

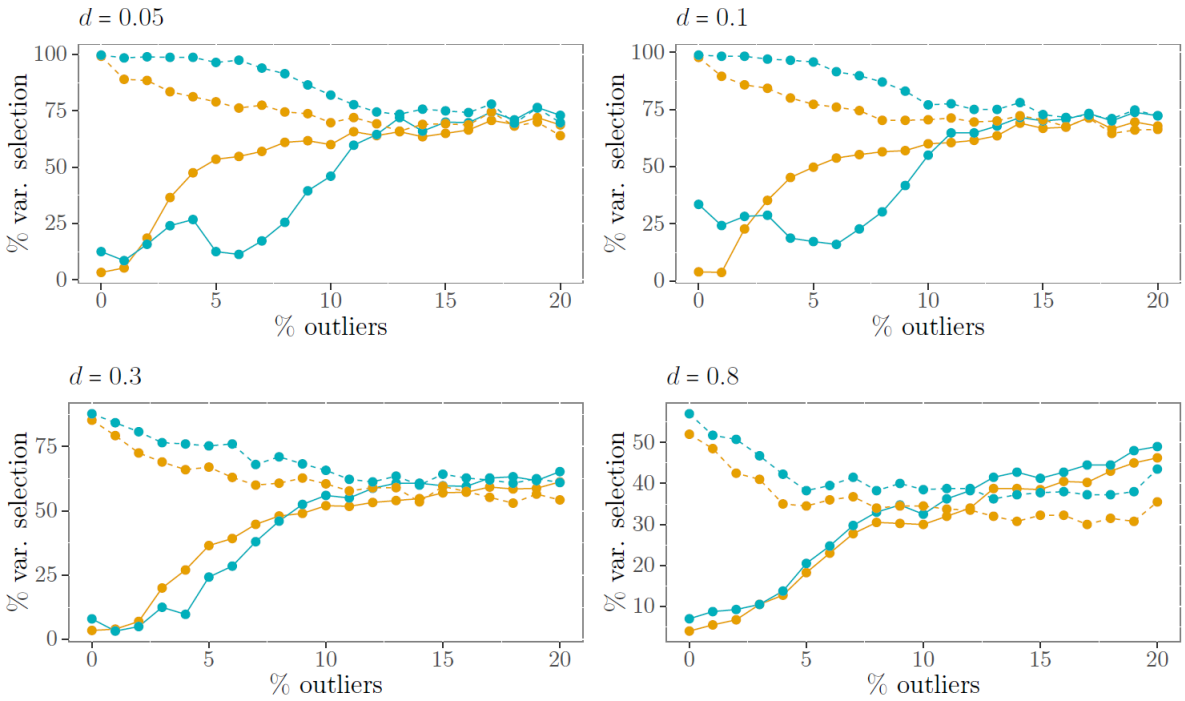

Figure 44: Average number of times the stepwise (backward) algorithm selects discriminant variables (dashed lines) and noisy variables (solid lines) where $\boldsymbol{\delta}^t = (0.8, 0.4, -0.4, 0.2, 0, 0, 0, 0)$. The blue and yellow colors respectively represent the Student and logistic links.

Appendix of chapter 3

B.1 Cumulative distribution function of the non-central t distribution

$$F_{\nu,\mu}(\eta) = \Phi(-\mu) + \frac{1}{2}\sum_{j=0}^{\infty}\left(p_j I_y\left(j+\frac{1}{2},\frac{\nu}{2}\right) + q_j I_y\left(j+1,\frac{1}{2}\right)\right),$$

where:

- Φ is the cdf of the standard normal distribution,
- $I_y(a,b)$ is the regularized incomplete beta function,
- $y = \dfrac{\eta^2}{\eta^2+\nu}$,
- $p_j = \exp\left(-\dfrac{\mu^2}{2}\right)\left(\dfrac{\mu^2}{2}\right)^j$, and
- $q_j = \dfrac{\mu}{\sqrt{2}\Gamma(j+3/2)}\exp\left(-\dfrac{\mu^2}{2}\right)\left(\dfrac{\mu^2}{2}\right)^j$.

B.2 Jacobian matrices

The Jacobian matrices $\partial\pi/\partial r$ used as part of the Fisher's scoring algorithm are presented for each ratio.

