# A new generalized linear model (GLM) framework for analysing categorical data; application to plant structure and development.

Jean Peyhardi

# THÈSE

## Pour obtenir le grade de Docteur

Délivré par l'**Université Montpellier II**

Préparée au sein de l'école doctorale **I2S**[*]
Et des unités de recherche **UMR 5149, UMR AGAP**

Spécialité: **Biostatistique**

Présentée par **Jean PEYHARDI**

# A new GLM framework for analysing categorical data.

### Application to plant structure and development.

Composition du jury :

| | | |
|---|---|---|
| M. Christophe BIERNACKI | Université Lille 1 | Rapporteur |
| M. Gerhard TUTZ | Université Munich | Rapporteur |
| Mme Hélène JACQMIN GADDA | INSERM Bordeaux | Examinatrice |
| M. Christian LAVERGNE | Université Montpellier 3 | Président |
| M. Yann GUÉDON | CIRAD Montpellier | Directeur de thèse |
| Mme Catherine TROTTIER | Université Montpellier 3 | Co-directrice de thèse |
| M. Pierre-Éric LAURI | INRA Montpellier | Membre invité |

[*] **I2S:** ÉCOLE DOCTORALE INFORMATION STRUCTURES SYSTÈMES

*À la mémoire de Damien, un papillon si vite envolé . . .*

# Acknowledgements

Au delà d'un travail de recherche de trois ans, cette thèse représente une période de transition, clôturant mes années d'études et m'ouvrant les portes du monde de la recherche . . . d'emploi ! Je passe donc d'une adolescence prolongée à l'âge adulte, accompagné durant ce périple par des sages autant que de grands enfants; je tiens ici à les remercier.

Tout d'abord, je tiens à remercier mes directeurs Yann Guédon et Catherine Trottier, pour m'avoir encadré durant ces trois années. Je vous remercie pour la confiance que vous m'avez accordée, laissant de coté certaines attentes présentes dans le sujet initial de thèse : je devais passer trois mois sur ces modèles pour données catégorielles, j'y ai finalement passé trois ans !

Yann, tu as su lire entre les lignes de mon CV, qui reflète un parcours quelque peu tumultueux, et me donner l'opportunité d'embrasser la carrière de chercheur. Merci aussi pour tous tes conseils avisés, ta rigueur et ta disponibilité. Tu m'as dit un jour qu'une bonne relation entre un doctorant et son directeur doit évoluer d'une situation élève-professeur vers une situation de collaborateurs ; je crois que nous y sommes parvenus.

Catherine, tu as su me mettre en confiance dès le départ et tu n'es pas pour rien dans ma décision d'entreprendre cette thèse. Je te remercie pour le temps et l'énergie que tu m'as accordés, et pour ta patience.

Quant à messieurs Biernacki et Tutz, ils ont accepté d'être les rapporteurs de cette thèse, et je les en remercie. Je remercie Hélène Jacqmin Gadda pour sa participation au jury de thèse ainsi qu'au jury des comités de suivi de thèse. Ma reconnaissance va également à Christian Lavergne, qui a accepté de présider le jury et à Pierre Éric Lauri pour sa participation au jury et son regard avisé en tant que botaniste.

Je tiens ensuite à remercier les différents collègues et collaborateurs qui ont croisé mon chemin. Je remercie d'abord Christophe Godin pour m'avoir accueilli dans l'équipe Virtual Plants. Elle réunit à mes yeux toutes les qualités requises pour constituer une très bonne équipe de recherche : le dynamisme, les différences au sein d'un même groupe et la bonne humeur. J'espère ne pas trop regretter cette équipe. Je remercie Christophe Pradal et Fred Boudon pour l'assistance informatique qu'ils m'ont apportée, ainsi qu'Évelyne Costes et Yves Caraglio, pour les notions de botanique qu'ils m'ont transmises. Merci enfin aux collègues qui m'ont accompagné autour de nombreux cafés, comme les Juliens, Jean-Philippe à la Galéra et Mickaël, Angelina, Christophe, Julien et Jojo le teufeur à l'I3M.

Je voudrais maintenant souligner toutes les amitiés qui sont nées durant ces trois années et qui ont rendu ce parcours plus sympathique. Je tiens d'abord à saluer Pierre pour toute l'aide qu'il a pu m'apporter au travail et son soutien moral tout au long des épreuves traversées. Merci à Léo qui est toujours à l'écoute et avec qui on peut échanger de manière constructive. Je remercie Vincent pour les nombreux chats tout aussi "constructifs" que nous avons échangés mais également pour son aide administrative. Je remercie aussi Yousri pour sa bonne humeur et son style qui ont bougé le bâtiment 9, Pierre le Grec pour son franc-parler ainsi que pour son aide pré-soutenance, et sa poupounette Val pour sa gentillesse et son aide. Je salue pour finir toute la bande à Zaza pour les nombreuses soirées passées à refaire le monde, notamment chez Julien puis l'Indien qui nous ont accueillis et nourris si souvent, quel qu'ai été notre état.

Je terminerai par ceux qui me sont les plus chers. Je remercie d'abord tous mes proches, Cheucheu, le ptit Mat', Nanou, les romanos Clément et Greg, Alex le gras, et la Truie, qui répondent présent chaque année à l'appel du 9 novembre. Je remercie en particulier Greg et Alex qui ont fait le déplacement depuis Bordeaux et Paris pour assister à ma soutenance.

Je remercie très chaleureusement ma famille pour son soutient sans limite. Je remercie ma mère, qui pense avoir mis au monde un deuxième Einstein, et mon père qui se demande si je vais finir par trouver un boulot. Je remercie aussi ma grande sœur qui croit encore que j'étudie les abeilles ! Je vous embrasse de tout mon cœur et vous remercie encore de croire en moi à chaque instant, sans vous poser de question.

Je terminerai en remerciant celle qui m'a suivi tout au long de cette aventure. C'est en elle que je puise ma force. Je te remercie pour toutes les concessions que tu as faites pour me supporter ; notamment dans cette période de fin de thèse avec comme bouquet final la date de soutenance qui n'est autre que celle du jour de ta naissance. Je ne voudrais pas te faire de l'ombre, alors je te souhaite un joyeux anniversaire !

# Summary in french

Depuis les années 60, de nombreux modèles et méthodes statistiques ont été proposés pour analyser des données catégorielles. On rencontre fréquemment ce type de donnée dans différents domaines, comme l'économétrie, la psychologie, la médecine ou encore la botanique par exemple. Deux échelles sont généralement distinguées pour les catégories : ordonnée et non ordonnée. Une variable avec une échelle catégorielle ordonnée est dite *ordinale*. Comme exemple de variables ordinale et ses catégories ordonnées on compte l'idéologie politique (avec les catégories gauche, centre, droite), l'évolution de la douleur après un traitement (avec les catégories pire, semblable, amélioration, rétablissement) ou encore la qualification des unités de croissance d'une plante (avec les catégories court, moyen, long). Une variable avec une échelle catégorielle non ordonnée est dite *nominale*. Par exemple on s'intéresse à la demande de transport urbain (avec les catégories bus, car, métro, vélo), le type de musique préférée (avec les catégories rock, classique, jazz, autre) ou encore la production axillaire d'une plante (avec les catégories bourgeon latent, branche épineuse, branche non épineuse, branche florifère). Mais beaucoup de variables catégorielles ne sont ni ordinales ni nominales ; on parle alors de variables partiellement ordonnées. Elles sont bien souvent le fruit du produit cartésien de plusieurs variables latentes, dont une au moins est ordinale. La classification de l'anxiété (avec les catégories pas d'anxiété, anxiété moyenne, anxiété aiguë, anxiété avec dépression) est par exemple une variable partiellement ordonnée ou encore la qualification des unités de croissance d'une plante (avec les catégories florifère, court, moyen, long).

Dans le contexte de la régression linéaire, la famille des modèles linéaires généralisés (GLM) a été introduite par Nelder and Wedderburn (1972) pour prendre en compte une variable réponse non gaussienne. Dans le cas d'une variable réponse nominale, le GLM le plus connu est le modèle logit multinomial. Il a été introduit par Luce (1959) comme un modèle de choix mais il est également appelé *baseline logit model* (Agresti, 2002). Il est aussi défini dans plusieurs domaines comme une extension du modèle logistique simple pour variable réponse binaire. Dans la théorie des modèles de choix probabilistes, il peut être vu comme une conséquence de l'axiome de choix de Luce (Luce, 1959) ou bien obtenu en maximisant l'utilité aléatoire de l'individu (Marschak, 1960; McFadden, 1973). On parle alors de modèle RUM (Randomize Utility Maximisation). D'autre modèles RUM ont été introduits comme le modèle logit conditionnel (McFadden, 1973) ou encore le modèle logit emboité (McFadden et al., 1978). Lorsque la variable réponse est ordinale, le modèle multinomial logit n'est plus approprié. En fait ce modèle n'utilise pas l'information d'ordre sur les catégories. Trois approches pour construire des modèles pour variable réponse ordinale prédominent : l'approche cumulative, séquentielle et adjacente (Tutz, 2012). Ces trois approches permettent de définir respectivement le modèle logit proportionnel (McCullagh, 1980), le modèle logit séquentiel (Tutz, 1990), et le modèle logit adjacent (Masters, 1982; Agresti, 2002). Beaucoup d'extensions du modèle logit proportionnel et du modèle logit séquentiel ont été considéré; voir Fahrmeir and Tutz (2001); Tutz (2012) et Agresti (2010). Enfin le cas d'une variable réponse partiellement ordonnée a été formellement traité par Zhang and Ip (2012), qui ont introduit la théorie des ensembles partiellement ordonnés dans le domaine des GLMs.

Dans le cadre de l'analyse de données catégorielles, on remarque que le cas de données nominales et ordinales a été traité en profondeur tandis que le cas de données partiellement ordonnées a été délaissé. Pourtant des données comprenant une structure hiérarchique sont souvent observées dans plusieurs domaines, en particulier celui de l'architecture des plantes. Le développement d'une plante est la somme d'événements qui contribuent à la mise en place progressive du **corps** d'un organisme (Steeves and Sussex, 1989). Le développement d'une plante est défini comme une série d'événements identifiables résultant d'une modification qualitative (germination, floraison ...) ou quantitative (nombre de feuilles, nombre de fleurs ...) de la structure de la plante (Gatsuk et al., 1980). La ramification est un processus clé de développement de la plante. Les données de ramification sont la plupart du temps collectées rétrospectivement et reflètent potentiellement une succession de phases de développement complexes et dépendantes telles que :

- ramification immédiate (c-a-d l'entité produite se développe la même année que l'entité parente),
- ramification différée (c-a-d ramification différée d'un an pour les espèces tempérées),
- transformation morphologique de l'entité produite comme la transformation de l'apex en épine ou en fleur interrompant la croissance,
- élongation ou non de l'entité produite.

La production axillaire (c-a-d entités produites et bourgeons latents) peut être codée en catégories bien définies et différentiées selon des critères morphologiques. Comme des phases de développement potentiellement complexes se succèdent, ces catégories ne sont bien souvent que partiellement ordonnées. Les approches hiérarchiques, prenant en compte des structures complexes sur les catégories, deviennent alors primordial pour l'analyse de la structure et du développement des plantes.

Dans le chapitre 1 un état de l'art de l'architecture des plantes et des modèles statistiques est proposé. Le contexte biologique est présenté, en introduisant quelques concepts basiques d'architecture des plantes. Le jeu de données du poirier, qui illustrera presque tous les modèles développés dans cette thèse, est décrit. Brièvement il contient un ensemble des séquences bivariées $(y_t, x_t)$ correspondant aux productions axillaires $y_t$ (bourgeon latent (l), branche courte non épineuse (u), branche longue non épineuse (U), branche courte épineuse (s), branche longue épineuse (S)) et aux longueurs d'entre-nœuds $x_t$. La production axillaire du poirier est illustrée dans la figure 1.4.

Puis nous revisitons l'ensemble des GLMs pour variable réponse catégorielles, en commençant par les GLMs pour réponse univariée (avec des précisions dans le cas d'une réponse binaire), et en élargissant ensuite le cadre au réponses multivariées. Le modèle logit multinomial est ensuite présenté comme un GLM pour variable réponse multivariée ainsi qu'un modèle de choix qualitatifs. Nous décrivons ensuite les approches cumulative, séquentielle et adjacente pour données ordinales, en donnant des interprétations à l'aide de variables latentes et aussi des détails sur l'estimation par maximum de vraisemblance. D'autre part le modèle stéréotype de Anderson (1984) est présenté comme une extension du modèle logit multinomial adaptée à une variable réponse ordinale. Nous présentons enfin trois modèles de régression pour données structurées hiérarchiquement. Ils ont tous une structure de partitionnement et conditionnement, utile pour différents types de variable réponse: nominale, ordinale et partiellement ordonnée. Ce chapitre conclu avec quelques définitions, notations et algorithmes autour des combinaisons semi-markoviennes de modèles linéaires généralisés (SMS-GLMs). Ces modèles intégratifs sont ensuite utilisés dans le chapitre 4 pour analyser conjointement les motifs de ramification et la croissance de la pousse, à partir de jeu de données sur pommiers et poiriers.

Le chapitre 2 est dédié à la manière de spécifier un GLM pour une variable réponse catégorielle. Depuis l'introduction des GLMs par Nelder and Wedderburn (1972), beaucoup de modèles de régression pour données catégorielles ont été développé. Ces modèles ont été introduit dans différents domaines tels que la médecine, l'économétrie, la psychologie et motivés par différents paradigmes. Cela implique un manque d'unification dans la manière de spécifier tous ces modèles. La plupart d'entre eux ont été développé pour traiter des données ordinales (Agresti, 2010), tandis qu'un seul a été développé pour traiter des données nominales: le modèle logit multinomial introduit par Luce (1959) (également appelé le *baseline-category logit model* (Agresti, 2002)). Les trois modèles pour données ordinales les plus représentatifs sont le modèle logit cumulatif (McCullagh, 1980), le modèle logit séquentiel (Tutz, 1990) (également appelé le *continuation ratio logit model* (Dobson, 2002)), et le modèle logit adjacent (Masters, 1982; Agresti, 2010). Chacun d'entre eux a été étendu en remplaçant la fonction de répartition logistique par d'autres fonctions de répartition (les fonctions de répartition de la loi normal ou la loi de Gumbel entre autres; voir le modèle de Cox par exemple, également appelé modèle de hasard), ou en modifiant la paramétrisation du prédicteur linéaire (c-a-d en changeant la matrice de design $Z$). Cependant, aucune extension de ce type n'ayant été proposée pour le modèle logit multinomial, un des buts du chapitre 2 est d'y remédier.

On remarque que les trois modèles pour données ordinales mentionnés précédemment et le modèle logit multinomial sont défini à partir de la fonction de répartition logistique. Ils se différencient donc par une autre partie dans la fonction de lien. En fait, les quatre fonctions de lien correspondantes peuvent être décomposées en deux parties : la fonction de répartition logistique et un ratio de probabilités $r$. Pour le modèle logit cumulatif de McCullagh (1980) par exemple, le ratio correspond aux probabilités cumulées $P(Y \leq j)$. Nous proposons alors de décomposer la fonction de lien de n'importe quel GLM pour données catégorielles en une fonction de répartition $F$ et un ratio de probabilités $r$. En utilisant cette décomposition, on remarque que toutes les extensions des modèles classiques pour variable réponse ordinale ont été défini en fixant le ratio $r$ et en changeant la fonction de répartition $F$ et la matrice de design $Z$. Par exemple, tous les modèles cumulatifs ont été obtenu en fixant les probabilités cumulées $P(Y \leq j)$ comme partie commune. De la même manière, les deux familles de modèles séquentiels et adjacents ont été défini à partir des ratios de probabilités $P(Y = j | \, Y \geq j)$ et $P(Y = j | \, j \leq Y \leq j+1)$. Nous proposons alors d'étendre de la même manière le modèle logit multinomial en fixant son ratio $r$ et en modifiant sa fonction de répartition $F$ et sa matrice de design $Z$.

La première contribution de ce chapitre (section 2.3) est d'unifier tous ces modèles, en introduisant une nouvelle spécification par le triplet $(r, F, Z)$. Les différences et les points communs entre les modèles sont ainsi mis en évidence, les rendant plus comparables (comme on peut le voir dans la table 3.1) . Dans ce nouveau cadre, le modèle logit multinomial est alors étendu en remplaçant la fonction de répartition logistique par d'autres fonctions de répartition. Nous obtenons ainsi une nouvelle famille de modèles pour données nominales, comparable aux trois autres familles de modèles pour données ordinales. On peut désormais comparer tous ces modèles selon les trois composantes : le ratio de probabilités $r$ pour la structure, la fonction de répartition $F$ pour l'ajustement, et la matrice de design $Z$ pour la paramétrisation.

Cette comparaison est étudiée en profondeur dans les sections 2.4 et 2.5, en s'intéressant aux équivalences entre modèles. Dans un premier temps nous rappelons trois équivalences entre modèles, démontrées par Läärä and Matthews (1985), Tutz (1991) et Agresti (2010). Ces équivalences sont décrites à l'aide du triplet $(r, F, Z)$, mettant en évidence des ratios différents. Nous proposons alors de généraliser deux équivalences à des égalités de familles de modèles. De plus nous démontrons certaines propriétés d'invariance et de stabilité sous permutation des catégories réponses. Comme l'a remarqué McCullagh (1978), les modèles pour catégories nominales devraient être invariant sous n'importe quelle permutation ta,ndis que les modèles pour données ordinales devraient être invariant uniquement sous la permutation qui renverse l'ordre.

En utilisant la famille étendue de modèles pour données nominales ainsi que leurs propriétés d'invariance, nous introduisons une famille de classificateurs supervisés dans la section 2.6. Dans la section finale 2.7, nous discutons la légitimité de certains modèles vis-à-vis de l'hypothèse d'ordre sue les catégories. Nous proposons alors une classification (représentée en figure 2.10) des différents GLMs sur une échelle nominale/ordinale, justifiée par les propriétés d'invariance précédemment démontrées. Enfin nous tempérons cette classification des modèles dans la pratique, en considérant certaines difficultés liées à l'interprétabilité ou l'inférence de ces modèles.

Le chapitre 3 est se concentre sur les GLMs adaptés à des données catégorielles reposant sur une structure hiérarchique des catégories. Même si cela semble naturel pour des données ordonnées ou partiellement ordonnées, on peut également l'observer pour des données nominales. Plusieurs modèles de partitionnement conditionnels ont été proposés dans différent domaines comme l'économétrie, la médecine ou bien encore la psychologie afin de prendre en compte cette nature hiérarchique des données. Le plus connu d'entre eux reste le modèle logit emboîté introduit par McFadden et al. (1978) en économétrie, pour des choix qualitatifs (c-a-d des catégories nominales). Toujours en économétrie, Morawitz and Tutz (1990) ont introduit le *two-step model* afin de prendre en compte la hiérarchie présente sur des choix ordonnés. Ce modèle a aussi été utilisé en médecine lorsque les catégories ordonnées peuvent être décomposées en une échelle grossière et échelle plus fine (Tutz, 1989). Enfin le *partitioned conditional model for partially-ordered set* (POS-PCM) a été introduit par Zhang and Ip (2012) pour traiter le cas de données partiellement ordonnées en médecine.

Contrairement aux modèles de régression simples pour données catégorielles, tels que le modèle logit multinomial ou le modèle logit adjacent par exemple, les modèles de partitionnement conditionnels captent plusieurs mécanismes latents. En effet, l'événement $\{Y = j\}$ est décomposé en plusieurs étapes correspondant à la structure hiérarchique latente, chaque étape pouvant être influencées par différentes variables explicatives. Cette approche permet d'obtenir des modèles plus flexible avec souvent un meilleur ajustement des données et une meilleure interprétation des phénomènes. Pour formaliser la specification de ces modèles, nous introduisons les arbres orientés qui résument bien la structure hiérarchique des catégories.

Jusqu'à présent, les modèles de partitionnement conditionnels n'ont été définis formellement que pour deux ou trois niveaux dans la hiérarchie. De plus, pour tous ces modèles la structure hiérarchique des catégories est supposée connue à priori. La première contribution de ce chapitre est d'utiliser les arbres orientés pour spécifier la structure hiérarchique. Cela permet de définir les modèles de partitionnement conditionnels pour un nombre quelconque de niveaux. De plus, en s'appuyant sur la généricité de notre spécification $(r, F, Z)$, nous développons une classe plus vaste de modèles de partitionnement conditionnels pour données nominales, ordinales mais également pour données partiellement ordonnées. Enfin, au lieu de considérer que la structure hiérarchique est connue à priori, nous proposons de la retrouver dans le cas de données ordinales.

Dans la section 3.2, la spécification $(r, F, Z)$ d'un GLM pour données catégorielles est brièvement rappelée et nous introduisons la définition d'un arbre de partition. A partir de ces deux briques de base, nous définissons la classe des GLMs de partitionnement conditionnels (voir la figure 3.13 avec l'exemple du poirier) et nous décrivons leur estimation.

Dans les sections 3.3, 3.4 et 3.5 nous généralisons trois modèles hiérarchiques de la littérature en les revisitant à partir de notre spécification. Nous nous intéressons respectivement au modèle logit emboîté pour données nominales, puis au *two-step model* pour données ordinales et enfin au *POS-PCM* pour données partiellement ordonnées. Dans la section 3.4 nous décrivons aussi aussi une procédure de sélection de modèle pour données ordinales, dérivée de la procédure d'indistinguabilité de Anderson (1984), qui sélectionne dans le même temps l'arbre de partition et les variables explicatives.

Cette procédure est illustrée dans la section 3.6 en utilisant l'exemple *back pain prognosis*, analysé précédemment par Anderson (1984). Notre méthodologie pour données partiellement ordonnées est ensuite illustrée en utilisant notre exemple du poirier.

Le chapitre 4 est dédiée à l'utilisation des combinaisons semi-markoviennes de modèles linéaires généralisés de partitionnement conditionnels (SMS-PCGLM) pour décrire les motifs de ramification chez le pommier et le poirier. Les motifs de ramification d'une plante prennent souvent la forme d'une succession de zones de ramification homogènes bien différentiées. Les types de productions axillaires ne changent pas réellement à l'intérieur de chaque zone mais changent significativement entre les zones. Ces motifs de ramification ont été mis en évidence à l'aide de modèles de segmentation, en particulier en utilisant des modèles semi-markoviens cachés (Guédon et al., 2001). La ramification est modulée par deux types de facteurs : ceux qui ont un effet global sur les motifs et ceux qui varient le long de la pousse et ont des effets différents sur les productions axillaires successives. L'influence de la position architectural d'une branche, qui peut être vue comme un facteur ayant un effet global, a déjà été étudié chez le pommier (Renton et al., 2006).

Dans ce chapitre, on s'intéresse en particulier aux facteurs qui varient le long du porteur et qui modulent sa ramification. Par exemple, il a été montré que la croissance du porteur module les motifs de ramification, en particulier la ramification immédiate (ou sylleptique); voir Lauri and Terouanne (1998) pour une illustration dans le cas du pommier. Il est également possible de prendre en compte l'effet de la courbure locale du porteur (Han et al., 2007). Suivant cette idée, nous introduisons une nouvelle famille de modèles statistiques intégratifs pour l'analyse conjointe des successions et longueurs de zones de ramification et la modulation de la production axillaire, dans chaque zone, par des facteurs variant le long du porteur. Ces modèles généralisent les modèles semi-markoviens cachés pour données catégorielles (Guédon et al., 2001) en rajoutant des variables explicatives et sont appelés combinaisons semi-markoviennes de modèles linéaires généralisés de partitionnement conditionnels (SMS-PCGLMs). D'autres combinaisons semi-markoviennes de modèles de régression ont déjà été introduites pour l'analyse de la croissance d'arbres forestiers. En effet des combinaisons semi-markoviennes de modèles linéaires mixtes ont permis d'identifier et de caractériser les trois principales composantes de la croissance : la composante ontogénique, la composante environnementale et la composante individuelle (Chaubert-Pereira et al., 2009).

Le chapitre 5 décrit les travaux en cours et les perspectives autour de la spécification $(r, F, Z)$. Dans un premier temps nous étudions la convergence de l'algorithme des scores de

Fisher pour certains modèles cumulatifs et références. La non-invariance des modèles (reference, $F$, $Z$) lorsque l'on transpose la catégorie référence est ensuite étudiée pour certaines fonctions de répartition analytiques $F$. Puis nous proposons de spécifier les ratios en utilisant des graphes orientés et nous l'illustrons avec les ratios reference, adjacent et sequential. Nous terminons en proposant une extension du modèle logit conditionnel (McFadden, 1974), dont l'estimation n'est pas encore totalement implémentée.

# Contents

# List of Figures

# List of Tables

# Introduction

Many statistical models and methods have been developed over the last 50 years for the analysis of categorical data. Such data is commonly encountered in fields, such as econometrics, psychology, medicine and botany. Two scales are usually distinguished for categories: ordered and unordered. A variable with an ordered categorical scale is called *ordinal*. Examples of ordinal variables and their ordered categorical scales include political ideology (with categories liberal, moderate, conservative), degree of suffering after a treatment (with categories worse, same, improvement, relief) or qualification of plant growth unit (with categories short, medium, long). A variable with an unordered categorical scale is called *nominal*. Examples of nominal variables include urban travel choice (with categories bus, car, metro, bicycle), favourite type of music (with categories rock, classic, jazz, other) and axillary production in plants (with categories latent bud, spiny shoot, unspiny shoot, flowering shoot). But many variables are intermediate between nominal and ordinal and are referred to as *partially-ordered* variables. They often result from a Cartesian product between two non-observable categorical variables, at least one of which is ordinal. Examples of partially-ordered variables include anxiety classification (with categories no anxiety, mild anxiety, anxiety with depression, and severe anxiety) and qualification of plant growth unit (with categories flowering, short, medium, long).

In a linear regression situation, the well-known family of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) to take account of non-normally distributed response variables. The well-known GLM for nominal response variables is the multinomial logit model introduced by Luce (1959), also referred to as the baseline logit model (Agresti, 2002). It is defined in many fields as an extension of the simple logit model for binary response variables. In probability choice theory, it may be viewed as a consequence of Luce's choice axiom (Luce, 1959) or obtained by maximising the random utility of a consumer (Marschak, 1960; McFadden, 1973). Other models based on stochastic utility maximisation have also been introduced such as the conditional logit model (McFadden, 1973) and the nested logit model (McFadden et al., 1978). When the response variable is ordinal, the multinomial logit model is no longer appropriate. In fact, the multinomial logit model does not utilize all information because the ordering of categories is ignored. Three approaches prevail when constructing models for ordinal response variables: cumulative, sequential and adjacent approaches (Tutz, 2012). These three approaches lead to the odds proportional logit model (McCullagh, 1980), the sequential logit model (Tutz, 1990) (also referred to as the continuation ratio logit model (Dobson, 2002)), and the adjacent logit model (Masters, 1982; Agresti, 2002), respectively. Many extensions of the odds proportional logit model and the sequential logit model have been considered; see Fahrmeir and Tutz (2001); Tutz (2012) and Agresti (2010). Finally, the case of a partially-ordered response variable has been formally investigated by Zhang and Ip (2012), who introduced the partially-ordered set theory into the GLM framework.

In the context of categorical data analysis, the case of nominal and ordinal data has been investigated in depth while that of partially ordered data has been comparatively neglected. But this type of hierarchically-structured data is often observed in many fields, especially in plant architecture. Development is the sum of events that contribute to the progressive elaboration of the body of an organism (Steeves and Sussex, 1989). Plant development is defined as a series of identifiable events resulting in a qualitative (germination, flowering ...) or quantitative (number of leaves, number of flowers ...) modification of plant structure (Gatsuk et al., 1980). Branching is a key developmental process in plants. Branching data are often collected retrospectively and potentially reflect a succession of complex but interrelated developmental phases such as:

- immediate branching (i.e. offspring shoots developed without delay with respect to the parent node establishment date),
- delayed branching (e.g. 1-year-delayed branching for temperate species),
- morphological transformation of offspring shoots such as transformation of the apex into spin or flower leading to growth interruption,
- elongation or not of the offspring shoots leading to short or long shoots.

Possible axillary production (i.e. offspring shoots and latent buds) can efficiently be coded as categories that are well defined and separated according to morphological criteria. Because of the potentially complex succession of developmental phases, these categories cannot in most cases be ordered but they are not unstructured. Hierarchical approaches that reflect complex structuring of categories thus constitute a very promising avenue for the analysis of plant structure and development.

This thesis aimed to propose a flexible class of GLMs for partially-ordered response variables. To this end it was first necessary to clarify differences and common threads between GLMs for nominal and ordinal response variables. We then propose a new approach that combines these two types of models in order to obtain the class of partitioned conditional GLMs. In our biological context, data take the form of sequences of axillary productions. Successions of branching patterns have already been analysed using hidden semi-Markov chains (HSMC) by Guédon et al. (2001). We propose to introduce explanatory variables that vary along the shoot eg, internode length, leaf surface or local curvature, that influence axillary productions. To this end we introduced semi Markov switching generalized linear models (SMS-GLMs) that incorporate partitioned conditional GLMs as observation models.

Chapter 1 describes some basic concepts of plant architecture and the pear tree dataset, which is used throughout the thesis. The GLM framework is then presented, focusing on binomial and multinomial distributions. The classical multinomial logit model for nominal data is first presented using different paradigms. Many regression models for ordinal data are then introduced. Finally, three hierarchically-structured models are presented, dedicated respectively to nominal, ordinal and partially-ordered data. The chapter ends with some definitions, notations and algorithms for SMS-GLMs.

In chapter 2 we propose to unify the classical GLMs for categorical data by means of a new specification. In this new framework the multinomial logit model can be extended and this led us to define a new family of models for nominal data, comparable to the three classic families for ordinal data (cumulative, sequential and adjacent families). Three equivalences between models are then reviewed and two are extended. Some properties of invariance and stability under permutation of the response variable categories are studied. We then propose a new method of supervised classification illustrated using three benchmark datasets. Finally, we propose a classification of the different models along a nominal/ordinal scale.

In chapter 3, some existing hierarchically-structured models are revisited in the proposed partitioned conditional GLM framework. We focus on the nested logit model (McFadden et al., 1978) for nominal data, the two-step model (Tutz, 1989) for ordinal data and the partitioned conditional model for partially-ordered set (Zhang and Ip, 2012). A new method of category partitioning and variable selection, based on the indistinguishability property of Anderson (1984), is then proposed. This method is illustrated with the back pain prognosis example, previously analysed by Anderson (1984). Finally, our methodology for partially-ordered data is illustrated using a pear tree dataset.

Chapter 4 corresponds to the application of the statistical models investigated in this thesis to plant architecture. The branching pattern of a shoot may be influenced by many factors that vary along the shoot eg, internode length, leaf surface or local curvature. We introduce a generalization of hidden semi-Markov chains for categorical response variables that incorporates explanatory variables which vary with the index parameter. Using this model, we demonstrate the influence of shoot growth pattern on its immediate branching.

Chapter 5 presents works in progress and perspectives. We first focus on the convergence of Fisher's scoring algorithm for some particular GLMs. A particular invariance property presented in chapter 2 is studied in depth. We then propose to represent the different GLMs for categorical data using graph theory. Finally we propose an extension of the conditional logit model (McFadden, 1973) which can be viewed as a family of qualitative choice models, whose implementation is not yet available.

Chapters 2, 3 and 4, corresponding to the original contribution of this thesis have been written as pre-publications which has led to some redundancy between these chapters (mainly between chapters 2 and 3). All the statistical models developed in this thesis were implemented in C++ with a Python interface. They will be available soon as a Python module within the OpenAlea software platform: *https://www.openalea.gforge.inria.fr*

Chapter 1

# State of the art

This chapter describes state of the art method for both plant architecture and statistical modelling. The biological context is presented, introducing some basic concepts of plant architecture. The pear tree dataset, which will illustrate almost all the models introduced in this thesis, is described. The GLM framework for categorical response variable is then revisited, starting from GLM for univariate response with a focus on binary response variable, and is then generalizing to the multivariate case. The multinomial logit model is then presented as a GLM for multivariate response and also as a qualitative choice model. The cumulative, sequential and adjacent approaches for ordinal data are described, along with underlying motivations and details on maximum likelihood estimation. The stereotype model is presented as an extension of the multinomial logit model for ordinal response variables. Finally, we present three regression models for hierarchically structured data. They share a partitioned conditional structure appropriate for different scales: nominal, ordinal and partially-ordered scales. This chapter ends with some definitions, notations and algorithms about semi-Markov switching generalized linear models (SMS-GLMs). These integrative models are used in chapter 4 to analyse branching patterns and shoot growth, in apple and pear tree datasets.

## Contents

## 1.1 Biological context

This section is largely based on Godin and Caraglio (1998) andBarthélémy and Caraglio (2007). The notion of plant topological structure is based on the idea of decomposing a plant into elementary constituents and describing their connections. To obtain natural decompositions, it is possible to take advantage of the fact that plants are modular organisms: plants can be decomposed into sets of constituents of identical nature, such as internodes, axes, etc. The topological structure stemming from a modular decomposition consists of a description of the connections between modules. The different modularities that can be observed in plants are the outcome of the plant growth process.

**The growth process** A meristem is a collection of embryogenic cells that creates new tissues by successive divisions. An apical meristem (or apex) is characterized by polar and apical activity, which produces either roots or shoots. In the following, we focus on shoots production. Shoot meristems generate tissues which, through repeated activity, form an oriented sequence of metamers (see figure 1.1, left), i.e. a leaf together with its insertion node, its axillary bud and the preceding internode (White, 1979). The apical growth process generates sequences of internodes connected one to another by a *succession relation* (figure 1.2, top). Plants make branching structures if the meristems located at leaf axils enter an apical growth process. The branching process generates a *branching relation* between internodes (figure 1.2, bottom).

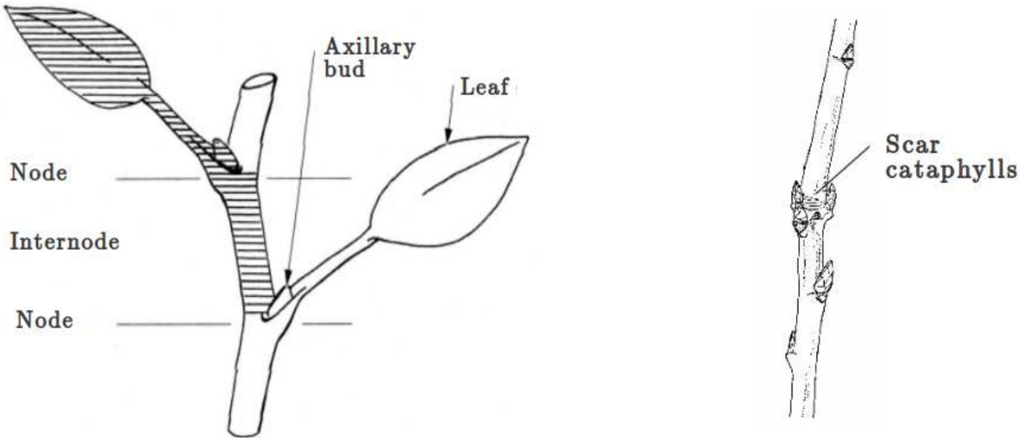

Figure 1.1: A metamer and scar cataphylls.

**Branching process** The topological distribution of sibling axes on a parent axis can take different forms. Depending on whether all the axillary meristems of a stem develop into lateral axes, or whether lateral axes are grouped as distinct tiers with an obvious regular alternation of a succession of unbranched and branched nodes on the parent stem, branching is respectively referred to as *continuous* or *rhythmic*. In some cases, none of the nodes of a parent axis are associated with a lateral axis and neither is there an obvious regular distribution of branches in tiers, and the branching pattern is then called *diffuse*. As revealed in Cupressaceae by qualitative observations (L., 1999) and, in recent years by sophisticated mathematical methods (Guédon et al., 2001; Heuret et al., 2002), a diffuse branching pattern may not mean an unorganized distribution of sibling shoots on a parent shoot, but may indicate a predictable, precise and subtle branching organization.

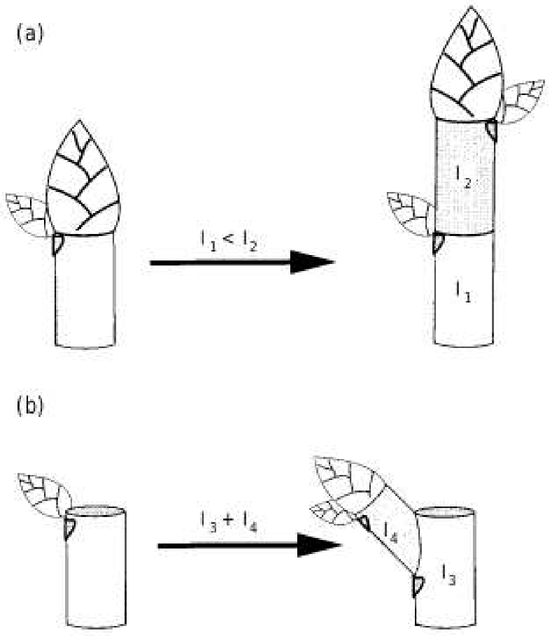

Figure 1.2: (a) Internode I1 precedes internode I2. (b) Internode I3 bears internode I4.

Shoot branching patterns often take the form of a succession of well-differentiated homogeneous branching zones where composition properties, in terms of axillary productions, do not change substantially within each zone, but change markedly between zones. These branching patterns have been analysed using segmentation models and in particular hidden semi-Markov chains (Guédon et al., 2001). Branching patterns are modulated by factors that have a global effect on the pattern and by factors that vary along the shoot and have differentiated effects on successive axillary productions. We previously investigated the influence of the architectural position of a shoot, which can be viewed as a factor that has a global effect, on apple tree branching patterns (Renton et al., 2006).

In this thesis, we focus on factors that vary along the shoot and modulate its branching pattern. For example, it has been shown that shoot growth modulates branching pattern, in particular immediate (or sylleptic) branching; see Lauri and Terouanne (1998) for an illustration in the apple tree case. Other potential factors include local curvature of the shoot. To this end, we introduce a new family of integrative models for analysing jointly the succession and length of branching zones and the modulation of the axillary productions within each zone by factors that vary along the shoot.

**Retrospective measurements** Plant growth is essentially a growth of the apical part of axes (shoots and roots). This means that new constituents are never inserted between two older constituents. From a topological perspective, plant growth may be considered as an aggregation of new constituents onto old ones (figure 1.3). The relative organization of the old structure is not modified by the appearance of new constituents. Plant topological structures grow incrementally. Also, growth occurs in such a manner that all constituents of the plant are linked to the base constituent by a single series of contiguous constituents. This property is characteristic of tree-like structures.

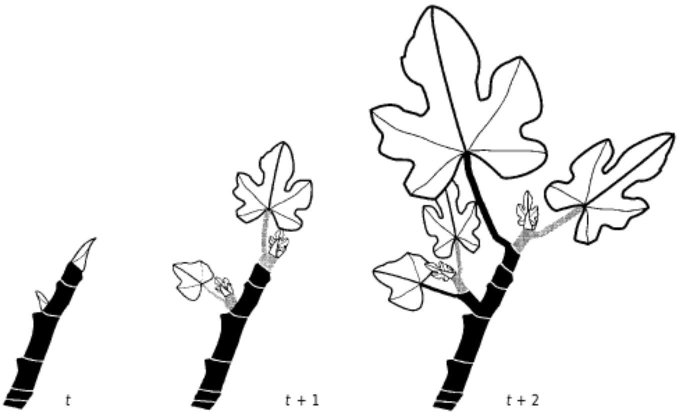

Figure 1.3: Growth process of a branching system in *Ficus carica*.

In our context, data were collected retrospectively, i.e. plant development was reconstituted at a given observation date from morphological markers (see figure 1.1, right) corresponding to past events; see Nicolini et al. (2001) for the use of pith markers. This ability to observe topological information retrospectively is a key property for plant structure analysis (plant topology is more conserved over time than plant geometry or plant biomechanical properties).

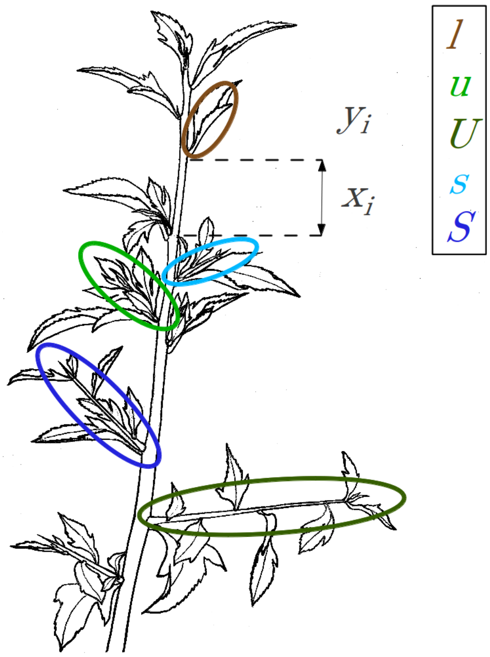

Figure 1.4: Pear tree axillary production.

**Pear tree data** Harvested seeds of Pyrus spinosa were sown and planted in January 2001 in a nursery located near Aix-en-Provence, southeastern France. Seedlings grew in 600cm3 WM containers grouped in plastic crates by 25. In winter 2001, the first annual shoots on the trunks of 50 one-year-old individuals were described by node. In this nursery context, individuals were able to grow twice a year, and the annual shoots were made up of one or two growth units (GU) - i.e. portion of the axis developed during an uninterrupted period of growth - referred to as GU1 or GU2 in the following. Seven monocyclic annual shoots (GU1 only) and 43 bicyclic annual shoots (GU1 and GU2) were observed.

The presence at each successive node of an immediate axillary shoot – i.e. developed without delay with respect to the parent node establishment date – was noted. Immediate shoots were classified in four categories according to length and transformation or not of the apex into spine (i.e. definite growth or not). The final dataset was thus made up of 50 bivariate sequences of cumulative length 3285 combining a categorical variable $Y$ (type of axillary production selected from among latent bud (l), unspiny short shoot (u), unspiny long shoot (U), spiny short shoot (s) and spiny long shoot (S)), with an interval-scaled variable $X$ (internode length). Axillary production of the pear tree is shown in figure 1.4.

## 1.2 Generalized Linear Models

A categorical response variable needs to be considered as multivariate in the GLM framework. This section revisits the basis of GLMs for univariate and multivariate response using two parametrizations of the exponential distribution family: the standard parametrization Nelder and Wedderburn (1972) and that described by Dobson (2002).

### 1.2.1 Generalized linear models for univariate response variables

The general parametrization of the linear predictor is first introduced in the context of the simple linear model. This linear model is then generalized using the enlarged exponential family of distributions and the link function is introduced. We chose to present the exponential distributions family with two parametrizations and give some hint concerning the link function. Fisher's scoring algorithm is described and a property of canonical models reviewed. Finally, GLMs for binary and binomial response variable are detailed.

Let us consider the situation of regression analysis, with the univariate response variable $Y$ and the vector of $Q$ explanatory variables $X = (X_1, \ldots, X_Q)$. We are interested in the conditional distribution of $Y|X$, observing values $(y_i, x_i)_{i=1,\ldots,n}$ of the pair $(Y, X)$. The vector of explanatory variables may be deterministic (e.g. fixed by experimental conditions) or stochastic. All the response variables $Y_i$ are supposed to be conditionally independent of each other, given $\{X_i = x_i\}$. The dependence on $x_i$ is expressed through the linear predictor $\eta_i$.

**The linear predictor** When all the explanatory variables are quantitative (discrete or continuous) the linear predictor is

$$\eta = \alpha + x^t \delta,$$

where $\alpha \in \mathbb{R}$ is the intercept and $\delta \in \mathbb{R}^Q$ is the vector of slopes.

When an explanatory variable is categorical, it has to be coded using dummy variables. A single categorical observed variable with $M$ different possible categories is transformed into an indicator vector $x$ of dimension $M - 1$. This means that the $m^{\text{th}}$ component of $x = (x_1, \ldots, x_{M-1})^t$ is defined by

$$x_m = \begin{cases} 1 & \text{if the category } m \text{ is observed,} \\ 0 & \text{else,} \end{cases}$$

and $x$ is the null vector if category $M$ is observed. The interaction between two explanatory variables $x_q$ and $x_h$ can be added using the product $x_q x_h$ (Cartesian product if both are categorical). The linear predictor $\eta$ can be written as the scalar product of the design vector $z^t = (1, x^t)$ and the parameter vector $\beta^t = (\alpha, \delta^t)$

$$\eta = z^t \beta,$$

where $\alpha \in \mathbb{R}$ is the intercept and $\delta \in \mathbb{R}^p$ is the vector of slopes.

Some equality constraints between the different slopes, called contrasts, can also be added. For example, considering the linear predictor $\eta = \alpha + \delta_1 x_1 + \delta_2 x_2$ with contrast $\delta_1 = 3\delta_2$, the reduced design vector $z = (1, 3x_1 + x_2)$ can be used instead of the design vector $z^t = (1, x_1, x_2)$. A contrast can be interpreted as a transformation of explanatory variables. It should be noted that in the case of $Q$ explanatory variables, with all the categorical explanatory variables being transformed and possible interactions and contrasts being added, the dimension $p$ of the vector $x$ is not necessarily equal to $Q$ (in fact $p \geq Q$).

**Linear model** For the classical linear model, the response variables $Y_i$ are normally distributed given $\{X_i = x_i\}$

$$Y_i | X_i = x_i \sim \mathcal{N}(\mu_i, \sigma^2),$$

where the mean parameter $\mu_i$ is a linear transformation of $x_i$, through the design vector $z_i$

$$\mu_i = z_i^t \beta,$$

where $\beta \in \mathbb{R}^{1+p}$ and $\sigma^2 \in \mathbb{R}_+$ are unknown parameters. Given a normal response variable $Y$ and a vector of explanatory variable $x$, a linear model is fully specified by the design vector $z$.

The linear model assumes that the response distribution is continuous. Generalized linear models were introduced by Nelder and Wedderburn (1972) to relax this assumption and in particular to take account of categorical and count response variables. In this framework, the distribution of the response variable is assumed to belong to the exponential family, which includes the normal distribution.

#### 1.2.1.1 Exponential family of distributions

The exponential family includes many well-known distributions such as the normal and gamma distributions for continuous variables, and the Poisson and binomial distributions for discrete variables. The density of a distribution, belonging to the exponential family, can be written in two different ways.

The first and most usual way (Nelder and Wedderburn, 1972) expresses the density function $f$ in terms of the natural parameter $\theta$

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

where

$\theta$ is the natural parameter,

$b$ and $c$ are specific functions corresponding to each distribution,

$\phi$ is the nuisance or dispersion parameter,

$\omega$ is a known weight.