Cumulative:

$$\frac{\partial\pi}{\partial r} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix}.$$

In the following, we present the form of the element corresponding to row i and column j of the Jacobian Matrix.

Adjacent:

$$\frac{\partial\pi_j}{\partial r_i} = \frac{1}{F(\eta_i)[1-F(\eta_i)]}\begin{cases}\pi_j(1-\gamma_i) & \text{if } i \geq j,\\ -\pi_j\gamma_i & \text{otherwise,}\end{cases}$$

where $\gamma_i = \Pr(Y \leq i) = \sum_{k=1}^{i} \pi_k$.

Sequential:

$$\frac{\partial \pi_j}{\partial r_i} = \begin{cases} \prod_{k=1}^{j-1} \{1 - F(\eta_k)\} & \text{if } i = j, \\ -F(\eta_j) \prod_{k=1, k\neq i}^{j-1} \{1 - F(\eta_k)\} & \text{if } i < j, \\ 0 & \text{otherwise.} \end{cases}$$

Reference:

$$\frac{\partial \pi_j}{\partial r_i} = \frac{\mathsf{COV}(Y_i, Y_j)}{F(\eta_i)[1 - F(\eta_i)]}.$$

Refer to the Supplementary Material of Peyhardi et al. (2015) for further details.

B.3 Proofs

B.3.1 Proof of Proposition 1

Consider the distribution of Y defined by the $(adjacent,\ F,\ Z)$ model. The adjacent ratio for category $J - j$ can be written as

$$r_{(J-j)}(\boldsymbol{\pi}) = \frac{\pi_{J-j}}{\pi_{J-j} + \pi_{J-j+1}} \tag{B.1}$$

for all $j \in \{1, ..., J-1\}$. Simultaneously, consider the distribution of $\tilde{Y}$ defined by the $(adjacent,$
$F,\ Z)_{\tilde{\sigma}}$ model (equivalent to $r(\boldsymbol{\pi}_\sigma) = F(Z\beta)$), where $\tilde{\sigma}$ is the reverse permutation, i.e., $\tilde{\sigma}(j) = J - j + 1$ for all $j \in \{1, ..., J-1\}$, we can prove the next equality

$$r_j(\boldsymbol{\pi}_{\tilde{\sigma}}) = 1 - r_{\tilde{\sigma}(j+1)}(\boldsymbol{\pi}) \quad (j = 1, ..., J-1) \tag{B.2}$$

through the ratio expressed for element $\tilde{\sigma}(j+1) = J - j$ in Equation B.1 where

$$\begin{aligned} 1 - r_{\tilde{\sigma}(j+1)}(\boldsymbol{\pi}) &= \frac{\pi_{J-j+1}}{\pi_{J-j} + \pi_{J-j+1}} \\ &= \frac{\pi_{\sigma(j)}}{\pi_{\sigma(j+1)} + \pi_{\sigma(j)}} \\ &= r_j(\boldsymbol{\pi}_{\tilde{\sigma}}). \end{aligned}$$

Given that $r_j(\boldsymbol{\pi}_{\tilde{\sigma}}) = F(\eta_j)$ and using Equality B.2, we obtain that $r_{J-j}(\boldsymbol{\pi}) = \tilde{F}(-\eta_j)$. If we denote $i = J - j$ the last equality becomes

$$r_i(\boldsymbol{\pi}) = \tilde{F}(-\eta_{J-i})$$

for all $j \in \{1, ..., J-1\}$. Hence $\tilde{y}$ follows the $(adjacent,\ \tilde{F},\ -PZ)$ model, where P is the restricted reverse permutation matrix of dimension $J-1$ defined in Equation 3.4. Since P has full rank, the design matrices Z and $-PZ$ are equivalent, meaning the $(adjacent,\ F,\ Z)$ model is equal to the $(adjacent,\ \tilde{F},\ -PZ)$ model. The above can be similarly demonstrated for the cumulative ratio, but not for the sequential ratio given that Equality B.2 is invalid for these models. To prove it by contradiction, the reader can assume that the statement is false, proceed from there, and at some point, a contradiction will result.