**Property 1.** *Let $Y$ be a random variable whose distribution belongs to the exponential family* (1.1)*. The function $b$ is assumed to be twice differentiable.*

$$\begin{array}{lrcl} (i) & \mathrm{E}(Y) & = & b'(\theta), \\ (ii) & \mathrm{Cov}(Y) & = & \dfrac{\phi}{\omega}\, b''(\theta). \end{array}$$

For each distribution of the exponential family, $\theta$ is a particular reparametrization of its mean $\mu$. The second way (Dobson, 2002) expresses the density function $f$ in terms of $\mu$

$$f(y;\mu) = \exp\left\{a(y)\theta(\mu) + b(\mu) + c(y)\right\}, \tag{1.2}$$

where $a$, $b$, $c$ and $\theta$ are known functions. If $a$ is the identity function, the distribution is said to be canonical, and $\theta(\mu)$ is called the natural parameter. Three parts can be identified in this writing: the first depends on $y$ and $\mu$, the second on $\mu$, and the third on $y$.

**Property 2.** *Let $Y$ be a random variable whose distribution belongs to the exponential family* (1.2)*. The functions $\theta$ and $b$ are assumed to be twice differentiable.*

$$\begin{array}{lrcl} (i) & \mathrm{E}[a(Y)] & = & -\dfrac{b'(\mu)}{\theta'(\mu)}, \\ (ii) & \mathrm{Cov}[a(Y)] & = & \dfrac{\theta''(\mu)b'(\mu) - \theta'(\mu)b''(\mu)}{[\theta'(\mu)]^3}. \end{array}$$

**Link function** For a simple linear model the conditional expectation $\mu$ and the linear predictor $\eta = z^t\beta$ are directly related. For a generalized linear model, they have to be related by a particular function $g$, called the *link function*

$$\begin{array}{rccc} g: & \mathcal{M} & \longrightarrow & \mathbb{R}, \\ & \mu & \longmapsto & \eta, \end{array}$$

because the space $\mathcal{M}$ is not necessarily $\mathbb{R}$. In fact, the linear predictor $\eta$ potentially lies between $-\infty$ and $+\infty$, while the mean parameter $\mu$ lies in a particular unidimensional space $\mathcal{M}$ depending on the response variable distribution. Thus the link function takes different forms according to the constraints on space $\mathcal{M}$. In the simple case of the normal distribution, there is no constraint on $\mathcal{M}$ ($\mu$ lies between $-\infty$ and $+\infty$) and therefore $g$ is the identity function. For each distribution of the exponential family, the natural parameter $\theta$ can be seen as a particular function of $\mu$; see parametrization (1.2). This function is called the *canonical link* function. All GLMs defined with the canonical link are easy to estimate because the likelihood is strictly concave (see next paragraph for details).

We have seen that the generalisation of the linear model, using the enlarged exponential family of distributions, is used to define the link function. Finally, a GLM for univariate response variables is fully specified by

- the response variable distribution belonging to the exponential family,
- the design vector $z$,
- the link function $g$.

**Maximum likelihood estimation** Parameter $\beta$ is estimated by maximizing the log-likelihood $l$. For the linear model, the equation $\partial l/\partial\beta = 0$ has an analytic solution. For other GLMs, the equation $\partial l/\partial\beta = 0$ is not linear with respect to $\beta$ because of the link function. Thus, optimisation algorithms, such as the Newton-Raphson algorithm or Fisher's scoring algorithm, are used to approximate $\hat{\beta}$. Fisher's scoring algorithm is given, at iteration $m+1$, by

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

For the sake of simplicity, the algorithm is detailed for only one observation $(y, x)$ and therefore $l = \log P(Y = y|X = x; \beta)$. Using the chain rule, the score is given by

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

Using Property 1 we obtain

$$\frac{\partial l}{\partial\beta} = z \left(\frac{\partial g}{\partial\mu}\right)^{-1} \frac{1}{\mathrm{V}(Y|x)} (y - \mu),$$

and Fisher's information matrix

$$\mathrm{E}_y\left(\frac{\partial^2 l}{\partial\beta^t \partial\beta}\right) = -z \left(\frac{\partial g}{\partial\mu}\right)^{-1} \frac{1}{\mathrm{V}(Y|x)} \left(\frac{\partial g}{\partial\mu}\right)^{-1} z^t. \tag{1.3}$$

It should be noted that the link function $g$ must be invertible and inverse $g^{-1}$ must be differentiable in order to obtain the score and Fisher's information matrix. Moreover, $g^{-1}$ must be strictly monotone to easily interpret explanatory effect through estimated parameter $\hat{\beta}$. The link function $g : \mathcal{M} \to \mathbb{R}$ is thus generally assumed to be a diffeomorphism.

This algorithm ensures convergence towards the global maximum for any initial parameter $\beta^{[0]}$, when the loglikelihood is strictly concave. For the canonical link function, the observed information matrix $\mathcal{J}(\theta) = \partial^2 l/\partial\beta^t\partial\beta$ and Fisher's information matrix $\mathcal{I}(\theta) = E_y[\mathcal{J}(\theta)]$ coincide. Therefore, for the canonical link function, the observed information matrix $\mathcal{J}(\theta)$ is negative definite; see (1.3). Thus, the log-likelihood is strictly concave.

#### 1.2.1.2 Bernoulli and binomial distributions

We focus here on GLMs for binary response variables.

**Bernoulli distribution as a member of the exponential family** Response variables are measured on a binary scale and coded by 0 or 1. The axillary production of a plant, for instance, may be qualified by the *presence* ($y = 1$) or *absence* ($y = 0$) of an axillary shoot. *Success* and *failure* are used as generic terms for the two categories. Let the binary random variable $Y$ follow the Bernoulli distribution with parameter $\pi \in [0, 1]$ with probability function

$$P(Y = y) = \pi^y (1 - \pi)^{1-y},$$

where $y \in \{0, 1\}$. This is denoted by $Y \sim \mathcal{B}er(\pi)$. This probability function can be rewritten as

$$P(Y = y) = \exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi) \right\},$$

which is of the form (1.1) with $b(\theta) = \log\{1 + \exp(\theta)\}$ and of the form (1.2) with $b(\pi) = \log(1 - \pi)$.

**Binomial distribution as a member of the exponential family** A binary variable is observed repeatedly $n$ times and focus is made on the number of successes, assuming independence between repetitions. For example, $y$ is the number of axillary shoots along $n$ successive nodes, and consequently, $n-y$ is the number of latent buds. Let the discrete variable $Y$ follow the binomial distribution with parameters $n \in \mathbb{N}^*$ and $\pi \in [0,1]$ with probability function

$$P(Y=y) = \binom{n}{y} \pi^y (1-\pi)^{n-y},$$

where $y \in \{0,\ldots,n\}$. This is denoted by $Y \sim \mathcal{B}(n,\pi)$. This probability function can be rewritten as

$$P(Y=y) = \exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y} \right\},$$

which is of the form (1.1) with

$$\theta = \log\left(\frac{\pi}{1-\pi}\right),$$

$$\omega = \phi = 1,$$

$$b(\theta) = n\log\{1+\exp(\theta)\},$$

$$c(y,\phi) = \log\binom{n}{y},$$

and of the form (1.2) with

$$\theta(\pi) = \log\left(\frac{\pi}{1-\pi}\right),$$

$$a(y) = y,$$

$$b(\pi) = n\log(1-\pi),$$

$$c(y) = \log\binom{n}{y}.$$

In the GLM framework, the mean is related to the linear predictor. For the binomial distribution, even if the mean is $n\pi$, the parameter of interest is just $\pi$. In fact, the parameter $n$ is involved in the function $c$ of (1.1) and (1.2). However this function must be independent of the mean $\mu$. The response variable has then to be slightly transformed. The binomial distribution is expressed in terms of proportion $\bar{y} = y/n$. The form of the distribution remains the same; only the support changes since for $n$ trials $y$ takes values in $\{0,\ldots,n\}$, whereas $\bar{y}$ takes values in $\{0, 1/n, \ldots, 1\}$. The distribution of $\bar{y}$ is called scaled binomial distribution and is noted $\bar{Y} \sim \mathcal{B}(n,\pi)/n$. The expectation of $\bar{Y}$ is now $\pi$. The scaled binomial distribution belongs to the exponential family since for $\bar{y} \in \{0, 1/n, \ldots, 1\}$ we have

$$P(\bar{Y} = \bar{y}) = \exp\left\{ n\bar{y} \log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{n\bar{y}} \right\}.$$

It should be noted that the number $n = n_x$ of observed data $(y,x)$ changes according to the different levels of explanatory variable $x$. Therefore the response variable $\bar{y}$ takes values in different sets $\{0, 1/n_x, \ldots, 1\}$. Therefore, $n_x$ independent response variables with Bernoulli distributions $\mathcal{B}er(\pi(x))$ are more appropriate in the GLM framework than a single response variable with a scaled binomial distribution $\mathcal{B}(n_x, \pi(x))/n_x$.

**Link function for binary response** For the Bernoulli distribution, $\pi$ lies within the unit interval $[0,1]$ and thus the identity link function is not suitable. The inverse of a cumulative distribution function $F$ is more appropriate. In fact, the link function $g$ must be a diffeomorfism between $\mathcal{M}$ and $\mathbb{R}$. This holds for the Bernoulli case if the inverse link function $g^{-1}$ is a strictly increasing and continuous cumulative distribution function $F$. Therefore $\mathcal{M}$ is the open unit interval $]0,1[$.

**Canonical link function** The canonical link is the logit function

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

The inverse canonical link is the cumulative logistic distribution function

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

Finally, the classical logit model has the following form

$$\log\left(\frac{\pi}{1-\pi}\right) = z^t\beta,$$

or equivalently

$$\pi = \frac{\exp(z^t\beta)}{1+\exp(z^t\beta)}.$$

**Alternative link functions** Common choices of link function are presented here, all defined by the inverse of classical cdfs. We consider models of the form

$$F^{-1}(\pi) = z^t\beta,$$

or equivalently

$$\pi = F(z^t\beta).$$

A widely used model, particularly in econometrics, is the probit model based on the standard normal distribution

$$\phi(\eta) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\eta} \exp(-t^2/2)dt.$$

In practice, the probit and logit models yield approximately the same results. Since $\phi$ is not analytically defined, the logit model is often preferred, parameters having simple interpretation in terms of log-odds. Another conventional model is the complementary log-log model, defined with the inverse cdf of the minimum extreme value distribution

$$F(\eta) = 1 - \exp\left\{-\exp(\eta)\right\},$$

also referred to as the Gumbel min distribution. Unlike the logistic and normal distributions, the Gumbel min distribution is not symmetric. Another model can therefore be directly obtained with the symmetric cdf $\tilde{F}(\eta) = 1 - F(-\eta)$ corresponding to the maximum extreme value distribution

$$\tilde{F}(\eta) = \exp\left\{-\exp(-\eta)\right\},$$

also referred to as the Gumbel max distribution. The corresponding model is called the log-log model. The Cauchy distribution characterized by its heavy tails can also be used

$$F(\eta) = \tan^{-1}(\eta)/\pi + 1/2,$$

where $\pi \simeq 3.14159$. Finally the exponential distribution can be used

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

bearing in mind that the linear predictor $\eta$ must be strictly positive in this case.

It should be noted for all these cdfs that if the location parameter $u$ and the scale parameter $s$ are modified we have

$$F_{u,s}(\eta) = F\left(\frac{\eta - u}{s}\right) = F\left(\frac{\alpha - u}{s} + x^t\frac{\delta}{s}\right),$$

and we obtain an equivalent model using the reparametrization $\alpha' = (\alpha - u)/s$ and $\delta' = \delta/s$. Finally, a GLM for a Bernoulli response variable is fully specified by:

- the design vector $z$,
- the cdf $F$.

**Fisher's scoring algorithm** If $f$ denotes the density function corresponding to cdf $F$, the score is given by

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

and Fisher's information matrix is given by

$$\mathrm{E}\left[\frac{\partial^2 l}{\partial \beta^T \partial \beta}\right] = -\frac{f^2(\eta)}{F(\eta)[1 - F(\eta)]}\,zz^t.$$

The log-likelihood is strictly concave for the logit canonical link. This is not the case for other links because the observed information matrix $\mathcal{I}(\theta)$ depends on observation $y$. Wedderburn (1976) has shown that the log-likelihood for the normal and the Gumbel min distributions is strictly concave. More generally, strict concavity of the log-likelihood holds if $F$ and $1 - F$ are strictly log-concave. It should be noted that concavity in $\beta$ is equivalent to concavity in $\eta$ because

$$\frac{\partial^2 l}{\partial \beta^t \partial \beta} = \frac{\partial^2 l}{\partial \eta^2} zz^t.$$

Finally, distinguishing the two cases $\{y = 1\}$ and $\{y = 0\}$, the loglikelihood is either $\ln(F)$ or $\ln(1 - F)$. Using results from convex analysis, the strict log-concavity of $F$ and $1 - F$ can be also shown for Gumbel max and Laplace distributions, but not for Student distributions (Bergstrom and Bagnoli, 2005).

### 1.2.2 Generalized linear models for multivariate response variable

The categorical distribution (with more than two categories), and the multinomial distribution cannot be written using the univariate exponential forms (1.1) and (1.2). The exponential family has to be defined in the multivariate case. Let us consider a random vector $Y$ of $\mathbb{R}^K$ whose distribution depends on a parameter $\theta \in \mathbb{R}^K$. The distribution belongs to the exponential family if it can be written as (generalization of the form (1.1))

$$f(y; \theta, \phi) = \exp\left\{\frac{y^t\theta - b(\theta)}{\phi}\omega + c(y, \phi)\right\}, \tag{1.4}$$

where $b$, $c$ are known functions, $\phi$ is the *dispersion parameter*, $\omega$ is a known weight and $\theta$ is the *natural parameter*. It should be noted that the product between $y$ and $\theta$ is a scalar product.

**Property 3.** *Let $Y$ be a random vector whose distribution belongs to the exponential family (1.4). The function $b$ is assumed to be twice differentiable with respect to $\theta$.*

$$\begin{array}{lrcl} (i) & \mathrm{E}(Y) & = & \nabla_b(\theta), \\ (ii) & \mathrm{Cov}(Y) & = & \dfrac{\phi}{\omega}\mathcal{H}_b(\theta). \end{array}$$

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

As with Dobson, we propose to generalize the parametrization (1.2) for the multivariate case

$$f(y;\theta) = \exp\left\{a(y)^t\theta(\mu) + b(\mu) + c(y)\right\}, \tag{1.5}$$

where $a$, $\theta$ are known functions from $\mathbb{R}^K$ to $\mathbb{R}^K$ and $b$, $c$ are known functions from $\mathbb{R}^K$ to $\mathbb{R}$. We also propose to generalize Property 2 in the multivariate case.

**Property 4.** *Let $Y$ be a random vector whose distribution belongs to the exponential family (1.5). The Jacobian matrix $\mathcal{J}_\theta(\mu)$ is assumed to be defined and invertible and the function $b$ is assumed to be twice differentiable.*

$$\begin{array}{lrcl} (i) & \mathrm{E}[a(Y)] & = & -\mathcal{J}_\theta^{-1}(\mu)\nabla_b(\mu) \\ (ii) & \mathrm{Cov}[a(Y)] & = & \mathcal{J}_\theta^{-1}(\mu)\left[\left\{\left(\dfrac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t\mathcal{J}_\theta^{-1}(\mu)\nabla_b(\mu)\right\}_{i,j} - \mathcal{H}_b(\mu)\right]\mathcal{J}_\theta^{-t}(\mu) \end{array}$$

See appendix A for the proof, which is a generalisation of Dobson's proof (Dobson, 2002).

#### 1.2.2.1 Multinomial distribution

Let $J \geq 2$ denote the number of categories of the response variable and $n \geq 1$ the number of trials. Let $\pi_1, \ldots, \pi_J$ denote the probabilities of each category, such that $\sum_{j=1}^J \pi_j = 1$. The discrete vector $\tilde{Y}$ follows the multinomial distribution

$$\tilde{Y} \sim \mathcal{M}(n, (\pi_1, \ldots, \pi_J)),$$

with $\sum_{j=1}^J \tilde{y}_j = n$. In the GLM framework, as only the probabilities $\pi_j$ are on interest, we focus on the case $n = 1$. Moreover, only $J-1$ probabilities $\pi_j$ are required to define the distribution (see chapter 2 for more details). Therefore, the truncated vector $Y = (Y_1, \ldots, Y_{J-1})^t$ and its expectation $\pi = (\pi_1, \ldots, \pi_{J-1})^t$ are introduced. One observation $y$ is an indicator vector of the observed category (the null vector corresponding to the last category). The distribution function is written in terms of $y$

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

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

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

and

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

Thus, the density function is

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

Using the weight $\omega = 1$, the dispersion parameter $\phi = 1$ and the null function $c(y,\lambda) = 0$, we see that this distribution function belongs to the exponential family of dimension $K = \dim(Y) = \dim(\theta) = \dim(\pi) = J - 1$.

#### 1.2.2.2 Canonical link function

The canonical link function for categorical GLMs is

$$\begin{array}{cccc} g: & \mathcal{M} & \longrightarrow & \mathbb{R}^{J-1} \\ & \pi & \longmapsto & \eta \end{array},$$

such that

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

for $j = 1, \ldots, J-1$, where $\mathcal{M} = \left\{\pi \in ]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\right\}$. The linear predictor is now a vector $\eta = (\eta_1, \ldots, \eta_{J-1})$ and thus we must use a design matrix $Z = Z(x)$ instead of a design vector $z$

$$\eta = Z\beta.$$

Finally, a GLM for a categorical response variable is fully specified by

- the design matrix $Z$,
- the link function $g = (g_1, \ldots, g_{J-1})$.

#### 1.2.2.3 Fisher's scoring algorithm

For categorical GLMs, since the mean parameter $\pi$ and the linear predictor are multivariate, the score is given by

$$\frac{\partial l}{\partial \beta} = Z^t \, \frac{\partial \pi}{\partial \eta} \, \mathrm{Cov}(Y|x)^{-1} \, [y - \pi], \tag{1.6}$$

where the Jacobian matrix $\partial\pi/\partial\eta$ depends on the link function. Therefore, in the following, we will simply detail the computation of this matrix for different link functions. It should be noted that the score, for the canonical link, is simplified as follows

$$\frac{\partial l}{\partial \beta} = Z^t[y - \pi].$$

## 1.3 Logit model for nominal data

The multinomial logit model is the most commonly used regression model for nominal response variables. The probability of category $j$ is given by

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

for $j = 1, \ldots, J$. A reference category, for example the last one, must be arbitrarily chosen and corresponding parameters $\alpha_J$ and $\delta_J$ are assumed to be zero in order to avoid identifiability problems. We thus obtain

$$P(Y = j|x) = \frac{\exp(\alpha_j + x^t\delta_j)}{1 + \sum_{k=1}^{J-1} \exp(\alpha_k + x^t\delta_k)}, \tag{1.7}$$

for $j = 1, \ldots, J-1$. This model has been introduced into 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.

### 1.3.1 GLM for nominal response

We have seen that a GLM for categorical response variables is fully specified by the design matrix $Z$ and the link function $g$. The multinomial logit model is defined by the canonical link function and the following design matrix

$$Z = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix},$$

with $J-1$ rows and $(J-1)(1+p)$ columns. See Fahrmeir and Tutz (2001) for more details about this definition of the multinomial logit model.

### 1.3.2 Baseline-category logit model

In this framework, log odds for all $\binom{J}{2}$ pairs of categories are described. Given a particular subset of $J-1$ log odds, the complementary subset is implicitly described. In fact, a baseline category must be chosen and the $J-1$ other proportions $\pi_j$ are related to the baseline proportion. For example, using the last category $J$ as baseline, we have

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

for $j = 1, \ldots, J-1$. The effects change according to the response paired with the baseline. These $J-1$ 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 $j, k \in \{1, \ldots, J-1\}$. See Agresti (2002) for more details about this definition of the multinomial logit model.

### 1.3.3 Qualitative choice model

In qualitative choice models the statistical individual $i$ is a consumer, the variable $Y$ is the choice of a consumer among $J$ alternatives, and $x$ is the vector of attributes. Two approaches lead to the multinomial logit model: Luce's choice axiom and the principle of random utility maximisation.

### Luce's choice axiom

Luce (1959) defined an axiom for qualitative choice models. This axiom is based on two fundamental properties: choice is probabilistic and the probability of choosing an alternative from one set is related to the probability of choosing the same alternative from a different set. Let $C$ denote a finite set of alternatives and $P_B(j)$ the conditional probability of choosing the alternative $j$ given the subset of alternatives $B \subset C$. Luce's choice axiom is defined by the two following parts
Part 1: if $P_{\{j,k\}} \in ]0,1[$ for all $j,k \in C$ with $j \neq k$, then for $A \subset B \subset C$

$$P_C(A) = P_B(A) P_C(B),$$

Part 2: if $P_{\{j,k\}} = 0$ for some $j,k \in C$ with $j \neq k$, then for $B \subset C$

$$P_C(B) = P_{C \setminus \{j\}}(B \setminus \{j\}).$$

The second part, which is useful when some alternatives are never chosen in pairwise choices, is often ignored. It should be noted that the first part is not formally a conditional probability because this axiom does not assume that $C$ is a universal set. Using Luce's choice axiom (part 1) for a subset of two alternatives $A = \{j,k\}$, we obtain the *independence of irrelevant alternative* (IIA) property

$$\frac{P_{\{j,k\}}(j)}{P_{\{j,k\}}(k)} = \frac{P_B(j)}{P_B(k)},$$

for all subsets $B$ such that $\{j,k\} \subset B$.

Luce (1959) showed that his axiom implies the existence of strictly positive values $v_j$ for each alternative $j \in C$ such that for all $B \subset C$ and all $j \in B$ we have

$$P_B(j) = \frac{v_j}{\sum_{k \in B} v_k}.$$

As the quantities $v_j$ are strictly positive, there exists real values $\eta_j$ such that

$$P_B(j) = \frac{\exp(\eta_j)}{\sum_{k \in B} \exp(\eta_k)},$$

for all $B \subset C$ and all $j \in B$.

### Principle of random utility maximisation

In probabilistic choice theory, it is often assumed that for each consumer $i$, an unobserved utility $U_{i,j}$ is associated with the $j^{\text{th}}$ alternative. In this framework, a rational consumer $i$ will choose the alternative $j$ that provides the highest utility $U_{i,j}$

$$\{Y_i = j\} \Leftrightarrow \{U_{i,j} = \max_{1 \leq k \leq J} U_{i,k}\}.$$

Although it is assumed that choices are made rationally, not every characteristic of the individual or choice situation that affects choice behaviour can be measured. Therefore, the random utility $U_{i,j}$ has the following form

$$U_{i,j} = \eta_{i,j} + \varepsilon_{i,j},$$

where $\eta_{i,j}$ is the structural part and $\varepsilon_{i,j}$ is the random part. The residuals $\varepsilon_{i,j}$ are independently identically distributed random variables with cdf $H$ and density function $h$. If this choice behaviour situation holds, $Y$ is determined by the principle of random utility maximisation and we obtain a random utility model (RUM).

Marschak (1960) and McFadden (1973) showed that the multinomial logit model is a RUM. In fact, the probability of alternative $j$ for one consumer is

$$\begin{aligned} P(Y=j) &= P\left(\bigcap_{k\neq j}\{U_j \geq U_k\}\right) \\ &= P\left(\bigcap_{k\neq j}\{\varepsilon_k \leq \eta_j - \eta_k + \varepsilon_j\}\right) \\ P(Y=j) &= \int_{-\infty}^{+\infty}\left(\prod_{k\neq j} H\left(\eta_j - \eta_k + e\right)\right) h(e)de. \end{aligned}$$

The last equality holds because the residuals $\varepsilon_j$ are independently and identically distributed. Finally, if we make the assumption of a Gumbel max distribution for residuals $\varepsilon_j$, we obtain

$$P(Y=j) = \frac{\exp(\eta_j)}{\sum_{k=1}^{J}\exp(\eta_k)},$$

for $j = 1, \dots, J$.

### Different parametrizations of logit models

Depending on the form of the linear predictors $\eta_j$, we obtain different logit models:

- Multinomial logit model: $\eta_j = \alpha_j + x^t\delta_j$. Here the attributes are the same for all alternatives and the parameters depend on each alternative.

- Conditional logit model: $\eta_j = \alpha + x_j^t\delta$. Here the attributes are dependent on each alternative and the parameters are the same for all alternatives.

- Universal logit model: $\eta_j = \alpha_j + x_j^t\delta_j$. Here the attributes and the parameters are dependent on each alternative.

These three types of logit models respect the principle of random utility maximisation. They also satisfy Luce's choice axiom and consequently share the IIA property. The ratio of probabilities for alternatives $j$ and $k$

$$\frac{P(Y=j)}{P(Y=k)} = \exp(\eta_j - \eta_k)$$

does not depend on other alternatives. This property is sometimes too restrictive to model individual choice behaviour, as explained by Debreu (1960) with the well known example of blue and red buses.

## 1.4 Generalized linear models for ordinal data

As noted by Agresti (2010), the results obtained with the multinomial logit model are invariant under permutations of the response variable categories. Therefore, the multinomial logit model does not utilize all information because the ordering of categories is ignored. Moreover often more parameters than are really needed are involved in the model. Models devoted to ordinal data are expected to be more parsimonious and have simpler interpretations than the multinomial logit model.

This section describes the cumulative, sequential and adjacent approaches for ordinal data. For each approach we present the model in its original logit form and in general form. We describe the latent regression model and give detail on the maximum likelihood estimation in the general case (except for the adjacent approach). The stereotype logit model (Anderson, 1984) is presented as an extension of the multinomial logit model which requires ordinal constraints on parameters. The indistinguishability procedure introduced by Anderson (1984) for an ordinal response variable is also described.

### 1.4.1 Cumulative models

#### 1.4.1.1 Cumulative logits

For $J$ categories with associated probabilities $\pi_1, \dots, \pi_J$, the cumulative logits are defined by

$$\begin{aligned}\text{logit}\left\{P(Y \leq j)\right\} &= \log\left\{\frac{P(Y \leq j)}{P(Y > j)}\right\} \\ &= \log\left\{\frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J}\right\},\end{aligned}$$

for $j = 1, \dots, J-1$. The cumulative proportional logit model is then defined by relating cumulative logits to proportional linear predictors

$$\text{logit}\left\{P(Y \leq j|x)\right\} = \alpha_j + x^t\delta,$$

for $j = 1, \dots, J-1$. It should be noted that the logit difference has the simple form

$$\begin{aligned}\text{logit}\left\{P(Y \leq j|x_1)\right\} - \text{logit}\left\{P(Y \leq j|x_2)\right\} &= \log\left\{\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)}\right\} \\ &= \delta^t(x_1 - x_2).\end{aligned}$$

We can see that the log odds ratio does not depend on category $j$ and is proportional to the distance between $x_1$ and $x_2$. Because of this proportional odds property, the cumulative logit model is also called the *proportional odds logit model* (McCullagh, 1980).

#### 1.4.1.2 Latent variable motivation

The proportional odds model can be defined using a latent variable to simplify its interpretation (McCullagh, 1980). Let $\tilde{Y}$ be a latent continuous variable and $a_1, \dots, a_{J-1}$ be strictly-ordered cut points. The response events can also be expressed in terms of the latent variable $\tilde{Y}$

$$\{Y = j\} \Leftrightarrow a_{j-1} < \tilde{Y} \leq a_j,$$

for $j = 1, \dots, J$ with $a_0 = -\infty$ and $a_J = +\infty$ by convention. The order is more easily interpretable using the latent continuous variable, where the ordered categories are now considered as successive intervals $]a_{j-1}, a_j]$. A linear regression model can then be defined, considering $\tilde{Y}$ as the response variable

$$\tilde{Y}_i = a + x_i^t b + \varepsilon_i, \tag{1.8}$$

where the residuals $\{\varepsilon_i\}_{i=1,\ldots,n}$ are independent and identically distributed random variables with cdf $F$. Finally, the cumulative probabilities for one individual are

$$\begin{aligned} P(Y \leq j|x) &= P(\tilde{Y} \leq a_j) \\ &= P\left(\varepsilon \leq a_j - a - x^t b\right) \\ P(Y \leq j|x) &= F(\alpha_j + x^t \delta) \end{aligned} \tag{1.9}$$

with $\alpha_j = a_j - a$, and $\delta = -b$. This latent variable motivation leads naturally to the parametrization $\eta_j = \alpha_j - x^t\delta$ (used for example in the *polr* package in R). Using the logistic cdf as residual distribution, we obtain the odds proportional logit model.

**Grouped Cox model.** Using the Gumbel min distribution for the residual $\varepsilon_i$ of latent model (1.8), we obtain the grouped Cox model. In this case (1.9) becomes

$$P(Y \leq j|x) = 1 - \exp\{-\exp(\alpha_j + x^t\delta)\}, \tag{1.10}$$

for $j = 1, \ldots, J-1$. This is equivalent to

$$\log\left[-\log\{P(Y > j|x)\}\right] = \alpha_j + x^t\delta,$$

for $j = 1, \ldots, J-1$, with the complementary log-log link.

Cumulative models can also be defined without the proportionality assumption. We can use any design matrix $Z$ such that the corresponding linear predictors $\eta_j(x)$ are strictly ordered for any observed values $x$. Finally, a cumulative model is fully specified by

- the design matrix $Z$,
- the cdf $F$.

#### 1.4.1.3 Fisher's scoring algorithm

Only the Jacobian matrix $\partial\pi/\partial\eta$ depends on the link function in the score equation of multinomial GLM (1.6). For cumulative models (1.9), we have $\pi_j = F(\eta_j) - F(\eta_{j-1})$, and thus the general term of the Jacobian matrix is

$$\frac{\partial \pi_j}{\partial \eta_i} = \begin{cases} f(\eta_j) & \text{if } i = j, \\ -f(\eta_{j-1}) & \text{if } i = j-1, \\ 0 & \text{otherwise,} \end{cases}$$

for row $i$ and column $j$, where $f$ denotes the density function $f = F'$.

It has been shown that concavity of the log-likelihood holds for cumulative proportional models if $F$, $1-F$ and $f$ are log-concave (Pratt, 1981; Burridge, 1981). Using results of convex analysis, the strict log-concavity of $F$, $1-F$ and $f$ can be shown for logistic, normal, Gumbel min, Gumbel max and Laplace distributions, but not for Student distributions (Bergstrom and Bagnoli, 2005).

## 1.4.2 Sequential models

### 1.4.2.1 Sequential logits

For $J$ categories with associated probabilities $\pi_1, \ldots, \pi_J$, the sequential logits are defined by

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

for $j = 1, \ldots, J-1$. The sequential proportional logit model is then defined by relating sequential logits to proportional linear predictors

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

for $j = 1, \ldots, J-1$.

### 1.4.2.2 Latent variables motivation

The sequential proportional model can also be motivated by latent random variables (Tutz, 1991). Let $(\tilde{Y}_{i,j})_{j=1,\ldots,J-1}$ be a random sequential continuous process for each individual $i$ such that for $j = 1, \ldots, J-1$

$$\tilde{Y}_{i,j} = a + x_i^t b + \varepsilon_{i,j}, \tag{1.11}$$

where the residuals $\{\varepsilon_{i,j}\}_{i=1,\ldots,n}^{j=1,\ldots,J-1}$ are independent and identically distributed random variables with cdf $F$ and corresponding density function $f$. For one individual, $(\tilde{Y}_j)_{j=1,\ldots,J-1}$ is a sequential binary mechanism. At each step $j$, if the continuous variable $\tilde{Y}_j$ exceeds cut point $a_j$, then the process continues, otherwise the process is stopped. The event $\{Y = j\}$ occurs if the latent process have been stopped at step $j$. Therefore, the event $\{Y = j\}$ can be decomposed into $j$ latent events

$$\{Y = j\} \Leftrightarrow \bigcap_{t=1}^{j-1}\{\tilde{Y}_t > a_t\}\bigcap\{\tilde{Y}_j \leq a_j\}.$$

The conditional event $\{Y = j|Y \geq j\}$ may also be expressed as

$$\{Y = j|Y \geq j\} \Leftrightarrow \{\tilde{Y}_j \leq a_j\},$$

leading to model

$$P(Y = j|Y \geq j; x) = F(\alpha_j + x^t\delta), \tag{1.12}$$

with $\alpha_j = a_j - a$, and $\delta = -b$. Using the logistic cdf as residual distribution, we obtain the odds proportional logit model.

**Proportional hazard model.** Using the Gumbel min distribution for the residual $\varepsilon_i$ of latent model (1.11), we obtain the proportional hazard model. In this case (1.12) becomes

$$P(Y = j|Y \geq j; x) = 1 - \exp\{-\exp(\alpha_j + x^t\delta)\}, \tag{1.13}$$

for $j = 1, \ldots, J-1$. This is equivalent to

$$\log\left[-\log\{P(Y > j|Y \geq j; x)\}\right] = \alpha_j + x^t\delta,$$

for $j = 1, \ldots, J-1$, with the complementary log-log link. This model is the categorical version of the continuous hazard model (Cox, 1972). It is also called the grouped Cox model.

**Equivalence between sequential and cumulative models.** Equivalence between the sequential form (1.13) and the cumulative form (1.10) of the grouped Cox model has been shown by Läärä and Matthews (1985), using the following parametrization

$$\begin{cases} \alpha'_j & = \log \left\{ \sum_{k=1}^{j} \exp(\alpha_j) \right\} \quad \text{for } j = 1, \ldots, J-1, \\ \delta' & = \delta. \end{cases}$$

Equivalence between the sequential and cumulative non-proportional models, with the exponential cdf $F(\eta) = 1 - \exp(-\eta)$, has been shown by Tutz (1991), using the following parametrization

$$\begin{cases} \alpha'_1 & = \alpha_1, \\ \delta'_1 & = \delta_1, \end{cases} \quad \text{and} \quad \begin{cases} \alpha'_j & = \alpha_j - \alpha_{j-1}, \\ \delta'_j & = \delta_j - \delta_{j-1}, \end{cases} \quad \text{for } j = 2, \ldots, J-1.$$

Caution should be exercised for this last model because it is defined only for positive values of the linear predictors $\eta_j$.

Sequential models can also be defined without the proportionality assumption. Any design matrix $Z$ may be used and there are no constraints on corresponding linear predictors. Finally, a sequential model is fully specified by:

- the design matrix $Z$,
- the cdf $F$.

#### 1.4.2.3 Fisher's scoring algorithm

Using the score equation of multinomial GLM (1.6), we see that only the Jacobian matrix $\partial\pi/\partial\eta$ must be computed. For sequential models (1.12), we have $\pi_j = F(\eta_j) \prod_{k=1}^{j-1}\{1 - F(\eta_k)\}$, and thus the general term of the Jacobian matrix is

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

for row $i$ and column $j$.

### 1.4.3 Adjacent models

#### 1.4.3.1 Adjacent logits

For $J$ categories with associated probabilities $\pi_1, \ldots, \pi_J$, the adjacent logits are defined by

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

for $j = 1, \ldots, J-1$. The adjacent proportional logit model is then defined by relating adjacent logits to proportional linear predictors

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

for $j = 1, \ldots, J-1$. The adjacent logit model is defined by relating adjacent logits to non-proportional linear predictors

$$\text{logit}\left\{P(Y = j | Y \in \{j, j+1\}; x)\right\} = \alpha_j + x^t \delta_j,$$

for $j = 1, \ldots, J-1$. Even if there is no latent variable motivation, it is possible to define an adjacent model for any cdf $F$

$$P(Y = j | Y \in \{j, j+1\}; x) = F(\eta_j),$$

with different possible parametrization of $\eta_j$. The Newton-Raphson algorithm is detailed for the adjacent logit models by Agresti (2010). In the same manner as for sequential models, the conditional form of adjacent models implies independence between all linear predictors $\eta_j$. No constraints are required on $\eta$ to obtain non-negative probabilities. Finally, an adjacent model is fully specified by

- the design matrix $Z$,
- the cdf $F$.

#### 1.4.3.2 Equivalence with baseline-category logit model

The connection between the adjacent logit model and the baseline-category logit model was described by Agresti (2010). If we assume that the distribution of $Y|X = x$ is defined by an adjacent logit model, then we have

$$\begin{aligned}
\log\left(\frac{\pi_j}{\pi_J}\right) &= \log\left(\frac{\pi_j}{\pi_{j+1}} \times \ldots \times \frac{\pi_{J-1}}{\pi_J}\right)\\
&= \sum_{k=j}^{J-1} \eta_k\\
&= \sum_{k=j}^{J-1} \alpha_k + x^t \left(\sum_{k=j}^{J-1} \delta_k\right)\\
&= \alpha'_j + x^t \delta'_j,
\end{aligned}$$

for $j = 1, \ldots, J-1$. Agresti (2010) noted also that if the adjacent logit has a proportional form, then we obtain a baseline-category logit model with the parametrization

$$\delta'_j = \sum_{k=j}^{J-1} \delta_k = \sum_{k=j}^{J-1} \delta = (J-j)\delta,$$

for $j = 1, \ldots, J-1$ .

### 1.4.4 Stereotype models

Instead of defining new logit ratios of probabilities, Anderson (1984) conserved the baseline-category logit structure but proposed a new predictor parametrization. Thus he used the model

$$\log \left( \frac{\pi_j}{\pi_J} \right) = \alpha_j + x^t \delta_j, \tag{1.14}$$

with additional constraints on $\delta_j$, for $j = 1, \ldots, J-1$. Using different parametrizations of predictors $\eta_j$, Anderson proposed a flexible logistic regression model useful for $d$-dimensional or ordered regression relationships. He defined a procedure to test the distinguishability of successive categories with respect to explanatory variables $x$.

#### 1.4.4.1 Dimensional regression relation

The multinomial logit model is defined only with the complete parametrization $\eta_j = \alpha_j + x^t \delta_j$. For a given explanatory vector $x$ of dimension $p$ in the most general form, this is the maximal possible parametrization with $(J-1)(1+p)$ parameters. We could define the multinomial logit model with the proportional parametrization $\eta_j = \alpha_j + x^t \delta$. Taking into account $x$, this is the minimal possible parametrization, with $J-1+p$ parameters. With this in mind, Anderson proposed a large set of parametrizations to cover the range between minimal and maximal parametrizations. Thus he used model (1.14) with a particular parametrization for the $J-1$ slopes $\delta_j$. First, Anderson proposed to relax slightly the proportional assumption $(\delta_1 = \ldots = \delta_{J-1} = \delta)$ using the parametrization

$$\delta_j = \phi_j \delta, \tag{1.15}$$

for $j = 1, \ldots, J-1$, where $\phi_j$ are scalars. To avoid identifiability problems, the number of parameters must be less than the number of parameters of the maximal model

$$(J-1)2 + p \leq (J-1)(1+p).$$

This is the case for $J > 2$ and $p > 1$. Otherwise, if $J = 2$ or $p = 1$, one scalar must be fixed ($\phi_1 = 1$ such as in Anderson (1984)). He referred to model (1.14) with parametrization (1.15) as the one-dimensional stereotype model. He then proposed the two-dimensional stereotype model using parametrization

$$\delta_j = \phi_j \delta + \psi_j \gamma,$$

for $j = 1, \ldots, J-1$. Finally, he defined the $d$-dimensional stereotype model using parametrization

$$\delta_j = \sum_{k=1}^{d} \phi_{k,j} \gamma_k,$$

for $j = 1, \ldots, J-1$, where $\phi_{k,j} \in \mathbb{R}$ and $\gamma_k \in \mathbb{R}^p$. The number of different slopes $\gamma_k$ is such that $d \leq J-1$ for identifiability issues. Therefore, the maximal number of identifiable scale parameters $\phi_{k,j}$ is $J-1-d$. Anderson proposed the maximal dimension $d = \min(J-1, p)$. Finally, using the flexibility of slope parametrization in two ways (scale parameters $\phi_{k,j}$ and slope parameters $\gamma_k$), Anderson proposed a multinomial logit model with any number of parameters between the minimal and the maximal number of parameters (i.e. between $(J-1+p)$ and $(J-1)(1+p)$).

#### 1.4.4.2 Ordered regression relation

Anderson proposed to order the parameters $(\phi_j)_{j=1,\ldots,J}$ of the one-dimensional stereotype model to obtain an ordered regression relation

$$1 = \phi_1 > \phi_2 > \ldots > \phi_J = 0.$$

It should be noted that Anderson did not use this ordering as an *a priori* constraint. If the estimated parameters $\hat{\phi}_j$ are ordered, then the order is considered as a consequence of the estimation. As noticed by Tutz (2012), the stereotype model cannot be considered to be an ordinal model, even though it is used for ordinal response variables in applications.

#### 1.4.4.3 Indistinguishability procedure

Anderson (1984) proposed a testing procedure - useful for ordinal data - to identify successive categories that can be clearly distinguished by the explanatory variables $x$. These categories are said to be indistinguishable with respect to $x$ when the explanatory variables $x$ do not have significantly different effects on them. He proposed to aggregate the corresponding successive slope parameters $\delta_j$ and use a deviance test. More precisely, he proposed an iterative procedure to find the best splitting point among categories $1, \ldots, J$ with respect to $x$. The minimal number of splitting point is zero, corresponding to the simple model without explanatory variables (null hypothesis $H_0$), and the maximal number of splitting points is $J-1$, corresponding to the classical multinomial logit model with $J-1$ different slopes.

The first step is to find the best split into two groups of categories. The hypothesis $H_{(2;r)}$ is then introduced

$$H_{(2;r)} : \delta_1 = \ldots = \delta_r; \quad \delta_{r+1} = \ldots = \delta_J = 0,$$

for $r = 1, \ldots, J-1$. Comparing the corresponding log-likelihood values $l_{(2;r)}$ yields the best splitting point $r^*$ such that $l_2 = l_{(2;r^*)} = \max_r l_{(2;r)}$. The hypothesis $H_{(2;r^*)}$ is then tested against $H_0$, using the deviance statistic $2(l_2 - l_0)$ which follows a $\chi^2_p$ distribution under $H_0$. Finally, if the splitting point $r^*$ is accepted, the procedure must be restarted in parallel for the two groups $\{1, \ldots, r^*\}$ and $\{r^*+1, \ldots, J\}$ in order to obtain the best partition into three groups.

This is a dichotomous partitioning procedure with at most $J(J-1)/2$ different parametrizations to test. It should be noted that this procedure is simplified for the one-dimensional stereotype model since the equality between slopes $\delta_1 = \ldots = \delta_r$ becomes equality between scalar parameters $\phi_1 = \ldots = \phi_r$. In practice, only this particular case of the procedure is used.

## 1.5 Partitioned conditional generalized linear models

The main idea is to recursively partition the $J$ categories then specify a conditional GLM at each step. This is why this type of models is called partitioned conditional GLMs. Such models have already been proposed, e.g. the nested logit model (McFadden et al., 1978), the two-step model (Tutz, 1989) and the partitioned conditional model for partially-ordered sets (POS-PCM) (Zhang and Ip, 2012). In a first level, the entire set of categories is partitioned into $L$ groups (or "nests") as follows

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

and in a second level, the groups are partitioned into categories. Thus, a basic event is decomposed as follows

$$\{Y = j\} \Leftrightarrow \{Y \in N_l\} \cap \{Y = j | Y \in N_l\}.$$

**Partitioning step** The groups $N_l$ are considered as categories and a categorical regression model is used to describe the probabilities $P(Y \in N_l | x)$ for $l = 1, \dots, L$.

**Conditioning step** For each group $N_l$, the response variable $Y$ is conditioned on $N_l$ and a GLM is used to describe the conditional probabilities $P(Y = j | Y \in N_l; x)$ for $j \in N_l$.

More generally, this partitioning/conditioning process may be reiterated for $N_l$. However, these models have been formally defined and used with only two or three levels. The first aim of these models is to aggregate the categories that are influenced by the same explanatory variable.This is very similar to the principle of distinguishability defined by Anderson. The nested logit model were introduced in this way. The categories are aggregated according to the explanatory variable. On the other hand, the two-step model takes account of the order assumption among the categories, aggregating only successive categories. A more recent aim was to use the partial ordering among the categories to define an adapted partitioned conditional model. For any partially-ordered structure, Zhang and Ip (2012) proposed a particular recursive partition which “conserves” the order relation.

### 1.5.1 Nested logit model

The most well known partitioned conditional model for nominal data is the nested logit model defined by McFadden et al. (1978) in the framework of individual choice behaviour. This model was introduced to avoid the inconsistency of the independence of irrelevant alternatives (IIA) property in some situations. Let us illustrate this inconsistency with the classical example of blue and red buses (Debreu, 1960). Assume we are interested in urban travel demand, with the simple situation of two alternatives: $A = \{\text{blue bus, car}\}$. Suppose that the consumer has no preference between the two alternatives; this means that $P_A(\text{blue bus}) = P_A(\text{car}) = 1/2$. Suppose now that the travel company adds some red buses and the consumer again has no preference between blue and red buses; this means that $P_B(\text{blue bus}) = P_B(\text{red bus})$ where $B = \{\text{blue bus, red bus, car}\}$. Using the IIA property we obtain

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

Finally, we obtain $P_B(\text{blue bus}) = P_B(\text{red bus}) = P_B(\text{car}) = 1/3$, whereas we expected the probabilities $P_B(\text{blue bus}) = P_B(\text{red bus}) = 1/4$ and $P_B(\text{car}) = 1/2$.

In this example the IIA property is not appropriate because two alternatives are very similar and also share many characteristics. The nested logit model captures the similarities between close alternatives by partitioning the choice set into “nests” (groups). Thus, the consumer chooses first between bus and car according to price, travel time, ... and secondly between the two buses according to preferred color. More generally, suppose that alternatives can 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. In the following, the nested logit model is presented with only two levels. Let $L$ be the number of nests obtained by partitioning the set of $J$ alternatives.

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

If $j$ denotes an alternative belonging to the nest $N_l$, then the probability of alternative $j$ is decomposed as follows

$$P(Y=j|x)=P(Y=j|Y\in N_l;x^l)P(Y\in N_l|x^0,IV), \tag{1.16}$$

where $IV=(IV_1,\dots,IV_L)$ denotes the vector of *inclusive values* described thereafter, $x^0$ are the attributes which influence only the first choice level between nests and $x=(x^0,x^1,\dots,x^L)$. Each probability of the product (3.3) is determined by a multinomial logit model as follows

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

and

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

where

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

The deterministic utilities (predictors) $\eta_j^l$ are function of attributes $x^l$ and $\eta_l^0$ are function of attributes $x^0$. In practice they are linear with respect to $x$. In some situations the attribute values depend on the alternative. For example, the travel price $x_j$ depends on the $J$ alternatives bus, car, metro, etc. In this case, the conditional logit model was introduced by McFadden (1974), using the linear predictors $\eta_j=\alpha_j+x_j^t\delta$ for $j=1,\dots,J$.

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 $\beta^l$ are different in each nest. The inclusive values $IV_l$ of each nest can then be computed and used, in a second step, to estimate the first level model.