B.3.2 Proof of $(reference,\ logistic,\ AZ) = (adjacent,\ logistic,\ Z)$

Assume that the distribution of Y is defined by the $(reference, logistic, Z)$ model. For $j = 1, \ldots, J$ we obtain

$$\ln\left(\frac{\pi_j}{\pi_J}\right) = \eta_j.$$

The adjacent ratio can be rewritten in terms of the reference ratio since

$$\ln\left(\frac{\pi_j}{\pi_{j+1}}\right) = \ln\left(\frac{\pi_j}{\pi_J}\right) + \ln\left(\frac{\pi_j}{\pi_{j+1}}\right),$$

therefore, using the reparametrization

$$= \begin{cases} \eta'_j = \eta_j - \eta_{j+1}, \text{ for } j = 1, \ldots, J-2, \\ \eta'_{J-1} = \eta_{J-1} \end{cases}$$

represented by the matrix

$$A = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix}$$

of dimension $J-1$, we obtain the equality $(reference, logistic, AZ) = (adjacent, logistic, Z)$, and, given that A is invertible, we obtain $(reference, logistic, Z) = (adjacent, logistic, A^{-1}Z)$.

Appendix C

Appendix of chapter 4

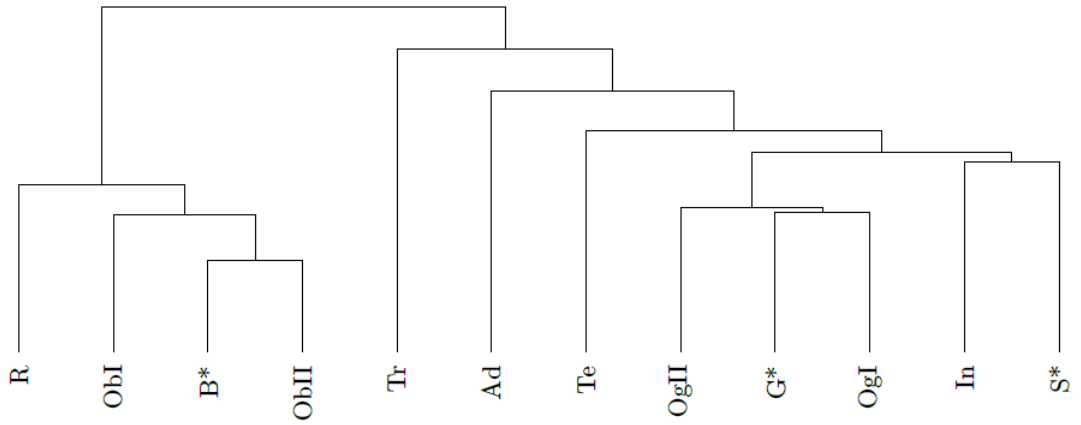

Figure 45: Tree structure found for the rice diversity data set in which missing subspecies labels are filled with the corresponding species labels.