### 1.5.2 Two-step model

The two-step model, or compound model, was defined by Tutz (1989) to decompose the latent mechanism of an ordinal response into two levels. Ordinal-scale response variables are commonly used in medicine and psychology for instance to assess a patient's condition. This ordinal scale is often built from a coarse and a fine scales. For the back pain prognosis dataset described by Doran and Newell (1975), the response variable $y$ is the assessment of back pain after three weeks of treatment using the six ordered categories: worse (1), same (2), slight improvement (3), moderate improvement (4), marked improvement (5), complete relief (6). Categories 3, 4 and 5 can be aggregated into a general category *improvement*. Thus, the coarse scale corresponds to the categories: worse, same, improvement, complete relief, and the fine scale corresponds to the categories: slight improvement, moderate improvement and marked improvement (see figure 3.4).

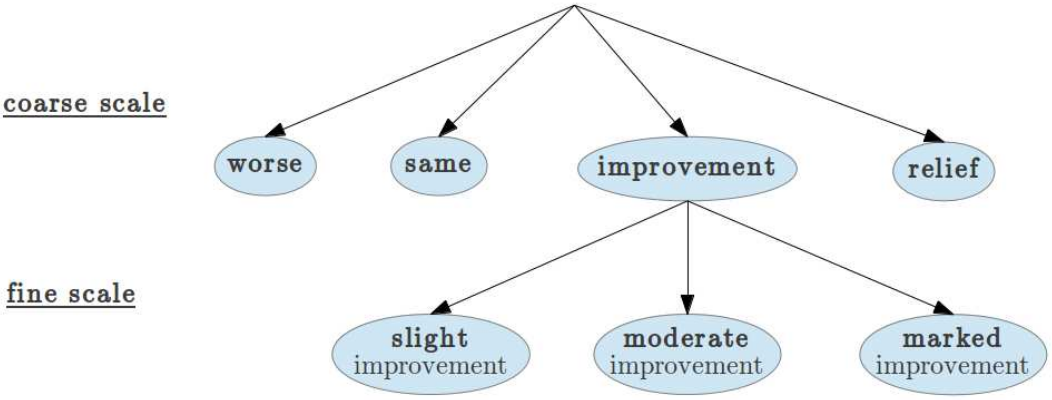

Figure 1.5: Two-scale back pain assessment.

Then, the model must be decomposed into two levels, corresponding to the coarse scale and the fine scale. More generally, let $N_1, \dots, N_L$ be a partition of the $J$ categories, which conserves the order relationship. The sets also have the form $N_l = \{m_{l-1}+1, \dots, m_l\}$ with $0 = m_0 < m_1 < \dots < m_L = J$. The probability of each category $j \in N_l$ is decomposed as follows

$$P(Y = j) = P(Y \in N_l)P(y = j|Y \in N_l).$$

An advantage of the model is that different parameters are involved in different steps:

$$P(Y = j|x, \beta) = P(Y \in N_l|x, \beta_1)P(y = j|Y \in N_l; x, \beta_2).$$

The ordinal scale of the response variable is used and must be assumed. Therefore, Tutz (1989) defined the cumulative compound model and the sequential compound model, appropriate for ordinal response variable.

### 1.5.2.1 Cumulative compound model

Here, we have two cumulative latent mechanisms for the two levels

$$\begin{array}{lll} \underline{\text{coarse scale}}: & P(Y \in \bigcup_{k=1}^{l} N_k|x) = F(\alpha_l + x^t\delta) & \text{for } l = 1, \dots, L \\ \underline{\text{fine scale}}: & P(Y \leq j|Y \in N_l; x) = F(\alpha_j^l + x^t\delta^l) & \text{for } j \in N_l \text{ and } l = 1, \dots, L \end{array}$$

### 1.5.2.2 Sequential compound model

Here, we have two sequential latent mechanisms for the two levels

$$\begin{array}{lll} \underline{\text{coarse scale}}: & P(Y \in N_l|Y \in \bigcup_{k=l}^{L} N_k; x) = F(\alpha_l + x^t\delta) & \text{for } l = 1, \dots, L \\ \underline{\text{fine scale}}: & P(Y = j|Y \geq j; Y \in N_l; x) = F(\alpha_j^l + x^t\delta^l) & \text{for } j \in N_l \text{ and } l = 1, \dots, L \end{array}$$

An advantage of the two-step model is its simple structure. After rearranging the data, the parameters of each step may be estimated in parallel using classic algorithm for generalized linear models. Thus, no two-step estimation procedure is necessary unlike, for example, for the nested logit model.

### 1.5.3 Partitioned conditional model for partially-ordered set

In categorical data analysis, the case of nominal and ordinal data has been investigated in depth while the case of partially-ordered data has been comparatively neglected. Zhang and Ip (2012) introduced the partitioned conditional model for partially-ordered set (POS-PCM). The main idea was to recursively partition the $J$ categories in order to lead the partial order back to degenerate cases: total order and no order. Thus, the odds proportional logit model was used for the total order case and the multinomial logit model was used for the no order case.

Zhang and Ip introduced the partially-ordered set theory into the GLM framework. A partial ordered set (poset) $(P, \preceq)$ is summarized by a Hasse diagram. The order relation $j \preceq k$ is represented by an edge between the two nodes (categories) and 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. Zhang and Ip described an algorithm of category partitioning which can be used to show the following proposition.

**Proposition 1.** *(Zhang and Ip, 2012) A finite poset (with one component) can always be partitioned into antichains that are totally weakly ordered.*

Note that the weakly order must be introduced because, in this approach, sets of categories are compared. Two subsets $N_1$ and $N_2$ in $P$ are weakly ordered if at least one element in $N_2$ is dominated by elements in $N_1$ and no element in $N_2$ dominates any element in $N_1$. One says that $N_1$ weakly dominates $N_2$.

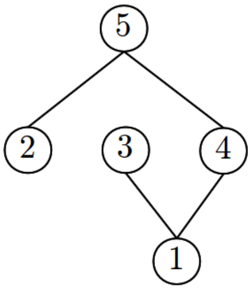

Figure 1.6: Hasse diagram among five categories.

Let $(P, \preceq)$ be the poset summarized by the Hasse diagram in figure 1.6. 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 $P(Y \in \bigcup_{k=1}^{l} N_k | x)$ for $l = 1, 2, 3$. Within each antichain $N_l$, the elements are not comparable, thus the multinomial logit model is used to describe the conditional probabilities $P(Y = j | Y \in N_l ; x)$ for $j \in N_l$. This POS-PCM is then defined with two levels.

## 1.6 Semi-Markov switching generalized linear models for categorical data

The branching pattern of a shoot may be influenced by many factors that vary along the shoot, e.g. internode length, leaf surface or local curvature. We introduce a generalization of hidden semi-Markov chains for categorical response variables that incorporates explanatory variables that vary with the index parameter. Using this model, we demonstrate in chapter 4 the influence of shoot growth pattern on its immediate branching.

We consider here a semi-Markov switching generalized linear model (SMS-GLM), which is a two-scale segmentation model. In this framework, the succession and length of branching zones (coarse scale) are represented by a non-observable semi-Markov chain while the axillary productions within a branching zone, modulated by factors that vary along the shoot (fine scale), are represented by generalized linear models attached to each state of the semi-Markov chain. A SMS-GLM combines three categories of variables: (i) "state" variable representing the non-directly observable branching zones, (ii) plant response categorical variable (types of axillary production), (iii) explanatory variables that vary with node rank (e.g., internode length). In this section, we first review the hidden semi-Markov chains definition and estimation for the general case of any observation process. The estimation procedure is then described in the case of a GLM as observation model.

### 1.6.1 Hidden semi-Markov chains

Let us first introduce semi-Markov chains by starting with simple Markov chains. The sequence of state variables $S_0 = s_0, \dots, S_t = s_t$ is denoted by $S_0^{t-1} = s_0^{t-1}$ and the entire sequence $S_0 = s_0, \dots, S_{T-1} = s_{T-1}$ by $\boldsymbol{S} = \boldsymbol{s}$. A first-order Markov chain is a discrete-time discrete-state-space stochastic process characterized by the dependency relation

$$P(S_t = s_t | S_1^{t-1} = s_1^{t-1}) = P(S_t = s_t | S_{t-1} = s_{t-1}).$$

**Definition 1.** *Let $\{S_t\}$ be a first-order time-homogeneous* ***Markov chain*** *with finite state space $\{0, \dots, A-1\}$ (simply referred to as Markov chain in the following). A Markov chain is defined by the following parameters:*

- *initial probabilities $\varphi_a = P(S_0 = a)$ with $\sum_a \varphi_a = 1$,*
- *transition probabilities $p_{a,b} = P(S_t = b | S_{t-1} = a)$ with $\sum_b p_{a,b} = 1$.*

A major drawback of the Markov chain is the inflexibility in describing the time spent in given state, which is geometrically distributed. The implicit occupancy (or sojourn time) distribution of a nonabsorbing state j is the 1-shifted geometric distribution with parameter $1 - p_{a,a}$

$$d_a(u) = (1 - p_{a,a}) p_{a,a}^{u-1} \tag{1.17}$$

This is the unique discrete memoryless distribution. A useful generalization of Markov chains lies in the class of semi-Markov chains in which the process moves out of a given state according to an embedded Markov chain with self-transition probability in non-absorbing states $p_{a,a} = 0$ and where the time spent in a given non-absorbing state is modelled by an explicit occupancy distribution. The possible parametric state occupancy distributions are binomial, negative binomial and Poisson distributions with an additional shift parameter $d \geq 1$ which defines the minimum sojourn time in a given state.

**Definition 2.** *(Guédon, 2003) A* ***semi-Markov chain*** *is constructed from an embedded first-order Markov chain. This A-state first-order Markov chain is defined by the following parameters:*

- *initial probabilities $\varphi_a = P(S_0 = a)$ with $\sum_a \varphi_a = 1$.*

- *transition probabilities*
  - *non absorbing state a:* $\forall b \neq a$, $\tilde{p}_{a,b} = P(S_t = b | S_t \neq a, S_{t-1} = a)$ *with* $\sum_{b \neq a} \tilde{p}_{a,b} = 1$ *and* $\tilde{p}_{a,a} = 0$,
  - *absorbing state a:* $p_{a,a} = P(S_t = a | S_{t-1} = a) = 1$ *and* $\forall b \neq a$, $p_{a,b} = 0$.

*This embedded first-order Markov chain represents transitions between distinct states except in the absorbing state case. An occupancy (or sojourn time) distribution is attached to each non absorbing state a of the embedded first-order Markov chain*

$$d_a(u) = P(S_{t+u+1} \neq a, S_{t+u-v} = a, v = 0, \ldots, u-2 | S_{t+1} = a, S_t \neq a)$$

*for* $u = 1, \ldots, U_a$*, where* $U_a$ *denotes the upper bound to the time spent in state a. Hence, we assume that the state occupancy distributions are concentrated on finite sets of time points. For the particular case of the last visited state, we need to introduce the survivor function of the sojourn time in state a,* $D_a(u) = \sum_{v \geq u} d_a(u)$*. The whole (first-order Markov chain + state occupancy distributions) constitutes a semi-Markov chain.*

Hidden Markov chains emerged in the 1970s in engineering and have since become a major tool for both pattern recognition applications, e.g. speech or handwriting recognition (see (Rabiner, 1989) for tutorial introductions), and biological sequence analysis (Churchill, 1989).

Hidden semi-Markov chains with non parametric state occupancy distributions were first proposed in the field of speech recognition by Ferguson (1980). After this pioneering work, the statistical inference problem related to hidden semi-Markov chains was further investigated by different authors (Russell and Moore, 1985; Levinson, 1986; Guédon, 1992) and different parametric hypotheses were put forward for the state occupancy distributions (Poisson, "discrete" gamma).

**Definition 3.** *A* ***hidden (semi-)Markov chain*** *can be viewed as a pair of stochastic processes* $\{S_t, Y_t\}$ *where the output process Yt is related to the state process* $\{S_t\}$*, which is a (semi-)Markov chain, by a probabilistic function or mapping denoted by f (hence* $Y_t = f(S_t)$*). Since the mapping f is such that a given output may be observed in different states, the state process* $\{S_t\}$ *is not observable directly but only indirectly through the output process* $\{Y_t\}$*. This output process* $\{Y_t\}$ *is related to the (semi-)Markov chain* $\{S_t\}$ *by the observation (or emission) probabilities. The definition of observation probabilities expresses the assumption that the output process at time t depends only on the underlying Markov chain at time t. For each state a the observation distribution is denoted by* $b_a$

$$b_a(y) = P(Y_t = y | S_t = a) \;\; \textit{with} \; \sum_y b_a(y) = 1.$$

It should be noted that an HMC can be interpreted as a finite mixture model with Markovian dependency. In the same way, an HSMC can be interpreted as a finite mixture model with semi-Markovian dependency.

### Conditional independence and likelihood of HSMC

For an HMC, both relations of conditional independence hold

- $S_t$ is independent of $\{S_0, Y_0, \ldots, S_{t-2}, Y_{t-2}, Y_{t-1}\}$ given $S_{t-1}$, for all $t > 0$.
- $Y_t$ is independent of $\{S_0, Y_0, \ldots, S_{t-1}, Y_{t-1}\}$ given $S_t$, for all $t \geq 0$.

Thus, for an HMC, the likelihood of the joint process $\{\boldsymbol{S}, \boldsymbol{Y}\}$ is

$$\begin{aligned} P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{S} = \boldsymbol{s}; \theta) &= P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{S} = \boldsymbol{s}) P(\boldsymbol{S} = \boldsymbol{s}) \\ &= \varphi_{s_0} \left\{ \prod_{t=1}^{T-1} p_{s_{t-1}, s_t} \right\} \left\{ \prod_{t=0}^{T-1} b_{s_t}(y_t) \right\}, \end{aligned} \tag{1.18}$$

where $\theta$ denotes the set of all HMC parameters (i.e. initial probabilities $\varphi_a$, transition probabilities $p_{a,b}$ and observed distributions $b_a$ for all $a \in \{0, \ldots, A-1\}$).

By replacing a first-order Markov chain by a semi-Markov chain, the Markovian property is transferred to the level of the embedded first-order Markov chain. In the semi-Markov chain case, the conditional independence between the past and the future is ensured only when the process moves from one state to another distinct state. This property holds at each time step in the case of a Markov chain. For an HSMC, the likelihood of the joint process $\{\boldsymbol{S}, \boldsymbol{Y}\}$ is

$$\varphi_{s_0} d_{s_0}(u_0) \left\{ \prod_{r=1}^{R-1} p_{s_{r-1}, s_r} d_{s_r}(u_r) \right\} p_{s_{R-1}, s_R} D_{s_R}(u_R) I \left( \sum_{r=0}^{R} u_r = T \right) \left\{ \prod_{t=0}^{T-1} b_{s_t}(y_t) \right\}, \tag{1.19}$$

where $s_r$ is the $(r+1)$th visited state, $u_r$ is the time spent in state $u_r$ and $I()$ denotes the indicator function; see for instance Russell and Moore (1985); Rabiner (1989).

### Maximum likelihood estimation

In the following, the algorithm computation is described for only one observed sequence $\boldsymbol{y}$. For an HMC, since the state sequence $\boldsymbol{s}$ is not observable, the likelihood is given by

$$L(\boldsymbol{y}; \theta) = \sum_{\boldsymbol{s}} P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{S} = \boldsymbol{s}; \theta),$$

where $\sum_{\boldsymbol{s}}$ means the sum on every possible state sequence of lenth $T$. This sum has exactly $A^T$ elements and can be written as a matrices product (Altman, 2007). Therefore the likelihood of an HMC can be directly maximized (Collings and Rydén, 1998; Turner, 2008).

The direct maximisation of the likelihood is more complex for an HSMC. In fact the likelihood cannot be written as a matrix product

$$L(\boldsymbol{y}; \theta) = \sum_{\boldsymbol{s}} \sum_{u} P(\boldsymbol{Y} = \boldsymbol{y}, S_0^{T-1+u} = s_0^{T-1+u}; \theta),$$

where $\sum_u$ means the sum on every supplementary duration from time $T$ spent in the state occupied at time $T-1$. Since the state sequence $\boldsymbol{s}$ is not observable, a standard solution for maximum likelihood estimation is the EM algorithm.

**EM algorithm for HSMC** The EM algorithm was introduced by Dempster et al. (1977) to find maximum likelihood estimates of parameters when the model depends on unobserved latent variables. Each iteration of the EM algorithm decomposes into two steps: the E step (Expectation) and the M step (Maximisation). In the following we describe these two steps at the iteration $k+1$ for an HSMC.

**E step** The focus here is on the conditional expectation of the completed data log-likelihood given the observed data. Using equality (1.19), the conditional expectation can be written as a sum of four terms that depend on initial probabilities, transition probabilities, sojourn time distributions and observed distributions

$$\begin{aligned} Q(\theta|\theta^{[k]}) &= E\left\{\log P(y_1^T, s_1^T; \theta)\boldsymbol{Y} = \boldsymbol{y}; \theta^{[k]}\right\} \\ &= Q_\varphi(\{\varphi_a\}_{a=0}^{A-1}|\theta^{[k]}) + \sum_{a=0}^{A-1} Q_{\tilde{p}}(\{\tilde{p}\}_{a=0}^{A-1}|\theta^{[k]}) \\ &+ \sum_{a=0}^{A-1} Q_d(\{d_a(u)\}|\theta^{[k]}) I(p_{a,a} = 0) + \sum_{a=0}^{A-1} Q_b(\{b_a(y)\}_{y=0}^{Y-1}|\theta^{[k]}) \end{aligned} \tag{1.20}$$

Guédon (2003) introduced a “forward-backward” algorithm for the E step, based on the following decomposition

$$\begin{aligned} L1_a(t) &= P(S_{t+1} \neq a, S_t = a|\boldsymbol{Y} = \boldsymbol{y}) \\ &= \frac{P(Y_{t+1}^T = y_{t+1}^T|S_{t+1} \neq a, S_t = a)}{P(Y_{t+1}^T = y_{t+1}^T|Y_1^t = y_1^t)} P(S_{t+1} \neq a, S_t = a|Y_1^t = y_1^t) \\ L1_a(t) &= B_a(t)F_a(t) \end{aligned}$$

which expresses the conditional independence between the past and the future of the process at state change times (1.19). Guedon and Cocozza-Thivent (1990) showed that the quantities $F_a(t)^{[k+1]}$ (respectively $L1_a(t)^{[k+1]}$) can be computed by a forward (respectively backward) pass through observed sequence $\boldsymbol{y}$, using the previous values of parameters $\theta^{[k]}$.

**M step** The reestimation formulas for the parameters of a HSMC are obtained by maximizing the different terms of $Q(\theta|\theta^{[k]})$ (see the decomposition (1.20)), each term depending on a given subset of $\theta$. For each parameter subset, the reestimation formula is given in Guédon (2003). This formula is directly deduced from the maximization of the first three terms of (1.20). Therefore we obtain the next parameter values $\varphi^{[k+1]}$ and $\tilde{P}^{[k+1]}$ using the quantities previously computed $F_a(t)^{[k+1]}$ and $L1_a(t)^{[k+1]}$. The computation of quantities $d^{[a+1]}$, corresponding to the sojourn time distribution in each state, is described by Guédon (2003). The update of observation model parameters (corresponding to quantities $b^{[k+1]}$ in the case of simple distribution) is detailed in the next subsection in the case of GLM for categorical data.

### 1.6.2 Semi-Markov switching generalized linear models

Here, the observation conditional $\{Y_t|X_t = x_t\}$ is defined by several GLMs for categorical response variables. In fact, one GLM is associated with each state $a \in \{0, \dots, A-1\}$ of the HSMC, corresponding to the observed distribution $b_a$. These $A$ distributions are characterized by the parameters $\beta_a$ of the corresponding GLM. As we have seen previously, the estimation of these parameters is not related to the semi-Markov parameters estimation and therefore only the M step differs.

In the M step, the term of $Q(\theta|\theta^{[k]})$ corresponding to an observation GLM is maximised using Fisher's scoring algorithm. Therefore there are two levels of iterations: Fisher's scoring iteration $m$ being nested in EM iteration $k$. For the M step, we must maximise the conditional expectation $Q_b(\{b_a(y)\}_{y=0}^{Y-1}|\theta^{[k]})$ with respect to $\beta_a$ for each state $a \in \{0, \dots, A-1\}$. On the iteration $m+1$ we obtain for state $a$

$$\beta_a^{[m+1]} = \beta_a^{[m]} - \left[ E \left\{ \frac{\partial^2 Q(b|\ \theta^{[k]})}{\partial \beta_a \partial \beta_a^t} \right\}_{\beta_a = \beta_a^{[m]}} \right]^{-1} \left\{ \frac{\partial Q(b|\theta^{[k]})}{\partial \beta_a} \right\}_{\beta_a = \beta_a^{[m]}}$$

For the EM algorithm, the conditional expectation $Q_b(\{b_a(y)\}_{y=0}^{Y-1}|\theta^{[k]})$ has the following form

$$Q_b(\{b_a(y)\}_{y=0}^{Y-1}|\theta^{[k]}) = \sum_{t=0}^{T-1} L1_a^{[k]}(t)\ b_a(y_t|x_t),$$

and thus we obtain

$$\left\{ \frac{\partial Q(b|\theta^{[k]})}{\partial \beta_a} \right\}_{\beta_a = \beta_a^{[m]}} = \sum_{t=0}^{T-1} L1_a^{[k]}(t) \left\{ \frac{\partial l(y_t, x_t)}{\partial \beta_a} \right\}_{\beta_a = \beta_a^{[m]}},$$

where Fisher's score is computed as detailed by (1.6).

Chapter 2

# A new specification of generalized linear models for categorical data

## Abstract

Many regression models for categorical data have been introduced in a variety of applied fields, motivated by different paradigms but their specification are not homogeneous. Therefore, these models are difficult to compare and their appropriateness with respect to category ordering assumptions is questionable. The first contribution of this chapter is to unify the classical regression models for categorical response variables, whether nominal or ordinal. This unification is based mainly on a decomposition of the link function into two parts: an inverse continuous cdf and a ratio of probabilities. This allows us to define a new family of models for nominal data, comparable to the cumulative, sequential and adjacent families of models used for ordinal data. We finally propose a classification of GLMs for categorical data along a nominal/ordinal scale.

**Keywords:** link function, nominal variable, ordinal variable, model reparametrization, invariance under permutation, stability under permutation.

## Contents

## 2.1 Introduction

Since GLMs were first introduced by Nelder and Wedderburn (1972), many regression models for categorical data have been developed and introduced into a variety of applied fields such as medicine, econometrics and psychology. They have been motivated by different paradigms and their specifications are not homogeneous. Most have been defined for ordinal data (Agresti, 2010), whereas only one has been defined for nominal data: the multinomial logit model introduced by Luce (1959) (also referred to as the baseline-category logit model (Agresti, 2002)). The three classical models for ordinal data are the odds proportional logit model (McCullagh, 1980), the sequential logit model (Tutz, 1990) (also referred to as the continuation ratio logit model (Dobson, 2002)), and the adjacent logit model (Masters, 1982; Agresti, 2010). They have been extended by either replacing the logistic cumulative distribution function (cdf) by other cdfs $F$ (e.g. normal or Gumbel cdfs; see the grouped Cox model for instance, also referred to as the proportional hazard model), or introducing different parametrizations of the linear predictor (i.e. changing the design matrix $Z$). No such developments have been undertaken for the multinomial logit case, and one of our goals is to fill this gap.

It should be noted that the three previously mentioned models for ordinal data, and the multinomial logit model, all rely on the logistic cdf. The difference between them stems from another element in the link function. In fact, the four corresponding link functions can be decomposed into the logistic cdf and a ratio of probabilities $r$. For the odds proportional logit model, the ratio corresponds to the cumulative probabilities $P(Y \leq j)$. Here, we propose to decompose the link function of any GLM for categorical data into a cdf $F$ and a ratio $r$. Using this decomposition it can be shown that all the models for ordinal data were defined by fixing the ratio $r$ and changing the cdf $F$ and the design matrix $Z$. For example, all the cumulative models were obtained by fixing the cumulative probabilities $P(Y \leq j)$ as the common structure. In the same way, the two families of sequential and adjacent models were defined with probability ratios $P(Y = j | Y \geq j)$ and $P(Y = j | j \leq Y \leq j + 1)$. Now, we can extend the multinomial logit model by fixing only its ratio $r$ and leaving $F$ and $Z$ unrestricted.

Our first contribution in this chapter (section 2.3) is to unify all these models by introducing the new $(r, F, Z)$-triplet specification. Differences and commonalities between models are thus highlighted, making them easily comparable. In this framework, the multinomial logit model is extended, replacing the logistic cdf by other cdfs. We thus obtain a new family of models for nominal data, comparable to the three classical families of models used for ordinal data. We can now compare all the models according to the three components: ratio $r$ for structure, cdf $F$ for fit, and design matrix $Z$ for parametrization.

Sections 2.4 and 2.5 investigate this comparison in depth by studying the equivalences between models. We first revisit three equivalences between models with different ratios, shown by Läärä and Matthews (1985), Tutz (1991) and Agresti (2010) then generalize two of them to obtain equalities between families of models. Some properties of invariance and stability under permutation of the response categories are then presented. "Models for nominal categories should be invariant under arbitrary permutations [. . . ]. On the other hand, models for ordinal data should be invariant only under the special reverse permutation" (McCullagh, 1978).

In section 2.6, using the extended family of models for nominal data, and their invariance property, we introduce a family of supervised classifiers. In final section 2.7 we discuss the appropriateness of certain models with respect to category ordering assumptions. A classification of the different models along a nominal/ordinal scale according to their invariance properties is then proposed. By also considering difficulties of interpretability and inference, we provide some advice concerning the practical choice of a model structure.

## 2.2 Exponential form of the categorical distribution

### Natural exponential family

The exponential family of distributions must first be introduced into the multivariate case. Consider a random vector $Y$ of $\mathbb{R}^K$ whose distribution depends on a parameter $\theta \in \mathbb{R}^K$. The distribution belongs to the exponential family if the density can be written as

$$f(y;\theta,\lambda) = \exp\left\{\frac{y^t\theta - b(\theta)}{\lambda} + c(y,\lambda)\right\},$$

where $b$, $c$ are known functions, $\lambda$ is the *nuisance parameter* and $\theta$ is the *natural parameter*. In this context, we recall the well-known property:

**Property 5.** *Let $Y$ be a random vector whose distribution belongs to the exponential family. The function $b$ is assumed to be twice differentiable and we obtain*

$$\begin{array}{llcl} (i) & E(Y) & = & \dfrac{\partial b}{\partial \theta}, \\ (ii) & Cov(Y) & = & \lambda \dfrac{\partial^2 b}{\partial \theta^t \partial \theta}. \end{array}$$

### Truncated multinomial distribution

Let $J \geq 2$ denote the number of categories for the variable of interest and $n \geq 1$ the number of trials. Let $\pi_1, \ldots, \pi_J$ denote the probabilities of each category, such that $\sum_{j=1}^{J} \pi_j = 1$. Let the discrete vector $\tilde{Y}$ follow the multinomial distribution

$$\tilde{Y} \sim \mathcal{M}(n, (\pi_1, \ldots, \pi_J)),$$

with $\sum_{j=1}^{J} \tilde{y}_j = n$. It should be remarked that the multinomial distribution is not exactly a generalization of the binomial distribution, just looking at the dimension of $\tilde{Y}$. In fact, the constraint $\sum_{j=1}^{J} \pi_j = 1$ expresses one of the probabilities in terms of the others. By convention we choose to put the last category aside: $\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j$. Finally, we must define the truncated multinomial distribution

$$Y \sim \mathcal{TM}(n, (\pi_1, \ldots, \pi_{J-1})),$$

where $Y$ is the truncated vector of dimension $J-1$ with the constraint $0 \leq \sum_{j=1}^{J-1} y_j \leq n$. The probabilities $\pi_j$ are strictly positive and $\sum_{j=1}^{J-1} \pi_j < 1$ to avoid degenerate cases. Let $\pi$ denote the truncated vector $(\pi_1, \ldots, \pi_{J-1})^t$ with $E(Y) = n\pi$. For $J = 2$ the truncated multinomial distribution is the Bernoulli distribution if $n = 1$ and the binomial distribution if $n > 1$ (see table 2.1). In the GLM framework, only $\pi$ is related to the explanatory variables thus we focus on the case $n = 1$. One observation $y$ is an indicator vector of the observed category (the null vector corresponding to the last category). The truncated multinomial distribution can be written as follows

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

| | $J=2$ | $J>2$ |
|---|---|---|
| $n=1$ | $\mathcal{B}er(\pi)$ | $\mathcal{TM}(\pi)$ |
| $n>1$ | $\mathcal{B}(n,\pi)$ | $\mathcal{TM}(n,\pi)$ |

Table 2.1: Truncated multinomial distribution according to $J$ and $n$ values.

The natural parameter $\theta = (\theta_1, \dots, \theta_{J-1})^t$ is defined by

$$\theta = \left( \ln \left\{ \frac{\pi_1}{1 - \sum_{j=1}^{J-1} \pi_j} \right\}, \dots, \ln \left\{ \frac{\pi_{J-1}}{1 - \sum_{j=1}^{J-1} \pi_j} \right\} \right)^t,$$

and

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

Then the distribution is

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

Using the nuisance parameter $\lambda = 1$ and the null function $c(y, \lambda) = 0$, we see that the truncated multinomial distribution $\mathcal{TM}(\pi)$ belongs to the exponential family of dimension $K = \dim(Y) = \dim(\theta) = \dim(\pi) = J - 1$.

## 2.3 Specification of generalized linear models for categorical data

Consider the situation of regression analysis, with the multivariate response variable $Y$ and the vector of $Q$ explanatory variables $X = (X_1, \dots, X_Q)$ in a general form (i.e. categorical variables being represented by indicator vectors). The dimension of $X$ is thus denoted by $p$ with $p \geq Q$. We are interested in the conditional distribution of $Y|X$, observing the values $(y_i, x_i)_{i=1,\dots,n}$ taken by the pair $(Y, X)$. All the response variables $Y_i$ are assumed to be conditionally independent of each other, given $\{X_i = x_i\}$. The variables $Y_i$ follow the conditional truncated multinomial distribution

$$Y_i|X_i = x_i \sim \mathcal{TM}(\pi(x_i)),$$

with at least $J = 2$. In the following we will misuse some notations for convenience. For example, we will often forget the conditioning on $X$ and the individual subscript $i$. Moreover, the response variable will sometimes be considered as a univariate categorical variable $Y \in \{1, \dots, J\}$ in order to use the univariate notation $\{Y = j\}$ instead of the multivariate notation $\{Y_1 = 0,\ \dots, Y_{j-1} = 0,\ Y_j = 1,\ Y_{j+1} = 0,\ \dots,\ Y_{J-1} = 0\}$.

### 2.3.1 Decomposition of the link function

The definition of a GLM for categorical data includes the specification of a link function $g$ which is a diffeomorphism from $\mathcal{M} = \{\pi \in \, ]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to an open subset $\mathcal{S}$ of $\mathbb{R}^{J-1}$, between the expectation $\pi = E[Y|X{=}x]$ and the linear predictor $\eta = (\eta_1, ..., \eta_{J-1})^t$. It also includes the parametrization of the linear predictor $\eta$ which can be written as the product of the design matrix $Z$ (as a function of $x$) and the vector of parameters $\beta$ (Fahrmeir and

Tutz, 2001). Given the vector of explanatory variables $x$, a GLM for a categorical response variable is characterized by the equation $g(\pi) = Z\beta$, where there are exactly $J-1$ equations $g_j(\pi) = \eta_j$. The model can also be represented by the following diagram

$$\begin{tikzcd} \mathcal{X} \arrow[r, "Z"] & \mathcal{S} \arrow[r, "g^{-1}"] & \mathcal{M}, \\ x \arrow[r, mapsto] & \eta \arrow[r, mapsto] & \pi, \end{tikzcd}$$

where $\mathcal{X}$ is the space of explanatory variables.

All the classical link functions (see Agresti (2002); Tutz (2012)) have the same structure which we propose to write as

$$g_j = F^{-1} \circ r_j, \quad j = 1, \ldots, J-1, \tag{2.1}$$

where $F$ is a continuous and strictly increasing cumulative distribution function and $r = (r_1, \ldots, r_{J-1})^t$ is a diffeomorphism from $\mathcal{M}$ to an open subset $\mathcal{P}$ of $]0,1[^{J-1}$. Finally, given $x$, we propose to summarize a GLM for categorical response variable by

$$r(\pi) = \mathcal{F}(Z\beta),$$

where $\mathcal{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$. The model can thus be presented by the following diagram

$$\begin{tikzcd} \mathcal{X} \arrow[r, "Z"] & \mathcal{S} \arrow[r, "\mathcal{F}"] & \mathcal{P} \arrow[r, "r^{-1}"] & \mathcal{M}, \\ x \arrow[r, mapsto] & \eta \arrow[r, mapsto] & r \arrow[r, mapsto] & \pi. \end{tikzcd}$$

### 2.3.2 *(r,F,Z)* specification of GLMs for categorical data

In the following, we will describe in more detail the components $Z$, $F$, $r$ and their modalities.

**Design matrix $Z$:** Each linear predictor has the form $\eta_j = \alpha_j + x^t\delta_j$ with $\beta = (\alpha_1, \ldots, \alpha_{J-1}, \delta_1^t, \ldots, \delta_{J-1}^t) \in \mathbb{R}^{(J-1)(1+p)}$ where $p$ is the dimension of the explanatory space $\mathcal{X}$. In general, the model is defined without constraints, like for the multinomial logit model. But linear equality constraints, called contrasts, can be added between different slopes $\delta_j$, for instance. The most common constraint is the equality of all slopes, like for the odds proportional logit model. The corresponding constrained space $\mathcal{C} = \{\beta \in \mathbb{R}^{(J-1)(1+p)} | \delta_1 = \ldots = \delta_{J-1}\}$ may be identified to $\tilde{\mathcal{C}} = \mathbb{R}^{(J-1)+p}$. Finally, the constrained space is represented by a design matrix, containing the vector of explanatory variable $x$. For example, the *complete* design $(J-1) \times (J-1)(1+p)$-matrix $Z_c$ (without constraint) has the following form

$$Z_c = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix}.$$

The *proportional* design $(J-1) \times (J-1+p)$-matrix $Z_p$ (common slope) has the following form

$$Z_p = \begin{pmatrix} 1 & & & x^t \\ & \ddots & & \vdots \\ & & 1 & x^t \end{pmatrix}.$$

The model without slopes ($\delta_1 = \ldots = \delta_{J-1} = 0$), considered as the minimal response model, is defined with different intercepts $\alpha_j$ (the design matrix is the identity matrix of dimension $J-1$). In most cases, the design matrix contains the identity matrix as minimal block, such as $Z_c$ and $Z_p$. These two matrices are sufficient to define all the classical models. It should be noted that for a given constrained space $\mathcal{C}$, there are an infinity of corresponding design matrices which will be considered as equivalent. For example

$$Z'_p = \begin{pmatrix} 1 & & & -x^t \\ \vdots & \ddots & & \vdots \\ 1 & \ldots & 1 & -x^t \end{pmatrix}$$

is equivalent to $Z_p$. In the following, the design matrices $Z_p$ and $Z_c$ are considered as the representative element of their equivalence class and the set of all possible design matrices $Z$, for a given vector of explanatory variables $x$, will be denoted by $\mathfrak{Z}$. This set $\mathfrak{Z}$ contains all design matrices between $Z_p$ and $Z_c$, with number of columns between $J-1+p$ and $(J-1)(1+p)$.

**Cumulative distribution function $F$:** The most commonly used symmetric distributions are the *logistic* and *normal* distributions, but *Laplace* and *Student* distributions may also be useful. The most commonly used asymmetric distribution is the *Gumbel min* distribution

$$F(\eta) = 1 - \exp\left\{-\exp(\eta)\right\}.$$

Let $\tilde{F}$ denote the symmetric of $F$ (i.e. $\tilde{F}(\eta) = 1 - F(-\eta)$). The symmetric of the Gumbel min distribution is the *Gumbel max* distribution

$$\tilde{F}(\eta) = \exp\left\{-\exp(-\eta)\right\}.$$

All these cdfs, being diffeomorphisms from $\mathbb{R}$ to $]0,1[$, ease the interpretation of estimated parameter $\hat{\beta}$ and computation of Fisher's scoring algorithm. The *exponential* distribution, which is a diffeomorphism from $\mathbb{R}_+^*$ to $]0,1[$, is also used but the positivity constraint on predictors may lead to divergence of estimates.

As noted by Tutz (1991), "distribution functions generate the same model if they are connected by a linear transformation". For instance, if the connexion is made through a location parameter $u$ and a scale parameter $s$ such that $F_{u,s}(w) = F\{(w-u)/s\}$, we have for $j = 1, \ldots, J-1$

$$F_{u,s}(\eta_j(x)) = F\left(\frac{\eta_j(x) - u}{s}\right) = F\left(\frac{\alpha_j - u}{s} + x^t \frac{\delta_j}{s}\right),$$

and obtain an equivalent model using the reparametrization $\alpha'_j = (\alpha_j - u)/s$ and $\delta'_j = \delta_j/s$ for $j = 1, \ldots, J-1$. This is the case for all distributions previously introduced. But Student distributions, with different degrees of freedom, are not connected by a linear transformation. Therefore they lead to different likelihood maxima. In applications, Student distributions will be used with few degrees of freedom. Playing on the symmetrical or asymmetrical character and the more or less heavy tails of distributions may markedly improve model fit. In the following, the set of all continuous cdf $F$ (respectively continuous and symmetric cdf $F$) will be denoted by $\mathfrak{F}$ (respectively by $\tilde{\mathfrak{F}}$).

| **Name** | $\boldsymbol{r_j(\pi)}$<br>for $j = 1, \ldots, J-1$ |
|---|---|
| *Cumulative* | $\pi_1 + \ldots + \pi_j$ |
| *Sequential* | $\dfrac{\pi_j}{\pi_j + \ldots + \pi_J}$ |
| *Adjacent* | $\dfrac{\pi_j}{\pi_j + \pi_{j+1}}$ |
| *Reference* | $\dfrac{\pi_j}{\pi_j + \pi_J}$ |

Table 2.2: The four ratios, diffeomorphisms between open subsets of $]0,1[^{J-1}$.

**Ratio of probabilities $\boldsymbol{r}$:** The linear predictor $\eta$ is not directly related to the expectation $\pi$, through the cdf $F$, but to a particular transformation $r$ of $\pi$ which we call the ratio.

In this context, the odds proportional logit model for instance relies on the *cumulative* ratio defined by

$$r_j(\pi) = \pi_1 + \ldots + \pi_j,$$

for $j = 1, \ldots, J-1$. If there is a total order among categories, cumulative models can be used and interpreted by introducing a latent continuous variable $V$ having cdf $F$ (McCullagh, 1980). The linear predictors $(\eta_j)_{j=1,\ldots,J-1}$ are then strictly ordered and we obtain for $j = 1, \ldots, J-1$

$$\{Y \leq j\} \Leftrightarrow \{V \leq \eta_j\}.$$

The continuation ratio logit model relies on the *sequential* ratio defined by

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \ldots + \pi_J},$$

for $j = 1, \ldots, J-1$. Total order may be interpreted in a different way with sequential models. A sequential model corresponds to a sequential independent latent continuous process $(V_t)_{t=1,\ldots,J-1}$ having the cdf $F$ (Tutz, 1990). This process is governed by

$$\{Y = j\} \Leftrightarrow \bigcap_{t=1}^{j-1} \{V_t > \eta_t\} \bigcap \{V_j \leq \eta_j\},$$

for $j = 1, \ldots, J-1$. The conditional event $\{Y = j | Y \geq j\}$ can be expressed by

$$\{Y = j | Y \geq j\} \Leftrightarrow \{V_j \leq \eta_j\}.$$

The adjacent logit model is based on the *adjacent* ratio defined by

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_{j+1}},$$

for $j = 1, \dots, J-1$. Adjacent models are not directly interpretable using latent variables.

Unlike these models for ordinal data, we propose to define a ratio that is independent of the category ordering assumption. Using the structure of the multinomial logit model, we define the *reference* ratio for each category $j = 1, \dots, J-1$ as

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J}.$$

Each category $j$ is then related to a reference category (here $J$ by convention) and thus no category ordering is assumed. Therefore, the reference ratio allows us to define new GLMs for nominal response variables.

| | |
|---|---|
| *Multinomial logit model*<br>$P(Y=j) = \dfrac{\exp(\alpha_j + x^T\delta_j)}{1 + \sum_{k=1}^{J-1}\exp(\alpha_k + x^T\delta_k)}$ | (reference, logistic, complete) |
| *Odds proportional logit model*<br>$\ln\left\{\dfrac{P(Y \leq j)}{1 - P(Y \leq j)}\right\} = \alpha_j + x^T\delta$ | (cumulative, logistic, proportional) |
| *Proportional hazard model*<br>*(Grouped Cox Model)*<br>$\ln\{-\ln P(Y > j \mid Y \geq j)\} = \alpha_j + x^T\delta$ | (sequential, Gumbel min, proportional) |
| *Adjacent logit model*<br>$\ln\left\{\dfrac{P(Y=j)}{P(Y=j+1)}\right\} = \alpha_j + x^T\delta_j$ | (adjacent, logistic, complete) |
| *Continuation ratio logit model*<br>$\ln\left\{\dfrac{P(Y=j)}{P(Y>j)}\right\} = \alpha_j + x^T\delta_j$ | (sequential, logistic, complete) |

Table 2.3: $(r, F, Z)$ specification of five classical GLMs for categorical data.

In the following, each GLM for categorical data will be specified by a $(r, F, Z)$ triplet. Table 2.3 shows $(r, F, Z)$ triplet specifications for classical models. This specification eases the comparison of GLMs for categorical data. Moreover, it enables to define an enlarged family of GLMs for nominal response variables (referred to as the reference family) using $(\text{reference}, F, Z)$ triplets, which includes the (reference, logistic, complete) multinomial logit model. GLMs for nominal and ordinal response variables are usually defined with different design matrices $Z$; see the first two rows in table 2.3. Fixing the design matrix $Z$ may ease the comparison between GLMs for nominal and ordinal response variables.

### 2.3.3 Compatibility of the three components $r$, $F$ and $Z$

A GLM for categorical data is specified by an $(r, F, Z)$ triplet but is it always defined? The condition $\pi(x) \in \mathcal{M}$ is required for all $x \in \mathcal{X}$. It should be noted that reference, adjacent and sequential ratios are defined with $J-1$ different conditioning. Therefore the linear predictors $\eta_j$ are not constrained one to another. Neither $\mathcal{P}$ nor $\mathcal{S}$ are constrained ($\mathcal{P} = ]0,1[^{J-1}$ and $\mathcal{S} = \mathbb{R}^{J-1}$) and thus no constraint on parameter $\beta$ is required.

The situation is different for the cumulative ratio, because the probabilities $r_j(\pi)$ are not conditional but linked ($r_{j+1}(\pi) = r_j(\pi) + \pi_{j+1}$). Both $\mathcal{P}$ and $\mathcal{S}$ are constrained ($\mathcal{P} = \{r \in ]0,1[^{J-1} | r_1 < \ldots < r_{J-1}\}$ and $\mathcal{S} = \{\eta \in \mathbb{R}^{J-1} | \eta_1 < \ldots < \eta_{J-1}\}$). Therefore the definition of a cumulative model entails constraints on $\beta = (\alpha_1, \ldots, \alpha_{J-1}, \delta_1^t, \ldots, \delta_{J-1}^t)$. Without loss of generality, we will work hereinafter with only one explanatory variable $x \in \mathcal{X}$. The constraints are different depending on the form of $\mathcal{X}$.

**Case 1: $x$ is categorical** then $\mathcal{X} = \{x \in \{0,1\}^{C-1} | \sum_{c=1}^{C-1} x_c \in \{0,1\}\}$. In this case, the form of the linear predictors is

$$\eta_j(x) = \alpha_j + \sum_{c=1}^{C-1} \mathbf{1}_{\{X=c\}} \, \delta_{j,c},$$

and the constraints $\eta_j(x) < \eta_{j+1}(x)$ $\forall x \in \mathcal{X}$ are equivalent to

$$\begin{cases} \alpha_j & < & \alpha_{j+1}, \\ \delta_{j,c} & \leq & \delta_{j+1,c}, \quad \forall c \in \{1, \ldots, C-1\}. \end{cases}$$

**Case 2: $x$ is continuous** then $\mathcal{X} \subseteq \mathbb{R}$. In this case, the form of the linear predictors is

$$\eta_j(x) = \alpha_j + \delta_j x.$$

Since the $\eta_j$ must be ordered on $\mathcal{X}$, three domains of definition $\mathcal{X}$ must be differentiated:

$\underline{\mathcal{X} = \mathbb{R}}$ $\eta_j$ are ordered and parallel straight lines

$$\begin{cases} \alpha_j & < & \alpha_{j+1}, \\ \delta_j & = & \delta_{j+1}. \end{cases}$$

This is the case of the odds proportional logit model.

$\underline{\mathcal{X} = \mathbb{R}_+}$ $\eta_j$ are ordered and non-intersected half-lines

$$\begin{cases} \alpha_j & < & \alpha_{j+1}, \\ \delta_j & \leq & \delta_{j+1}. \end{cases}$$

This is the case of a positive continuous variable $X$, such as a size or a dosage for instance. Moreover, if $X$ is strictly positive, the intercepts $\alpha_j$ can be common.

$\underline{\mathcal{X} = [a, b]}$ $\eta_j$ are ordered and non-intersected segments. The constraints cannot be simply rewritten in terms of intercept and slope constraints.

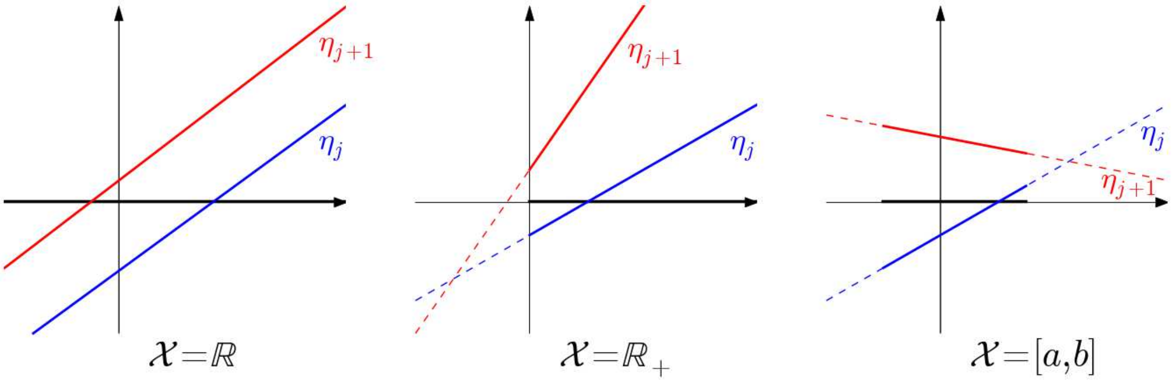

Figure 2.1: Linear predictors for different configurations of the continuous space $\mathcal{X}$.

For the last two cases a vector of probabilities $\pi(x)$ for $x$ out of $\mathcal{X}$ cannot always be predicted (see figure 2.1).

### 2.3.4 Fisher's scoring algorithm

For maximum likelihood estimation, the iteration of Fisher's scoring algorithm is given by

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

For the sake of simplicity, the algorithm is detailed for only one observation $(y, x)$ with $l = \ln P(Y = y | X = x; \beta)$. Using the chain rule we obtain the score

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

Using Property 5, we obtain

$$\frac{\partial l}{\partial \beta} = Z^t \frac{\partial \pi}{\partial \eta} \operatorname{Cov}(Y|x)^{-1} [y - \pi].$$

Then using decomposition (3.1) of the link function we obtain

$$\frac{\partial l}{\partial \beta} = Z^t \frac{\partial \mathcal{F}}{\partial \eta} \frac{\partial \pi}{\partial r} \operatorname{Cov}(Y|x)^{-1} [y - \pi]. \tag{2.2}$$

Again using Property 5 and decomposition (3.1) of the link function, we obtain Fisher's information matrix

$$\begin{aligned}\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta^t\partial\beta}\right) &= -\frac{\partial\pi}{\partial\beta}\,\mathrm{Cov}(Y|x)^{-1}\,\frac{\partial\pi}{\partial\beta^t}\\ &= -Z^t\,\frac{\partial\pi}{\partial\eta}\,\mathrm{Cov}(Y|x)^{-1}\,\frac{\partial\pi}{\partial\eta^t}\,Z\\ \mathrm{E}\left(\frac{\partial^2 l}{\partial\beta^t\partial\beta}\right) &= -Z^t\,\frac{\partial\mathcal{F}}{\partial\eta}\,\frac{\partial\pi}{\partial r}\,\mathrm{Cov}(Y|x)^{-1}\,\frac{\partial\pi}{\partial r^t}\,\frac{\partial\mathcal{F}}{\partial\eta^t}\,Z. \qquad (2.3)\end{aligned}$$

We only need to evaluate the associated density function $\{f(\eta_j)\}_{j=1,\dots,J-1}$ to compute the diagonal Jacobian matrix $\partial\mathcal{F}/\partial\eta$. For details on computation of the Jacobian matrix $\partial\pi/\partial r$ according to each ratio, see appendix B.

## 2.4 Properties of GLMs for categorical data

This section focuses on equivalences between GLMs for categorical data. All the following properties strongly depend on the link function, especially the ratio. It should be noted that for the case $J = 2$, all four ratios (see table 2.2) are the same, leading to the Bernoulli case. Hence we focus only on the case $J > 2$.

The truncated multinomial distribution $\mathcal{TM}(\pi)$ is fully specified by parameter $\pi$ of dimension $J-1$. Therefore, the distribution of $Y|X = x$ is fully specified by the $(r, F, Z)$ triplet for a fixed value of $\beta \in \tilde{\mathcal{C}}$

$$\pi = r^{-1} \circ \mathcal{F}\{Z(x)\beta\}.$$

Equality and equivalence between two models are differentiated here, using the $(r, F, Z)$ specification.

**Remark 1.** *In this thesis we employ* ***model*** *when the three component $r$, $F$ and $Z$ are determined, whereas we employ* ***family of models*** *when at least one of the three components is undetermined. For example (reference, logistic, complete) is a model, whereas $\{$(reference, $F$, $Z$)$|\ F \in \mathfrak{F},\ Z \in \mathfrak{Z}\}$ is a family of models.*

**Definition 4.** *Two models $(r, F, Z)$ and $(r', F', Z')$ are said to be* ***equal*** *if the corresponding distributions of $Y|X = x$ are equal for all $x$ and all $\beta$*

$$r^{-1} \circ \mathcal{F}\{Z(x)\beta\} = r'^{-1} \circ \mathcal{F}'\{Z'(x)\beta\},\ \forall x \in \mathcal{X},\ \forall \beta \in \tilde{\mathcal{C}}.$$

**Definition 5.** *Two models $(r, F, Z)$ and $(r', F', Z')$ are said to be* ***equivalent*** *if one is a reparametrization of the other, and conversely. Hence, there exists a bijection $h$ from $\tilde{\mathcal{C}}$ to $\tilde{\mathcal{C}}'$ such that*

$$r^{-1} \circ \mathcal{F}\{Z(x)\beta\} = r'^{-1} \circ \mathcal{F}'\{Z'(x)h(\beta)\},\ \forall x \in \mathcal{X},\ \forall \beta \in \tilde{\mathcal{C}}.$$

It should be noted that equality between two models necessarily means that they are also equivalent.

### 2.4.1 Equivalence between GLMs for categorical data

In this subsection we first compare sequential and cumulative ratios, both of which are used for ordinal data, then compare reference and adjacent ratios, used for nominal and ordinal data, respectively.

#### 2.4.1.1 Comparison between sequential and cumulative models

We first remind two equivalences between models for ordinal response variables and extend the second one for other design matrices. Thus, we obtain equality between families of sequential and cumulative models using a transformation between design matrices. Finally, we focus on particular models for the *complete* and *proportional* design matrices.

**Equivalences between models** Läärä and Matthews (1985) showed the equivalence between (sequential, Gumbel min, proportional) and (cumulative, Gumbel min, proportional) models. Tutz (1991) noted that "equivalence holds only for the simple version of the models where the thresholds are not determined by the explanatory variables". In our framework, this means that (sequential, Gumbel min, $Z$) and (cumulative, Gumbel min, $Z$) models are equivalent only for the *proportional* design matrix $Z_p$. Consider the bijection between the two predictor spaces $\mathcal{S} = \mathbb{R}^{J-1}$ and $\mathcal{S}' = \{\eta \in \mathbb{R}^{J-1} | \eta_1 < \ldots < \eta_{J-1}\}$ for $j = 1, \ldots, J-1$ defined by

$$\eta'_j = \ln\left\{\sum_{k=1}^{j} \exp(\eta_k)\right\},$$

which is not linear with respect to $\eta$. The predictor $\eta'$ must be linear at least with respect to $x$. Rewriting $\eta'_j$ as

$$\eta'_j = \ln\left\{\sum_{k=1}^{j} \exp(\alpha_k)\exp(x^t\delta_k)\right\},$$

we see that linearity with respect to $x$ holds if and only if $\delta_1 = \ldots = \delta_{J-1}$. This corresponds to the proportional design matrix. Finally the equivalence of Läärä and Matthews (1985) holds with the bijection $h$ between $\tilde{\mathcal{C}} = \mathbb{R}^{J-1+p}$ and $\tilde{\mathcal{C}}' = \{\beta \in \mathbb{R}^{J-1+p} | \alpha_1 < \ldots < \alpha_{J-1}\}$ defined by

$$\begin{cases} \alpha'_j &= \ln\left\{\sum_{k=1}^{j} \exp(\alpha_k)\right\}, \text{ for } j = 1, \ldots, J-1, \\ \delta' &= \delta. \end{cases}$$

The equivalence shown by Tutz (1991)

$$\text{(cumulative, exponential, complete)} \Leftrightarrow \text{(sequential, exponential, complete)},$$

is quite different because the reparametrization is linear with respect to $\eta$. This result can therefore be generalized for any design matrix.

**Equality between families of models** It should be noted that the exponential cdf is defined only for strictly positive values. Therefore, the following property holds only for parameter values $\beta$ such that $\eta = Z\beta \in \mathbb{R}_+^{*J-1}$.

**Property 6.** *The two families of models* $\{$*(cumulative, exponential,* $Z$*)*; $Z \in \mathfrak{Z}\}$ *and* $\{$*(sequential, exponential,* $Z$*)*; $Z \in \mathfrak{Z}\}$ *are equal.*

*Proof.* The equality between the two families is shown using double inclusion method. Assume that the distribution of $Y|X = x$ is defined by the (cumulative, exponential, $Z$) model with an unknown design matrix $Z \in \mathfrak{Z}$. For $j = 1, \ldots, J-1$ we obtain

$$\pi_1 + \ldots + \pi_j = 1 - \exp(-\eta_j). \tag{2.4}$$

The sequential ratio can be rewritten in terms of cumulative ratio

$$\frac{\pi_j}{\pi_j + \ldots + \pi_J} = \frac{(\pi_1 + \ldots + \pi_j) - (\pi_1 + \ldots + \pi_{j-1})}{1 - (\pi_1 + \ldots + \pi_{j-1})},$$

for $j = 2, \ldots, J-1$. Using (2.4) it becomes

$$\frac{\pi_j}{\pi_j + \ldots + \pi_J} = 1 - \exp\{-(\eta_j - \eta_{j-1})\},$$

for $j = 2, \ldots, J-1$. Therefore, we consider the reparametrization

$$\begin{cases} \eta'_1 & = & \eta_1, \\ \eta'_j & = & \eta_j - \eta_{j-1} \text{ for } j = 2, \ldots, J-1, \end{cases}$$

between the two predictor spaces $\mathcal{S} = \{\eta \in \mathbb{R}^{J-1} | 0 \leq \eta_1 < \ldots < \eta_{J-1}\}$ and $\mathcal{S}' = \mathbb{R}_+^{*J-1}$. As this transformation is linear with respect to $\eta$, we introduce the following square matrix

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

of dimension $J-1$ and then

$$\eta' = A\eta,$$
$$\eta' = AZ\beta.$$

Hence $Y|X = x$ follows the (sequential, exponential, $AZ$) model with the same parameter $\beta$. This means that

$$\text{(cumulative, exponential, } Z) = \text{(sequential, exponential, } AZ),$$

and we thus obtain first inclusion. Finally, noting that $A$ is invertible we obtain

$$\text{(cumulative, exponential, } A^{-1}Z) = \text{(sequential, exponential, } Z),$$

and we thus obtain second inclusion. □

**Proportional and complete design matrices** It should be remarked that $A$ changes in general the constraints on space $\mathcal{C}$ and thus the likelihood maximum. For example, the design matrix $Z_p$ corresponds to the constrained space $\mathcal{C} = \{\beta \in \mathbb{R}^{(J-1)(1+p)} | \delta_1 = \ldots = \delta_{J-1}\}$, whereas the design matrix

$$AZ_p = \begin{pmatrix} 1 & & & & x^t \\ -1 & 1 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & -1 & 1 & 0 \end{pmatrix}$$

corresponds to the constrained space $\mathcal{C} = \{\beta \in \mathbb{R}^{(J-1)(1+p)} | \delta_2 = \ldots = \delta_{J-1} = 0\}$. The design matrices $Z_p$ and $AZ_p$ are not equivalent and thus the (cumulative, exponential, $Z_p$) and (sequential, exponential, $Z_p$) models are not equivalent, whereas the (cumulative, exponential,

$Z_p$) and (sequential, exponential, $AZ_p$) models are equal. In the same way, the (cumulative, exponential, $A^{-1}Z_p$) and (sequential, exponential, $Z_p$) models are equal with

$$A^{-1}Z_p = \begin{pmatrix} 1 & & & & x^t \\ 1 & 1 & & & 2x^t \\ \vdots & & \ddots & & \vdots \\ 1 & 1 & \dots & 1 & (J-1)x^t \end{pmatrix}.$$

For the particular complete design there is no constraint on $\mathcal{C}$, thus $A$ cannot change it. We must simply check that $A$ does not reduce the dimension of $\mathcal{C}$. Since $A$ has full rank, the matrices $Z_c$ and $AZ_c$ are equivalent. Therefore the equality between the (cumulative, exponential, $Z_c$) and (sequential, exponential, $AZ_c$) models becomes

$$\text{(cumulative, exponential, } Z_c) \Leftrightarrow \text{(sequential, exponential, } Z_c),$$

and we recover the equivalence described by Tutz (1991).

### Comparison between reference and adjacent models

Here we follow the same approach as previously with sequential and cumulative ratios, and thus provide fewer details. We start from the equivalence

$$\text{(reference, logistic, complete)} \Leftrightarrow \text{(adjacent, logistic, complete)}, \tag{2.5}$$

shown by Agresti (2010) and then generalize this equivalence for any design matrix. Before starting, the family of canonical models must be introduced using the $(r, F, Z)$ specification. In the GLM framework, a model is said to be canonical if the equality between the natural parameter $\theta$ and the linear predictor $\eta$ holds. This equality can be rewritten in terms of link function $g = \mathcal{F}^{-1} \circ r$.

$$\begin{array}{rrcll} & \theta & = & \eta & \\ \Leftrightarrow & \ln\left(\dfrac{\pi_j}{\pi_J}\right) & = & \eta_j & \forall j = 1, \dots, J-1 \\ \Leftrightarrow & \dfrac{\pi_j}{\pi_j + \pi_J} & = & \dfrac{\exp(\eta_j)}{1+\exp(\eta_j)} & \forall j = 1, \dots, J-1. \end{array}$$

The reference ratio $r$ can be recognized on the left hand side and the logistic cdf $F$ on the right hand side. Therefore, the family of canonical models is $\{$(reference, logistic, $Z$); $Z \in \mathfrak{Z}\}$. The multinomial logit model is a particular canonical model with the complete design matrix. Finally, we generalize the equivalence, described by Agresti (2010), between two models to the family of canonical models.

**Property 7.** *The family of canonical models* $\{$*(reference, logistic,* $Z$*)*; $Z \in \mathfrak{Z}\}$ *is equal to the family of models* $\{$*(adjacent, logistic,* $Z$*)*; $Z \in \mathfrak{Z}\}$*.*

*Proof.* Following the same approach as for Property 6 proof, we can see that (reference, logistic, $Z$) and (adjacent, logistic, $A^tZ$) models are equal for every design matrix $Z \in \mathfrak{Z}$, where $A^t$ turns out to be the transpose of the matrix $A$ previously defined. Noting that $A^t$ is invertible ($(A^t)^{-1} = (A^{-1})^t$), we obtain the desired result. □

As previously, $A^t$ generally changes the constraints on space $\mathcal{C}$. For example, the design matrices $Z_p$ and $A^tZ_p$ are not equivalent. The particular equality

$$\text{(reference, logistic, } (A^{-1})^t Z_p) = \text{(adjacent, logistic, } Z_p),$$

where

$$(A^{-1})^t Z_p = \begin{pmatrix} 1 & 1 & \dots & 1 & (J-1)x^t \\ & \ddots & & \vdots & \vdots \\ & & 1 & 1 & 2x^t \\ & & & 1 & x^t \end{pmatrix},$$

corresponds to a particular reparametrization described by Agresti (2010). Noting that the design matrices $Z_c$ and $A^t Z_c$ are equivalent because $A^t$ has full rank, we recover the equivalence (2.5) described by Agresti (2010).

### 2.4.2 Permutation invariance and stability

Property 7 shows equality between a family defined for nominal data and a family defined for ordinal data since the adjacent ratio uses the category ordering assumption whereas the reference ratio does not. The two families of reference and adjacent models overlap for the particular case of the logistic cdf. Thus, we need to determine whether this subfamily of canonical models is more appropriate for nominal or ordinal data. More generally, we want to classify each $(r, F, Z)$ triplet as a nominal or an ordinal model. “It is proposed that models for nominal categories should be invariant under arbitrary permutations [...] On the other hand, models for ordinal data should be invariant only under the special reverse permutation” (McCullagh, 1978). We therefore propose to investigate permutation invariances of each model and permutation stabilities of some families of models; especially the four families defined with the four ratios.

Let us first introduce the notion of permutation invariance. Each index $j \in \{1, \dots, J\}$ is associated with a particular category. Modifying this association potentially changes the model. Such a modification is characterized by a permutation $\sigma$ of $\{1, \dots, J\}$. The permuted vector parameter $\pi_\sigma = (\pi_{\sigma(1)}, \dots, \pi_{\sigma(J-1)})$ is thus introduced and the permuted model is summarized by

$$r(\pi_\sigma) = \mathcal{F}(Z\beta),$$

and denoted by the indexed triplet $(r, F, Z)_\sigma$. In this subsection we focus on the permutations that preserve the model, or at least preserve the link function, or even the ratio. The main idea is to find $\sigma$ such that the ratio of permuted probabilities and the permuted ratio of probabilities are equal (i.e. such that $r(\pi_\sigma) = r_\sigma(\pi)$). We will see that these permutations are not the same depending on the link function, especially because of the ratio. We need to introduce the following definition.

**Definition 6.** *Let $\sigma$ be a permutation of $\{1, \dots, J\}$. A model $(r, F, Z)$ is said to be **invariant** under $\sigma$ if the models $(r, F, Z)$ and $(r, F, Z)_\sigma$ are equivalent. A family of models $\mathfrak{M}$ is said to be **stable** under $\sigma$ if $\sigma(\mathfrak{M}) = \mathfrak{M}$, where $\sigma(\mathfrak{M}) = \{(r, F, Z)_\sigma | \, (r, F, Z) \in \mathfrak{M}\}$.*

Compared to the previous subsection, we focus here on equivalences between models that have the same ratio. It could be said that equivalences between models become invariances of models and equalities between families become stabilities of families. Following the same approach, we will show the stability of families under permutations and then focus on invariant models for the complete and proportional design matrices.

#### Models for nominal data

Unlike the adjacent, cumulative and sequential ratios, the reference ratio is built without the category ordering assumption. The probability of each category is connected only with the probability of reference category $J$. Thus, changing the reference category could change the fit. Contrary to all the other permutations we have the following property.

**Property 8.** *The family of models* $\{$*(reference,* $F$*,* $Z$*)*; $F \in \mathfrak{F}$, $Z \in \mathfrak{Z}\}$ *is stable under the* $(J-1)!$ *permutations that fix the reference category.*

*Proof.* Let $\sigma$ denote a permutation of $\{1, \dots, J\}$ such that $\sigma(J) = J$. For the reference ratio it can be shown that

$$r_j(\pi_\sigma) = r_{\sigma(j)}(\pi),$$

for $j \in \{1, \dots, J-1\}$. Thus we need to permute the linear predictors $\eta_j$ using the restricted permutation matrix of dimension $J-1$, defined as follows

$$(P_\sigma)_{i,j} = \begin{cases} 1 & \text{if } i = \sigma(j), \\ 0 & \text{otherwise,} \end{cases}$$

for $i, j \in \{1, \dots, J-1\}$. Noting that $\eta' = P_\sigma \eta = P_\sigma Z \beta$ we obtain

$$(\text{reference}, F, Z)_\sigma = (\text{reference}, F, P_\sigma Z),$$

for all $F \in \mathfrak{F}$ and all $Z \in \mathfrak{Z}$. Noting that $P_\sigma$ is invertible ($P_\sigma^{-1} = P_{\sigma^{-1}}$) we obtain the desired result. □

Furthermore, the design matrices $Z_c$ and $P_\sigma Z_c$ are equivalent because $P_\sigma$ has full rank. Therefore, (reference, $F$, $Z_c)_\sigma$ and (reference, $F$, $Z_c$) models are equivalent, which means that (reference, $F$, $Z_c$) model is invariant under $\sigma$. Moreover, the design matrices $Z_p$ and $P_\sigma Z_p$ are also equivalent. In fact, permutation $\sigma$ does not change the contrast of common slope

$$\delta_1 = \ldots = \delta_{J-1} \Leftrightarrow \delta_{\sigma(1)} = \ldots = \delta_{\sigma(J-1)}.$$

Finally, noting that $Z_c$ and $P_\sigma Z_c$ (respectively $Z_p$ and $P_\sigma Z_p$) are equivalent, we obtain the following property of invariance.

**Property 9.** *Let* $F \in \mathfrak{F}$*. The two (reference,* $F$*, complete) and (reference,* $F$*, proportional) models are invariant under the* $(J-1)!$ *permutations that fix the reference category.*

But what happens if we transpose the reference category? The family of canonical models has the following property.

**Property 10.** *The family of canonical models* $\{$*(reference, logistic,* $Z$*)*; $Z \in \mathfrak{Z}\}$ *is stable under all permutations.*

*Proof.* Let $\tau$ denote a non identical transposition of the reference category $J$. Using Property 8, we need to show stability under $\tau$. Assume that the distribution of $Y|X = x$ is defined by the transposed canonical (reference, logistic, $Z)_\tau$ model. Thus we obtain

$$\begin{cases} \dfrac{\pi_j}{\pi_{\tau(J)}} = \exp(\eta_j) & \text{for } j \neq J \text{ and } j \neq \tau(J), \\ \dfrac{\pi_J}{\pi_{\tau(J)}} = \exp(\eta_{\tau(J)}), \end{cases}$$

or equivalently

$$\begin{cases} \dfrac{\pi_j}{\pi_J} = \dfrac{\pi_j}{\pi_{\tau(J)}} \dfrac{\pi_{\tau(J)}}{\pi_J} = \exp(\eta_j - \eta_{\tau(J)}) & \text{for} j \neq J \text{ and } j \neq \tau(J), \\ \dfrac{\pi_{\tau(J)}}{\pi_J} = \exp(-\eta_{\tau(J)}). \end{cases}$$

Hence $Y|X = x$ follows the canonical (reference, logistic, $B_\tau Z$) model, where $B_\tau$ is the $(J-1)$-identity matrix, whose $\tau(J)^{\text{th}}$ column is replaced by a column of $-1$. Noting that $B_\tau$ is invertible ($B_\tau^{-1} = B_\tau$) we obtain the desired result. □

It should be remarked that the design matrices $Z_p$ and $B_\tau Z_p$ are not equivalent. In fact the contrast $\delta_1 = \ldots = \delta_{J-1}$ becomes $\delta_j = 0$ for all $j \neq \tau(J)$. Thus the (reference, logistic, proportional) model is not invariant under $\tau$ but preserves its link function. By contrast, the canonical (reference, logistic, complete) model is invariant under all permutations because $P_\sigma$ and $B_\tau$ have full rank. This last result is just another way of saying that for the multinomial logit model, the choice of the reference category has no impact on model's fit. This holds only for this model and its adjacent equivalent. Finally, the (adjacent, logistic, complete) model, which is invariant under all permutations, is inappropriate for ordinal data. More generally, the family of models $\{(\text{adjacent, logistic}, Z);\ Z \in \mathfrak{Z}\}$, which is stable under all permutations (properties 7 and 10), is inappropriate for ordinal data.

### Models for ordinal data

Adjacent, cumulative and sequential ratios are defined with the category ordering assumption and thus are naturally devoted to ordinal data. We therefore expect a permutation which changes the order to change also the corresponding models. But what happens if we simply reverse the order? We would like the model to be invariant. McCullagh (1980) noted that the three models (cumulative, logistic, proportional), (cumulative, normal, proportional) and (cumulative, Cauchy, proportional) are invariant under the reverse permutation. More generally, we demonstrate stability of the cumulative family of models under this permutation. The adjacent ratio turns out to have the same property, unlike the sequential ratio.

**Property 11.** *The two families of models $\{$(adjacent, $F$, $Z$); $F \in \mathfrak{F}$, $Z \in \mathfrak{Z}\}$ and $\{$(cumulative, $F$, $Z$); $F \in \mathfrak{F}$, $Z \in \mathfrak{Z}\}$ are stable under the reverse permutation.*

*Proof.* Let $\tilde{\sigma}$ denote the reverse permutation (i.e. $\tilde{\sigma}(j) = J - j + 1$, $\forall j \in \{1, \ldots, J-1\}$). For adjacent and cumulative ratios, it can be shown that

$$r_j(\pi_{\tilde{\sigma}}) = 1 - r_{J-j}(\pi), \tag{2.6}$$

for $j \in \{1, \ldots, J-1\}$. Assume that the distribution $Y|X = x$ is defined by the permuted $(r, F, Z)_{\tilde{\sigma}}$ model with $r =$ *adjacent* or *cumulative*, i.e.

$$r_j(\pi_{\tilde{\sigma}}) = F(\eta_j),$$

for $j \in \{1, \ldots, J-1\}$. Using equality (2.6), we obtain equivalently

$$r_{J-j}(\pi) = \tilde{F}(-\eta_j), \tag{2.7}$$

for $j \in \{1, \ldots, J-1\}$. Now we denote $i = J - j$ and (2.7) becomes

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

for $i \in \{1, \ldots, J-1\}$. Hence $Y|X = x$ follows the $(r, \tilde{F}, -\tilde{P}Z)$ model, where $\tilde{P}$ is the restricted reverse permutation matrix of dimension $J-1$

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

Finally we have

$$(\text{adjacent}, F, Z) = (\text{adjacent}, \tilde{F}, -\tilde{P}Z),$$

and

$$(\text{cumulative}, F, Z) = (\text{cumulative}, \tilde{F}, -\tilde{P}Z).$$

Noting that $\tilde{P}$ is invertible ($\tilde{P}^{-1} = \tilde{P}$) we obtain the desired result. □

The design matrices $Z_p$ and $-\tilde{P}Z_p$ are equivalent because

$$\delta_1 = \ldots = \delta_{J-1} \Leftrightarrow -\delta_J = \ldots = -\delta_1.$$

The design matrices $Z_c$ and $-\tilde{P}Z_c$ are also equivalent because $\tilde{P}$ has full rank. Finally, we obtain the following property of invariance.

**Property 12.** *Let $F \in \tilde{\mathfrak{F}}$. The four (adjacent, F, complete), (adjacent, F, proportional), (cumulative, F, complete) and (cumulative, F, proportional) models are invariant under the reverse permutation.*

For example, the (cumulative, Laplace, proportional) model is invariant under the reverse permutation, whereas the (cumulative, Gumbel min, proportional) model is not. But the (cumulative, Gumbel min, proportional)$_{\tilde{\sigma}}$ and (cumulative, Gumbel max, proportional) models are equivalent.

The situation is quite different for sequential models because equality (2.6) is no longer valid. The reverse permutation changes the structure of a sequential model. Using the equivalence between sequential and cumulative models shown by Läärä and Matthews (1985), and the previous result, we obtain the following equivalence under the reverse permutation

$$(\text{sequential, Gumbel min, proportional})_{\tilde{\sigma}} \Leftrightarrow (\text{cumulative, Gumbel max, proportional}).$$

But for other sequential models, if we introduce the elementary transposition $\tilde{\tau}$ of the last two categories, we obtain the following property.

**Property 13.** *The family of models $\{$(sequential, $F$, $Z$); $F \in \tilde{\mathfrak{F}}$, $Z \in \mathfrak{Z}\}$ is stable under the transposition $\tilde{\tau}$ of the last two categories.*

*Proof.* Assume that the distribution of $Y|X = x$ is defined by the transposed (sequential, $F$, $Z$)$_{\tilde{\tau}}$ model with a symmetric cdf $F$. Thus we obtain

$$\frac{\pi_{\tilde{\tau}(j)}}{\pi_{\tilde{\tau}(j)} + \ldots + \pi_{\tilde{\tau}(J-1)} + \pi_{\tilde{\tau}(J)}} = F(\eta_j),$$

for $j \in \{1, \ldots, J-1\}$. The last equation can be isolated

$$\begin{cases} \dfrac{\pi_j}{\pi_j + \ldots + \pi_J} = F(\eta_j) \quad \forall j \in \{1, \ldots, J-2\}, \\ \dfrac{\pi_J}{\pi_J + \pi_{J-1}} = F(\eta_{J-1}). \end{cases}$$

This is equivalent to

$$\begin{cases} \dfrac{\pi_j}{\pi_j + \ldots + \pi_J} = F(\eta_j), \qquad \forall j \in \{1, \ldots, J-2\}, \\ \dfrac{\pi_{J-1}}{\pi_{J-1} + \pi_J} = \tilde{F}(-\eta_{J-1}). \end{cases}$$

Since $F$ is symmetric (i.e. $F = \tilde{F}$), then $Y|X = x$ follows the (sequential, $F$, $A_{\tilde{\tau}}Z$) model, where $A_{\tilde{\tau}}$ is the following squared matrix of dimension $J-1$

$$A_{\tilde{\tau}} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & -1 \end{pmatrix}.$$

Noting that $A_{\tilde{\tau}}$ is invertible ($A_{\tilde{\tau}}^{-1} = A_{\tilde{\tau}}$) we obtain the desired result. □

Furthermore, noting that $A_{\tilde{\tau}}Z_c$ and $Z_c$ are equivalent because $A_{\tilde{\tau}}$ has full rank, we obtain the following property of invariance:

**Property 14.** *Let $F \in \tilde{\mathfrak{F}}$. The (sequential, $F$, complete) model is invariant under the transposition $\tilde{\tau}$ of the last two categories.*

However the design matrices $A_{\tilde{\tau}}Z$ and $Z$ are not equivalent in general. For example, the particular design matrices $A_{\tilde{\tau}}Z_p$ and $Z_p$ are not equivalent. Therefore, the (sequential, $F$, proportional) model is not invariant under $\tilde{\tau}$, even if $F$ is symmetric.

## 2.5 Investigation of invariance properties using benchmark datasets

Models for ordinal data should be invariant **only** under the reverse permutation (McCullagh, 1980). According to Property 12, we still have to show that (cumulative, $F$, complete) and (cumulative, $F$, proportional) models, with $F \in \tilde{\mathfrak{F}}$, are not invariant under other permutations. But invariance under a permutation is easier to show than the contrary. Thus, we proposed to investigate the $J!$ permutations on a dataset. To highlight the possible equivalences between models, all the models were ordered according to their log-likelihood value. Each plateau of log-likelihood therefore likely corresponds to an invariance under a particular permutation. For cumulative models, we thus expect to obtain exactly $J!/2!$ plateaus. If cumulative models are invariant under other permutations, we expect some merging of plateaus.

We investigated also invariances of (adjacent, $F$, complete) and (adjacent, $F$, proportional) models, with $F \in \tilde{\mathfrak{F}}$ and $F \neq$ logistic. Using Property 9 (respectively 14), a similar approach was then applied to (reference, $F$, complete) and (reference, $F$, proportional) models with permutations $\sigma$ fixing the reference category (respectively to (sequential, $F$, complete) models with the transposition $\tilde{\tau}$ of the last two categories).

We used the *boy's disturbed dreams* benchmark dataset (given in table 2.4) drawn from a study that cross-classified boys by their age $x$ and the severity of their disturbed dreams $y$ (Maxwell, 1961). The explanatory variable $x$ was assigned mid-point values for each stratum and was used as a continuous variable.

**Reference models** The (reference, $F$, complete) and (reference, $F$, proportional) models are invariant under the $(J-1)!$ permutations that fix the reference category (Property 9). But are they still invariant under other permutations? The canonical (reference, logistic, complete) model, being invariant under all permutations, is excluded. Non-invariance of models may be shown when $F$ is analytically defined (see chapter 5). This is more complex for normal or Student distributions, and figure 2.2 therefore investigates these two cases. All the models are ordered according to their log-likelihood value and we obtain $J!/(J-1)! = J = 4$ plateaus as

| $x \setminus y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 5-7 | 7 | 4 | 3 | 7 |
| 8-9 | 10 | 15 | 11 | 13 |
| 10-11 | 23 | 9 | 11 | 7 |
| 12-13 | 28 | 9 | 12 | 10 |
| 14-15 | 32 | 5 | 4 | 3 |

Table 2.4: Degree of suffering from disturbed dreams of boys by age.

expected. Each plateau corresponds to a particular reference category with the $(J-1)! = 6$ permutations that fix this category.

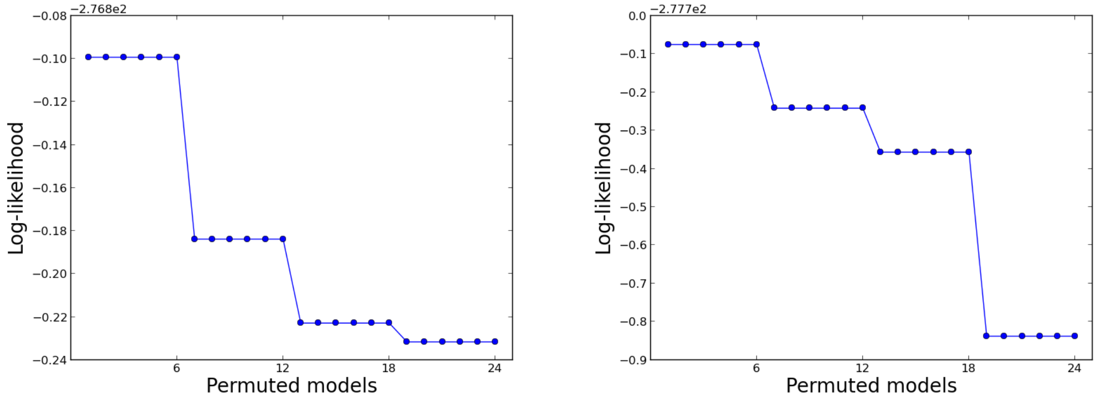

Figure 2.2: Ordered log-likelihood of (reference, normal, complete)$_\sigma$ and (reference, Student(1), proportional)$_\sigma$ models for all permutations $\sigma$.

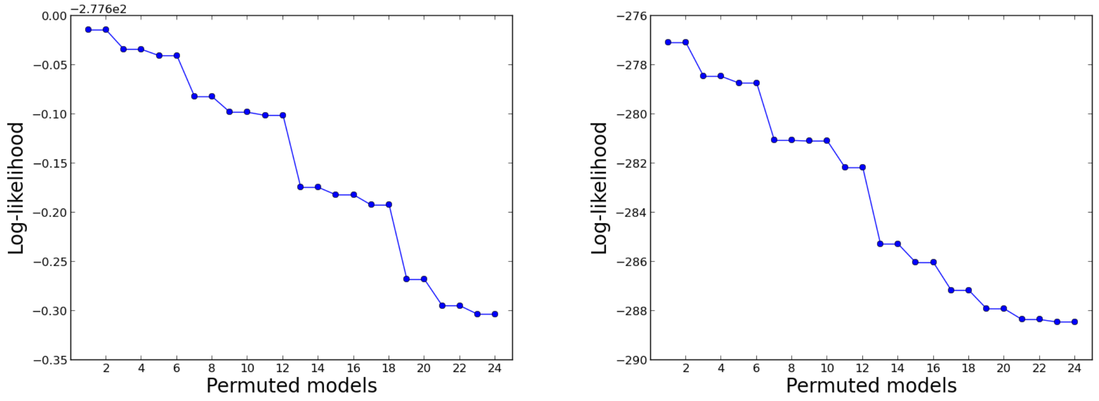

Figure 2.3: Ordered log-likelihood of (adjacent, Laplace, complete)$_\sigma$ and (cumulative, logistic, proportional)$_\sigma$ models for all permutations $\sigma$.

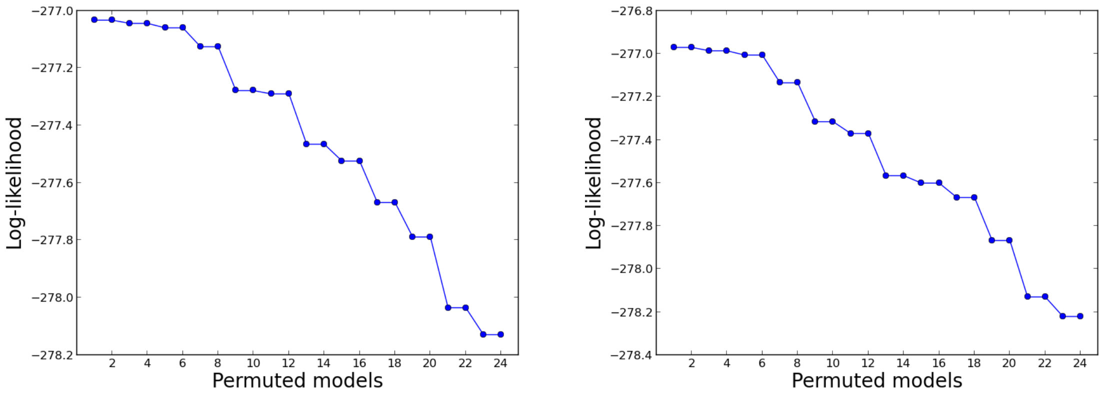

Figure 2.4: Ordered log-likelihood of (sequential, logistic, complete)$_\sigma$ and (sequential, Student(4), complete)$_\sigma$ models for all permutations $\sigma$.

**Adjacent and cumulative models** The (adjacent, $F$, complete) and (cumulative, $F$, proportional) models are invariant under the reverse permutation when $F$ is symmetric (Property 12). But are they still invariant under other permutations? The (adjacent, logistic, complete) model, being invariant under all permutations, is excluded. Figure 2.3 investigates the case of (adjacent, Laplace, complete) and (cumulative, logistic, proportional) models for all the $J! = 24$ permutations. All the models are ordered according to their log-likelihood value and we obtain $J!/2! = 12$ plateaus as expected. Each plateau corresponds to a particular permutation and its associated reverse permutation.

**Sequential models** The (sequential, $F$, complete) models are invariant under the transposition of the last two categories when $F$ is symmetric (Property 14). But are they still invariant under other permutations? Figure 2.4 investigates the case of (sequential, symmetric $F$, complete) models for all the $J! = 24$ permutations, with the logistic and Student(4) cdfs. All the models are ordered according to their log-likelihood value and we obtain $J!/2! = 12$ plateaus as expected. Each plateau corresponds to a particular permutation and its associated transposition of the last two categories.

## 2.6 Applications

### 2.6.1 Supervised classification

Linear, quadratic and logistic discriminant analyses are three classical methods used for supervised classification. The logistic discriminant analysis often outperforms other classical methods (Lim et al., 2000). In our context, we prefer to consider the logistic regression rather than the logistic discriminant analysis, these two methods being very similar (Hastie et al., 2005). In this subsection, we propose a family of classifiers that includes the logistic regression as a particular case. We then compare the classification error rates on two benchmark datasets (available on UCI), using 10-fold cross validation. The logistic regression is fully specified by the (reference, logistic, complete) triplet. This model is more suitable for non-ordered classes, even though it can also be used for ordered classes.

We propose to use the entire set of reference models with a complete design, which have all the same number of parameters. We can change the cdf $F$ to obtain a better fit. For the application, we use ten different cdf $F$:

$$\mathfrak{F}_0 = \{\text{normal, Laplace, Gumbel min, Gumbel max, Student}(1), \ldots, \text{Student}(6)\},$$

from which ten classifiers are built

$$\mathfrak{C}^* = \{(\text{reference}, F, \text{complete});\ F \in \mathfrak{F}_0\}.$$

All these classifiers are compared with the logistic regression, using 10-fold cross validation. For each classifier, the mean error rate is computed on the ten sub-samples and compared with the logistic regression error rate (represented in blue in figures 2.5, 2.6 and 2.7). The impact of changing the cdf can be seen (the corresponding minimal error rate is represented in green).

In the previous section, we saw that (reference, $F$, complete) models, with $F \neq$ logistic, do not seem to be invariant under transpositions of the reference category. This means that changing the reference category potentially changes the fit. Therefore, we propose to extend the set of classifiers $\mathfrak{C}^*$ to obtain

$$\mathfrak{C} = \{(\text{reference}, F, \text{complete})_\tau;\ F \in \mathfrak{F}_0,\ \tau \in \mathcal{T}_J\},$$

where $\mathcal{T}_J$ contains all transpositions of the reference category $J$. Finally, the set $\mathfrak{C}$ contains exactly $10 \times J$ classifiers. All these classifiers are then compared with the logistic regression. The impact of changing the reference category can be seen (the corresponding minimal error rate is represented in red). The three following original datasets are drawn from the UCI machine learning repository and the datasets already partitioned by means of a 10-fold cross validation procedure are drawn from the KEEL dataset repository. They contain respectively 3, 4 and 5 classes.

**Thyroid** This dataset is one of the several thyroid databases available in the UCI repository. The task is to detect if a given patient is normal (1) or suffers from hyperthyroidism (2) or hypothyroidism (3). This dataset contains $n = 7200$ individuals and all 21 attributes are numeric.

For the (reference, logistic, complete) model, the mean error rate of misclassification (in blue) was 6.12%. Using all the classifiers of $\mathfrak{C}^*$, the best classifier was the (reference, Student(3), complete) model with a misclassification mean error rate (in green) of 2.32% (see figure 2.5 on the left side). Finally, using all the classifiers of $\mathfrak{C}$, the best classifier was the (reference, Student(2), complete)$_\tau$ model (where $\tau(J) = 2$) with a misclassification mean error rate (in red) of 1.61%. The gain appeared to be mainly due to the change in cdf $F$ (see figure 2.5 on the right side).

**Vehicle** The purpose is to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles. The four types of vehicle are: bus (1), opel (2), saab (3) and van (4). This dataset contains $n = 846$ instances and all 18 attributes are numeric.

For the (reference, logistic, complete) model, the misclassification mean error rate (in blue) was 19.03%. All classifiers of $\mathfrak{C}^*$ or $\mathfrak{C}$ obtained an upper error rate (see figure 2.6).

**Pages blocks** The problem consists in classifying all the blocks of page layout in a document detected by a segmentation process. This is an essential step in document analysis to separate text from graphic areas. The five classes are: text (1), horizontal line (2), picture (3), vertical line (4) and graphic (5). The $n = 5473$ examples are drawn from 54 distinct documents. Each observation concerns one block. All 10 attributes are numeric.

For the (reference, logistic, complete) model, the misclassification mean error rate (in blue) was 5.55%. Using all the classifiers of $\mathfrak{C}^*$, the best classifier was the (reference, Student(3), complete) model with a misclassification mean error rate (in green) of 3.67% (see figure 2.7 on the left side). Finally, using all the classifiers of $\mathfrak{C}$, the best classifier was the (reference, Student(1), complete)$_\tau$ model (where $\tau(J) = 1$) with a misclassification mean error rate (in red) of 2.94% (see figure 2.7 on the right side).

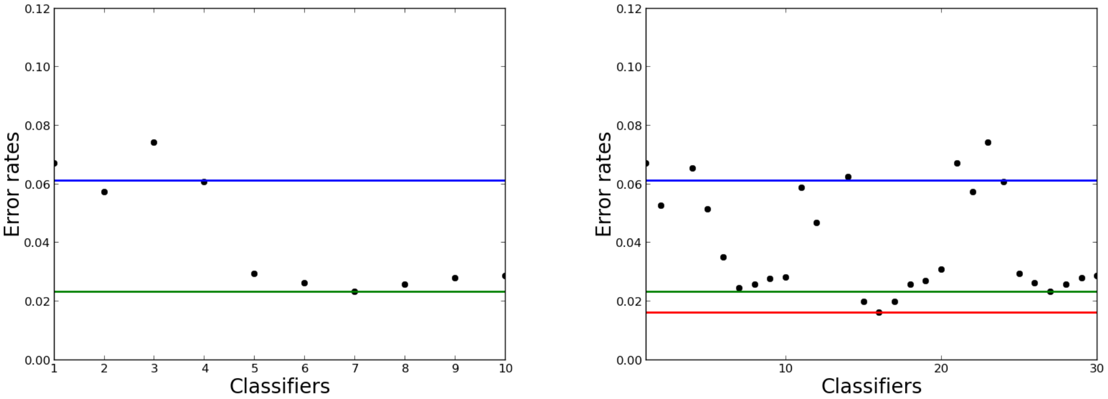

Figure 2.5: Error rates for the classifiers of $\mathfrak{C}^*$ and $\mathfrak{C}$ on the thyroid dataset.

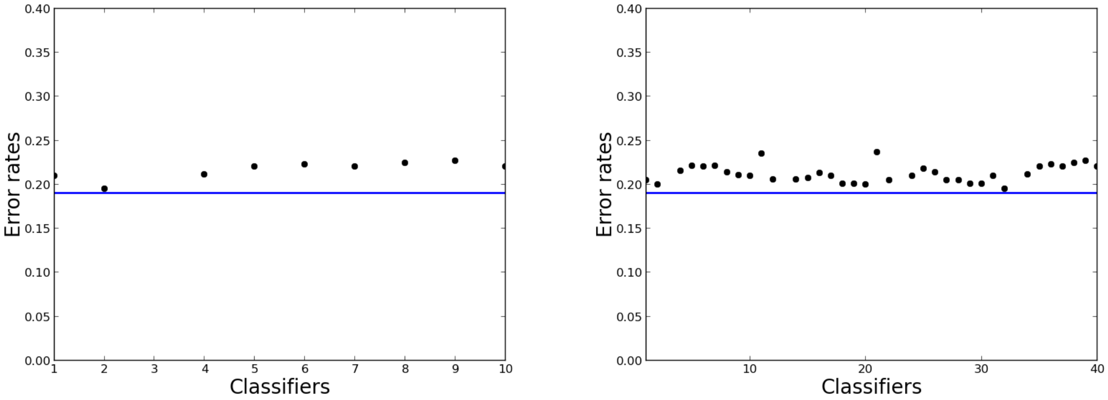

Figure 2.6: Error rates for the classifiers of $\mathfrak{C}^*$ and $\mathfrak{C}$ on the vehicle dataset.

Given the results, the Gumbel min distribution seems to define the worst classifiers. The Normal, Laplace and Gumbel max distributions seem to be comparable to the logistic distribution. Finally, Student distributions seem to outperform the other, perhaps due to their heavy tails.

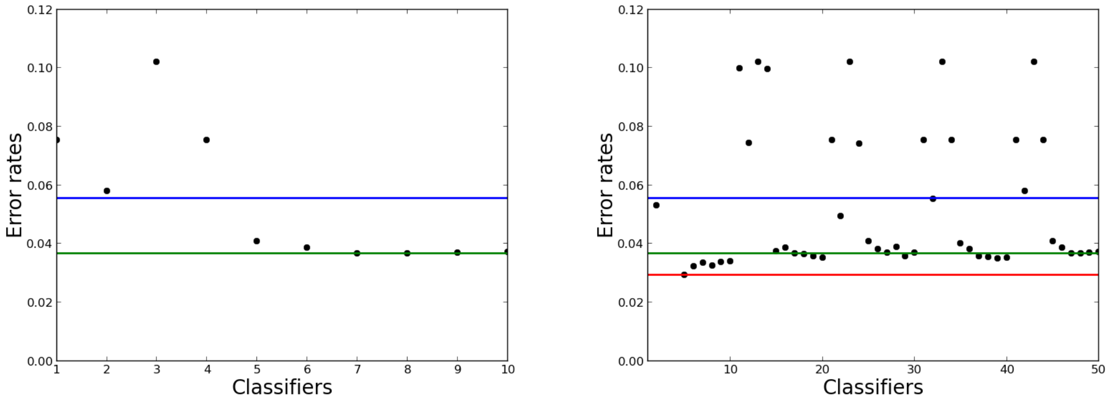

Figure 2.7: Error rates for the classifiers of $\mathfrak{C}^*$ and $\mathfrak{C}$ on the pages-blocks dataset.