C.0.1 Accuracies Cleveland data set

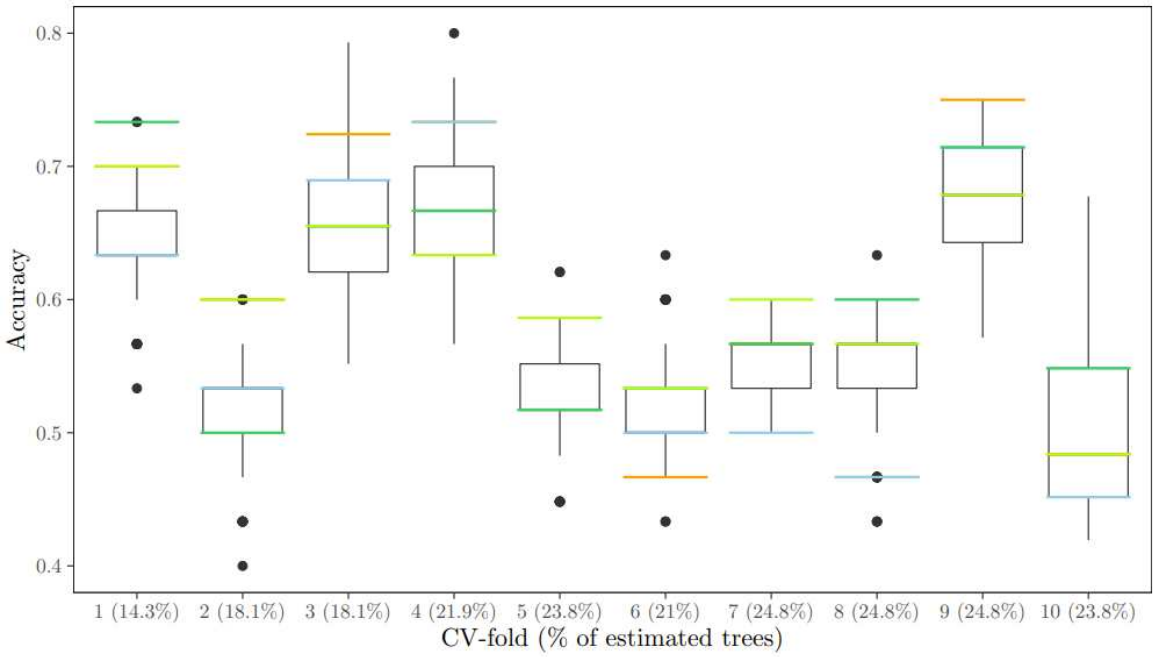

Figure 46: Box plots of the 105 B-PCGLMs classification accuracies corresponding to the 10 samples resulting from the 10-folds cross-validation procedure (x-axis). The orange, blue, and dark green lines correspond respectively to the accuracies of the multinomial model, the initial partition tree, and the best tree found using algorithm 2. The above partition structures used the logistic link function at all non-terminal vertices. The light green line is the accuracy of the model with the same partition tree structure found in the neighborhood search, but the cdfs at each non-terminal vertex were selected using algorithm 1.

Bibliography

Abreu, M. N. S., Siqueira, A. L., Cardoso, C. S. and Caiaffa, W. T. (2008), 'Ordinal logistic regression models: application in quality of life studies', *Cadernos de saude publica* **24 Suppl 4**, s581—91.

Agresti, A. (1981), 'Measures of nominal-ordinal association', *Journal of the American Statistical Association* **76**(375), 524–529.

Agresti, A. (1989), 'Tutorial on modeling ordered categorical response data.', *Psychological Bulletin* **105**, 290–301.

Agresti, A. (2010), *Analysis of ordinal categorical data*, Vol. 656, John Wiley & Sons.

Agresti, A. (2018), *An introduction to categorical data analysis*, John Wiley & Sons.

Al-Tam, F., Adam, H., Anjos, A. d., Lorieux, M., Larmande, P., Ghesquière, A., Jouannic, S. and Shahbazkia, H. R. (2013), 'P-trap: a panicle trait phenotyping tool', *BMC plant biology* **13**(1), 1–14.

Albert, A. and Anderson, J. A. (1984), 'On the existence of maximum likelihood estimates in logistic regression models', *Biometrika* **71**(1), 1–10.

Ananth, C. V. and Kleinbaum, D. G. (1997), 'Regression models for ordinal responses: a review of methods and applications.', *International journal of epidemiology* **26**(6), 1323–1333.

Andersen, E. B. (1995), Polytomous rasch models and their estimation, *in* 'Rasch Models', Springer, pp. 271–291.

Armstrong, B. G. and Sloan, M. (1989), 'Ordinal regression models for epidemiologic data', *American Journal of Epidemiology* **129**(1), 191–204.

Bagnoli, M. and Bergstrom, T. (2005), 'Log-concave probability and its applications', *Economic theory* **26**(2), 445–469.

Bouscasse, H., Joly, I. and Peyhardi, J. (2019), 'A new family of qualitative choice models: An application of reference models to travel mode choice', *Transportation Research Part B: Methodological* **121**, 74–91.

Brent, R. P. (2013), *Algorithms for minimization without derivatives*, Courier Corporation.

Brian S. Everitt, Dr Sabine Landau, D. M. L. D. D. S. (2011), *Cluster Analysis, Fifth Edition (Wiley Series in Probability and Statistics)*, Wiley Series in Probability and Statistics, 5th edn, Wiley.

Christensen, R. H. B. (2019), 'ordinal—regression models for ordinal data'. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Christmann, A. and Rousseeuw, P. (2001), 'Measuring overlap in binary regression', *Computational Statistics Data Analysis* **37**, 65–75.