### 2.6.2 Partially-known total ordering

Let us first briefly introduce the pear tree dataset. Harvested seeds of Pyrus spinosa were sown and planted in January 2001 in a nursery located near Aix-en Provence, southeastern France. In winter 2001, the first annual shoots on the trunk of 50 one-year-old individuals were described by node. The presence of an immediate axillary shoot was noted at each successive node. Immediate shoots were classified in four categories according to length and transformation or not of the apex into spine (i.e. definite growth or not). The final dataset was thus constituted of 50 bivariate sequences of cumulative length 3285 combining a categorical variable $Y$ (type of axillary production selected from among latent bud (l), unspiny short shoot (u), unspiny long shoot (U), spiny short shoot (s) and spiny long shoot (S)) with an interval-scaled variable $X$ (internode length).

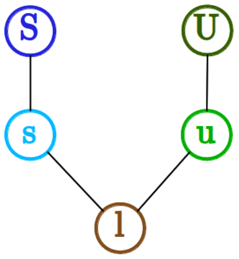

Figure 2.8: Hasse diagram of order relationships $l < u < U$ and $l < s < S$.

We sought to explain the type of axillary production depending on the internode length, using only partial information about category ordering. In fact, the three categories $l$, $u$ and $U$ (respectively $l$, $s$ and $S$) are ordered. But the two mechanisms of elongation and transformation into spine are difficult to compare. Thus the order among the five categories was only partially known, summarized by the Hasse diagram in figure 2.8. However, total ordering among the five categories was assumed and we attempted to recover it. We therefore tested all permutations of the five categories such that $l < u < U$ and $l < s < S$ (i.e. only $4!/2!2! = 6$ permutations). Since axillary production may be interpreted as a sequential mechanism, we chose to use the sequential ratio.

The design matrix was first selected using BIC rather than AIC since we sought to explain

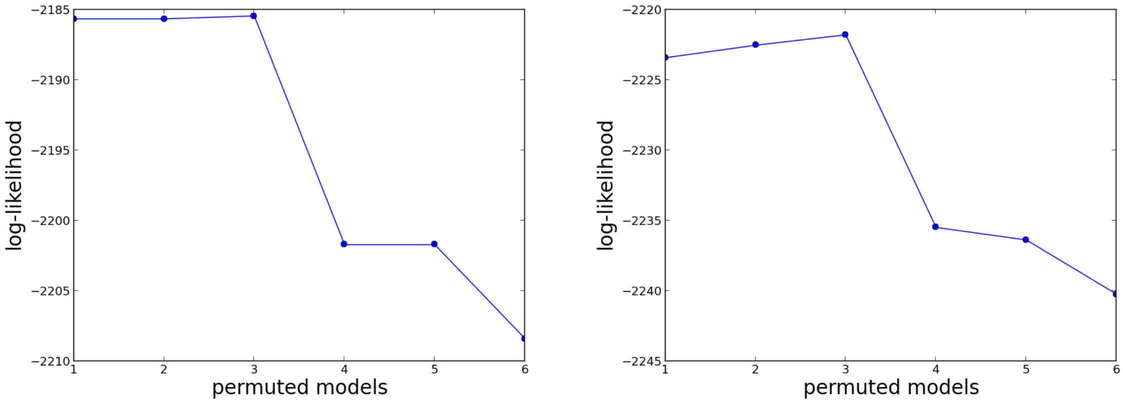

Figure 2.9: Log-likelihood of the (sequential, logistic, complete)$_\sigma$ models on the left and the (sequential, Gumbel max, complete)$_\sigma$ models on the right for the six permutations $\sigma$ that preserve the order relationships.

axillary production rather than predict it. We compared the (sequential, logistic, complete)$_\sigma$ and (sequential, logistic, proportional)$_\sigma$ models for the six permutations $\sigma$: {l, u, s, U, S}, {l, u, s, S, U}, {l, u, U, s, S}, {l, s, u, S, U}, {l, s, u, U, S} and {l, s, S, u, U}. The complete design matrix was selected in all cases. We compared the six permuted (sequential, logistic, complete)$_\sigma$ models using the log-likelihood as criterion (see figure 2.9 on the left side). The third permutation $\sigma^*$ was the best, but the corresponding log-likelihood was very similar to the first two. Since models 1 and 2 (respectively 4 and 5) had exactly the same log-likelihood (illustrating Property 14), they could not be differentiated. To differentiate all the permuted models we used a non-symmetric cdf $F$ (such as Gumbel min or Gumbel max), because Property 14 of invariance is no longer valid. The best result was obtained with the Gumbel max cdf, summarized in figure 2.9 on the right side. The third permutation $\sigma^*$ was still the best: {l, u, U, s, S}. Furthermore, a huge difference appeared between the first three permuted models and the last three. Therefore, the unspiny short shoot (u) seems to occur before the spiny short shoot (s).

## 2.7 Discussion

GLMs for categorical data are better characterized using the $(r, F, Z)$ specification. The differences and commonalities between models can in this way be highlighted. In fact models are easily compared using the three components: ratio $r$ for structure, cdf $F$ for fit, and design matrix $Z$ for parametrization. Moreover, using the proposed decomposition of the link function, Fisher's scoring algorithm is directly derived for any $(r, F, Z)$ triplet. As a by-product of the proposed framework we extended the multinomial logit model and defined a new family of models for nominal data, comparable to the three classical families of models for ordinal data. Finally, we highlighted the properties of models that incorporate the logistic cdf, and also the practical relevance of Student cdfs.

Using the stability properties we propose to classify models along a nominal/ordinal scale (see figure 2.10). We have seen that the nominal/ordinal nature of a model is given by its link function, especially its ratio. The cdf $F$ only affects model fit, except for the logistic distribution which induces some particular properties. We detail in the following how each family of models is positioned along the nominal/ordinal scale.

Focusing on each ratio definition, we see that adjacent, cumulative and sequential ratios are defined using the category ordering assumption, whereas the reference ratio is defined without this assumption. Thus, we obtain a coarse classification with reference models on the nominal side and the others on the ordinal side. We will focus now on the properties of the different models in order to refine this classification. We have seen that the family of canonical models $\{(\text{reference, logistic}, Z); Z \in \mathfrak{Z}\}$ is stable under all permutations. This family is consequently the most appropriate for nominal data. Without the logistic distribution, the choice of the reference category has a marked incidence on model behaviour since one of the categories plays a specific role. The complementary family $\{(\text{reference}, F, Z); F \in \mathfrak{F}, F \neq \text{logistic}, Z \in \mathfrak{Z}\}$ is thus more distant from the nominal end. The family $\{(\text{adjacent, logistic}, Z); Z \in \mathfrak{Z}\}$ is *a priori* defined for ordinal data but is exactly the family of canonical models and is thus appropriate for nominal data. Therefore this family cannot be positioned precisely along the nominal/ordinal scale. By contrast, the complementary family $\{(\text{adjacent}, F, Z); F \in \mathfrak{F}, F \neq \text{logistic}, Z \in \mathfrak{Z}\}$, being stable (apparently only) under the reverse permutation, is thus appropriate for ordinal data. For the same reason, the family of cumulative models is positioned close to the ordinal end. Finally, the family of sequential models do not share the stability under the reverse permutation and thus is more distant from the ordinal end.

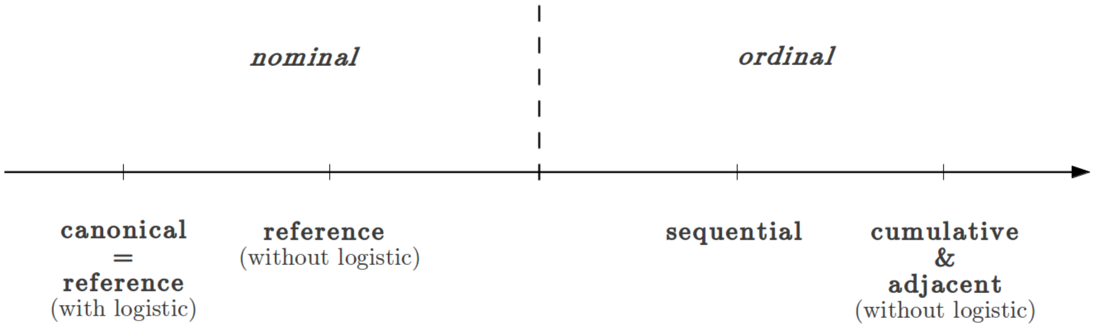

Figure 2.10: Classification of GLMs along a nominal/ordinal scale based on their stability properties under category permutations.

Given this classification, we would *a priori* recommend cumulative and adjacent (without logistic distribution) models rather than sequential models, for ordinal data. But cumulative models are not always defined and thus some problems may occur in the application of Fisher's scoring algorithm. Adjacent models are always defined but are difficult to interpret. From this point of view, sequential models should be favoured since they are always defined and are easy to interpret. The difference in invariance properties between sequential and cumulative models might be explained by the different ordering interpretations. Sequential models correspond to process ordering, whereas cumulative models correspond to scale ordering. An ordinal scale can be reversed whereas reversing a process may change its nature.

Finally, the systematic use of the logistic cdf is not really justified since adjacent models are not appropriate for ordinal data with this cdf. Likewise alternative cdfs can be used to define new models for nominal data and this gives importance to the choice of reference category.

Chapter 3

# Partitioned conditional generalized linear models for categorical data

## Abstract

In categorical data analysis, several regression models have been proposed for hierarchically-structured response variables, e.g. the nested logit model (McFadden et al., 1978). But they have been formally defined for only two or three levels in the hierarchy. Here, we introduce the class of partitioned conditional generalized linear models (PCGLMs) defined for any numbers of levels. The hierarchical structure of these models is fully specified by a partition tree of categories. Using the genericity of the $(r, F, Z)$ specification, the PCGLM can handle nominal, ordinal but also partially-ordered response variables.

**Keywords:** hierarchically-structured categorical variable, partition tree, nominal variable, ordinal variable, partially-ordered variable, GLM specification.

## Contents

## 3.1 Introduction

Categorical data are often based on a hierarchical structure. Although this may seem natural for partially-ordered or even ordinal data, it still makes sense for nominal data. Several partitioned conditional regression models have been proposed in different applied fields, including econometrics, medicine and psychology. The most well-known is the nested logit model, introduced by McFadden et al. (1978) in econometrics for qualitative choice (i.e. nominal categories). In the same field, Morawitz and Tutz (1990) introduced the two-step model to take account of hierarchy among ordinal choices. This model is also used in medicine when ordered categories can be decomposed into coarse and fine scales (Tutz, 1989). The partitioned conditional model for partially-ordered set (POS-PCM) was introduced in medicine by Zhang and Ip (2012).

Compared to simple regression models for categorical data, e.g. the multinomial logit and the odds proportional logit models, partitioned conditional models capture several latent mechanisms. The event $\{Y = j\}$ is decomposed into several steps corresponding to the latent hierarchical structure, with these steps being potentially influenced by different explanatory variables. This approach leads to more flexible models with often a better fit and an easier step-by-step interpretation. In this chapter we introduce the directed trees as the main tool used to formalize the hierarchical structure among categories.

Until now, partitioned conditional models have been formally defined only for two or three levels in the hierarchy. Furthermore, they all assume that the hierarchical structure among the categories is *a priori* known. Our first contribution is to use directed trees to specify the hierarchical structure. This enables us to define partitioned conditional models for an arbitrary number of levels. Moreover, using the genericity of the $(r, F, Z)$ specification (see chapter 2), we develop an extended class of partitioned conditional models for nominal, ordinal but also partially-ordered data. Finally, in the case of ordinal data, instead of considering that the hierarchical structure is known *a priori*, we propose to recover it.

The $(r, F, Z)$ specification of a GLM for categorical data is reviewed in section 3.2, and partition trees are defined. We use these two main building blocks to define and estimate the class of partitioned conditional GLMs.

Sections 3.3, 3.4 and 3.5 extend three existing hierarchically-structured models, revisiting them with the proposed partitioned conditional GLM framework. These three sections focus respectively on the nested logit model for nominal data, the two-step model for ordinal data and the POS-PCM for partially-ordered data. Section 3.4 also describes a model selection procedure for ordinal data, derived from the indistinguishability procedure described by Anderson (1984), which selects the partition tree and at the same time the explanatory variables.

This procedure is illustrated in section 3.6 using the back pain prognosis example previously analysed by Anderson (1984). Our methodology for partially-ordered data is then illustrated using the pear tree example.

## 3.2 Partitioned conditional GLMs

This section briefly outlines the $(r, F, Z)$ specification of a GLM for categorical data and its estimation. The partition tree is then defined in order to specify the hierarchical structure among categories. Finally, we introduce the class of partitioned conditional GLMs and describe the estimation of such models.

### 3.2.1 *(r,F,Z)* specification of GLM for categorical data

The definition of a GLM includes the specification of a link function $g$ which is a diffeomorphism from $\mathcal{M} = \{\pi \in ]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to an open subset $\mathcal{S}$ of $\mathbb{R}^{J-1}$. This function links the expectation $\pi = E[Y|X{=}x]$ and the linear predictor $\eta = (\eta_1, ..., \eta_{J-1})^t$. It also includes the parametrization of the linear predictor $\eta$, which can be written as the product of the design matrix $Z$ (as a function of $x$) and the vector of parameters $\beta$ (Fahrmeir and Tutz, 2001). All the classical link functions $g = (g_1, \dots, g_{J-1})$ described in the literature (Agresti, 2002; Tutz, 2012), rely on the same structure which we propose to write as

$$g_j = F^{-1} \circ r_j, \quad j = 1, \dots, J-1. \tag{3.1}$$

where $F$ is a continuous and strictly increasing cumulative distribution function (cdf) and $r = (r_1, \dots, r_{J-1})^t$ is a diffeomorphism from $\mathcal{M}$ to an open subset $\mathcal{P}$ of $]0,1[^{J-1}$. Finally, given $x$, we propose to summarize a GLM for a categorical response variable by the $J-1$ equations

$$r(\pi) = \mathcal{F}(Z\beta),$$

where $\mathcal{F}(\eta) = (F(\eta_1), \dots, F(\eta_{J-1}))^T$. In the following, we describe in more detail the components $r$, $F$ and $Z$.

**Ratio $r$ of probabilities:** The linear predictor $\eta$ is not directly related to the expectation $\pi$ but to a particular transformation $r$ of the vector $\pi$ which we call the ratio. In the following we will consider four particular diffeomorphisms. The *adjacent*, *sequential* and *cumulative* ratios are respectively defined by $r_j(\pi) = \pi_j/(\pi_j + \pi_{j+1})$, $r_j(\pi) = \pi_j/(\pi_j + \ldots + \pi_J)$ and $r_j(\pi) = \pi_1 + \ldots + \pi_j$ for $j = 1, \dots, J-1$. They all include an order assumption among categories, corresponding to different motivations. On the other hand the *reference* ratio, defined by $r_j(\pi) = \pi_j/(\pi_j + \pi_J)$ for $j = 1, \dots, J-1$, is devoted to nominal response variables.

**Cumulative distribution function $F$:** The *logistic* and *normal* cdfs are the symmetric cdfs most commonly used, but *Laplace* and *Student* cdfs may also be useful. The *Gumbel min* and *Gumbel max* cdfs are the asymmetric cdfs most commonly used. Playing on the symmetrical or asymmetrical character, and the more or less heavy tails, may markedly improve model fit. In applications, Student distributions are used with small degrees of freedom.

**Design matrix $Z$:** Each linear predictor has the form $\eta_j = \alpha_j + x^t\delta_j$ and the vector of parameters is $\beta = (\alpha_1, \dots, \alpha_{J-1}, \delta_1^t, \dots, \delta_{J-1}^t) \in \mathbb{R}^{(J-1)(1+p)}$ where $p$ is the dimension of the explanatory space $\mathcal{X}$. The model is generally defined without constraint, as this is the case for the multinomial logit model. However some linear equality constraints, called contrasts, may be added for instance between different slopes $\delta_j$. The most common contrast is the equality of all slopes, as in the odds proportional logit model. The corresponding constrained space $\mathcal{C} = \{\beta \in \mathbb{R}^{(J-1)(1+p)} | \delta_1 = \ldots = \delta_{J-1}\}$ may be identified to $\mathbb{R}^{(J-1)+p}$. Finally the contrast space is represented by a design matrix. For example, the *complete* design matrix $Z_c$ (without constraint) of dimension $(J-1) \times (J-1)(1+p)$ has the following form

$$Z_c = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix}.$$

The *proportional* design matrix $Z_p$ (common slope) of dimension $(J-1) \times (J-1+p)$ has the following form:

$$Z_p = \begin{pmatrix} 1 & & & x^t \\ & \ddots & & \vdots \\ & & 1 & x^t \end{pmatrix}.$$

| | |
|---|---|
| *Multinomial logit model*<br>$P(Y=j) = \dfrac{\exp(\alpha_j + x^T\delta_j)}{1+\sum_{k=1}^{J-1}\exp(\alpha_k + x^T\delta_k)}$ | (reference, logistic, complete) |
| *Odds proportional logit model*<br>$\log\left\{\dfrac{P(Y \le j)}{1-P(Y \le j)}\right\} = \alpha_j + x^T\delta$ | (cumulative, logistic, proportional) |
| *Proportional hazard model*<br>*(Grouped Cox Model)*<br>$-\log P(Y > j \mid Y \ge j) = \exp(\alpha_j + x^T\delta)$ | (sequential, Gumbel min, proportional) |
| *Adjacent logit model*<br>$\log\left\{\dfrac{P(Y = j)}{P(Y = j+1)}\right\} = \alpha_j + x^T\delta_j$ | (adjacent, logistic, complete) |

Table 3.1: $(r, F, Z)$ specification of four classical GLMs for categorical data.

The triplet $(r, F, Z)$ will play a key role in the following since each GLM for categorical data will be specified by one of these triplets. Table 3.1 shows the specification of four classical models. This specification eases the comparison of GLMs for categorical response variables. Moreover, it can be used to define an extended set of GLMs for nominal response variables by (reference, $F$, $Z$) triplets, which includes the multinomial logit model.

Finally, a single estimation procedure based on Fisher's scoring algorithm can be applied to all the GLMs specified by an $(r, F, Z)$ triplet. The score function can be decomposed into two parts, where the first, unlike the second, depends on the $(r, F, Z)$ triplet.

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T \frac{\partial \mathcal{F}}{\partial \eta} \frac{\partial \pi}{\partial r}}_{(r,F,Z) \text{ dependent part}} \underbrace{\mathrm{Cov}(Y|X=x)^{-1}\,[y-\pi]}_{(r,F,Z) \text{ independent part}}. \tag{3.2}$$

We only need to evaluate the associated density function values $\{f(\eta_j)\}_{j=1,\ldots,J-1}$ to compute the diagonal Jacobian matrix $\partial\mathcal{F}/\partial\eta$. For details on computation of the Jacobian matrix $\partial\pi/\partial r$ for each ratio, see appendix B.

### 3.2.2 Definition of partitioned conditional GLMs

The main idea is to recursively partition the $J$ categories then specify a conditional GLM at each step. This type of model is therefore referred to as partitioned conditional GLM. Models of this class have already been proposed, e.g. the nested logit model (McFadden et al., 1978), the two-step model (Tutz, 1989) and the partitioned conditional model for partially-ordered set (POS-PCM) (Zhang and Ip, 2012). Our proposal can be seen as a generalization of these three models that benefits from the genericity of the $(r, F, Z)$ specification. In particular, our objective is not only to propose GLMs for partially-ordered response variables but also to differentiate the role of explanatory variables for each partitioning step using different design matrices and different explanatory variables. We are seeking also to formally define the partitioned conditional GLMs for any number of levels in the hierarchy. Hence we need to introduce definitions and notations for directed trees.

**Definition 7.** *A directed tree $\mathfrak{T}$ is said to be a* ***partition tree*** *of $\{1,\ldots,J\}$ if*

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

In the following, $\mathcal{V}^*$ is the set of non-terminal vertices of $\mathfrak{T}$ and for each $v \in \mathcal{V}^*$, $Ch(v) = \{\Omega_1^v, \ldots, \Omega_{J_v}^v\}$ is the set of indexed children of $v$. The children must be indexed because the GLMs are not necessarily invariant under permutation of the response categories (see chapter 2). 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. Also, for each vertex $v$ (except the root), $Pa(v)$ denotes the parent of $v$ and $An^*(v)$ denotes the ancestors set of $v$ except the root.

**Definition 8.** *Let $J \geq 2$ and $1 \leq k \leq J-1$. A* ***k-partitioned conditional GLM of categories $\{1,\ldots,J\}$*** *(k-PCGLM) is specified by*

- *a* ***partition tree*** *$\mathfrak{T}$ of $\{1,\ldots,J\}$ with $card(\mathcal{V}^*) = k$,*
- *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 $v$.*

With this definition, the probability of each category $j$ is then obtained by

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

where $P(Y \in v|Y \in Pa(v), x^{Pa(v)})$ is described by the GLM of $\mathfrak{C}$ associated with vertex $Pa(v)$.

The class of PCGLMs for categorical response variables is the set of $k$-PCGLMs for $1 \leq k \leq J-1$. The boundary cases are classical GLMs. For instance for $k = 1$ (see figure 3.1 on the left), the root is the only non-terminal vertex of $\mathcal{T}$, thus we have a classical GLM for categories $\{1,\ldots,J\}$. For $k = J-1$ (see figure 3.1 on the right), $\mathcal{T}$ is a binary tree and

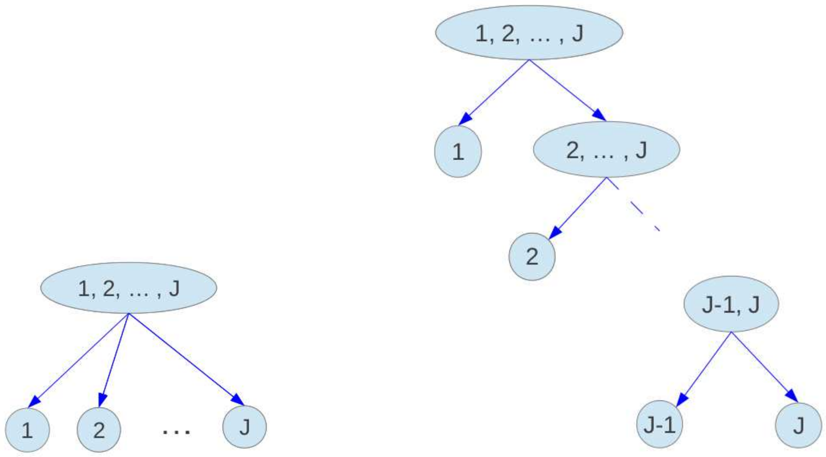

Figure 3.1: 1-partition tree and $(J-1)$-partition tree.

thus $\mathfrak{C}$ is a collection of $J-1$ GLMs for binary response variables. In this case all the ratios are the same. With common cdf $F$ and explanatory variables $x$ for each vertex $v \in \mathcal{V}^*$, the $(J-1)$-PCGLM is exactly the (sequential, $F$, complete) GLM.

There are exactly $J-1$ independent equations to define a simple GLM for categorical data. As noticed by Zhang and Ip (2012), we must check the identifiability of the PCGLMs.

**Proposition 2.** *Let $J \geq 2$ and $1 \leq k \leq J-1$. There are exactly $J-1$ independent equations for any $k$-PCGLM of categories $\{1, \dots, J\}$.*

*Proof.* The cardinal of a set $v \in \mathcal{V}$ is denoted by $|v|$. For each vertex $v \in \mathcal{V}^*$, $\mathcal{M}^v$ denotes the associated GLM and $\mathcal{M}_v$ the PCGLM associated with the sub-tree pruned at vertex $v$. Finally $|\mathcal{M}|$ denotes the number of independent equations of $\mathcal{M}$. Here we are reasoning recursively on $k$, the cardinal of $\mathcal{V}^*$.

- **Initialisation** For $k = 1$, the 1-PCGLM of categories $\{1, \dots, J\}$ turns out to be a simple GLM for categorical data and we obtain the desired result.

- **Recursion** For $k < J-1$, let us assume that, considering any subset $v$ of $\{1, \dots, J\}$, all the $m$-PCGLMs of $v$, such that $m \leq k$, contain exactly $|v|-1$ independent regression equations.

  Now, let $\mathcal{M}$ be a $(k+1)$-PCGLM of $\{1, \dots, J\}$. Noting $r$ the root node, we obtain the following decomposition:

  $$|\mathcal{M}| = |\mathcal{M}^r| + \sum_{v \in Ch(r) \cap \mathcal{V}^*} |\mathcal{M}_v|$$

  Since the root model $\mathcal{M}^r$ is a GLM of the root's children, then $|\mathcal{M}^r| = |Ch(r)| - 1$. Since each model $\mathcal{M}_v$ is a $m$-PCGLM of $v$ such that $m \leq k$, we can use the recursive assumption and obtain $|\mathcal{M}_v| = |v| - 1$. Therefore, the number of independent equations of $\mathcal{M}$ is

$$\begin{aligned}
|\mathcal{M}| &= |Ch(r)| - 1 + \sum_{v \in Ch(r) \cap \mathcal{V}^*} (|v| - 1) \\
&= |\mathcal{C}(v^*)| - 1 + \sum_{v \in Ch(r)} (|v| - 1) \\
&= -1 + \sum_{v \in Ch(r)} |v| \\
|\mathcal{M}| &= J - 1.
\end{aligned}$$

□

A PCGLM is fully specified by the partition tree $\mathfrak{T}$ and the associated collection $\mathfrak{C}$ of GLM(s) and a GLM for categorical data by the $(r, F, Z)$ triplet. Thus, we will specify a PCGLM by its graphical representation, with each non-terminal vertex being labelled by an $(r, F, Z)$ triplet (see figure 3.10 for example). In the case of a minimal response model (i.e. without explanatory variables), the component $r$ and $F$ do not play any role and therefore no label is given.

### 3.2.3 Estimation of PCGLMs

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

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

where $l^v$ represents the log-likelihood of $\mathcal{M}^v$. The maximisation of the log-likelihood with respect to $\{\beta^v\}_{v \in \mathcal{V}^*}$ depends on possible constraints on parameters $\beta^v$ for each vertex $v \in \mathcal{V}^*$. We can differentiate two kinds of model hypothesis, the first being the most common.

**First hypothesis:** $\beta^v \neq \beta^{v'}$ $\forall (v, v') \in \mathcal{V}^* \times \mathcal{V}^*$
Each component $l^v$ can be maximised individually since GLMs attached to non-terminal vertices do not share common regression coefficients. Thus, each $(r^v, F^v, Z^v(x^v))$ model, corresponding to the sub-dataset $\{(y, x^v) | \; y \in v\}$, can be estimated separately using the procedure described in subsection 3.2.1 (see chapter 2 for more details). The score $\partial l / \partial \beta = \partial \eta / \partial \beta \; \partial l / \partial \eta$ has a block structure, as illustrated considering only the two vertices $v$ and $v'$

$$\left[\begin{array}{c|c} \{Z^v(x^v)\}^t & \\ \hline & \{Z^{v'}(x^{v'})\}^t \end{array}\right] \left[\begin{array}{c} \dfrac{\partial l^v}{\partial \eta^v} \\ \hline \dfrac{\partial l^{v'}}{\partial \eta^{v'}} \end{array}\right].$$

**Second hypothesis:** $\exists v \neq v' \in \mathcal{V}^* |\ \beta^v = \beta^{v'}$
In this case we assume not only that explanatory variables are the same for these two nodes, but also that $|Ch(v)| = |Ch(v')|$. This corresponds to particular models that are appropriate in very few practical situations. Such a situation is shown in section 3.5.2.1. Score computation is almost the same as in the previous case, only the design matrix has to be changed and is no longer defined as a diagonal block matrix, as illustrated considering only the two vertices $v$ and $v'$

$$\begin{bmatrix} \hline \{Z^v(x^v)\}^t \\ \hline \{Z^v(x^v)\}^t \\ \hline \end{bmatrix} \begin{bmatrix} \hline \dfrac{\partial l^v}{\partial \eta^v} \\ \hline \dfrac{\partial l^{v'}}{\partial \eta^{v'}} \\ \hline \end{bmatrix}.$$

## 3.3 PCGLMs for nominal data

### 3.3.1 PCGLM specification of the nested logit model

The most well known partitioned conditional model for nominal data is the nested logit model defined by McFadden et al. (1978) in the framework of individual choice behaviour. This model was introduced in order to avoid the inconsistency of the independence of irrelevant alternatives (IIA) property in some situations. Let us illustrate this inconsistency using the classical example of blue and red buses (Debreu, 1960). Assume we are interested in the urban travel demand, with the simple situation of two alternatives: $A = \{\text{blue bus, car}\}$. Suppose that the consumer has no preference between the two alternatives; this means that $P_A(\text{blue bus}) = P_A(\text{car}) = 1/2$. Suppose now that the travel company adds some red buses and the consumer again has no preference between blue and red buses; this means that $P_B(\text{blue bus}) = P_B(\text{red bus})$ where $B = \{\text{blue bus, red bus, car}\}$. Using the IIA property we obtain

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

Finally we obtain $P_B(\text{blue bus}) = P_B(\text{red bus}) = P_B(\text{car}) = 1/3$, whereas we expected the probabilities $P_B(\text{blue bus}) = P_B(\text{red bus}) = 1/4$ and $P_B(\text{car}) = 1/2$.

In this example the IIA property is not appropriate because two alternatives are very similar and also share many characteristics. The nested logit model captures the similarities between close alternatives by partitioning the choice set into “nests” (groups). Thus, the consumer chooses first between bus and car according to price, travel time, ... and secondly between the two buses according to preferred color. More generally, suppose that alternatives can 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. In the following, the nested logit model is presented with only two levels. Let $L$ be the number of nests obtained by partitioning the set of $J$ alternatives.

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

If $j$ denotes an alternative belonging to the nest $N_l$, then the probability of alternative $j$ is decomposed as follows

$$P(Y=j|x)=P(Y=j|Y\in N_l;x^l)P(Y\in N_l|x^0,IV), \tag{3.3}$$

where $IV=(IV_1,\ldots,IV_L)$ denotes the vector of *inclusive values* described thereafter, $x^0$ are the attributes which influence only the first choice level between nests and $x=(x^0,x^1,\ldots,x^L)$. Each probability of the product (3.3) is determined by a multinomial logit model as follows

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

and

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

where

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

The deterministic utilities (predictors) $\eta_j^l$ are function of attributes $x^l$ and $\eta_l^0$ are function of attributes $x^0$. In practice they are linear with respect to $x$. In some situations the attribute values depend on the alternative. For example, the travel price $x_j$ depends on the $J$ alternatives bus, car, metro, etc. In this case, the conditional logit model was introduced by McFadden (1974), using the linear predictors $\eta_j=\alpha_j+x_j^t\delta$ for $j=1,\ldots,J$.

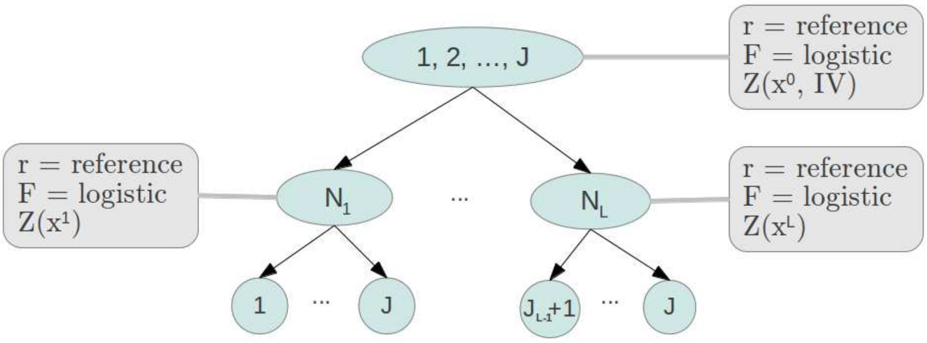

Figure 3.2: PCGLM specification of the nested logit model.

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 $\beta^l$ are different in each nest. The inclusive values $IV_l$ of each nest can then be computed and used, in a second step, to estimate the first level model. More precisely, the parameter $\beta^0$ of the first level is estimated using the design matrix

$$Z(x^0,IV)=\begin{pmatrix} 1 & & x^{0t} & & IV_1 & \\ & \ddots & & \ddots & & \ddots \\ & 1 & & x^{0t} & & IV_{L-1} \end{pmatrix}$$

for a multinomial logit model and the design matrix

$$Z(x^0, IV) = \left( \begin{array}{ccccccc} 1 & & & \tilde{x}_1^{0t} & IV_1 & & \\ & \ddots & & \vdots & & \ddots & \\ & & 1 & \tilde{x}_{L-1}^{0t} & & & IV_{L-1} \end{array} \right)$$

for a conditional logit model, where $\tilde{x}_l^0 = x_l^0 - x_L^0$ for $l = 1, \ldots, L-1$. Finally, the nested logit model is fully specified by the PCGLM in figure 3.2.

### 3.3.2 PCGLMs for qualitative choices

It has been shown that the nested logit model can be considered as a random utility model (RUM) if and only if $0 < \lambda_l \leq 1$ for $l = 1, \ldots, L$ (McFadden et al., 1978). The particular case of $\lambda_l = 1$ leads to the simple multinomial logit model. If the random utility maximisation assumption is relaxed, the model becomes more flexible. The case $\lambda_l = 0$ leads to a particular PCGLM for nominal data with different explanatory variables for each node. Therefore we propose a flexible PCGLM for qualitative choices (see figure 3.3), similar to the nested logit model (without IIA property) but which is not a RUM. We thus avoid difficulties of parameter $\lambda_l$ interpretation and estimation. Moreover, different link functions can be used for each node. The reference ratio must be used because the data are nominal (see chapter 2) whereas any cdf $F$ can be chosen. The reference category can also be changed to obtain a better fit (see chapter 2). Finally, a PCGLM for qualitative choice is specified by

- a partition tree $\mathfrak{T}$ such that the alternatives are aggregated when they share attributes (like for the nested logit model),

- a collection $\mathfrak{C}$ of reference models.

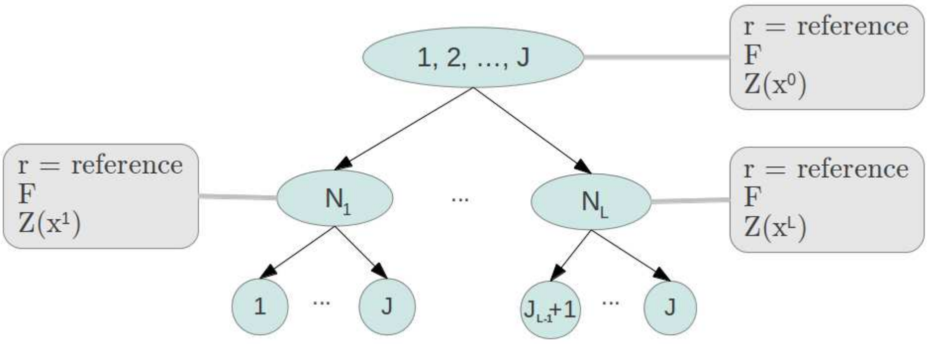

Figure 3.3: PCGLM for qualitative choices.

## 3.4 PCGLMs for ordinal data

### 3.4.1 PCGLM specification of the two-step model

The two-step model, or compound model, was defined by Tutz (1989) in order to decompose the latent mechanism of an ordinal response into two levels. Ordinal-scale response variables

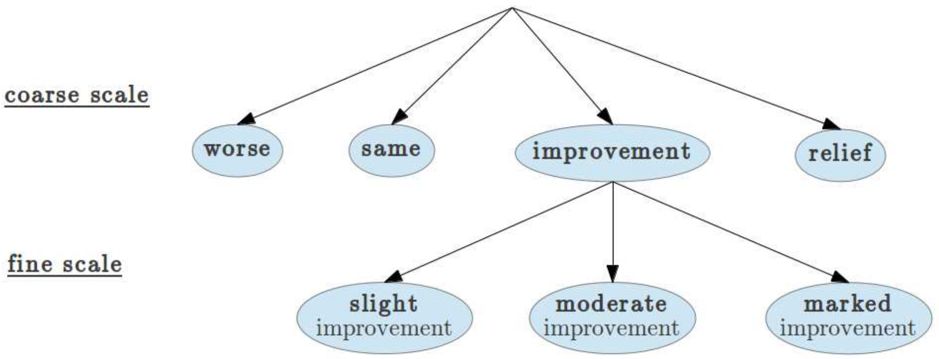

Figure 3.4: Two-scale back pain assessment.

are commonly used in medicine and psychology for instance, to assess a patient's condition. This ordinal scale is often built from a coarse and a fine scale.

For the back pain prognosis dataset described by Doran and Newell (1975), the response variable $y$ is the assessment of back pain after three weeks of treatment using the six ordered categories: worse (1), same (2), slight improvement (3), moderate improvement (4), marked improvement (5), complete relief (6). Categories 3, 4 and 5 can be aggregated into a general category *improvement*. Thus, the coarse scale corresponds to the categories: worse, same, improvement, complete relief, and the fine scale corresponds to the categories: slight improvement, moderate improvement and marked improvement (see figure 3.4).

The model can be decomposed into two levels. More precisely, the cumulative (respectively sequential) two-step model is exactly a $k$-PCGLM (see figure 3.5) with

- a partition tree $\mathcal{T}$ of depth 2 which respects the order assumption,
- a collection of $k$ (cumulative, $F_0$, proportional) models with common cdf $F_0$ (respectively (sequential, $F_0$, proportional)).

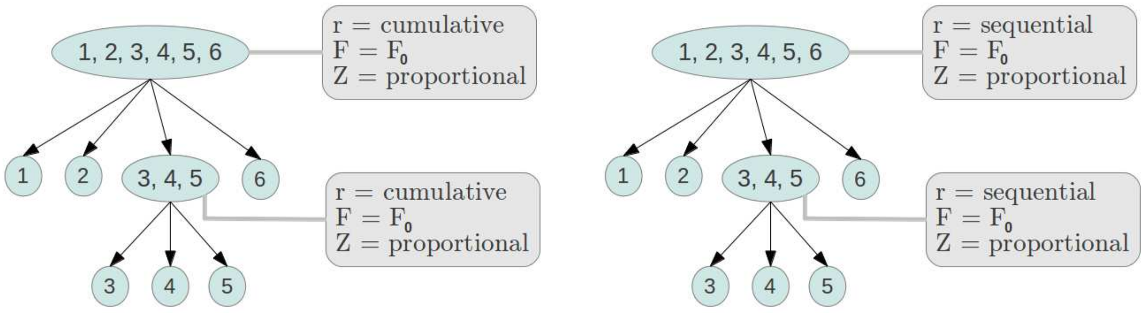

Figure 3.5: PCGLM specification of cumulative and sequential two-step models for the back pain prognosis example.

The two-step model can be extended in different ways. A partition tree with a depth of more than two can be used, providing that ordering among categories is conserved. Furthermore different link functions can be used for each non-terminal node, providing they are appropriate for ordinal data. The (adjacent, $F$, $Z$) models, with $F \neq$ logistic, can be used (see chapter 2 for details).

### 3.4.2 Indistinguishability of response categories

Anderson (1984) introduced the stereotype model derived from the classical multinomial logit model

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

using different parametrizations for the slopes $\delta_j$. For instance, he defined the one-dimensional stereotype model using the particular parametrization of slopes

$$\delta_j = \phi_j\delta,$$

for $j = 1, \ldots, J-1$, where $\phi_j$ are scalars and $\delta$ is a vector.

#### 3.4.2.1 Original Anderson's indistinguishability procedure

Anderson (1984) proposed a testing procedure - useful for ordinal data - to identify successive categories that can be clearly distinguished by the explanatory variables $x$. These categories are said to be indistinguishable with respect to $x$ when the explanatory variables $x$ do not have significantly different effects on them. He proposed to aggregate the corresponding successive slope parameters $\delta_j$ and use a deviance test. More precisely he proposed an iterative procedure to locate the best splits between the categories $1, \ldots, J$ with respect to $x$. The minimal number of splits is zero, corresponding to the simple model without explanatory variable (null hypothesis $H_0$), and the maximal number of splits is $J-1$, corresponding to the classical multinomial logit model with $J-1$ different slopes.

The first step is to locate the best partition into two groups of categories. The hypothesis $H_{(2;r)}$ is then introduced

$$H_{(2;r)} : \delta_1 = \ldots = \delta_r; \quad \delta_{r+1} = \ldots = \delta_J = 0,$$

for $r = 1, \ldots, J-1$. Comparing the corresponding log-likelihood values $l_{(2;r)}$ yields the best splitting point $r^*$ such that $l_2 = l_{(2;r^*)} = \max_r l_{(2;r)}$. The hypothesis $H_{(2;r^*)}$ is tested against $H_0$, using the deviance statistic $2(l_2 - l_0)$ which follows a $\chi^2_p$ distribution under $H_0$. Finally, if the splitting point $r^*$ is accepted, the procedure must be restarted in parallel for the two groups $\{1, \ldots, r^*\}$ and $\{r^*+1, \ldots, J\}$ in order to obtain the best partition into three groups. For example, the procedure is restarted on group $\{r^*+1, \ldots, J\}$ and the hypothesis $H_{(3;r^*,s)}$ is tested

$$H_{(3;r^*,s)} : \delta_1 = \ldots = \delta_{r^*}; \quad \delta_{r^*+1} = \ldots = \delta_s; \quad \delta_{s+1} = \ldots = \delta_J = 0.$$

By comparing the corresponding log-likelihood values of the two procedures in parallel, we obtain the best second splitting point $s^*$ (or respectively $t^*$) such that $l_3 =_{(3;r^*,s^*)}= \max_s l_{(3;r^*,s)}$ (respectively $l_3 =_{(3;t^*,r^*)}= \max_t l_{(3;t,r^*)}$). The hypothesis $H_{(3;r^*,s^*)}$ (or $H_{(3;t^*,r^*)}$) is then tested against $H_{(2;r^*)}$, using the deviance statistic $2(l_3 - l_2)$ which follows a $\chi^2_p$ distribution under $H_{(2;r^*)}$.

This is a dichotomous partitioning procedure with at most $J(J-1)/2$ different parametrizations to test. It should be noted that this procedure is simplified for the one-dimensional stereotype model since the equality between slopes $\delta_1 = \ldots = \delta_r$ becomes equality between scalar parameters $\phi_1 = \ldots = \phi_r$. In practice, only this particular case of the procedure is used.

#### 3.4.2.2 Indistinguishability procedure with $(r, F, Z)$ specification

Here we express the indistinguishability procedure in terms of canonical models by simply changing the design matrix. In fact, the hypothesis $H_{(2;r)}$ corresponds to the canonical (reference, logistic, $Z_r$) model (see chapter 2) with

$$Z_r = \left[\begin{array}{cccc|c} 1 & & & & x^t \\ & \ddots & & & \vdots \\ & & \ddots & & x^t \\ \cline{5-5} & & & \ddots & \\ & & & 1 & \end{array}\right],$$

the design matrix with $r$ repetitions of $x^t$, whereas the null hypothesis $H_0$ corresponds to the $(J-1)$-identity design matrix. If the first splitting point $r^*$ is accepted, the procedure is restarted to test the hypothesis $H_{(3;r^*,s)}$ which corresponds to the (reference, logistic, $Z_{r^*,s}$) model with

$$Z_{r^*,s} = \left[\begin{array}{ccccccc|c|c} 1 & & & & & & & x^t & \\ & \ddots & & & & & & \vdots & \\ & & \ddots & & & & & x^t & \\ \cline{8-9} & & & \ddots & & & & & x^t \\ & & & & \ddots & & & & \vdots \\ & & & & & \ddots & & & x^t \\ \cline{8-9} & & & & & & \ddots & & \\ & & & & & & 1 & & \end{array}\right],$$

the design matrix with $r^*$ repetitions of $x^t$ for the first block and $s - r^*$ repetitions of $x^t$ for the second block. The indistinguishability procedure, specified in terms of the $(r, F, Z)$ triplet, can be seen as a design matrix selection procedure.

#### 3.4.2.3 Indistinguishability procedure with PCGLM specification

Here we express the indistinguishability procedure in terms of PCGLM by simply changing the partition tree. In fact any canonical (reference, logistic, $Z$) model with a block structured design matrix $Z$ is equivalent to a PCGLM of depth 2 with the canonical (reference, logistic, complete) model for the root and minimal response models for other non-terminal nodes. Let us describe this result in detail using the block structured design matrix $Z_{r,s}$.

**Lemma 1.** *The canonical model (reference, logistic, $Z_{r,s}$) is equivalent to the PCGLM specified in figure 3.6.*

*Proof.* Assume that the distribution of $Y|X = x$ is defined by the canonical (reference, logistic, $Z_{r,s}$) model. We thus obtain

$$\frac{\pi_j}{\pi_J} = \begin{cases} \exp(\alpha_j + x^t\delta_1), & 1 \leq j \leq r, \\ \exp(\alpha_j + x^t\delta_2), & r < j \leq s, \\ \exp(\alpha_j), & s < j \leq J-1. \end{cases} \tag{3.4}$$

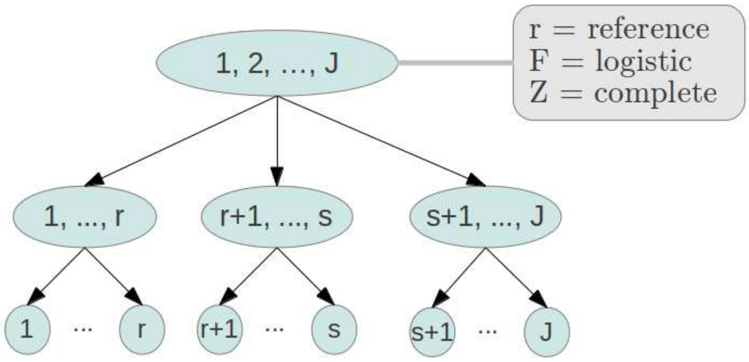

Figure 3.6: PCGLM specification of indistinguishability hypothesis $H_{(3,r,s)}$.