Copas, J. B. (1988), 'Binary regression models for contaminated data', *Journal of the Royal Statistical Society. Series B (Methodological)* **50**(2), 225–265.
URL: *http://www.jstor.org/stable/2345763*

Coull, B. A. and Agresti, A. (2000), 'Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution', *Biometrics* **56**(1), 73–80.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society Series B (Methodological)* **34**, 187–220.

Darlu, P. and Tassy, P. (1993), *La reconstruction phylogénétique*, Concepts et méthodes.

Debreu, G. (1960), 'Review of rd luce, individual choice behavior: A theoretical analysis', *American Economic Review* **50**(1), 186–188.

Dua, D. and Graff, C. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Bates, D. and Chambers, J. (2020), *Rcpp: Seamless R and C++ Integration*. R package version 1.0.5.
URL: *https://CRAN.R-project.org/package=Rcpp*

Engel, J. (1988), 'Polytomous logistic regression', *Statistica Neerlandica* **42**(4), 233–252.

Fahrmeir, L. and Tutz, G. (2001), *Multivariate statistical modelling based on generalized linear models*, Springer.

Fienberg, S. E. (1980), *The analysis of cross-classified categorical data*, MIT press, Massachusetts.

Finney, D. (1947), 'The estimation from individual records of the relationship between dose and quantal response', *Biometrika* **34**(3/4), 320–334.

Forinash, C. V. and Koppelman, F. S. (1993), 'Application and interpretation of nested logit models of intercity mode choice', *Transportation research record* (1413).

Friedman, J., Hastie, T., Tibshirani, R. et al. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Geweke, J., Keane, M. and Runkle, D. (1994), 'Alternative computational approaches to inference in the multinomial probit model', *The Review of Economics and Statistics* **76**(4), 609–632.

Gianola, D. (2008), *Inferences from Mixed Models in Quantitative Genetics*, Vol. 1, pp. 678 – 717.

Goodman, L. A. (1983), 'The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds.', *Biometrics* **39 1**, 149–60.

Gordon, A. and Vichi, M. (2001), 'Fuzzy partition models for fitting a set of partitions', *Psychometrika* **66**(2), 229–247.

Gower, J. C. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* pp. 857–871.

Greene, W. (2003), *Econometric Analysis*, Pearson Education.

Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernández-Navarro, F. and Hervás-Martínez, C. (2016), 'Ordinal regression methods: Survey and experimental study', *IEEE Transactions on Knowledge and Data Engineering* **28**(1), 127–146.

Hampel, F. R. (1974), 'The influence curve and its role in robust estimation', *Journal of the american statistical association* **69**(346), 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011), *Robust statistics: the approach based on influence functions*, Vol. 196, John Wiley & Sons.

Harrell, F. (2015), *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer Series in Statistics, Springer International Publishing.

Harrell Jr, F. E. (2021), *rms: Regression Modeling Strategies*. R package version 6.2-0. **URL:** *https://CRAN.R-project.org/package=rms*

Hartzel, J., Agresti, A. and Caffo, B. (2001), 'Multinomial logit random effects models', *Statistical Modelling* **1**(2), 81–102.

Hothorn, T. (2020), 'Most likely transformations: The mlt package', *Journal of Statistical Software* **92**(1), 1–68.

Hothorn, T., Barbanti, L., Ripley, B., Venables, B., Bates, D. M. and Klein, N. (2021), *tram: Transformation Models*. R package version 0.6-1. **URL:** *https://CRAN.R-project.org/package=tram*

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., Zhao, Y., Liu, K., Lu, H., Li, W., Guo, Y., Lu, Y., Congcong, Z., Fan, D., Weng, Q., Zhu, C., Huang, T., Zhang, L., Wang, Y. and Han, B. (2012), 'A map of rice genome variation reveals the origin of cultivated rice', *Nature* **490**, 497–501.

IBM Corporation (2017), *IBM SPSS Statistics 25*, Armonk, NY: IBM Corporation.

Izenman, A. J. (2008), 'Modern multivariate statistical techniques', *Regression, classification and manifold learning* **10**, 978–0.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995), *Continuous univariate distributions*, Wiley Series in Probability and Statistics, 2 edn, Wiley-Interscience.

Kaufman, L. and Rousseeuw, P. J. (2009), *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons.

Koenker, R. and Yoon, J. (2009), 'Parametric links for binary choice models: A fisherian–bayesian colloquy', *Journal of Econometrics* **152**(2), 120–130.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989), 'Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models', *Journal of the American Statistical Association* **84**(406), 460–466.

Lall, R; Campbell, M. W. S. M. K. M. C. T. (2002), 'A review of ordinal regression models applied on health-related quality of life assessments', *Statistical Methods in Medical Research* **11**.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989), 'Robust statistical modeling using the t distribution', *Journal of the American Statistical Association* **84**(408), 881–896.

Lapointe, F.-J. and Cucumel, G. (1997), 'The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa', *Systematic Biology* **46**(2), 306–312.

León, L., Peyhardi, J. and Trottier, C. (2021), *GLMcat: Generalized Linear Models for Categorical Responses*. R package version 0.2.4.
URL: *https://CRAN.R-project.org/package=GLMcat*

Liddell, T. and Kruschke, J. (2018), 'Analyzing ordinal data with metric models: What could possibly go wrong?', *Journal of Experimental Social Psychology* **79**, 328–348.

Likert, R. (1932), 'A technique for the measurement of attitudes.', *Archives of psychology* .

Liu, C. (2005), *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*, pp. 227 – 238.

Liu, I. and Agresti, A. (2005), 'The analysis of ordered categorical data: An overview and a survey of recent developments', *Test* **14**(1), 1–73.

Liu, X. (2009), 'Ordinal regression analysis: Fitting the proportional odds model using stata, sas and spss', *Journal of Modern Applied Statistical Methods* **8**(2), 30.

Louviere, J. J., Hensher, D. A., Swait, J. D. and Adamowicz, W. (2000), *Stated Choice Methods: Analysis and Applications*, Cambridge University Press.

Lovelace, R., Morgan, M., Hama, L., Padgham, M., Ranzolin, D. and Sparks, A. (2019), 'stats 19: A package for working with open road crash data', *The Journal of Open Source Software* **4**(33), 1181.

Lucas, J. M. (1987), 'The rotation graph of binary trees is hamiltonian', *Journal of Algorithms* **8**(4), 503–535.

Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical analysis*, Wiley, New York, NY, USA.

Ludwig Fahrmeir, Thomas Kneib, S. L. B. M. (2013), *Regression: Models, Methods and Applications*, Springer.
URL: *http://gen.lib.rus.ec/book/index.php?md5=fec75bf6762d7ef0954d2709c596fa99*

Maxwell, A. (1961), *Analysing Qualitative Data*, Methuen.

McCullagh, P. (1980), 'Regression models for ordinal data', *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(2), 109–142.
URL: *http://www.jstor.org/stable/2984952*

McFadden, D. (1973), Conditional logit analysis of qualitative choice behaviour, *in* P. Zarembka, ed., 'Frontiers in Econometrics', Academic Press New York, New York, NY, USA, pp. 105–142.

McFadden, D. et al. (1978), 'Modelling the choice of residential location'.

Mellenbergh, G. J. (1995), 'Conceptual notes on models for discrete polytomous item responses', *Applied Psychological Measurement* **19**(1), 91–100.

Michel Marie Deza, E. D. a. (2016), *Encyclopedia of Distances*, 4 edn, Springer-Verlag Berlin Heidelberg.

Morawitz, B. and Tutz, G. (1990), 'Alternative parameterizations in business tendency surveys', *Zeitschrift für Operations Research* **34**(2), 143–156.

Mouchet, M., Guilhaumon, F., Villéger, S., Mason, N. W., Tomasini, J.-A. and Mouillot, D. (2008), 'Towards a consensus for calculating dendrogram-based functional diversity indices', *Oikos* **117**(5), 794–800.

Mudholkar, G. S. and George, E. O. (1978), 'A remark on the shape of the logistic distribution', *Biometrika* **65**(3), 667–668.

Murtagh, F. (1984), 'Counting dendrograms: a survey', *Discrete Applied Mathematics* **7**(2), 191–199.

Nelder, J. A. and Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

O'Connell, A. A. (2006), *Logistic regression models for ordinal response variables*, Vol. 146, Sage.

Peyhardi, D. J. (2020), 'Robustness of student link function in multinomial choice models', *Journal of Choice Modelling* **36**, 100228.