Let $\mathfrak{T}$ denote the partition tree of figure 3.6 and $\Omega_1$, $\Omega_2$ and $\Omega_3$ the children of the $\mathfrak{T}$'s root. We thus obtain

$$\frac{\pi_{\Omega_1}}{\pi_{\Omega_3}} = \frac{\pi_1 + \ldots + \pi_r}{\pi_{s+1} + \ldots + \pi_J}.$$

Using equalities (3.4), we obtain

$$\frac{\pi_{\Omega_1}}{\pi_{\Omega_3}} = \frac{\left\{\sum_{j=1}^{r} \exp(\alpha_j + x^t\delta_1)\right\}\pi_J}{\left\{1 + \sum_{j=s+1}^{J-1} \exp(\alpha_j)\right\}\pi_J},$$

and thus

$$\frac{\pi_{\Omega_1}}{\pi_{\Omega_3}} = \exp(\alpha'_1 + x^t\delta'_1),$$

using the following parametrization

$$\begin{cases} \alpha'_1 = \log\left\{\dfrac{\sum_{j=1}^{r}\exp(\alpha_j)}{1 + \sum_{j=s+1}^{J-1}\exp(\alpha_j)}\right\}, \\ \delta'_1 = \delta_1. \end{cases}$$

Similarly, we obtain $\pi_{\Omega_2}/\pi_{\Omega_3} = \exp(\alpha'_2 + x^t\delta'_2)$ with the parametrization

$$\begin{cases} \alpha'_2 = \log\left\{\dfrac{\sum_{j=r+1}^{s}\exp(\alpha_j)}{1 + \sum_{j=s+1}^{J-1}\exp(\alpha_j)}\right\}, \\ \delta'_2 = \delta_2. \end{cases}$$

Therefore, the root model is exactly the canonical (reference, logistic, complete) model. We want to ensure that we have a minimal response model for each non-terminal vertex of the second level. For the non-terminal vertex $\Omega_1 = \{1, \ldots, r\}$, we have

$$\frac{\pi_j}{\pi_r} = \frac{\pi_j}{\pi_J}\frac{\pi_J}{\pi_r} = \exp(\alpha_j + x^t\delta_1)\exp(-\alpha_r - x^t\delta_1) = \exp(\alpha_j - \alpha_r),$$

for $j < r$. These $r-1$ ratios do not depend on $x$ and therefore correspond exactly to the minimal response model. Similarly we have $\pi_j/\pi_s = \exp(\alpha_j - \alpha_s)$ for $r < j < s$ and $\pi_j/\pi_J = \exp(\alpha_j)$ for $s < j < J$. Then, $Y|X = x$ follows exactly the expected PCGLM. As the parametrization is invertible, we obtain the equivalence. □

Using this equivalence, the canonical (reference, logistic, $Z_{r,s}$) model is easily estimated. In fact, we need to transform the data, aggregating the response categories according to the partitioning sets $\Omega_1 = \{1, \ldots, r\}$, $\Omega_2 = \{r+1, \ldots, s\}$ and $\Omega_3 = \{s+1, \ldots, J\}$. We then simply need to estimate the canonical (reference, logistic, complete) model using this new dataset (and also the three minimal response models of vertices $\Omega_1$, $\Omega_2$ and $\Omega_3$).

#### 3.4.2.4 Extended indistinguishability procedure with PCGLM

The indistinguishability procedure specified with PCGLM can be viewed as a partitioning procedure. With this form, we see that the procedure uses the ordering assumption to partition the categories (only successive categories are aggregated) but the root model does not use the ordering assumption among the groups of categories. The canonical (reference, logistic, complete) model is appropriate for nominal categories (see chapter 2). Thus, we can define the same procedure with an ordinal model for the root, such as an adjacent (without logistic cdf), a cumulative, or a sequential model (see chapter 2). Some convergence problems of the Fisher's scoring algorithm may appear for cumulative models because the constraints $\eta_j(x) < \eta_{j+1}(x)$ are more difficult to check with a complete design matrix. Thus, we propose to use the indistinguishability procedure with the (cumulative, logistic, proportional) model to avoid these difficulties. Our procedure is more comparable to Anderson's procedure since he used the stereotype logit model which is often more parsimonious than the multinomial logit model (between proportional and complete design matrices).

Assume that we apply this procedure and we determine the best root partition for the vector $x$ of explanatory variables. We can say that categories of the same non-terminal vertex are indistinguishable with respect to $x$. But what about indistinguishability with respect to a subset of $x$? We therefore propose to select the best subset of $x$ for each non-terminal node. If this subset is non-empty, the procedure is restarted, otherwise the procedure is stopped. A final refinement step is then used to select $F$ in each non-terminal vertex to obtain a better fit. We illustrate this procedure with the back pain prognosis example in section 3.6.1.

## 3.5 PCGLMs for partially-ordered data

### 3.5.1 PCGLM specification of the POS-PCM

In categorical data analysis, the case of nominal and ordinal data has already been investigated in depth while the case of partially-ordered data has been comparatively neglected. Zhang and Ip (2012) introduced the partitioned conditional model for partially-ordered set (POS-PCM). The main idea was to recursively partition the $J$ categories in order to obtain either ordinal or nominal models at each step. Zhang and Ip (2012) then used the odds proportional logit model for the total order case and the multinomial logit model for the no order case.

Zhang and Ip introduced the partially-ordered set theory into the GLM framework. A partially-ordered set (poset) $(P, \preceq)$ is summarized by a Hasse diagram. The order relation $j \preceq k$ is represented by an edge between the two vertices (categories) and vertex $k$ is above vertex $j$. 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. Zhang and Ip defined an algorithm for categories partitioning which gave the following result:

**Property 15.** *(Zhang and Ip, 2012) A finite poset can always be partitioned into antichains that are totally weakly ordered.*

For any poset (with one component), there exists a partition tree $\mathfrak{T}$ of depth 2 such that the siblings of the first level are totally weakly ordered and the siblings of the second level are not comparable. The categories are partitioned according to each level of the Hasse diagram (each level is an antichain). Since the antichains are totally (weakly) ordered between them, Zhang and Ip proposed using the odds proportional logit model. Within each antichain, the categories are not comparable, thus they proposed using the multinomial logit model.

It should be noted that Property 15 holds only if the poset has one component. If there are two or more components, they must first be partitioned. Since these components are not comparable, they form an antichain. Thus, a previous level must be added to separate each component, using the multinomial logit model, and Property 15 must be used for each component. The depth of the partition tree is exactly 2 if the poset has exactly one component, otherwise it is 3. Finally, for any poset, Zhang and Ip (2012) proposed to associate a particular partitioned conditional model. This model is a particular PCGLM with

- A partition tree $\mathfrak{T}$ built from the Hasse diagram.
- A collection $\mathfrak{C}$ which alternates between the ordinal (cumulative, logistic, proportional) model and the nominal (reference, logistic, complete) model.

Figure 3.7 illustrates this association between a poset (equivalently an Hasse diagram) and the POS-PCM.

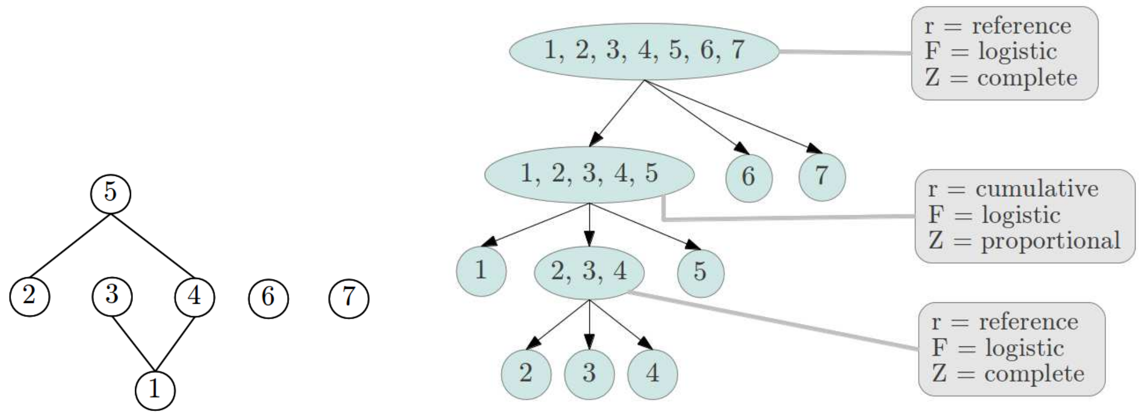

Figure 3.7: Association between an Hasse diagram and a POS-PCM (specified in the PCGLM framework).

## 3.5.2 Inference of PCGLMs for partially-ordered data

### 3.5.2.1 Poset structure and partition tree

Zhang and Ip (2012) used poset structure information to define the POS-PCM. But how is this poset obtained? It is usual to have a nominal or ordinal response variable, but what does a partially-ordered variable mean? In fact, every partially-ordered variable $Y$ can be expressed in terms of elementary ordinal or nominal variables $Y_i$ (with at least one ordinal variable). For example, let $Y = (Y_1, Y_2)$ be a pair of ordinal variables. Let $a$, $b$, $c$ be the ordered categories of $Y_1$, and 1, 2, 3 be the ordered categories of $Y_2$. The ordering relationship for $Y$ depends on the relation between $Y_1$ and $Y_2$.

**$Y_1$ and $Y_2$ are not comparable** In this case the Cartesian product order is used. Let $y$ and $y'$ be two observed responses. The Cartesian product order $\preceq_C$ is defined by

$$y \preceq_C y' \quad \text{if} \quad \left(y_1 \preceq y_1' \text{ and } y_2 \preceq y_2'\right).$$

In this case we can use the Property 15 to obtain the partition tree from the Hasse diagram in figure 3.8.

**$Y_1$ and $Y_2$ are ordered** In this case the lexicographic order has to be used. Assume that $Y_1 \preceq Y_2$ and let $y$ and $y'$ be two observed responses. The lexicographic order $\preceq_L$ is defined by

$$y \preceq_L y' \quad \text{if} \quad \left(y_1 \preceq y_1'\right) \text{ or } \left(y_1 = y_1' \text{ and } y_2 \preceq y_2'\right).$$

In this case the order among the response categories is total. But a 2-partition tree seems to be appropriated: with a first level for $Y_1$ and a second level for $Y_2|Y_1$ (see figure 3.9). The order among latent variables (shown in red) seems to have priority over the order among categories (shown in blue). A common slope $\delta$ can be considered (see section 3.2.3 for parameter estimation) because the same response variable $Y_2$ is involved in all the non-terminal vertices of the second level.

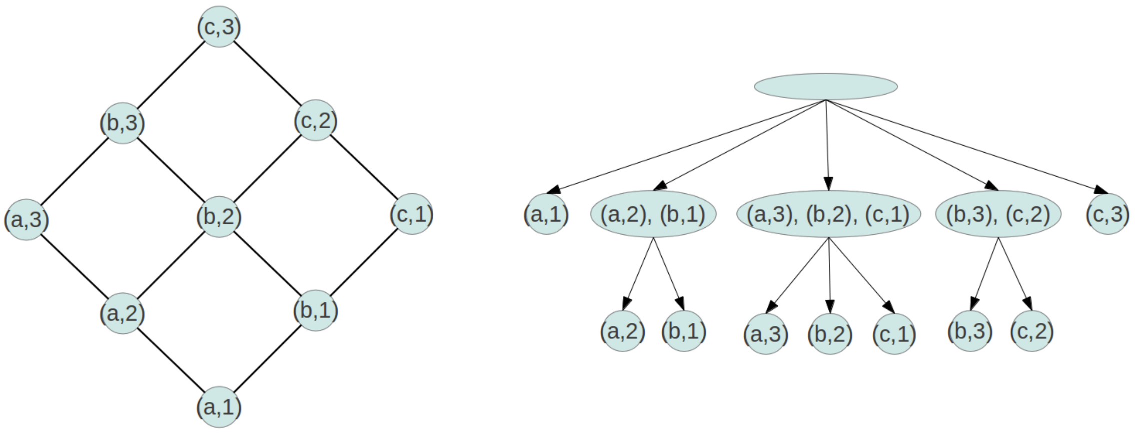

Figure 3.8: Hasse diagram of Cartesian product order and corresponding partition tree.

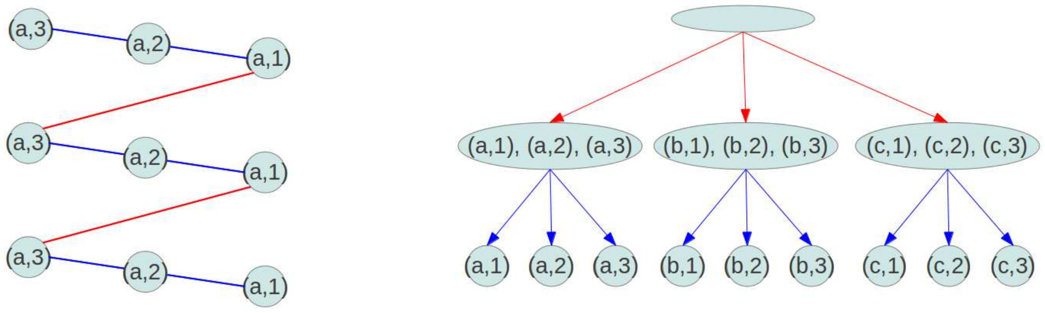

Figure 3.9: Hasse diagram of lexicographic order and corresponding partition tree.

#### 3.5.2.2 Collection of models $\mathfrak{C}$

Given a non-terminal vertex $v$ of $\mathfrak{T}$, we choose an ordinal model if the children of $v$ are totally ordered. Otherwise we choose a nominal model. In this way Zhang and Ip (2012) used the odds proportional logit model in the ordinal case and the multinomial logit model in the nominal case. More generally, we propose to use the families of cumulative, sequential and adjacent (without logistic distribution) models for ordinal data and the family of reference models for nominal data (see chapter 2).

## 3.6 Applications

### 3.6.1 Totally ordered data: back pain prognosis example

Doran and Newell (1975) described a back pain study involving 101 patients. The response variable $y$ was the assessment of back pain after three weeks of treatment using the six ordered categories: *worse* (1), *same* (2), *slight improvement* (3), *moderate improvement* (4), *marked improvement* (5), *complete relief* (6). The three selected explanatory variables observed at the beginning of the treatment period were $x_1$ = length of previous attack (1=short, 2=long), $x_2$ = pain change (1=getting better, 2=same, 3=worse) and $x_3$ = lordosis (1=absent/decreasing, 2=present/increasing).

Here, the response categories are defined by the experimentalist and this ordinal scale may thus not be the most efficient to describe the back pain prognosis of a patient. Firstly, we will use the hierarchy among the categories shown in figure 3.4 and select the best regression model for it. Secondly, we will select the hierarchy and at the same time the explanatory variables, using our extended indistinguishability procedure. We will thus compare the two results and the result obtained by Anderson (1984).

**The case of known partition tree $\mathfrak{T}$**

Here, the partition tree is *a priori* defined with {worse, same, improvement, complete relief} at the first level, with improvement being partitioned into {slight improvement, moderate improvement, marked improvement} at the second level; see figure 3.4. We must select the best GLM for the root of $\mathfrak{T}$ and the non-terminal vertex {slight improvement, moderate improvement, marked improvement}. For these two vertices we have an ordinal scale, thus the most appropriate ratios are adjacent and cumulative. We chose the adjacent ratio in order to avoid algorithm difficulties with the complete design matrix, and the symmetric normal cdf, appropriate for ordinal data (see chapter 2 for details). Since there were at most $K = 3$ explanatory variables, we compared all $2^3$ combinations. Complete and proportional design matrices were tested for each combination. The variable $x_1$ was the only one selected for the two vertices with the complete design matrix. Since this explanatory variable was categorical, the model was exactly the saturated model. Therefore all the link functions were equivalent. Finally, the maximised log-likelihood was $l = -161.14$ for 10 parameters. This partition tree does not seem to be appropriate for the data as only $x_1$ was selected for it, whereas $x_1$, $x_2$, $x_3$ were selected for the canonical 1-partition tree. More precisely, the simple (cumulative, logistic, proportional) model had a log-likelihood of $-159.045$ for 8 parameters, using the three explanatory variables.

### The case of unknown partition tree $\mathfrak{T}$

We will use this dataset to illustrate the extended indinguishability procedure which corresponds to a partition tree and variable selection procedure. Since $\mathfrak{T}$ must respect category ordering, the space of possible partition trees is reduced. During the procedure, only the ordinal (cumulative, logistic, proportional) model and the minimal response model (i.e. without explanatory variable) will be used in the collection $\mathfrak{C}$.

**First level** Note that every PCGLM with only a root proportional model (and minimal response models for other non-terminal nodes) have exactly the same number of parameters: $J-1+p=8$. Thus, we simply use the log-likelihood to compare these models. We begin the procedure with the the simple model $\mathfrak{M}_0$ = (cumulative, logistic, proportional) which can be seen to be a 1-PCGLM. The corresponding log-likelihood is $l_0=-159.046$.

Here we are looking for the best splitting point $r \in \{1,2,3,4,5\}$ for explanatory variables $x_1$, $x_2$ and $x_3$. Note that model $\mathfrak{M}_0$ corresponds exactly to the splitting point $r=J-1=5$ since all the $J-1$ slopes are common in this case. The best model is obtained for $r^*=4$ with log-likelihood $l_{r^*}=-158.132$. Since $l_{r^*}>l_0$, the splitting point $r^*$ is selected. We now look for the best splitting point $s \in \{1,2,3\} \cup \{5\}$ that gives three nodes. The best model is obtained for $s^*=1$ with log-likelihood $l_{s^*,r^*}=-155.756$. Since $l_{s^*,r^*}>l_{r^*}$, the second splitting point $s^*$ is also selected. As every partitions in four groups are rejected, the best root partition is $\{1\} \cup \{2,3,4\} \cup \{5,6\}$ for explanatory variables $x_1$, $x_2$,$x_3$.

**Second level** We now focus on the non-terminal vertices $v_1=\{2,3,4\}$ and $v_2=\{5,6\}$. We first select the subset of influential explanatory variables for these two nodes, using again the simple (cumulative, logistic, proportional) model with the Bayesian Information Criteria (BIC). As previously seen, the different models of collection $\mathfrak{C}$ can be estimated separately because the parameter $\beta_v$ is different for each non-terminal vertex $v$. The explanatory variable $x_2$ is selected for vertex $v_1$ and no variable is selected for vertex $v_2$. For vertex $v_2$, the minimal response model has a log-likelihood $l^{v_2}=-28.841$. Thus, we simply focus on vertex $v_1$ and obtain a log-likelihood $l_0^{v_1}=-54.561$ with the simple (cumulative, logistic, proportional) model, using only $x_2$.

We now look for the best splitting point $t \in \{2,3,4\}$ of vertex $v_1$ for the explanatory variable $x_2$. The best model is obtained for $t^*=3$ with a log-likelihood $l_{t^*}^{v_1}=-54.31$. Since $l_{t^*}^{v_1}>l_0^{v_1}$, the splitting point $t^*$ is selected. This is the last possible partition for the second level of the partition tree.

**Last level and refinement step** There is only the vertex $v_3=\{2,3\}$ at the third level with only the explanatory variable $x_2$. Since we have to reduce the set of explanatory variables, the minimal response model is estimated for this vertex with log-likelihood $l^{v_3}=-21.93$. The selection procedure of the partition tree and the explanatory variables is then stopped. The corresponding log-likelihood is $l=-153.418$ for 9 parameters, with the logistic cdf for each node. We then execute a refinement step by selecting the best cdf $F$ for each vertex and determine the model $\mathcal{M}_*$ (see figure 3.10) with log-likelihood $l_*=-152.727$ for 9 parameters.

Looking at the results obtained in the first part, it can be seen that the categories do not appear to be appropriate for describing back pain. In fact, in the second part, our results are similar to those of Anderson for the first step of the model: i.e. the partition {worse}, {same, little imp., moderate imp.}, {slight imp., complete relief} for the three explanatory variables. He obtained a log-likelihood of $-154.39$ for 9 parameters. Our methodology allows us to go a

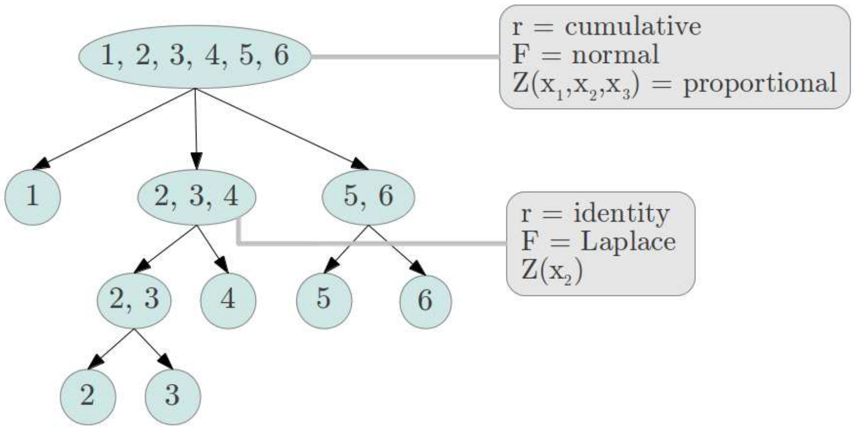

Figure 3.10: PCGLM for back pain prognosis.

step further and find a separation between {same, little imp.} and {moderate imp.} according to pain change $x_2$. Looking at the partition tree in figure 3.10, we propose a new ordinal scale of four categories: worse $= \{1\}$, same $= \{2, 3\}$, improvement $= \{4\}$ and relief $= \{5, 6\}$, which seems to be better suited.

### 3.6.2 Partially-ordered data: pear tree example

The class of PCGLMs for categorical data is so vast that we need a method to determine the structure of the model. We first propose to select the partition tree $\mathfrak{T}$ and then the collection $\mathfrak{C}$ of models. We illustrate this methodology using the pear tree example.

#### Selection of the partition tree $\mathfrak{T}$

Axillary production of the pear tree can be decomposed using three binary unobservable variables $Y_1$, $Y_2$ and $Y_3$. Firstly, the bud either stays in the latent state or becomes a branch ($Y_1 \in$ {latent bud, branching}). If branching occurs, then $Y_2$ denotes the branch elongation ($Y_2 \in$ {short, long}) and $Y_3$ denotes the spiny character of the branch ($Y_3 \in$ {unspiny, spiny}). The variables $Y_2$ and $Y_3$ are clearly conditioned with respect to $Y_1$ because if we have a latent bud, axillary production is over. We chose to use the order relationship to build a partition tree. The variables $Y_1$ and $Y_2$ are naturally ordered, whereas it is not manifest for $Y_3$. Using the Cartesian product order among $(Y_1, Y_2, Y_3)$ we obtain a partial order among $Y$. Depending on whether $Y_3$ is considered as a nominal or an ordinal variable, we obtain two posets structure and thus two Hasse diagrams $\mathfrak{D}_1$ and $\mathfrak{D}_2$ (see figure 3.11). Using Property 15 described by Zhang and Ip (2012), we obtain two corresponding partition trees $\mathfrak{T}_1$ and $\mathfrak{T}_2$ (see figure 3.12).

We now need to select the best partition tree. “It should be noted that the assumption of a logit model on both levels yields a model that is not equivalent to a one-step logit model” (Tutz, 2012). Therefore, we compare these two partition trees with the simple 1-partition tree $\mathfrak{T}_0$. For now, we simply want to compare the different partitioned conditional structure without modelling assumption for each non-terminal node. We therefore use the canonical (reference, logistic, complete) model for the three partition trees since this model is invariant under all permutation (see chapter 2). Thus it is not necessary to test different permuted partition trees. Moreover, the log-likelihood is globally concave for all canonical models and thus we avoid algorithm convergence difficulties. Finally, the three models $\mathfrak{M}_i$, corresponding to each

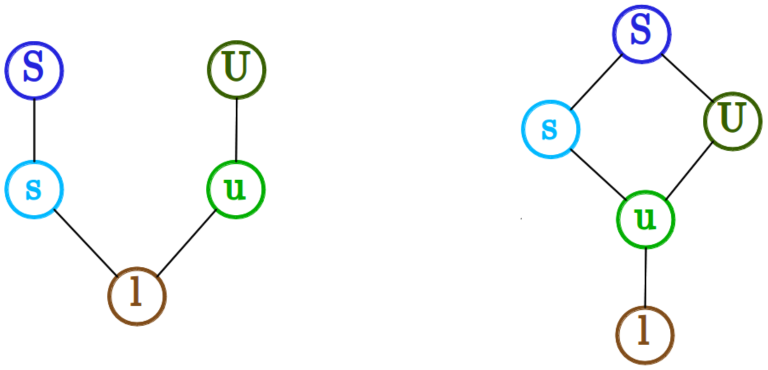

Figure 3.11: Two Hasse diagrams $\mathfrak{D}_1$ and $\mathfrak{D}_2$ for the response categories of the pear tree example.

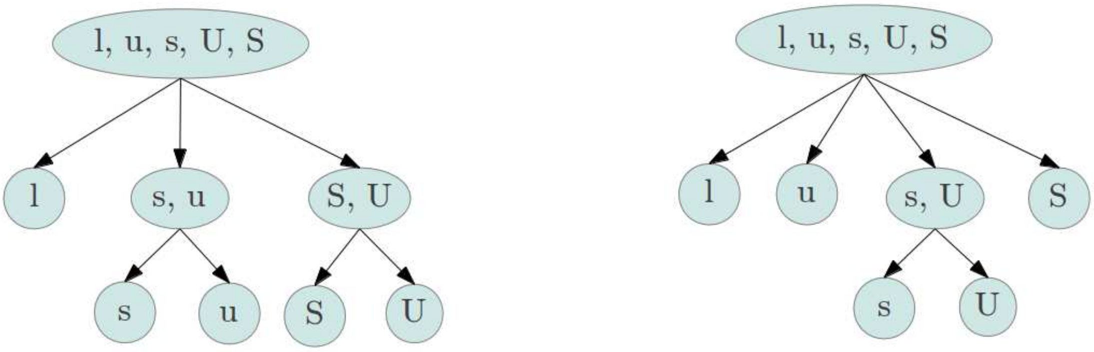

Figure 3.12: Two partition trees $\mathfrak{T}_1$ and $\mathfrak{T}_2$ for the response categories of the pear tree example.

partition tree $\mathfrak{T}_i$ $(i = 0, 1, 2)$, have exactly the same number of parameters $((J-1)(1+p) = 12)$, thus we can use the log-likelihood as criteria. We obtain respectively $l_0 = -2087.42$, $l_1 = -2083.20$, $l_2 = -2089.61$ , selecting $\mathfrak{T}_1$ as the best partition tree.

### Selection of the models collection $\mathfrak{C}$

As the partition tree is fixed ($\mathfrak{T} = \mathfrak{T}_1$), we must select one model for each non-terminal vertex of $\mathfrak{T}$. We first select the explanatory variables for each non-terminal node, using BIC. For each explanatory variable $x_k$, we estimate the model with $x_k$ (using the complete design) or without. Thus, we must test $2^K$ models for each non-terminal node, where $K$ is the number of explanatory variables. In our example, $K = 2$, thus all combinations are tested, again using the canonical (reference, logistic, complete) model for the same reasons as previously. The $2^2 = 4$ combinations are: no effect ($\emptyset$), effect of the first variable ($x_1$), effect of the second variable ($x_2$) and effect of both variables ($x_1, x_2$). As the parameters $\beta_v$ for each vertex $v \in \mathcal{V}^*$ are different, the collection models can be estimated separately. BIC values for the root vertex are respectively: $\text{BIC}_\emptyset = -1497.79$, $\text{BIC}_{x_1} = -1339.45$, $\text{BIC}_{x_2} = -1449.22$ and $\text{BIC}_{x_1,x_2} = -1329.97$. Thus, for the root node, $x_1$ and $x_2$ are selected but we note that internode length ($x_1$) is more important than distance to growth unit end ($x_2$) when distinguishing between latent bud ($y = l$), short shoot ($y \in \{u, s\}$) and long shoot ($y \in \{U, S\}$).

Following the same approach, only $x_2$ is selected for the two others GLMs of the collection. This means that the transformation into spine is influenced by growth unit end, and not by the internode length.

We must now select the $(r, F, Z)$ model for each non-terminal vertex of $\mathfrak{T}$. First, we select the ratio, using the order relationship among the partition tree $\mathfrak{T}$. The siblings of the first level are totally (weakly)-ordered, thus we must use an adjacent, cumulative or sequential ratio. Axillary production is well represented by a sequential mechanism, and therefore we use the sequential ratio. The complete design matrix is preferred to the proportional design matrix using BIC. Finally, we select the best cdf $F$ in a refinement step. For the second level of $\mathfrak{T}$, the siblings are not comparable. We could use the reference ratio, but there are only two siblings for each node, thus all the ratios are equivalent. In fact, in the Bernoulli situation, given the vector of explanatory variable $x$, a GLM is fully specified by the cdf $F$ only. After selecting the cdf $F$ for the two last non-terminal nodes, we obtain the model $\mathfrak{M}^*$ (summarized in figure 3.13) with BIC value: $BIC_* = -2109.58$. Finally, the selected model $\mathfrak{M}^*$ has a better log-likelihood than the classical multinomial logit model ($l_* = -2072.19$ versus $l_0 = -2087.42$) with fewer parameters (10 versus 12).

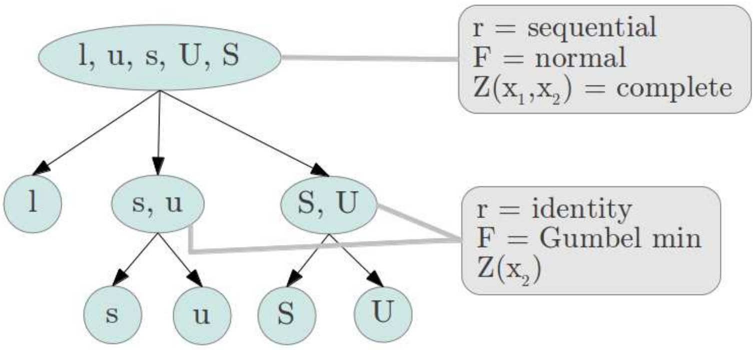

Figure 3.13: PCGLM for pear tree data.

We also obtain a better interpretation with this model. The axillary production of the pear tree can be decomposed into two levels. Production first follows a sequential mechanism, choosing between latent bud, short shoot and long shoot, which is strongly influenced by internode length (the longer the internode, the longer the axillary branching). The axillary shoot then differentiates into unspiny or stays spiny shoot depending on distance to growth unit end.

## 3.7 Discussion

PCGLMs constitute a flexible and interpretable framework for analysing categorical data. Explanatory variables can be selected at each non-terminal node. An explanatory variable may thus have an effect on one partition of categories, not on another. It should be borne in mind that the non effect of a variable is as interesting as the effect. PCGLMs are thus more parsimonious than simple GLMs. Regarding other regression models, various variable selection procedures can be applied to PCGLMs. Because of the small number of explanatory variables ($K = 2, 3$), we used BIC and tested all the combinations in our examples. With a higher number of explanatory variables, methods to reduce of the predictor space, or regularization methods, should be used (Tutz, 2012). Moreover, the decomposition into several steps makes the interpretation easier, using a sequential latent mechanism approach, and also leads to a better fit. If the underlying sequential process can be interpreted as conditioning of latent variables ($Y_2|Y_1 \in v$ and $Y_2|Y_1 \in v'$), a common effect on two vertices ($\beta^v = \beta^{v'}$) can be considered.

Except in this last case, PCGLMs can be easily estimated. After rearranging the data by partitioning and conditioning, classical algorithms can be applied for each data subset. For the simplest and most common case $\beta^v \neq \beta^{v'}$, the different algorithms may be parallelized and running time is thus reduced. Moreover, a canonical GLM with a block structured design matrix can be written as a PCGLM with simple design matrices (see lemma 1) that is easier to estimate.

An important issue with PCGLMs is selecting the partition tree. The tree may be determined *a priori*, as in classical approaches. The two-step model, for instance, relies on an *a priori* known hierarchy among ordered categories. The nested logit model aggregates categories that are similar (i.e. influenced by the same variables). Finally, a POS-PCM associates a partition tree to a Hasse diagram (poset). But defining this poset from the corresponding latent process is not an easy task. In most applications the partition tree is not *a priori* known and should thus be selected. The proposed approach for selecting the partition tree and the variables could be used in the supervised classification context with ordered classes. The indistinguishability procedure selects the best splitting between categories, starting from the entire set $\{1, \ldots, J\}$. Alternatively, we may aggregate adjacent categories, starting from singletons $\{j\}$.

Finally, caution should be exercised to penalize log-likelihood. Let us consider the context of BIC penalization. The total number of observations $n$ should *a priori* be used. But if an explanatory variable influences a non terminal vertex $v$ associated with a small proportion of the observations ($n_v \ll n$), should we incorporate a term related to $n_v$ in the penalty?

Chapter 4

# Integrative models for jointly analyzing shoot growth and branching patterns

## Abstract

*Background and Aims:* It has long been known that shoot growth has an effect on its branching patterns, but the characterization of the corresponding patterning mechanism is still an open issue.
*Methods:* Dedicated statistical models, called semi-Markov switching partitioned conditional generalized linear models, were applied to apple and pear tree data sets. In the semi-Markov switching partitioned conditional generalized linear models estimated from these data sets, the underlying semi-Markov chain represents both the succession and lengths of branching zones, while the partitioned conditional generalized linear models represent the influence of growth explanatory variables on axillary productions within each branching zone.
*Key results:* On the basis of these integrative statistical models, we show that smoothed and delayed growth explanatory variables influence specific branching events.
*Conclusions:* The partitioned conditional generalized linear model selected for each branching zone can be used to identify which developmental event (e.g. shoot initiation, shoot elongation or apex transformation into spine) is affected by a given explanatory variable. The proposed integrative statistical modelling approach could incorporate other explanatory variables such as local curvature of the parent shoot or maximum growth rate of the internode or leaf.

**Keywords:** branching pattern, categorical data, generalized linear model, growth pattern, semi-Markov switching regression model.

## Contents

## 4.1 Introduction

Soot branching patterns often take the form of a succession of well-differentiated homogeneous branching zones where composition properties, in terms of axillary productions, do not change substantially within each zone, but change markedly between zones. These branching patterns have been analysed using segmentation models and in particular hidden semi-Markov chains (Guédon et al., 2001). Branching patterns are modulated by factors that have an overall effect on the pattern, and by factors that vary along the shoot and have differentiated effects on successive axillary productions. We previously investigated the influence of the architectural position of a shoot, which can be viewed as a factor that have an overall effect, on apple tree branching patterns (Renton et al., 2006).

Here, we focus on factors that vary along the shoot and modulate its branching pattern. For example, it has been shown that shoot growth modulates branching pattern, in particular immediate (or sylleptic) branching; see Lauri and Terouanne (1998) for an illustration in the apple tree case. Other potential factors include local curvature of the shoot (Han et al., 2007). To this end, we introduced a new family of integrative statistical models for analysing jointly the succession and length of branching zones and the modulation of the axillary productions within each zone by factors that vary along the shoot. These models generalize hidden semi-Markov chains for categorical data (Guédon et al., 2001) by incorporating explanatory variables and are called semi-Markov switching partitioned conditional generalized linear models (SMS-PCGLMs). It should be noted that another family of semi-Markov switching regression models has been previously introduced for analysing forest tree growth components. More precisely, semi-Markov switching linear mixed models have been used to identify and characterize ontogenetic, environmental and individual growth components on the basis of tree main stems described by annual shoot and climatic data (Chaubert-Pereira et al., 2009).

## 4.2 Materials and methods

### 4.2.1 Tree data sets

The proposed approach is illustrated by the analysis of immediate branching patterns in apple and pear trees.

#### 4.2.1.1 Apple tree (*Malus domestica* Borkh)

Twenty two one-year-old apple trees, “Fuji” cultivar, grafted on M9 (Pajam 1) rootstock and planted at the DiaScope experimental at INRA Montpellier were analysed in this study. Distances between trees corresponded to 3 m between rows and 0.7 m between trees in the same row. Agricultural practices - including irrigation with micro-sprinklers, fertilization and spraying against pests and diseases - were done according to standard practices in the area.

The one-year-old main axis developed from the graft was described by node. The presence of an immediate axillary shoot - i.e. developed without delay with respect to the parent node establishment date - was noted at each successive node,. Immediate shoots were classified into two categories, short and long, according to length ($\leq$ 5cm or $>$ 5cm, respectively). Successive internode lengths along the main axis were measured with a tape ruler that was precise to within 0.5 cm. This dataset was thus constituted of 22 bivariate sequences of cumulative length 1494 (length between 63 and 73 nodes) associating a categorical variable (type of axillary production selected from among latent bud, short or long immediate shoot) with an interval-scaled variable (internode length).

#### 4.2.1.2 Pear tree (*Pyrus spinosa*)

Harvested seeds of *Pyrus spinosa* were sown and planted in January 2001 in a nursery located near Aix-en-Provence, southeastern France. Seedlings grew in 600cm3 WM containers grouped in plastic crates by 25. In winter 2001, the first annual shoot on the main axis of 50 one-year-old individuals was described by node. In this nursery context, individuals were able to grow twice a year, and the annual shoots were made up of one or two growth units (GU) - i.e. portion of the axis built up during an uninterrupted period of growth - referred to as GU1 or GU2 in the following. Seven monocyclic annual shoots (only GU1) and 43 bicyclic annual shoots (GU1 and GU2) were observed.

The presence at each successive node of an immediate axillary shoot was noted. Immediate shoots were classified in four categories according to length ($\leq$ 1cm or $>$ 1cm, with internodes not distinguishable for short shoots), and to transformation or not of the apex into spine (i.e. definite growth or not). This dataset was thus made up of 50 bivariate sequences of cumulative length 3285 associating a categorical variable (type of axillary production selected from among latent bud, unspiny short, unspiny long, spiny short and spiny long immediate shoot), with an interval-scaled variable (internode length).

### 4.2.2 Models

A semi-Markov switching partitioned conditional generalized linear model, which is a two-scale segmentation model, was built on the basis of each data set. In this framework, the succession and length of branching zones (coarse scale) are represented by a non-observable semi-Markov chain while the types of axillary productions within each branching zone (fine scale) modulated by explanatory variables that vary along the parent shoot are represented by partitioned conditional generalized linear models attached to each state of the semi-Markov chain. Hence, each state of the semi-Markov chain represents a branching zone. In our application context, the explanatory variables reflect the growth of the parent shoot, but the statistical framework is general.

The overall model thus combines an $A$-state semi-Markov chain with $A$ partitioned conditional generalized linear models and is referred to as a semi-Markov switching partitioned conditional generalized linear model (SMS-PCGLM). A SMS-PCGLM combines three categories of variables: (i) “state” variable representing non-directly observable branching zones, (ii) plant response categorical variable (types of axillary production), (iii) explanatory variables that vary with node rank (e.g. internode length). This family of statistical models broadens the family of Markov switching models; see Frühwirth-Schnatter (2006) for an overview of Markov switching models.

An $A$-state semi-Markov chain is defined by three subsets of parameters:

- initial probabilities ($\varphi_a$; $a = 0, \ldots, A-1$) to model which is the first branching zone in the parent shoot,

- transition probabilities ($p_{a,b}$; $a, b = 0, \ldots, A-1$) to model the succession of branching zones along a parent shoot,

- occupancy distributions attached to non-absorbing states (a state is said to be absorbing if, after entering this state, it is impossible to leave it) to model lengths of branching zones in number of nodes. We used binomial distributions $\mathcal{B}(d, n, p)$, Poisson distributions $\mathcal{P}(d, \lambda)$ and negative binomial distributions $\mathcal{NB}(d, r, p)$ as possible parametric state occupancy distributions, with an additional shift parameter $d \geq 1$; see Appendix C for formal definitions of these distributions.

In the context of regression models for categorical data analysis, the focus was mainly on models for nominal response variables (unordered categories) or ordinal response variables (totally ordered categories). Here we adopt a more general framework where the categories can be represented as a tree of nested partitions of categories (e.g. latent but versus immediate shoots at the first level, immediate shoots partitioned into short and long at the second level for the apple tree example); see Peyhardi et al. (2013b). This framework means we can tackle the case of partially-ordered response variables and also differentiate the role of explanatory variables at different levels of the partition tree (e.g. an explanatory variable influencing the occurrence of immediate shoots but not their subsequent growth as short or long shoots).

In partitioned conditional generalized linear models, the partition tree of categories and the explanatory variables relevant for each non-terminal vertex of the partition tree need to be selected. We thus adopted a two-stage approach for inference where in a first stage, a simple hidden semi-Markov chain (i.e. without explanatory variables) was estimated for a given data set. This estimated hidden semi-Markov chain was used to segment the observed sequences into homogeneous branching zones and for each branching zone (except unbranched zones where only latent buds were observed), a PCGLM was selected by identifying the partition tree and the explanatory variables. In a second stage a SMS-PCGLM was estimated on the basis of the observed sequences.

## 4.3 Results

### 4.3.1 Apple tree

Superimposition of the proportions of immediate shoots (short and long shoots are not distinguished) and pointwise average internode lengths showed that an unbranched zone corresponding to short internodes spreading over the first few nodes was followed by a zone where internode length increased abruptly. Shoot occurrence also increased abruptly but with a shift of a few nodes with respect to the increase in internode length (Figure 4.1). Internode length then decreased gradually along the parent shoot while the decrease in shoot occurrence was more irregular.

We first estimated a 3-state HSMC made up of two transient states (corresponding to an unbranched and a branching zone) followed by a final absorbing end state modelling growth cessation (Renton et al., 2006) on the basis of the observed bivariate sequences, including the internode length variable. It should be noted that the segmented branching zones obtained using only the types of axillary production variable were very similar (98.5% match between the segmentations obtained using the estimated univariate and bivariate HSMCs). We then extracted the sub-sequences corresponding to the branching zone (state 1). The posterior probabilities of the optimal segmentations (i.e. weight of the optimal segmentation among all the possible segmentations of a given observed sequence) were very high (always above 0.995 for the 22 individuals).

As a second step, we selected the explanatory variable using canonical GLMs (multinomial logit model for nominal data) and then identified the partition tree of categories using the selected explanatory variable on the basis of the data extracted from the branching zone. We tested two types of transformation of the measured internode length variable to build potential explanatory variables:

- shift of the variable, in particular backward shifts, since the internode elongation lasts about 12 days in apple tree and temporally overlaps the initiation of immediate shoots a few nodes below the apex.

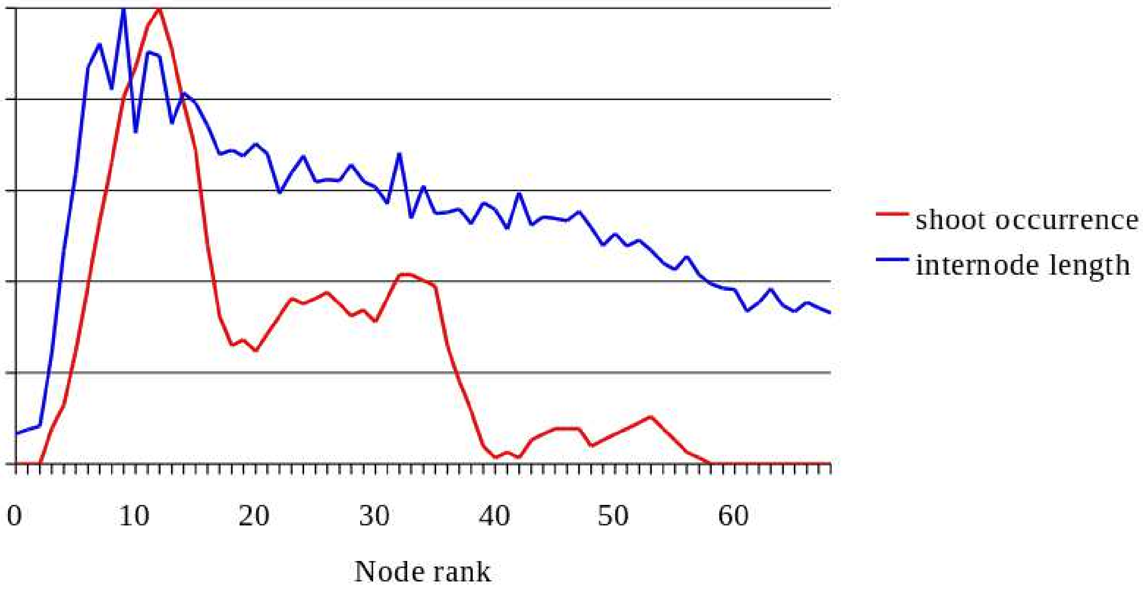

Figure 4.1: *Proportions of immediate shoots (short and long shoots are not distinguished) and pointwise average lengths of internodes as a function of node rank represented on a common y scale whose maximum corresponds to the maximum proportion of immediate shoots and the maximum average length of internodes.*

- smoothing of the variable to remove fluctuations and extract the local internode length trend. This smoothing can be interpreted as an averaging over internodes that elongate at a given time $t$.

To this end, canonical GLMs were estimated for each possible explanatory variable. The best GLM according to the Bayesian information criterion (BIC) was obtained with no shift and smoothing of the internode length was obtained using a symmetric smoothing filter corresponding to the probability mass function of the binomial distribution of parameters 32 and 0.5 (95% of the mass concentrated on the 11 central values). This smoothing width appears to be consistent with the order of magnitude of the number of internodes that elongate at a given time t. Having selected the explanatory variable, we then identified the partition tree of categories. We obtained a first partition into latent bud and immediate shoots, then a subsequent partition of immediate shoots into short and long shoot; see Figure 4.2.

Finally, a 3-state SMS-PCGLM was estimated using the partition tree and explanatory variable previously selected for the PCGLM associated with state 1 (branching zone); see Figure 4.2. The deterministic succession of states resulted from the iterative estimation procedure. The unbranched zone corresponding to short internodes at the base of the shoot (state 0) corresponds to the preformed part of the shoot. The fact that the highest probability of branching and the longest internodes, were found near the shoot base (around rank 10) likely resulted from the propagation mode of the observed young plants derived from bud grafting on one-year-old rootstock. The locally smoothed internode length markedly influences immediate shoot initiation (first level of the partition tree corresponding to latent bud versus immediate shoot) but only slightly influences the subsequent growth of these immediate shoots (second level of the partition tree corresponding to short versus long shoot); see Figure 4.3.

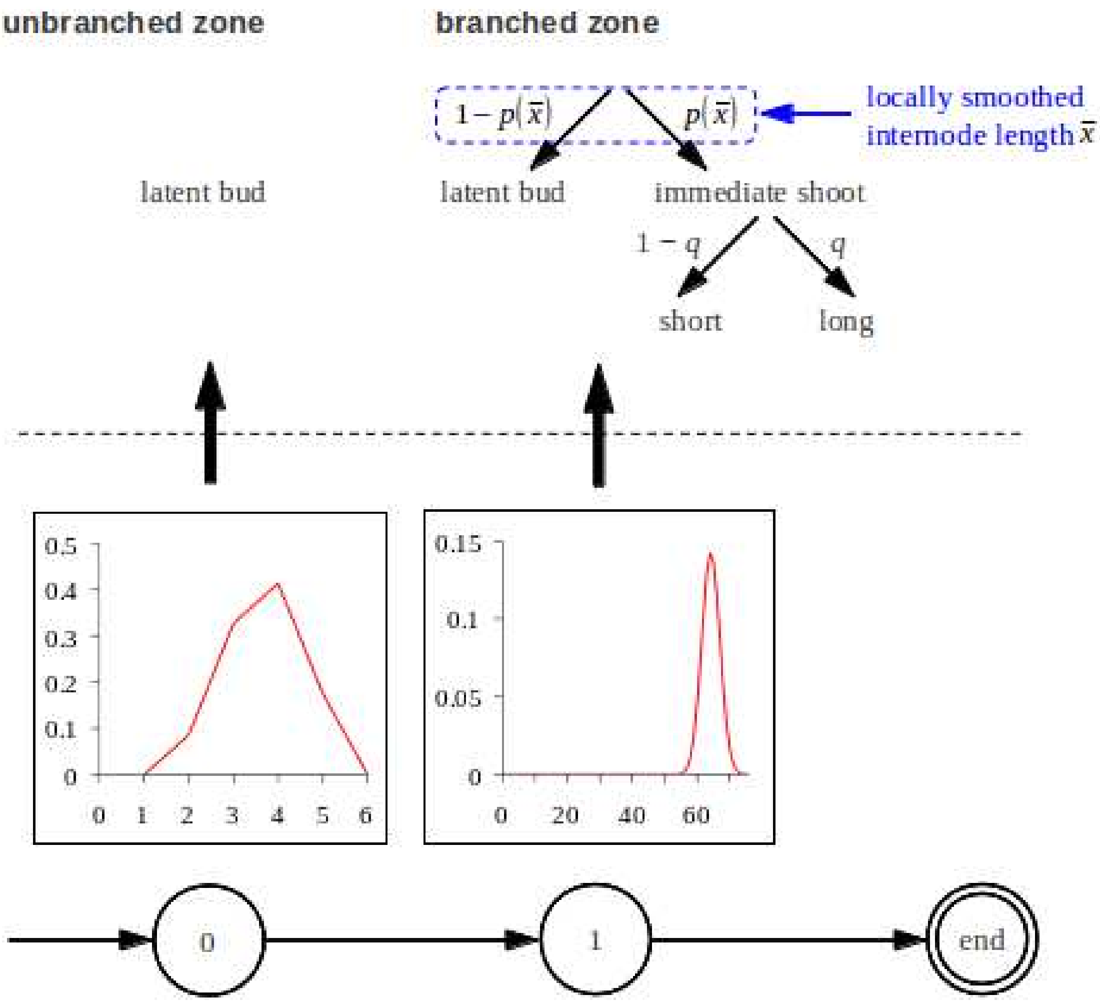

Figure 4.2: *Semi-Markov chain: each state is represented by a vertex which is numbered. Vertices representing transient states are edged by a single line while the vertex representing the absorbing end state is edged by a double line. The possible transitions between states are represented by arcs (the attached probabilities are always* 1*). The arc entering in state 0 indicates that it is the only possible initial state. The occupancy distributions of the transient states are shown above the corresponding vertices. Observation models: for state* 0 *(unbranched zone), the observation model is degenerate since the only possible observation is latent bud. For state* 1 *(branched zone), the estimated branching probability p decreases with internode length along the parent shoot. Locally smoothed internode length markedly influences immediate shoot initiation but only slightly influences the subsequent growth of these immediate shoots (and the influence of the explanatory variable is only shown for the partition latent bud/immediate shoot).*

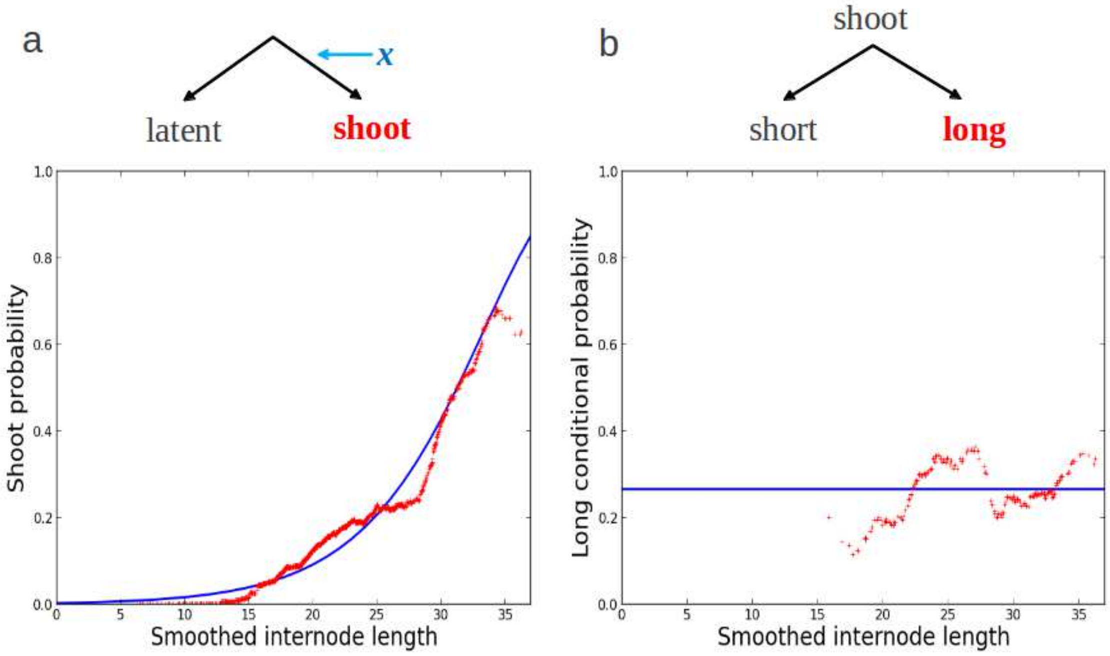

Figure 4.3: *Fit of generalized linear models for each non-terminal vertex of the partition tree (influence of smoothed internode length): (a) first level (latent bud versus immediate shoot), (b) second level (short versus long shoot).*

### 4.3.2 Pear tree

Superimposition of the proportions of the immediate shoots (unspiny short, unspiny long, spiny short and spiny long shoots are not distinguished) and pointwise average internode lengths showed that the shoot occurrence increased very abruptly at the beginning of GU1 while the increase in internode length was more gradual than that in shoot occurrence (Figure 4.4.a). The internode length trend was more similar to the shoot proportion trend in the GU2 case than in the GU1 case (Figure 4.4).

We first estimated a 6-state HSMC made up of five transient states (GU1 bottom unbranched zone, GU1 branching zone, unbranched zone intermediate between GU1 and GU2, GU2 branching zone, GU2 top unbranched zone) followed by a final absorbing end state modelling growth cessation on the basis of the trivariate sequences (types of axillary production, internode length and GU rank); see Figure 4.5. It should be noted that the segmented branching zones obtained using only the types of axillary production and the GU rank variables were very similar (99.3% match between the segmentations obtained using the estimated bivariate and trivariate HSMCs). We then extracted the sub-sequences corresponding to the GU1 and GU2 branching zones (states 1 and 3). The posterior probabilities of the optimal segmentations were most often high: 78% above 0.5, 62% above 0.75 and 32% above 0.9 to be related to an average number of possible segmentations around 300.

We then selected explanatory variables for GU1 and GU2. Concerning GU1, we selected the backward-5-shifted smoothed internode length (symmetric smoothing filter such that 95% of the mass is concentrated on the 15 central values) and the distance to GU end. Concerning this second explanatory variable, the assumption was made that transformation of the offspring

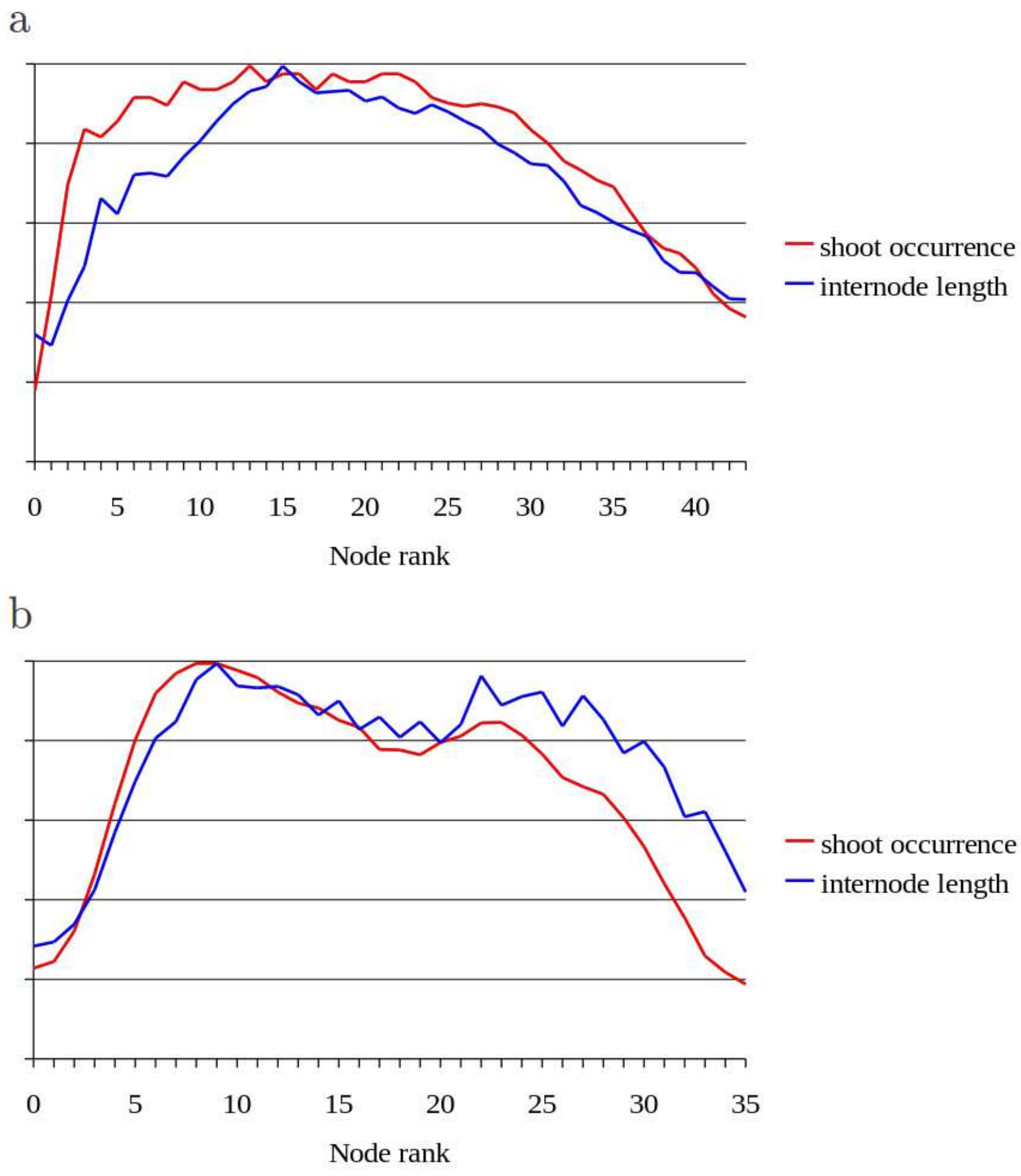

Figure 4.4: *Proportions of immediate shoots (unspiny short, unspiny long, spiny short and spiny long shoots are not distinguished) and pointwise average lengths of the internodes as a function of node rank represented on a common y scale whose maximum corresponds to the maximum proportion of immediate shoots and the maximum average length of internodes. (a) GU1, (b) GU2.*

shoot apex into spine is related to the growth end of the parent shoot. Concerning tree partition, we investigated in particular the partition corresponding to successive developmental events: shoot initiation, shoot growth and transformation of the shoot apex into spine (i.e. partition into latent bud and immediate shoots at the first level, partition of immediate shoots into short and long at the second level, short and long each being partitioned into unspiny and spiny at the third level). We found that the "sequential" partition (latent bud and immediate shoots at the first level, unspiny short and other shoots at the second level, unspiny long and spiny shoots at the third level, short and long spiny shoots at the fourth level) was favoured by BIC; see Figure 4.6.a. This partition tree stays consistent with the succession of developmental events - shoot initiation, shoot growth and transformation of the shoot apex into spine - but can be interpreted by also considering a trend along the branching zone where the transformation into spine affects mainly shoots near the end of the branching zone initiated later than shoots at the beginning of the branching zone (Figure 4.7). The internode length explanatory variable mainly influences the first (latent bud versus immediate shoots), second (unspiny short versus other shoots) and fourth (short versus long spiny shoots) levels of the partition tree, while the distance to GU end mainly influences the third level (unspiny long versus spiny shoots); see Figure 4.8.

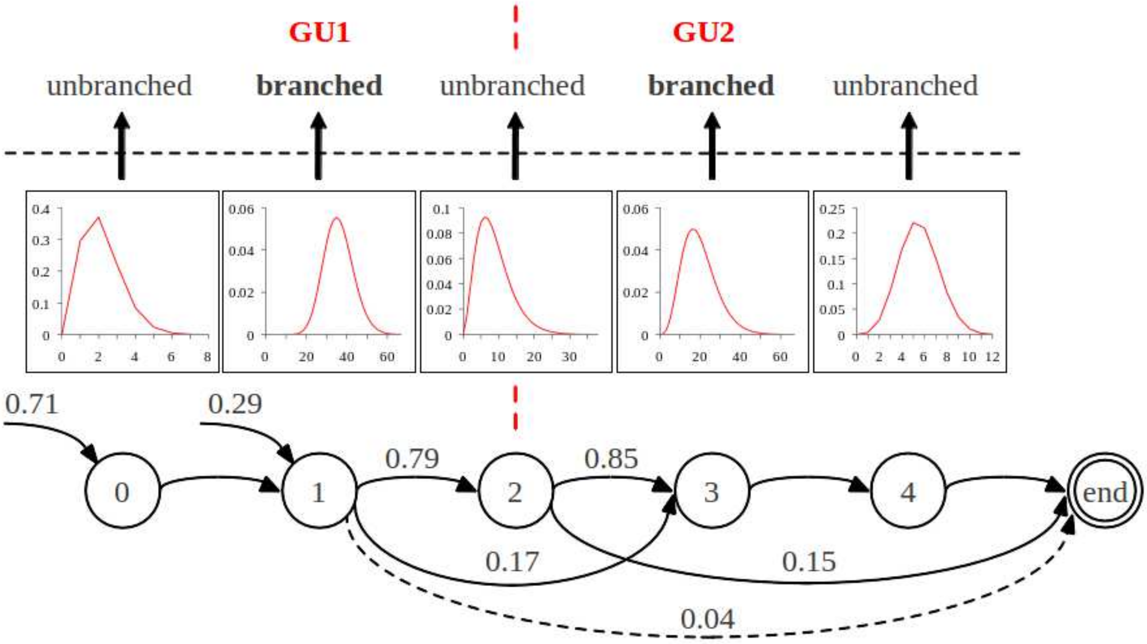

Figure 4.5: *Non-observable semi-Markov chain of the semi-Markov switching partitioned conditional generalized linear models: each state is represented by a vertex which is numbered. Vertices representing transient states are edged by a single line while the vertex representing the absorbing end state is edged by a double line. Possible transitions between states are represented by arcs with the attached probabilities noted nearby when* $< 1$. *Arcs entering into states indicate initial states. The attached initial probabilities are noted nearby. The occupancy distributions of the transient states are shown above the corresponding vertices.*

Concerning GU2, we selected the backward-1-shifted smoothed internode length (symmetric smoothing filter such that 95% of the mass is concentrated on the 15 central values) and the distance to GU end. The partition tree was similar to that of GU1 except that the second and third levels were aggregated into a single level (unspiny short versus spiny shoots) since unspiny long shoots were very rare in GU2 (17 out of 794 immediate shoots) and were thus not considered in the identification of the partition tree. The internode length explanatory variable influences all three levels (latent bud versus immediate shoots, unspiny short versus spiny shoots, short versus long spiny shoots) of the partition tree, while the distance to GU end mainly influences the last two levels; see Figure 4.6.b.

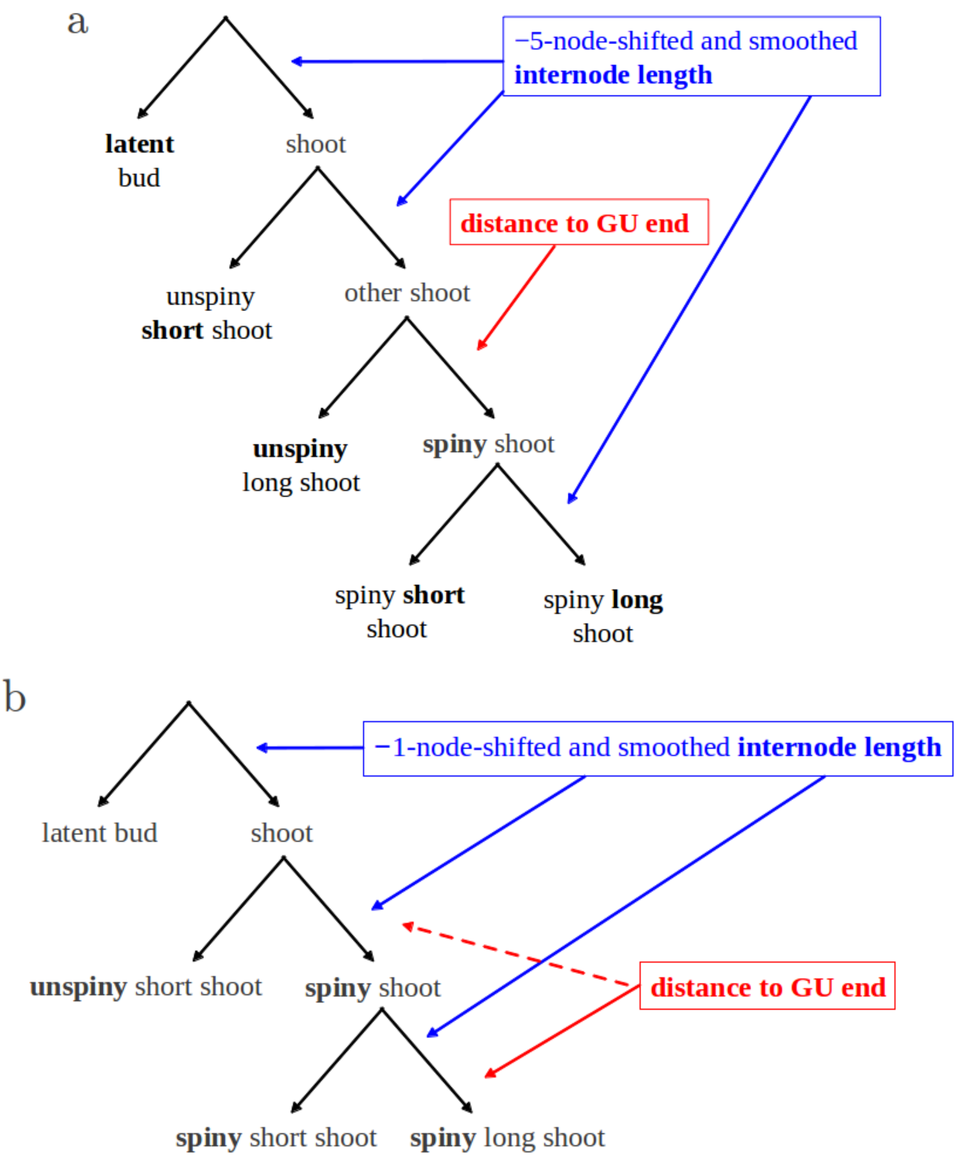

Figure 4.6: *Partitioned conditional generalized linear models for (a) GU1 branching zone and (b) GU2 branching zone.*

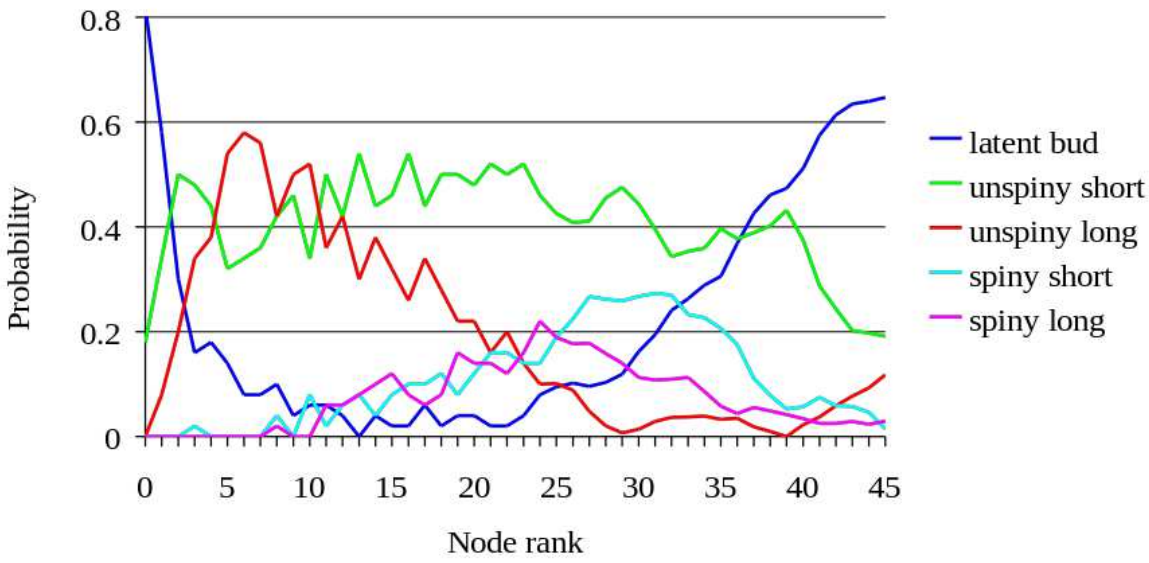

Figure 4.7: *GU1: Proportions of the different types of axillary productions (latent bud, unspiny short, unspiny long, spiny short and spiny long immediate shoots) as a function of node rank.*

## 4.4 Discussion

The examples given above illustrate how the definition of appropriate explanatory variables is a crucial step in retrospective measurements. We are currently investigating the extraction of explanatory variables based on growth data follow up (e.g. leaf expansion). In this context, the extraction of explanatory variables requires two steps, (i) extraction of growth parameters using for instance nonlinear regression models (e.g. the maximum absolute growth rate deduced from the fit of a sigmoidal function), (ii) the shifting and smoothing of growth parameters deduced from nonlinear regression models. This transformation of explanatory variables is an important issue in the analysis of tree structure development based on retrospective measurements for at least two main reasons:

- Trees are large organisms with potentially great inertia in their development. Modulation of the branching process by the growth process is therefore not instantaneous. In the same way, tree responses to changes in climatic or local environment conditions are not instantaneous; see illustrations in Chaubert-Pereira et al. (2009) and Taugourdeau et al. (2011).
- The temporal dimension of growth and immediate branching is only partially reflected by the topological indexing using node ranks, and is better represented by the transformation of explanatory variables.

Concerning observation regression models, a standard solution would have consisted of assuming that the categorical response variable was ordinal in the apple tree case (with the following category order: latent bud, immediate short shoot and immediate long shoot) and estimating a classic generalized linear model for ordinal data for the branching zones. We chose to develop the more general framework of partitioned conditional generalized linear models

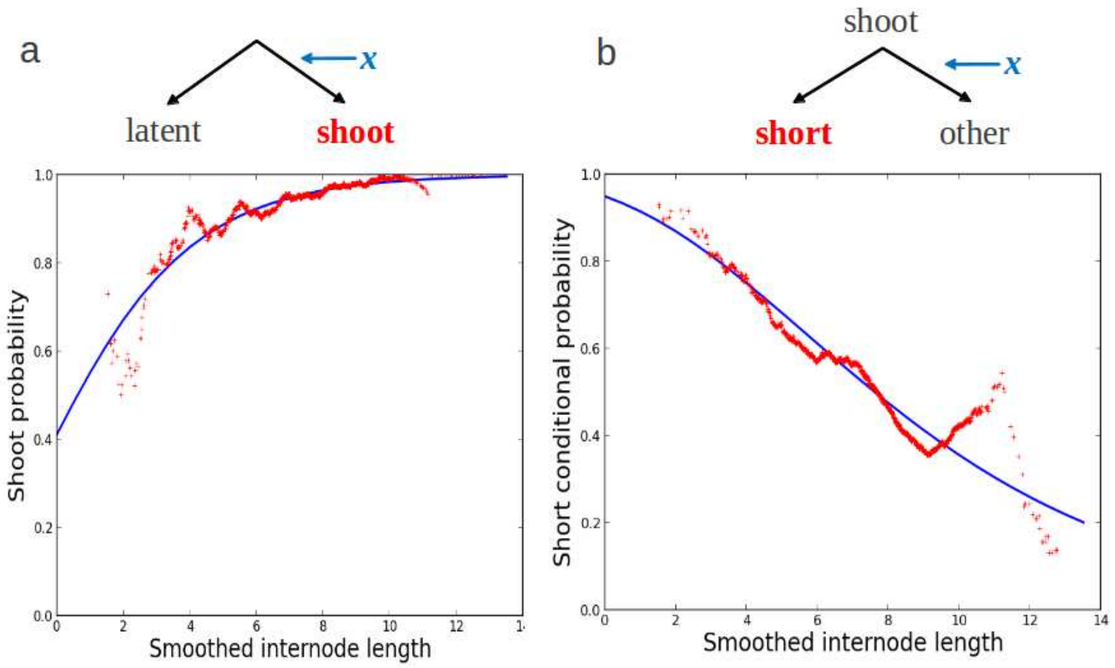

Figure 4.8: *Pear tree GU1, fit of generalized linear models for non-terminal vertices of the partition tree (influence of the backward-5-shifted and smoothed internode length): (a) first level (latent bud versus immediate shoot), (b) second level (immediate unspiny short shoot versus other shoots).*

that can be used to tackle not only the classical cases of nominal and ordinal categorical response variables, but also the case of partially-ordered categorical response variables (Peyhardi et al., 2013b). Using this hierarchical modelling, it was possible to show in the apple tree case - applying model selection criteria - that locally smoothed internode length markedly influences immediate shoot initiation but far less the subsequent growth of these immediate shoots which likely depend on environmental factors at this time. Combining the transformation of explanatory variables and recursive partitioning of the axillary productions using the proposed hierarchical modelling, it was possible to test many assumptions concerning the influence of growth patterns on the immediate branching pattern in our examples.

Development is the sum of events that contribute to the progressive elaboration of the body of an organism (Steeves and Sussex, 1989). Plant development is defined as a series of identifiable events resulting in a qualitative (germination, flowering . . . ) or quantitative (number of leaves, number of flowers . . . ) modification of plant structure (Gatsuk et al., 1980). Branching is a key developmental process in plants. Branching data are most of the time collected retrospectively and potentially reflect a succession of complex but interrelated developmental phases such as:

- immediate branching i.e. offspring shoots developed without delay with respect to the parent node establishment date,
- delayed branching (e.g. 1-year-delayed branching for temperate species),
- elongation or not of the offspring shoots leading to short or long shoots,

- morphological transformation of offspring shoots such as transformation of the apex into spine or flower leading to growth interruption.

Possible axillary productions can efficiently be coded as categories that are well defined and separated according to morphological criteria. Because of the potentially complex succession of developmental phases, these categories cannot in most cases be ordered, but they are not unstructured. Hierarchical approaches that reflect complex structuring of categories thus constitute a very promising avenue for the analysis of plant structure and development.

This study together with that of Chaubert-Pereira et al. (2009) illustrate the versatility of semi-Markov switching regression models where a semi-Markov chain can represent homogeneous branching zones at the node scale as well as growth phases at the annual shoot scale, and where all the panoply of regression models can be incorporated depending on the type of plant response variable (categorical variable for type of axillary production and interval-scaled variable for annual shoot length).

**Acknowledgements** The authors wish to thank Michael Renton for his participation in apple tree data collection.

This chapter extends the work presented in the abstract Peyhardi et al. (2013a).

Chapter 5

# Works in progress and perspectives

In this thesis, we propose a new GLM framework for the analysis of categorical data. In chapter 2 we introduced a unifying specification of GLMs for categorical data using the $(r, F, Z)$ triplet. We then used this specification to define the family of reference models for nominal data. Some classical equivalences between models and invariance properties were extended. And using these properties we proposed a classification of different models along a nominal/ordinal scale. In chapter 3 we introduced the class of PCGLMs based on the $(r, F, Z)$ specification. These models capture the hierarchical structure among categories for nominal, ordinal and partially-ordered response variables. Using these models for the observation process of a hidden semi-Markov chain, we developed in chapter 4 a methodology for jointly analysing shoot growth and branching patterns of plants. This methodology was applied to two datasets in this thesis and will be applied to others that possess a more complex hierarchical structure.

In this chapter, we describe works in progress and perspectives for the $(r, F, Z)$ specification. We first focus on the convergence of Fisher's scoring algorithm for a number of cumulative and reference models. The non-invariance of (reference, $F$, $Z$) models under transposition of the reference category is shown for some analytic cdfs $F$. We then propose to specify the ratio using directed graph and illustrate this with reference, adjacent and sequential ratios. Finally, we propose an extension of the conditional logit model (McFadden, 1974), whose implementation is not available yet.

## Contents

## 5.1 Convergence of Fisher's scoring algorithm

The convergence of an iterative algorithm, such as the Newton-Raphson algorithm or Fisher's scoring algorithm, depends on the concavity of the log-likelihood and the connexity of parameter space $\mathcal{C}$. Convergence has already been shown for some reference and cumulative models. Our objective here is to extend these results.

### 5.1.1 Convergence for cumulative models

It has been shown by Pratt (1981) and Burridge (1981) that concavity of the log-likelihood holds for (cumulative, $F$, proportional) models, if $F$, $1-F$ and $f$ are log-concave. Using results of convex analysis, strict log-concavity of $F$, $1-F$ and $f$ can be shown for logistic, normal, Gumbel min, Gumbel max and Laplace distributions, but not for Student distributions (Bergstrom and Bagnoli, 2005).

But the demonstrations by Pratt (1981) and Burridge (1981) did not utilize the proportionality of the model. Therefore, concavity still holds with other design matrices. But a non-proportional design matrix may lead to non-positive probabilities since the sequence of linear predictors $\{\eta_j(x)\}_{j=1,\ldots,J-1}$ must be strictly increasing, for any $x \in \mathcal{X}$ to define strictly positive probabilities $\{\pi_j(x)\}_{j=1,\ldots,J}$.

In the proportional design, the identified space $\tilde{\mathcal{C}} = \{\beta = (\alpha_1, \ldots, \alpha_{J-1}, \delta^t) \in \mathbb{R}^{J-1+p} | \ \alpha_1 < \ldots < \alpha_{J-1}\}$ does not depend on explanatory space $\mathcal{X}$ and is an open convex. Convexity of the parameter space (and concavity of the log-likelihood) implies convergence of Fisher's scoring algorithm.

**Discussion** In the non-proportional design, the problem of convergence is quite different because contrast space $\mathcal{C}$ depends on $\mathcal{X}$. The possible non-convexity of $\mathcal{C}$ may also lead to local maxima. The strict increase in linear predictors $\{\eta_j(x)\}_{j=1,\ldots,J-1}$ seems difficult to preserve from one algorithm iteration to another. We could add constraint on explanatory space $\mathcal{X}$ to relax the constraints on $\mathcal{C}$. In the case of a strictly positive explanatory variable (see chapter 2) the contrast space is

$$\mathcal{C} = \{\beta = (\alpha_1, \ldots, \alpha_{J-1}, \delta_1^t, \ldots, \delta_{J-1}^t) \in \mathbb{R}^{(J-1)(1+p)} | \ \alpha_1 < \ldots < \alpha_{J-1} \&, \delta_1^t \leq \ldots \leq \delta_{J-1}^t\},$$

which is convex. We could also add a projection of $\beta^{[t]}$ on $\mathcal{C}$ at each iteration $t$, using results of convex analysis.

### 5.1.2 Convergence for reference models

Let us first recall that the reference ratio does not *a priori* constrain space $\mathcal{C}$. In fact, $\mathcal{C}$ may be identified to a $\mathbb{R}$-vector space. Therefore, $\mathcal{C}$ is convex and only concavity of the log-likelihood is needed. Noting that concavity holds for all canonical (reference, logistic, $Z$) models, we now focus on other reference models. We can write a reference model in a more general form

$$\pi_j = \frac{h_j}{1 + \sum_{k=1}^{J-1} h_k},$$

for $j = 1, \ldots, J-1$, with a non-negative and twice differentiable function $h$ such that $h_j = h(\eta_j)$ (in fact for a reference model $h_j = F(\eta_j)/[1 - F(\eta_j)]$). The probability of reference category $J$ is

$$\pi_J = \frac{1}{1 + \sum_{k=1}^{J-1} h_k}.$$

For one observation $(y, x)$, the log-likelihood is

$$l = \sum_{j=1}^{J} y_j \ln \pi_j.$$

We must here recall that concavity of $l$ with respect to $\beta$ is equivalent to concavity of $l$ with respect to $\eta$ since

$$\frac{\partial^2 l}{\partial \beta^t \partial \beta} = Z^t \frac{\partial^2 l}{\partial \eta^t \partial \eta} Z.$$

According to the $J$ possible observations of $y$, there are only two cases: either the reference category is observed or not

$$l = \begin{cases} \ln \pi_J & \text{if } y = J, \\ \ln h_j + \ln \pi_J & \text{if } y \neq J. \end{cases}$$

**First case: $\boldsymbol{y = J}$** Here we are looking for conditions on $h$ such that $\ln \pi_J$ is concave, or equivalently such that the Hessian $\mathcal{H}$ of $-\ln \pi_J$ is positive definite. The first derivative is

$$-\frac{\partial \ln \pi_J}{\partial \eta_j} = h'_j \pi_J,$$

for column $j$. The second derivative is

$$\mathcal{H}_{i,j} = -\frac{\partial^2 (\ln \pi_J)}{\partial \eta_i \partial \eta_j} = \frac{\partial h'_j}{\partial \eta_i} \pi_J - h'_i h'_j \pi_J^2,$$

for row $i$ and column $j$. Differentiating the cases $i = j$ and $i \neq j$ we obtain

$$\mathcal{H}_{i,j} = \begin{cases} h''_j \pi_J - h'^2_j \pi_J^2, & \text{if } i = j, \\ -h'_i h'_j \pi_J^2, & \text{if } i \neq j. \end{cases}$$

Again using the equality $\pi_J = \pi_j / h_j$ we obtain

$$\mathcal{H}_{i,j} = \begin{cases} \dfrac{h''_j}{h_j} \pi_j - \dfrac{h'^2_j}{h_j^2} \pi_j^2, & \text{if } i = j, \\ -\dfrac{h'_i}{h_i} \dfrac{h'_j}{h_j} \pi_i \pi_j, & \text{if } i \neq j. \end{cases}$$

Noting $a_j = h''_j / h_j$ and $b_j = h'_j / h_j$ the Hessian matrix is

$$\mathcal{H} = \text{diag}\{a_j \pi_j\}_j - (b_i b_j \pi_i \pi_j)_{i,j}.$$

whose form is similar to that of the covariance matrix

$$\text{Cov}(Y) = \text{diag}\{\pi_j\}_j - (\pi_i \pi_j)_{i,j}$$

which is positive definite. The objective is thus to find sufficient conditions on $a_j$ and $b_j$ to preserve a positive definite matrix $\mathcal{H}$. Then we should rewrite these sufficient conditions in terms of $F(\eta_j)$, $F'(\eta_j)$ and $F''(\eta_j)$ and obtain a differential inequality.

**Second case: $y \neq J$** Assume that concavity is obtained for the first case. We simply must show that $\ln h$ is concave (because the sum of concave functions is concave) but this condition is not necessary. The Hessian of $\ln h_j$ is a diagonal matrix $\Delta$ with the general term

$$\Delta_{j,j} = \frac{h_j'' h_j - h_j'^2}{h_j^2}.$$

We must show that the difference $\mathcal{H} - \Delta$ remains positive definite, or equivalently that $\Delta \preceq \mathcal{H}$ where $\preceq$ denotes Loewner ordering. My intuition is that a sufficient and necessary condition is $\Delta_{j,j} < \lambda$, where $\lambda$ is the minimal eigenvalue of the matrix $\mathcal{H}$.

### Numerical investigation of concavity

We expect to obtain concavity of the log-likelihood for reference models if certain properties of cdf $F$ are fulfilled. Meanwhile, we explored concavity numerically in the simple case $J = 3$ (because a 3D representation can be used) using Gumbel min and max cdfs. As seen previously, we can split this into two cases.

**First case: $y = 3$** The log-likelihood has the form

$$l(\eta_1, \eta_2) = -\ln\left\{1 + \frac{F(\eta_1)}{1 - F(\eta_1)} + \frac{F(\eta_2)}{1 - F(\eta_2)}\right\}.$$

The log-likelihood shown in figure 5.1 seems to be globally concave for the two different link functions.

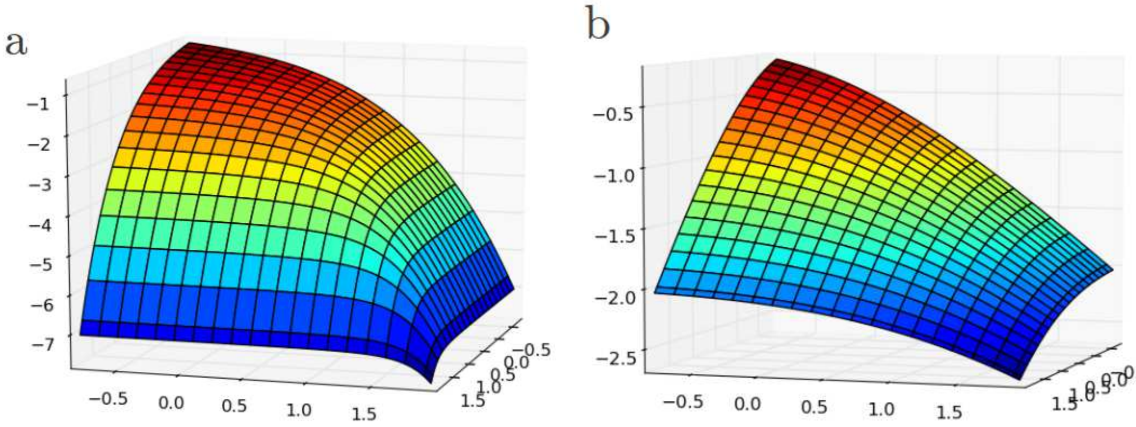

Figure 5.1: Log-likelihood of observed reference category given (a) the (reference, Gumbel min, $Z$) model and (b) the (reference, Gumbel max, $Z$) model.

**Second case: $y \neq 3$** The log-likelihood has the same form for $y = 1$ and $y = 2$. Focusing only on the case $y = 1$, the log-likelihood is

$$l(\eta_1, \eta_2) = \ln\left\{\frac{F(\eta_1)}{1 - F(\eta_1)}\right\} - \ln\left\{1 + \frac{F(\eta_1)}{1 - F(\eta_1)} + \frac{F(\eta_2)}{1 - F(\eta_2)}\right\}.$$

The log-likelihood shown in figure 5.2 seems to be globally concave with the Gumbel max cdf, but not with the Gumbel min cdf.

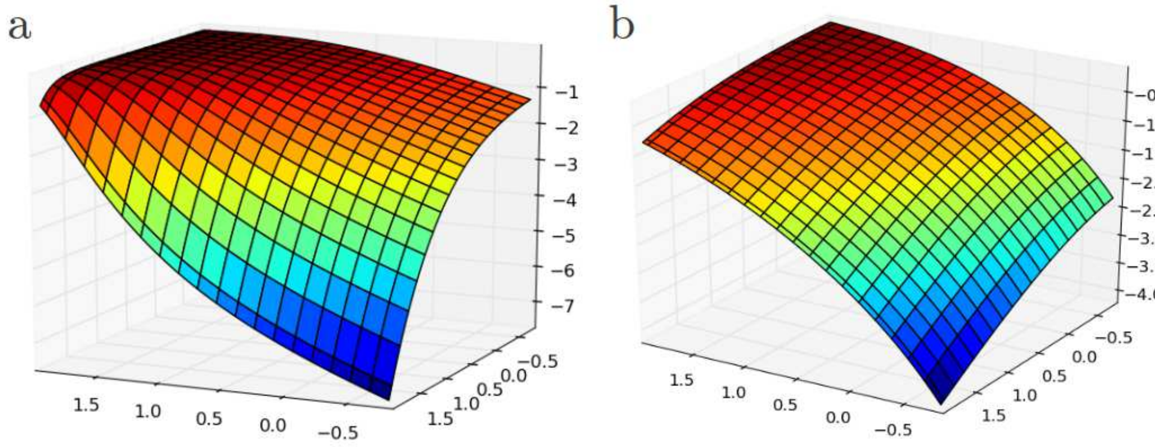

Figure 5.2: Log-likelihood of observed non-reference category given (a) the (reference, Gumbel min, $Z$) model and (b) the (reference, Gumbel max, $Z$) model.

**Discussion** Finally, the demonstration of Fisher's scoring algorithm convergence for (cumulative, $F$, $Z$) models, with non proportional design matrix $Z$, is based on convex analysis results whereas that of convergence for (reference, $F$, $Z$) models, with non logistic cdf $F$, is based on linear algebra results.

## 5.2 Non-invariance of GLMs under permutations

In chapter 2 we showed the invariance of certain GLMs under particular permutations. But are these models still invariant under other permutations?

This question is illustrated using the (reference, $F$, complete) and (reference, $F$, proportional) models. They are invariant under the $(J-1)!$ permutations that fix the reference category (Property 9). But are they still invariant under other permutations? Invariance under a permutation is easier to show than the contrary. The canonical (reference, logistic, complete) model is invariant under all permutations unlike other canonical models $\{$(reference, logistic, $Z$); $Z \in \mathfrak{Z} \setminus \{Z_c\}\}$. But a non identical transposition $\tau$ of $J$ simply changes the design matrix of a canonical model

$$\text{(reference, logistic, } Z)_\tau = \text{(reference, logistic, } B_\tau Z).$$

Hence, the canonical link function (reference, logistic) is invariant under all permutations. But what about other link functions (reference, $F$) when $F$ is not the logistic cdf? Non-invariance of models may be shown when $F$ is analytically defined.

**Property 16.** *Let $\sigma$ be a permutation of $\{1,\ldots,J\}$, $F \in \{$Gumbel min, Gumbel max, exponential$\}$ and let $Z$ be a design matrix depending on $x$. The (reference, $F$, $Z$) model is invariant under $\sigma$ if and **only if** $\sigma(J) = J$.*

*Proof.* We have already shown that $\sigma(J) = J$ is a sufficient condition for the invariance of reference models under $\sigma$. Now we must show the necessity of this condition. Let $\tau$ be a non identical transposition of the reference category $J$ and $F \in \{$Gumbel min, Gumbel max, exponential$\}$. Using a *reductio ad absurdum*, assume that the (reference, $F$, $Z$) model is invariant under $\tau$ or equivalently that

$$(\text{reference}, F, Z)_\tau \Leftrightarrow (\text{reference}, F, Z).$$

There is a bijection $h : \tilde{\mathcal{C}} \rightarrow \tilde{\mathcal{C}}'$ such that

$$\left\{ \begin{array}{lcl} \pi & = & r^{-1} \circ \mathcal{F}\{Z\beta\}, \\ \pi_\tau & = & r^{-1} \circ \mathcal{F}\{Zh(\beta)\}. \end{array} \right.$$

Noting $\eta' = Zh(\beta)$ we obtain

$$\left\{ \begin{array}{ll} \dfrac{\pi_j}{\pi_J} = \dfrac{F(\eta_j)}{1-F(\eta_j)}, & \forall j \neq J, \\ \dfrac{\pi_j}{\pi_{\tau(J)}} = \dfrac{F(\eta'_j)}{1-F(\eta'_j)}, & \forall j \neq \tau(J), \end{array} \right.$$

and thus

$$\left\{ \begin{array}{ll} \dfrac{\pi_j}{\pi_{\tau(J)}} = \dfrac{\pi_j}{\pi_J}\,\dfrac{\pi_J}{\pi_{\tau(J)}} = \dfrac{F(\eta_j)}{1-F(\eta_j)}\,\dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})}, & \forall j \neq J, \tau(J), \\ \dfrac{\pi_J}{\pi_{\tau(J)}} = \dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})}, & \end{array} \right.$$

or equivalently

$$\left\{ \begin{array}{ll} \dfrac{F(\eta'_j)}{1-F(\eta'_j)} = \dfrac{F(\eta_j)}{1-F(\eta_j)}\,\dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})}, & \forall j \neq J, \tau(J), \\ \dfrac{F(\eta'_J)}{1-F(\eta'_J)} = \dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})}. & \end{array} \right.$$

Noting that $F/(1-F)$ is invertible (because $\{F/(1-F)\}' = f/(1-F)^2 > 0$) we obtain

$$\left\{ \begin{array}{ll} \eta'_j = \left(\dfrac{F}{1-F}\right)^{-1} \left\{ \dfrac{F(\eta_j)}{1-F(\eta_j)}\,\dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})} \right\}, & \forall j \neq J, \tau(J), \\ \eta'_J = \left(\dfrac{F}{1-F}\right)^{-1} \left\{ \dfrac{1-F(\eta_{\tau(J)})}{F(\eta_{\tau(J)})} \right\}. & \end{array} \right.$$

To obtain a contradiction we must show that $\eta'$ is not linear with respect to $x$. With the Gumbel min cdf we obtain

$$\left\{ \begin{array}{ll} \eta'_j = \ln\left[ \ln\left\{ \dfrac{\exp(\exp(\eta_j)) + \exp(\exp(\eta_{\tau(J)})) - 2}{\exp(\exp(\eta_{\tau(J)})) - 1} \right\} \right], & \forall j \neq J, \tau(J), \\ \eta'_J = \ln\left[ \ln\left\{ \dfrac{\exp(\exp(\eta_{\tau(J)}))}{\exp(\exp(\eta_{\tau(J)})) - 1} \right\} \right]. & \end{array} \right.$$

Since $Z$ depends on $x$, there is at least one $\eta_j$ (for $j \in \{1, \ldots, J-1\}$) that is linear in $x$ whereas the corresponding $\eta'_j$ (or $\eta'_J$) is not. The same argument holds for the Gumbel max and exponential cdfs. □

**Discussion** We may obtain equivalent properties for adjacent, cumulative and sequential models, following the same idea. For the particular cases of non-analytic cdfs $F$, such as normal and Student cdfs, it is more difficult to show the non-linearity of $\eta'$ with respect to $x$.

## 5.3 Graph representation

Reference, adjacent and sequential ratios are defined using $J-1$ different conditionings. Therefore, the linear predictors $\eta_j$ are unconstrained one to another. Neither $\mathcal{P}$ nor $\mathcal{S}$ is constrained ($\mathcal{P} = ]0,1[^{J-1}$ and $\mathcal{S} = \mathbb{R}^{J-1}$) and thus no constraint is required on parameter $\beta$. This is an advantage compared to the cumulative ratio.

We propose to define an unconstrained ratio by a directed graph. Each probability $\pi_j$ is represented by a vertex $j$ and the relation $r_j(\pi)$ is represented by directed edges. The graph is made up of $J$ vertices and for each vertex $j = 1, \ldots, J-1$, the ratio is defined by

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \displaystyle\sum_{k \in \mathrm{Ch}(j)} \pi_k},$$

where $\mathrm{Ch}(j)$ denotes the children of vertex $j$. The reference category $J$ has no child.

We must find the necessary and sufficient conditions for sets $\mathrm{Ch}(1), \ldots, \mathrm{Ch}(J-1)$ such that the corresponding ratio is a diffeomorphism from $M = \{\pi \in ]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to $\mathcal{P} = ]0,1[^{J-1}$.