Peyhardi, J., Trottier, C. and Guédon, Y. (2015), 'A new specification of generalized linear models for categorical responses', *Biometrika* **102**(4), 889–906.

Peyhardi, J., Trottier, C. and Guédon, Y. (2016), 'Partitioned conditional generalized linear models for categorical responses', *Statistical Modelling* **16**(4), 297–321.

Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001), 'Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution', *Journal of Computational and Graphical Statistics* **10**(2), 249–276.

Pratt, J. W. (1981), 'Concavity of the log likelihood', *Journal of the American Statistical Association* **76**(373), 103–106.

Pregibon, D. (1981), 'Logistic regression diagnostics', *The annals of statistics* **9**(4), 705–724.

Pregibon, D. (1982), 'Resistant fits for some commonly used logistic models with medical applications', *Biometrics* pp. 485–498.

Stata Corp. (2015), *Stata Statistical Software: Release 14*, College Station, TX.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rasch, G. (1961), On general laws and the meaning of measurement in psychology, *in* 'Proceedings of the fourth Berkeley symposium on mathematical statistics and probability', Vol. 4, pp. 321–333.

Reinhard, R., Rutrecht, H. M., Hengstenberg, P., Tutulmaz, E., Geissler, B., Hecht, H. and Muttray, A. (2017), 'The best way to assess visually induced motion sickness in a fixed-base driving simulator', *Transportation Research Part F: Traffic Psychology and Behaviour* **48**, 74–88.

Renk, T., Iankov, D. and Jondral, F. K. (2009), Adaptive resource allocation in wireless relay networks, *in* 'VTC Spring 2009-IEEE 69th Vehicular Technology Conference', IEEE, pp. 1–5.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A. and Firth, D. (2021), *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-54.
URL: *https://CRAN.R-project.org/package=MASS*

Ripley, B. and Venables, W. (2021), *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-16.
URL: *https://CRAN.R-project.org/package=nnet*

Salsas, P., Guillen, M. and Alemany, R. (1999), 'Perfect value and outlier detection in logistic binary choice models', *Communications in Statistics-Theory and Methods* **28**(6), 1447–1460.

Sánchez-Monedero, J., Pérez-Ortiz, M., Saez, A., Gutiérrez, P. A. and Hervás-Martínez, C. (2018), 'Partial order label decomposition approaches for melanoma diagnosis', *Applied Soft Computing* **64**, 341–355.

Sankey, S. S. and Weissfeld, L. A. (1998), 'A study of the effect of dichotomizing ordinal data upon modeling', *Communications in Statistics-Simulation and Computation* **27**(4), 871–887.

SAS Institute Inc. (2020), *SAS/STAT Software, Version 9.4*, Cary, NC.
URL: *http://www.sas.com/*

Scott, S. C., Goldberg, M. S. and Mayo, N. E. (1997), 'Statistical assessment of ordinal outcomes in comparative studies', *Journal of Clinical Epidemiology* **50**(1), 45–55.

Silvapulle, M. J. (1981), 'On the existence of maximum likelihood estimators for the binomial response models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 310–313.

Smithson, M. and Verkuilen, J. (2006), 'A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables.', *Psychological methods* **11**(1), 54.

Tutz, G. (1989), 'Compound regression models for ordered categorical data', *Biometrical Journal* **31**(3), 259–272.

Tutz, G. (1991), 'Sequential models in categorical regression', *Computational Statistics Data Analysis* **11**(3), 275–295.

Tutz, G. (2011), *Regression for Categorical Data*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Tutz, G. (2021), 'Hierarchical models for the analysis of likert scales in regression and item response analysis', *International Statistical Review* **89**(1), 18–35.

Tutz, G. and Berger, M. (2015), 'Tree-structured modelling of categorical predictors in regression', *arXiv preprint arXiv:1504.04700* .

Tutz, G. and Hennevogl, W. (1996), 'Random effects in ordinal regression models', *Computational Statistics Data Analysis* **22**(5), 537–557.

Tutz, G. and Petry, S. (2012), 'Nonparametric estimation of the link function including variable selection', *Statistics and Computing* **22**(2), 545–561.

Van der Paal, B. (2014), 'A comparison of different methods for modelling rare events data', *Master of Statistical Data Analysis thesis, Universiteit Gent* .