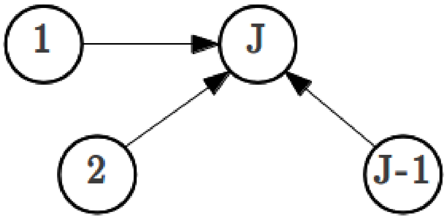

Figure 5.3: Graph representation of reference ratio.

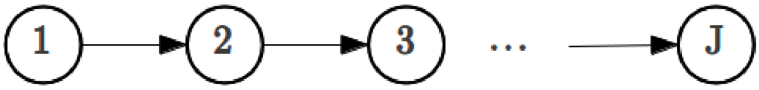

Figure 5.4: Graph representation of adjacent ratio.

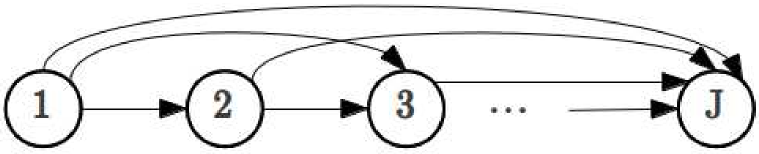

Figure 5.5: Graph representation of sequential ratio.

For the reference ratio (see figure 5.3), the children are defined for $j = 1, \ldots, J-1$ by

$$\mathrm{Ch}(j) = \{J\}.$$

For the adjacent ratio (see figure 5.4), the children are defined for $j = 1, \ldots, J-1$ by

$$\mathrm{Ch}(j) = \{j+1\}.$$

For the sequential ratio (see figure 5.5) , the children are defined for $j = 1, \ldots, J-1$ by

$$\mathrm{Ch}(j) = \{j+1, \ldots, J\}.$$

This graph representation allows us to identify equivalence between models more easily. Let us focus for instance on graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, corresponding respectively to the reference ratio with 2 as reference category, and the adjacent ratio with $J = 3$ categories (see figure 5.6). We remark that we can switch from $\mathcal{G}_1$ to $\mathcal{G}_2$ simply by changing the direction of the edge $(3, 2)$. In this case, the directed edge is reversible if $F$ is symmetric. More precisely, we have the following property.

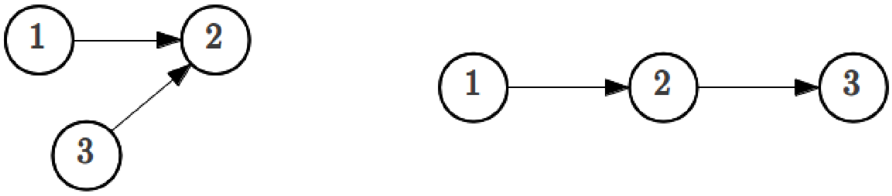

Figure 5.6: Graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ of reference and adjacent ratios.

**Property 17.** *Let $J = 3$ and $\tau$ be the transposition between 2 and 3. Let $A_\tau$ be the matrix*

$$A_\tau = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

*Then the (reference, symmetric $F$, $Z)_\tau$ and (adjacent, symmetric $F$, $A_\tau Z$) models are equivalent.*

*Proof.* Assume that the distribution of $Y|X = x$ is defined by the transposed model (reference, $F$, $Z)_\tau$. The first equations of the (reference, symmetric $F$, $Z)_\tau$ and (adjacent, symmetric $F$, $A_\tau Z$) models are the same. Hence, we focus only on the second equation of the (reference, symmetric $F$, $Z)_\tau$ model

$$\frac{\pi_3}{\pi_3 + \pi_2} = F(\eta_2),$$

thus

$$\frac{\pi_2}{\pi_2 + \pi_3} = \tilde{F}(-\eta_2).$$

This means that $Y|X = x$ follows the (adjacent, $F$, $A_\tau Z$) model if $F = \tilde{F}$. □

**Discussion** If we find the necessary and sufficient conditions on the graph to obtain a well defined ratio, we expect to be able to define new ratios more easily. We expect in this way to find other equivalences between GLMs for categorical data.

## 5.4 Qualitative choice models

In this framework, statistical unit $i$ is a consumer, $J$ categories are different alternatives and $x$ are attributes influencing the consumer's choice. In this section we first present a number of qualitative choice models using a classical predictive approach, then propose to extend these models using the $(r, F, Z)$ specification.

### 5.4.1 Classical approach

#### 5.4.1.1 Model specification

We saw in chapter 1 that Luce's choice axiom Luce (1959) and the principle of random utility maximisation lead to the model

$$P(Y = j) = \frac{\exp(\eta_j)}{1 + \sum_{k=1}^{J-1} \exp(\eta_k)} \tag{5.1}$$

for $j = 1, \ldots, J-1$. Depending on the form of the linear predictors $\eta_j$, we obtain different logit models:

- Multinomial logit model: $\eta_j = \alpha_j + x^t \delta_j$. The attributes are common for all alternatives and the parameters depend on each alternative.
- Conditional logit model: $\eta_j = \alpha + x_j^t \delta$. The attributes depend on each alternative and the parameters are common for all alternatives.
- Universal logit model: $\eta_j = \alpha_j + x_j^t \delta_j$. The attributes and the parameters depend on each alternative.

#### 5.4.1.2 Independence of irrelevant alternatives

Since these three logit models satisfy Luce's choice axiom, they share the IIA property. The ratio of probabilities for alternatives $j$ and $k$

$$\frac{P(Y = j)}{P(Y = k)} = \exp(\eta_j - \eta_k)$$

does not depend on other alternatives. As noted by McFadden (1986), "the IIA axiom is a blessing and a curse". On the negative side, it is inconsistent with the heterogeneous patterns of similarities often encountered in economics and marketing problems. This is well illustrated by Marschak (1960) through the blue/red bus example. On the positive side, "it makes forecasting the demand for a new alternative an easy calculation" (McFadden, 1986).

#### 5.4.1.3 Prediction for a new alternative

In the econometrics framework, qualitative choice regression models are often used in a predictive manner. With the multinomial logit model, the probability of each alternative $P(Y = j|x)$ can be predicted for any value $x$ which is not observed in the dataset. For the conditional logit model (McFadden, 1974), the situation is quite different because linear predictors have a different form: $\eta_j = \alpha + x_j^t \delta$. Here, the attribute value $x_j$ is related to the alternative $j$, which is why the model is called the *conditional* logit model. The probability $P(Y = 0|x_0)$ of a new alternative 0 can be predicted because the intercept $\alpha$ and the slope $\delta$ are common for all alternatives.

Let us illustrate the predictive approach using the classical situation of urban travel demand. The consumer has $J = 3$ alternatives: car, bus or bicycle. He makes a decision according to travel cost $c$ and travel time $t$. It should be noted that these two attributes are quantitative and are related to the different alternatives: $c_j$ and $t_j$ are travel cost and travel time of alternative $j$. The conditional logit model may be summarized by

$$P(Y = j) = \frac{\exp(\alpha + \tilde{x}_j^t \delta)}{1 + \sum_{k=1}^{J-1} \exp(\alpha + \tilde{x}_k^t \delta)},$$

where $\tilde{x}_j = x_j - x_J$ for $j = 1, \ldots, J-1$. For urban travel demand, we obtain

$$P(Y = j) = \frac{\exp(\alpha + \tilde{c}_j^t \delta_1 + \tilde{t}_j^t \delta_2)}{1 + \sum_{k=1}^{2} \exp(\alpha + \tilde{c}_k^t \delta_1 + \tilde{t}_k^t \delta_2)},$$

where $\tilde{c}_j = c_j - c_J$, $\tilde{t}_j = t_j - t_J$, for $j = 1, 2$. The estimated parameter $\hat{\beta} = (\hat{\alpha}, \hat{\delta})$ is obtained by likelihood maximisation, using Fisher's scoring algorithm.

Now, imagine that the construction of a tram system is planned. The probability of each individual using it can be predicted to assess the resident's demand. Since the tram system (new alternative 0) does not exist, the attributes values $c_0$ and $t_0$ are not yet available. They must therefore be simulated or estimated using observed data in another city where a tram is already used. Let $\bar{c}_0$, $\bar{t}_0$ be the corresponding values and $\tilde{c}_0$, $\tilde{t}_0$ the translated values with respect to the reference alternative $\tilde{c}_0 = \bar{c}_0 - c_3$, $\tilde{t}_0 = \bar{t}_0 - t_3$. The predicted probability of choosing the tramway is

$$\hat{P}(Y = 0) = \frac{\exp(\hat{\alpha} + \tilde{c}_0^t \hat{\delta}_1 + \tilde{t}_0^t \hat{\delta}_2)}{1 + \sum_{k=0}^{2} \exp(\hat{\alpha} + \tilde{c}_k^t \hat{\delta}_1 + \tilde{t}_k^t \hat{\delta}_2)}.$$

More generally, the predicted probability of choosing the new alternative 0 is

$$\hat{P}(Y = 0) = \frac{\exp(\hat{\alpha} + \tilde{x}_0^t \hat{\delta})}{1 + \sum_{k=0}^{J-1} \exp(\hat{\alpha} + \tilde{x}_k^t \hat{\delta})}.$$

It should be noted that adding the new alternative 0 does not change the ratio of probabilities $\pi_j / \pi_k$ of other alternatives $j, k$ because of the IIA property. For example, the number of bus users is still three times the number of bicycle users after construction of the tram system. But the new alternative needs to be different from other alternatives (reminiscent of the blue/red bus paradox).

### 5.4.2 *(r,F,Z)* approach

#### 5.4.2.1 Model specification

Qualitative choice models of form (5.1) are exactly canonical (reference, logistic, $Z$) models. The three previously mentioned logit models correspond to different forms of the design matrix $Z$:

- Multinomial logit model: $Z = Z_c$ with

$$Z_c = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix}.$$

This is the complete design matrix.

- Conditional logit model: $Z = \tilde{Z}$ with

$$\tilde{Z} = \begin{pmatrix} 1 & \tilde{x}_1^t \\ \vdots & \vdots \\ 1 & \tilde{x}_{J-1}^t \end{pmatrix},$$

where $\tilde{x}_j = x_j - x_J$ (translation with respect to reference category $J$).

- Universal logit model: $Z = \tilde{Z}_c$ with

$$\tilde{Z}_c = \begin{pmatrix} 1 & & & \tilde{x}_1^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & \tilde{x}_{J-1}^t \end{pmatrix}.$$

These three types of qualitative choice models can be easily extended using other cdfs $F$. They are useful for qualitative choices, especially because of the reference ratio. Also the reference category can be changed (with a transposition $\tau$) to obtain a better fit (changing also the translated variables $\tilde{x}_j = x_j - x_{\tau(J)}$ for $j \neq \tau(J)$).

#### 5.4.2.2 Independence of irrelevant alternatives

**Property 18.** *All the reference models share the IIA property.*

*Proof.* All the reference models have the form (reference, $F$, $Z)_\tau$, where $\tau$ is a transposition of the reference category $J$. Without loss of generality, we simply consider the case of the reference category $J$ (i.e. $\tau$ is the identical transposition). Assume that the distribution of $Y|X = x$ is defined by a reference model

$$\frac{\pi_j}{\pi_j + \pi_J} = F(\eta_j),\ j = 1, ..., J-1,$$

$$\Leftrightarrow \pi_j = \frac{F(\eta_j)}{1 - F(\eta_j)} \pi_J,\ j = 1, ..., J-1.$$

Thus, for two alternatives $j, k \in \{1, ..., J-1\}$

$$\frac{\pi_j}{\pi_k} = \frac{F(\eta_j)/[1 - F(\eta_j)]}{F(\eta_k)/[1 - F(\eta_k)]},$$

and

$$\frac{\pi_j}{\pi_J} = F(\eta_j)/[1 - F(\eta_j)].$$

Finally, the ratio of probabilities of two alternatives $\pi_j/\pi_k$ does not depend on other alternatives. □

In fact, the IIA property is related to the ratio, not the logistic cdf of the multinomial logit model. By contrast, the adjacent, cumulative and sequential models do not share the IIA property because they use the category ordering assumption (see Appendix D for details). It should be noted that the situation of new alternative is not easily interpretable for ordered alternatives.

#### 5.4.2.3 Prediction for a new alternative

We propose to extend the conditional logit model specified by the (reference, logistic, $\tilde{Z}$) triplet to (reference, $F$, $\tilde{Z}$) models

$$P(Y=j)=\frac{F(\alpha+\tilde{x}_j^t\delta)/(1-F(\alpha+\tilde{x}_j^t\delta))}{1+\sum_{k=1}^{J-1}F(\alpha+\tilde{x}_k^t\delta)/(1-F(\alpha+\tilde{x}_k^t\delta))},$$

for $j=1,\ldots,J-1$. The cdf $F$ must first be selected using error of misclassification or log-likelihood criteria. The log-likelihood does not need to be penalized because all the proposed models have the same design matrix and thus the same number of parameters $1+p$. After a comparison with classical cdfs (logistic, normal, Laplace, Gumbel min, Gumbel max and Student($d$)), the best model is obtained with cdf $\bar{F}$ and parameters $\hat{\beta}=(\hat{\alpha},\hat{\delta})$. The predicted probability of a new alternative 0 is

$$\hat{P}(Y=0)=\frac{\bar{F}(\hat{\alpha}+\tilde{x}_0^t\hat{\delta})/(1-\bar{F}(\hat{\alpha}+\tilde{x}_0^t\hat{\delta}))}{1+\sum_{k=0}^{J-1}\bar{F}(\hat{\alpha}+\tilde{x}_k^t\hat{\delta})/(1-\bar{F}(\hat{\alpha}+\tilde{x}_k^t\hat{\delta}))}.$$

Furthermore the reference alternative can be changed to obtain a better fit. The best model is obtained with cdf $\bar{F}$, reference alternative $\bar{j}$ and parameters $\hat{\beta}=(\hat{\alpha},\hat{\delta})$. In this case the predicted probability of a new alternative 0 is

$$\hat{P}(Y=0)=\frac{\bar{F}(\hat{\alpha}+\tilde{x}_0^t\hat{\delta})/(1-\bar{F}(\hat{\alpha}+\tilde{x}_0^t\hat{\delta}))}{1+\sum_{k=0,k\neq\bar{j}}^{J}\bar{F}(\hat{\alpha}+\tilde{x}_k^t\hat{\delta})/(1-\bar{F}(\hat{\alpha}+\tilde{x}_k^t\hat{\delta}))},$$

where $\tilde{x}_j=x_j-x_{\bar{j}}$ for $j\neq\bar{j}$.

**Discussion** The theoretical work conducted on this perspective has almost been completed but remains questionable. On the one hand, the proposed qualitative choice models do not respect the principle of random utility maximisation. On the other hand, does a better fit or a smallest error of misclassification mean better predictions?

With regard to applications, the proposed qualitative choice models could be estimated using benchmark datasets and could be compared with the conditional logit model. Ideally, we would like to obtain the real observed proportions of the new alternatives and compare predicted and true values. For Fisher's scoring algorithm, only the design matrix has to be modified. But the explanatory variables $x=\{x_j\}_{j=1,\ldots,J}$ must be considered with dependencies on alternatives.

Appendix A

# Proof of Property 4

The following proof is a generalisation of that described by Dobson (2002) for the multivariate case (i.e. $K > 1$). From the definition of a probability density function, the area under the curve is unity

$$\int f(y, \mu)\, dy = 1, \tag{A.1}$$

where integration is over all possible values of $y$ (if the random variable $Y$ is discrete then integration is replaced by summation). If we differentiate both sides of (A.1) with respect to $\mu$ we obtain

$$\frac{\partial}{\partial \mu}\left\{\int f(y,\mu)dy\right\} = \frac{\partial}{\partial \mu} 1 = 0_K,$$

where $0_K$ is the null vector of dimension $K$. Assuming differentiability of functions $\theta$ and $b$ with respect to $\mu$, we can apply the Leibniz integral rule and obtain

$$\int \frac{\partial}{\partial \mu} f(y,\mu)dy = 0_K. \tag{A.2}$$

For distribution of the exponential family we have

$$\begin{aligned}\frac{\partial}{\partial \mu} f(y,\mu) &= \left\{\frac{\partial}{\partial \mu}\left(a(y)^t\theta(\mu)\ + b(\mu) + c(y)\right)\right\} \cdot f(y,\mu)\\ \frac{\partial}{\partial \mu} f(y,\mu) &= \{\mathcal{J}_\theta(\mu) a(y) + \nabla_b(\mu)\} \cdot f(y,\mu) \end{aligned} \tag{A.3}$$

Therefore (A.2) becomes

$$\begin{aligned}\int \{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\} \cdot f(y,\theta)dy &= 0_K\\ \int \{\mathcal{J}_\theta(\mu) * a(y)\} \cdot f(y,\theta)dy &= -\nabla_b(\mu) \cdot \int f(y,\theta)dy\\ \mathcal{J}_\theta(\mu) * \mathrm{E}[a(Y)] &= -\nabla_b(\mu)\end{aligned} \tag{A.4}$$

and thus the desired result *(i)*.

A similar method can be used to obtain $\mathrm{Var}[a(Y)]$.

$$\frac{\partial^2}{\partial \mu^t \partial \mu}\left\{\int f(y,\mu)dy\right\} = \frac{\partial^2}{\partial \mu^t \partial \mu} \cdot 1 = 0_{K\times K},$$

where $0_{K\times K}$ is the null matrix of dimension $K \times K$. Thus

$$\int \frac{\partial^2}{\partial \mu^t \partial \mu} f(y,\mu)\, dy = 0_{K\times K} \tag{A.5}$$

Using (A.3), we obtain

$$\frac{\partial^2}{\partial\mu^t\partial\mu}f(y,\mu) = \frac{\partial}{\partial\mu^t}\left[\{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\} \cdot f(y,\mu)\right]$$

$$\frac{\partial^2}{\partial\mu^t\partial\mu}f(y,\mu) = \left[\frac{\partial}{\partial\mu^t}\{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\}\right] \cdot f(y,\mu) + \{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\}\frac{\partial}{\partial\mu^t}f(y,\mu).$$

Using (A.3) again, we obtain

$$\frac{\partial^2}{\partial\mu^t\partial\mu}f(y,\mu) = \left[\frac{\partial}{\partial\mu^t}\{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\}\right] \cdot f(y,\mu) \tag{A.6}$$

$$+ \{\mathcal{J}_\theta(\mu) * a(y) + \nabla_b(\mu)\}\{\mathcal{J}_\theta(\mu)a(y) + \nabla_b(\mu)\}^t \cdot f(y,\mu). \tag{A.7}$$

Computation of part (A.6)

$$\begin{aligned}\frac{\partial}{\partial\mu^t}\{\mathcal{J}_\theta(\mu)a(y) + \nabla_b(\mu)\} &= \frac{\partial}{\partial\mu^t}\{\mathcal{J}_\theta(\mu)a(y)\} + \mathcal{H}_b(\mu)\\ &= \left\{\left(\frac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t a(y)\right\}_{i,j} + \mathcal{H}_b(\mu).\end{aligned}$$

Computation of part (A.7)

$$\begin{aligned}\mathcal{J}_\theta(\mu)a(y) + \nabla_b(\mu) &= \mathcal{J}_\theta(\mu)\{a(y) - \mathrm{E}[a(Y)] + \mathrm{E}[a(Y)]\} + \nabla_b(\mu)\\ &= \mathcal{J}_\theta(\mu)\{a(y) - \mathrm{E}[a(Y)]\} + \underbrace{\mathcal{J}_\theta(\mu) * \mathrm{E}[a(Y)] + \nabla_b(\mu)}_{=0_K \text{ according to (A.2)}}.\end{aligned}$$

Finally (A.3) becomes

$$\begin{aligned}\frac{\partial^2}{\partial\mu^t\partial\mu}f(y,\mu) = &\left[\left\{\left(\frac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t a(y)\right\}_{i,j} + \mathcal{H}_b(\mu)\right] \cdot f(y,\mu)\\ &+ \mathcal{J}_\theta(\mu)\{a(y) - \mathrm{E}[a(Y)]\}\{a(y) - \mathrm{E}[a(Y)]\}^t\mathcal{J}_\theta^t(\mu).\end{aligned}$$

By integrating with respect to $y$ and using (A.5), we obtain

$$\begin{aligned}&\left[\int\left\{\left(\frac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t a(y)\right\}_{i,j} \cdot f(y,\mu)dy\right] + \mathcal{H}_b(\mu)\underbrace{\left[\int f(y,\mu)dy\right]}_{=1}\\ &+\mathcal{J}_\theta(\mu)\underbrace{\left[\int\{a(y) - \mathrm{E}[a(Y)]\}\{a(y) - \mathrm{E}[a(Y)]\}^t \cdot f(y,\theta)dy\right]}_{=\mathrm{Cov}[a(Y)]}\mathcal{J}_\theta^t(\mu) = 0_{K\times K},\end{aligned}$$

or equivalently

$$\left[\int\left\{\left(\frac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t a(y)\right\}_{i,j} \cdot f(y,\mu)dy\right] + \mathcal{H}_b(\mu) + \mathcal{J}_\theta(\mu)\mathrm{Cov}[a(Y)]\mathcal{J}_\theta^t(\mu) = 0_{K\times K}.$$

The scalar product and expectation are linear functions and we thus obtain

$$\left\{\left(\frac{\partial^2\theta}{\partial\mu_j\partial\mu_i}\right)^t \mathrm{E}[a(y)]\right\}_{i,j} + \mathcal{H}_b(\mu) + \mathcal{J}_\theta(\mu)\mathrm{Cov}[a(Y)]\mathcal{J}_\theta^t(\mu) = 0_{K\times K}.$$

Finally, using the first result *(i)*, we obtain the second result *(ii)*.

Appendix B

# Details on Fisher's scoring algorithm

Below are details on computation of the Jacobian matrix $\frac{\partial \pi}{\partial r}$ for four different ratios.

**Reference** For the reference ratio we have for $j = 1, \ldots, J-1$

$$\pi_j = \frac{r_j}{1 - r_j} \pi_J. \tag{B.1}$$

Summing on $j$ from 1 to $J-1$ we obtain

$$\pi_J = \frac{1}{1 + \sum_{j=1}^{J-1} \frac{r_j}{1-r_j}}.$$

The derivative of the product (B.1) with respect to $r_i$ is

$$\frac{\partial \pi_j}{\partial r_i} = \frac{\partial}{\partial r_i} \left( \frac{r_j}{1 - r_j} \right) \pi_J + \frac{r_j}{1 - r_j} \frac{\partial \pi_J}{\partial r_i}. \tag{B.2}$$

For the term of the sum part we obtain

$$\frac{\partial}{\partial r_i} \left( \frac{r_j}{1 - r_j} \right) = \begin{cases} \dfrac{1}{(1 - r_i)^2} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

For the second term of the sum we obtain

$$\frac{\partial \pi_J}{\partial r_i} = -\frac{1}{(1 - r_i)^2} \pi_J^2.$$

Then (B.2) becomes

$$\frac{\partial \pi_j}{\partial r_i} = \begin{cases} \dfrac{1}{r_i(1 - r_i)} \left[ \dfrac{r_i}{1 - r_i} \pi_J - \left( \dfrac{r_i}{1 - r_i} \pi_J \right)^2 \right] & \text{if } i = j, \\ -\dfrac{1}{r_i(1 - r_i)} \dfrac{r_i}{1 - r_i} \dfrac{r_j}{1 - r_j} \pi_J^2 & \text{otherwise.} \end{cases}$$

Using (B.1) again we obtain

$$\frac{\partial \pi_j}{\partial r_i} = \frac{1}{r_i(1 - r_i)} \begin{cases} \pi_i(1 - \pi_i) & \text{if } i = j, \\ -\pi_i \pi_j & \text{otherwise.} \end{cases}$$

Finally we have

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

for row $i$ and column $j$ of the Jacobian matrix.

**Adjacent** For the adjacent ratio we have for $j = 1, \dots, J-1$

$$\pi_j = \frac{r_j}{1-r_j}\pi_{j+1}.$$

and thus

$$\pi_j = \left( \prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) \pi_J. \tag{B.3}$$

Summing on $j$ from 1 to $J-1$ we obtain

$$\pi_J = \frac{1}{1 + \sum_{j=1}^{J-1} \prod_{k=j}^{J-1} \frac{r_k}{1-r_k}}.$$

The derivative of the product (B.3) with respect to $r_i$ is

$$\frac{\partial \pi_j}{\partial r_i} = \frac{\partial}{\partial r_i} \left( \prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) \pi_J + \left( \prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) \frac{\partial \pi_J}{\partial r_i}. \tag{B.4}$$

For the first term of the sum we obtain

$$\frac{\partial}{\partial r_i} \left( \prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) = \begin{cases} \dfrac{1}{r_i(1-r_i)} \left( \displaystyle\prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) & \text{if } i \geq j, \\ 0 & \text{otherwise.} \end{cases}$$

For the second term of the sum we obtain

$$\frac{\partial \pi_J}{\partial r_i} = -\frac{1}{r_i(1-r_i)} \left( \sum_{k=1}^{i} \prod_{k=j}^{J-1} \frac{r_k}{1-r_k} \right) \pi_J^2.$$

Using (B.3) it becomes

$$\frac{\partial \pi_J}{\partial r_i} = -\frac{\pi_J}{r_i(1-r_i)} \sum_{k=1}^{i} \pi_k.$$

Then (B.4) becomes

$$\frac{\partial \pi_j}{\partial r_i} = \begin{cases} \dfrac{1}{r_i(1-r_i)} \left( \pi_j - \pi_j \displaystyle\sum_{k=1}^{i} \pi_k \right) & \text{if } i \geq j, \\ -\dfrac{1}{r_i(1-r_i)} \pi_j \displaystyle\sum_{k=1}^{i} \pi_k & \text{otherwise.} \end{cases}$$

Finally we have

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

for row $i$ and column $j$ of the Jacobian matrix, where $\gamma_i = P(Y \leq i) = \sum_{k=1}^{i} \pi_k$.

**Sequential** For the sequential ratio we have for $j = 1, \dots, J-1$

$$\pi_j = r_j \prod_{k=1}^{j-1} (1 - r_k),$$

with the convention $\prod_{k=1}^{0}(1 - r_k) = 1$. Hence we obtain directly

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

for row $i$ and column $j$ of the Jacobian matrix.

**Cumulative** For the cumulative ratio we have for $j = 1, \dots, J-1$

$$\pi_j = r_j - r_{j-1},$$

with the convention $r_0 = 0$. Hence we obtain directly

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

# Appendix C
# Details on SMS-PCGLM

## Definition of hidden semi-Markov chains (HSMCs) and of semi-Markov switching partitioned conditional generalized linear models (SMS-PCGLMs) and associated statistical methods

### Semi-Markov chain

Let $\{S_t\}$ be a semi-Markov chain with finite-state space $\{0, \ldots, A-1\}$. This $A$-state semi-Markov chain $\{S_t\}$ is defined by the following parameters:

- initial probabilities $\varphi_a = P(S_0 = a)$ with $\sum_a \varphi_a = 1$.
- transition probabilities
  - non absorbing state $a$: $\forall b \neq a$, $\tilde{p}_{a,b} = P(S_t = b | S_t \neq a, S_{t-1} = a)$ with $\sum_{b \neq a} \tilde{p}_{a,b} = 1$ and $\tilde{p}_{a,a} = 0$,
  - absorbing state $a$: $p_{a,a} = P(S_t = a | S_{t-1} = a) = 1$ and $\forall b \neq a$, $p_{a,b} = 0$.

An explicit occupancy (or sojourn time) distribution is attached to each non absorbing state $a$:

$$d_a(u) = P(S_{t+u+1} \neq a, S_{t+u-v} = a, v = 0, \ldots, u-2 | S_{t+1} = a, S_t \neq a), \quad u = 1, 2, \ldots$$

Since $t = 0$ is assumed to correspond to a state entering, the following relation is verified

$$P(S_t \neq a, S_{t-v} = a, v = 1, \ldots, t) = d_a(t)\varphi_a.$$

We define as possible parametric state occupancy distributions binomial distributions, Poisson distributions and negative binomial distributions with an additional shift parameter $d$ ($d \geq 1$) which defines the minimum sojourn time in a given state.

The binomial distribution with parameters $d$, $n$ and $p$ ($q = 1 - p$), $\mathcal{B}(d, n, p)$ where $0 \leq p \leq 1$, is defined by

$$d_a(u) = \binom{n-d}{u-d} p^{u-d} q^{n-u}, \quad u = d, d+1, \ldots, n,$$

with $\mu = d + (n-d)p$ and $\sigma^2 = (n-d)pq$.

The Poisson distribution with parameters $d$ and $\lambda$, $\mathcal{P}(d, \lambda)$, where $\lambda$ is a real number ($\lambda > 0$), is defined by

$$d_a(u) = \frac{\exp(-\lambda)\lambda^{u-d}}{(u-d)!}, \quad u = d, d+1, \ldots$$

with $\mu = d + \lambda$ and $\sigma^2 = \lambda$.

The negative binomial distribution with parameters $d$, $r$ and $p$, $\mathbb{NB}(d, r, p)$, where $r$ is a real number ($r > 0$) and $0 < p \leq 1$, is defined by

$$d_a(u) = \binom{u-d+r-1}{r-1} p^r q^{u-d}, \quad u = d, d+1, \ldots$$

with $\mu = d + rq/p$ and $\sigma^2 = rq/p^2$.

### Partitioned conditional generalized linear models for categorical response variables

A PCGLM is defined by a partition tree of categories $\{1, \ldots, J\}$ and by a triplet specifying the GLM associated with each non-terminal node of the partition tree:

- Ratio of category probabilities i.e. reference ratio for non-ordered subset of categories defined by $\pi_j/(\pi_j + \pi_J)$ for the reference category $J$ ($\pi_j$ is the probability of category $j$), adjacent, sequential and cumulative ratios for ordered subset of categories defined respectively by $\pi_j/(\pi_j + \pi_{j+1})$, $\pi_j/(\pi_j + \ldots + \pi_J)$ and $\pi_1 + \ldots + \pi_j$.

- Cumulative distribution function of a continuous distribution (e.g. symmetric distributions such as the logistic, Gaussian or Student distributions, or non-symmetric distributions such as the Gumbel min or max distributions) whose inverse defined the mapping between the category probability ratios and the linear predictor.

- Design matrix for the parameterization of the linear predictor.

### Hidden semi-Markov chains

A hidden semi-Markov chain can be viewed as a pair of stochastic processes $\{S_t, Y_t\}$ where the "output" process $\{Y_t\}$ is related to the "state" process $\{S_t\}$, which is a finite-state semi-Markov chain, by a probabilistic function or mapping denoted by $f$ (hence $Y_t = f(S_t)$). Since the mapping $f$ is such that a given output may be observed in different states, the state process $\{S_t\}$ is not observable directly but only indirectly through the output process $\{Y_t\}$. This output process $\{Y_t\}$ is related to the semi-Markov chain $\{S_t\}$ by the observation (or emission) probabilities $b_a(y) = P(Y_t = y | S_t = a)$. The definition of observation probabilities expresses the assumption that the output process at time t depends only on the underlying semi-Markov chain at time t. In the case of a categorical observed variable such as the types of axillary productions, the observation probabilities are directly estimated (categorical observation distribution). In the case of a quantitative observed variable such as the internode length, discrete parametric observation distributions are estimated (in this context, binomial, Poisson or negative binomial distributions may be unshifted). In the multivariate case, the elementary observed variables at time t are assumed to be conditionally independent given the state $S_t = s_t$.

### Semi-Markov switching partitioned conditional generalized linear models

Compared to a simple HSMC, the difference concerns the output process $\{Y_t\}$ which is now modulated by explanatory variables $\{X_t\}$ that vary with the index parameter (in our case node rank). Observation distributions are thus replaced by regression models, in our case by PCGLMs, but the conditional independence assumption between state and output processes is unchanged.

### Statistical methods and algorithms for hidden semi-Markov chains and semi-Markov switching partitioned conditional generalized linear models

The maximum likelihood estimation of the parameters of a HSMC requires an iterative optimization technique, which is an application of the expectation-maximization (EM) algorithm. The maximum likelihood estimation of the parameters of a SMS-PCGLM requires a variant of the EM algorithm called the gradient EM algorithm. Compared to the estimation of hidden semi-Markov chains using the simple EM algorithm, the main difference concerns the M-step of the algorithm where the direct maximization of observation distributions is replaced by the iterative Fisher's scoring algorithm.

Once a HSMC or a SMS-PCGLM has been estimated, the most probable state sequence $\boldsymbol{s}^*$ with its associated posterior probability $P(\boldsymbol{S} = \boldsymbol{s}^* | \boldsymbol{Y} = \boldsymbol{y})$ can be computed for each observed sequence y using the so-called Viterbi algorithm. In our application context, the most probable state sequence can be interpreted as the optimal segmentation of the corresponding observed sequence in successive branching zones; see Guédon (2003, 2005); Guedon et al. (2007) for statistical methods for hidden semi-Markov chains (with exception of the iterative M-step for the estimation, all the other statistical methods and algorithms for HSMCs directly apply to SMS-PCGLM).

Appendix D

# Details on IIA property

All the reference models share the IIA property. By contrast we show that adjacent, sequential and cumulative models do not share this property because they are defined using the category ordering assumption.

**Adjacent** For adjacent models we have

$$\frac{\pi_j}{\pi_j + \pi_{j+1}} = F(\eta_j),\ \forall j \in \{1, \ldots, J-1\},$$

or equivalently

$$\pi_j = \frac{F(\eta_j)}{1 - F(\eta_j)} \pi_{j+1},\ \forall j \in \{1, \ldots, J-1\}.$$

Thus for $j, k \in \{1, \ldots, J\}$ with $j < k$ we have

$$\frac{\pi_j}{\pi_k} = \prod_{m=j}^{k-1} \frac{F(\eta_m)}{1 - F(\eta_m)}.$$

Finally the ratio of probabilities of two alternatives $\pi_j / \pi_k$ depends on intermediate alternatives.

**Sequential** For sequential models we have

$$\frac{\pi_j}{\pi_j + \ldots + \pi_J} = F(\eta_j),\ \forall j \in \{1, \ldots, J-1\},$$

or equivalently

$$\pi_j = F(\eta_j) \prod_{k=1}^{j-1} [1 - F(\eta_k)],\ \forall j \in \{1, \ldots, J-1\}.$$

Then for $j, k \in \{1, \ldots, J\}$ with $j > k$ we have

$$\frac{\pi_j}{\pi_k} = \frac{F(\eta_j)}{F(\eta_k)} \prod_{m=k}^{j-1} [1 - F(\eta_m)],$$

with the convention $\eta_J = \infty$. Finally, the ratio of probabilities of two alternatives $\pi_j / \pi_k$ depends on intermediate alternatives.

**Cumulative** For cumulative models we have

$$\pi_1 + \ldots + \pi_j = F(\eta_j),\ \forall j \in \{1, \ldots, J-1\},$$

or equivalently

$$\pi_j = F(\eta_j) - F(\eta_{j-1}),\ \forall j \in \{1, \ldots, J\},$$

using the convention $\eta_0 = -\infty$ and $\eta_J = \infty$. Then for $j, k \in \{1, \ldots, J\}$ with $j \neq k$ we have

$$\frac{\pi_j}{\pi_k} = \frac{F(\eta_j) - F(\eta_{j-1})}{F(\eta_k) - F(\eta_{j-k})}.$$

# Bibliography

A. Agresti. *Categorical data analysis*, volume 359. John Wiley and Sons, 2002. 7, 9, 19, 36, 56, 59, 83

A. Agresti. *Analysis of ordinal categorical data*, volume 656. Wiley, 2010. 7, 9, 10, 19, 39, 43, 56, 68, 69

R. M. K. Altman. Mixed hidden markov models. *Journal of the American Statistical Association*, 102(477):201–210, 2007. 52

J. A. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–30, 1984. 8, 11, 20, 21, 39, 44, 45, 82, 92, 98

D. Barthélémy and Y. Caraglio. Plant architecture: a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Annals of Botany*, 99(3):375–407, 2007. 24

T. Bergstrom and M. Bagnoli. Log-concave probability and its applications. *Economic theory*, 26:445–469, 2005. 33, 40, 120

J. Burridge. A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–45, 1981. 40, 120

F. Chaubert-Pereira, Y. Caraglio, C. Lavergne, and Y. Guédon. Identifying ontogenetic, environmental and individual components of forest tree growth. *Annals of botany*, 104(5): 883–896, 2009. 11, 106, 116, 118

G. A. Churchill. Stochastic models for heterogeneous dna sequences. *Bulletin of mathematical biology*, 51(1):79–94, 1989. 51

I. B. Collings and T. Rydén. A new maximum likelihood gradient algorithm for on-line hidden markov model identification. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 4, pages 2261–2264. IEEE, 1998. 52

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972. 41

G. Debreu. Review of rd luce, individual choice behavior: A theoretical analysis. *American Economic Review*, 50(1):186–188, 1960. 38, 46, 88

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977. 52

A. J. Dobson. *An introduction to generalized linear models*. CRC Pr I Llc, 2002. 9, 19, 27, 29, 34, 56, 131

D. M. Doran and D. J. Newell. Manipulation in treatment of low back pain: a multicentre study. *British Medical Journal*, 2(5964):161, 1975. 47, 91, 98

L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models.* Springer Verlag, 2001. 7, 19, 36, 58, 83

J. D. Ferguson. Variable duration models for speech. In *Proc. Symposium on the application of hidden Markov models to text and speech*, pages 143–179, 1980. 51

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models.* 2006. 107

L. E. Gatsuk, O. V. Smirnova, L. I. Vorontzova, L. B. Zaugolnova, and L. A. Zhukova. Age states of plants of various growth forms: a review. *The Journal of Ecology*, pages 675–696, 1980. 8, 19, 117

C. Godin and Y. Caraglio. A multiscale model of plant topological structures. *Journal of theoretical biology*, 191(1):1–46, 1998. 24

Y. Guédon. Review of several stochastic speech unit models. *Computer Speech & Language*, 6(4):377–402, 1992. 51

Y. Guédon. Estimating hidden semi-markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3):604–639, 2003. 50, 53, 139

Y. Guédon. Hidden hybrid markov/semi-markov chains. *Computational statistics & Data analysis*, 49(3):663–688, 2005. 139

Y. Guedon and C. Cocozza-Thivent. Explicit state occupancy modelling by hidden semi-markov models: application of derin's scheme. *Computer Speech & Language*, 4(2):167–192, 1990. 53

Y. Guédon, D. Barthélémy, Y. Caraglio, and E. Costes. Pattern analysis in branching and axillary flowering sequences. *Journal of theoretical biology*, 212(4):481–520, 2001. 11, 20, 24, 25, 106

Y. Guedon, Y. Caraglio, P. Heuret, E. Lebarbier, and C. Meredieu. Analysing growth components in trees. *Journal of Theorical Biology*, pages 418–447, 2007. 139

H-H. Han, C. Coutand, H. Cochard, C. Trottier, and P-E. Lauri. Effects of shoot bending on lateral fate and hydraulics: invariant and changing traits across five apple genotypes. *Journal of experimental botany*, 58(13):3537–3547, 2007. 11, 106

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005. 75

P. Heuret, D. Barthélémy, Y. Guédon, X. Coulmier, and J. Tancre. Synchronization of growth, branching and flowering processes in the south american tropical tree cecropia obtusa (cecropiaceae). *American Journal of Botany*, 89(7):1180–1187, 2002. 24

Baillaud L. Structures répétitives spatiales et spatio-temporelles des plantes. *Phytomorphology*, 49:377–404, 1999. 24

E. Läärä and J. N. S. Matthews. The equivalence of two models for ordinal data. *Biometrika*, 72(1):206–207, 1985. 10, 42, 56, 66, 72

P-E. Lauri and E. Terouanne. The influence of shoot growth on the pattern of axillary development on the long shoots of young apple trees (malus domestica borkh.). *International journal of plant sciences*, pages 283–296, 1998. 11, 25, 106

S. E. Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45, 1986. 51

T-S. Lim, W-Y. Loh, and Y-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, 40(3): 203–228, 2000. 75

R. D. Luce. Individual choice behavior. 1959. 7, 9, 19, 37, 56, 127

J. Marschak. Binary-choice constraints and random utility indicators. In *Proceedings of a Symposium on Mathematical Methods in the Social Sciences*, 1960. 7, 19, 38, 127

G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982. 7, 9, 19, 56

A. E. Maxwell. *Analysing qualitative data*. Methuen London, 1961. 73

P. McCullagh. A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika*, 65(2):413–418, 1978. 10, 56, 69

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142, 1980. 7, 9, 19, 39, 56, 61, 71, 73

D. McFadden. Conditional logit analysis of qualitative choice behavior. 1973. 7, 19, 21, 38

D. McFadden. Conditional logit analysis of qualitative choice analysis. *Frontiers in Econometrics*, pages 105–142, 1974. 12, 47, 89, 119, 127

D. McFadden. The choice theory approach to market research. *Marketing science*, 5(4): 275–297, 1986. 127

D. McFadden et al. *Modelling the choice of residential location*. Institute of Transportation Studies, University of California, 1978. 7, 10, 19, 20, 45, 46, 81, 82, 85, 88, 90

B. Morawitz and G. Tutz. Alternative parameterizations in business tendency surveys. *Mathematical Methods of Operations Research*, 34(2):143–156, 1990. 10, 82

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384, 1972. 7, 9, 19, 27, 28, 56

E. Nicolini, B. Chanson, and F. Bonne. Stem growth and epicormic branch formation in understorey beech trees (fagus sylvatica l.). *Annals of Botany*, 87(6):737–750, 2001. 26

J. Peyhardi, E. Costes, Y. Caraglio, P-E. Lauri, C. Trottier, and Y. Guédon. Integrative models for analyzing jointly shoot growth and branching patterns. In *7th International Workshop on Functional Structural Plant Models*, 2013a. 118

J. Peyhardi, C. Trottier, and Y. Guédon. A unifying framework for specifying generalized linear models for categorical data. In *28th International Workshop on Statistical Modeling*, 2013b. 108, 117

J. W. Pratt. Concavity of the log likelihood. *Journal of the American Statistical Association*, 76(373):103–106, 1981. 40, 120

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 51, 52

M. Renton, Guédon Y., C. Godin, and E. Costes. Similarities and gradients in growth unit branching patterns during ontogeny in 'fuji'apple trees: a stochastic approach. *Journal of Experimental Botany*, 57(12):3131–3143, 2006. 11, 25, 106, 108

M. Russell and R. Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 5–8. IEEE, 1985. 51, 52

T. A. Steeves and I. M. Sussex. *Patterns in plant development*. Cambridge University Press, 1989. 8, 19

O. Taugourdeau, F. Chaubert-Pereira, S. Sabatier, and Y. Guédon. Deciphering the developmental plasticity of walnut saplings in relation to climatic factors and light environment. *Journal of experimental botany*, 62(15):5283–5296, 2011. 116

R. Turner. Direct maximization of the likelihood of a hidden markov model. *Computational Statistics & Data Analysis*, 52(9):4147–4160, 2008. 52

G. Tutz. Compound regression models for ordered categorical data. *Biometrical Journal*, 31 (3):259–272, 1989. 10, 20, 45, 47, 48, 82, 85, 90

G. Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43(1):39–55, 1990. 7, 9, 19, 56, 61

G. Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11(3):275–295, 1991. 10, 41, 42, 56, 60, 66, 68

G. Tutz. *Regression for categorical data*, volume 34. Cambridge University Press, 2012. 7, 19, 45, 59, 83, 100, 103

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32, 1976. 33

J. White. The plant as a metapopulation. *Annual Review of Ecology and Systematics*, 10: 109–145, 1979. 24

Q. Zhang and E. H. Ip. Generalized linear model for partially ordered data. *Statistics in Medicine*, 2012. 7, 10, 19, 20, 45, 46, 49, 82, 85, 86, 95, 96, 98, 100

**Titre:** Une nouvelle famille de modèles linéaires généralisés (GLMs) pour l'analyse de données catégorielles ; application à la structure et au développement des plantes.

**Résumé:** Le but de cette thèse est de proposer une nouvelle classe de GLMs pour une variable réponse catégorielle structurée hiérarchiquement, comme une variable partiellement ordonnée par exemple. Une première étape a été de mettre en évidence les différences et les point communs entre les GLMs pour variables réponses nominale et ordinale. Sur cette base nous avons introduit une nouvelle spécification des GLMs pour variable réponse catégorielle, qu'elle soit ordinale ou nominale, basée sur trois composantes : le ratio de probabilitées $r$, la fonction de répartition $F$ et la matrice de design $Z$. Ce cadre de travail nous a permis de définir une nouvelle famille de modèles pour données nominales, comparable aux familles de modèles cumulatifs, séquentiels et adjacents pour données ordinales. Puis nous avons défini la classe des modèles linéaires généralisés partitionnés conditionnels (PCGLMs) en utilisant des arbres orientés et la specification $(r, F, Z)$. Dans notre contexte biologique, les données sont des séquences multivariées composées d'une variable réponse catégorielle (le type de production axillaire) et de variables explicatives (longueur de l'entre-noeud par exemple). Dans les combinaisons semi-markoviennes de modèles linéaires généralisés partitionnés conditionnés (SMS-PCGLM) estimées sur la base de ces séquences, la semi-chaîne de Markov sous-jacente représente la succession et les longueurs des zones de ramification, tandis que les PCGLMs représentent, l'influence des variables explicatives de croissance sur les productions axillaires dans chaque zone de ramification. En utilisant ces modèles statistiques intégratifs, nous avons montré que la croissance de la pousse influençait des événements de ramification particuliers.

**Mots clés:** fonction de lien, variable nominale, variable ordinale, variable structurée hiérarchiquement, reparamétrisation de modèle, motif de ramification.

**Title:** A new generalized linear model (GLM) framework for analysing categorical data; application to plant structure and development.

**Abstract:** This thesis aims at proposing a new class of GLMs for a hierarchically-structured categorical response variable such as a partially-ordered variable for instance. A first step consisted of clarifying differences and commonalities between GLMs for nominal and ordinal response variables. On this basis we introduced a new specification of GLM for categorical response variable, weather ordinal or nominal, based on three components: the ratio of probabilities $r$, the cumulative distribution function $F$ and the design matrix $Z$. This framework allowed us to define a new family of models for nominal data, similar to the cumulative, sequential and adjacent families of models for ordinal data. Then we defined the class of partitioned conditional GLMs (PCGLMs) using directed trees and $(r, F, Z)$ specification. In our biological context, data takes the form of multivariate sequences associating a categorical response variable (type of axillary production) with explanatory variables (e.g. internode length). In the semi-Markov switching partitioned conditional generalized linear models (SMS-PCGLM) estimated on the basis of these sequences, the underlying semi-Markov chain represents both the succession and lengths of branching zones, while the PCGLMs represent the influence of growth explanatory variables on axillary productions within each branching zone. On the basis of these integrative statistical models, we showed that shoot growth influences specific branching events.

**Keywords:** link function, nominal variable, ordinal variable, hierarchically-structured variable, model reparametrization, branching pattern.