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4/*

Vit, K. (1985), 'An efficient interfacing of golden-section and quadratic searches', *Journal of the Franklin Institute* **320**(3-4), 151–160.

Wang, X., Roy, V. and Zhu, Z. (2018), 'A new algorithm to estimate monotone non-parametric link functions and a comparison with parametric approach', *Statistics and Computing* **28**(5), 1083–1094.

Wickham, H. (2015), *R Packages: Organize, Test, Document, and Share Your Code*, O'Reilly Media.

Wooldridge, J. M. (2012), *Introductory Econometrics: A Modern Approach*, 5 edn, Cengage Learning.

Wurm, M., Rathouz, P. and Hanlon, B. (2020), *ordinalNet: Penalized Ordinal Regression*. R package version 2.9.
URL: *https://CRAN.R-project.org/package=ordinalNet*

Yang, Y. (2017), Chapter 3 - temporal data clustering, *in* Y. Yang, ed., 'Temporal Data Mining Via Unsupervised Ensemble Learning', Elsevier, pp. 19–34.
URL: *https://www.sciencedirect.com/science/article/pii/B9780128116548000038*

Yee, T. W. (2021), *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-5.
URL: *https://CRAN.R-project.org/package=VGAM*

Zhang, Q. and Ip, E. H. (2012), 'Generalized linear model for partially ordered data', *Statistics in Medicine* **31**(1), 56–68.

Title: About the link function in generalized linear models for categorical responses

Abstract: The logit, proportional-odds logit, and multinomial logit models are the most common models for binary, ordinal, and nominal responses, respectively. Although these models have outstanding properties, they are too sensitive to the presence of outliers, and they do not capture specific characteristics of categorical data, such as the order type or the potential grouping relationships among categories. The link function is a key component of GLMs to address these particularities. The purpose of this thesis is precisely to study this link function in various forms for categorical regression models. We first investigate the robustness of the Student link function in the case of binary outcomes according to different data separation settings. For the case of more than two categories, we then propose in the framework of a unified R-package, a practical guide to identify the most suitable model for ordered categories according to the nature of the data and the properties of the model. Finally, when assuming a binary hierarchical structure among categories, we elaborate a two-step methodology to infer it. The first step is to construct a partition tree based on the agglomerative hierarchical clustering algorithm. The second step consists of a search algorithm based on rotations to efficiently visit the space of partition trees. Overall, this thesis aims to explore, popularize, and extend the range of regression models for categorical responses.

Keywords: GLM for categorical response, Link function, Robustness, Data separation, Binary partition tree.

Titre : À propos de la fonction de lien dans les modèles linéaires généralisés pour réponses catégorielles

Résumé : Les modèles logit, logit à côtes proportionnelles et multinomial logit sont les plus classiques pour modéliser respectivement les réponses binaires, ordinales et nominales. Même si ces modèles ont des propriétés remarquables, ils sont sensibles à la présence de valeurs aberrantes, et ne permettent pas de tenir compte de caractéristiques spécifiques aux données catégorielles, comme le type d'ordre ou les possibles groupements de catégories. La fonction de lien est une composante clé des GLMs pour prendre en compte ces particularités. L'objet de cette thèse est précisément l'étude de cette fonction de lien sous diverses formes pour les modèles de régression catégorielle. Nous nous intéressons d'abord à la robustesse de la fonction de lien Student dans le cas d'observations binaires selon différentes situations de séparation des données. Avec plus de deux catégories, nous proposons ensuite, dans le cadre d'un package R unifié, un guide pratique permettant de choisir le modèle ordinal le plus adapté selon la nature des données et les propriétés des modèles. Enfin, lorsque l'on suppose une structure hiérarchique binaire des catégories, nous définissons une méthodologie en deux étapes pour l'inférer. La première étape construit un arbre de partition en se basant sur l'algorithme de classification ascendante hiérarchique. La deuxième consiste en un algorithme de recherche basé sur des rotations pour visiter efficacement l'espace des arbres de partition. De manière générale, cette thèse vise à explorer, populariser et étendre l'ensemble des modèles de régression pour données catégorielles.

Mots clés: GLM pour réponse catégorielle, Fonction de lien, Robustesse, Séparation de données, Arbre de partition binaire.