# Partitioned conditional generalized linear models for categorical data

Jean Peyhardi, Catherine Trottier, Yann Guédon

# Partitioned conditional generalized linear models for categorical data

Jean Peyhardi[1,2], Catherine Trottier[1], Yann Guédon[2]

[1] Université Montpellier 2, I3M, Montpellier, France
[2] CIRAD, UMR AGAP and Inria, Virtual Plants, Montpellier, France

E-mail for correspondence: jean.peyhardi@math.univ-montp2.fr

**Abstract:** In categorical data analysis, several regression models have been proposed for hierarchically-structured response variables, such as the nested logit model. 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 an arbitrary number 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 of GLMs for categorical data, PCGLMs can handle nominal, ordinal but also partially-ordered response variables.

**Keywords:** hierarchically-structured categorical variable; partition tree; partially-ordered variable; GLM specification.

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

The triplet $(r, F, Z)$ will play a key role in the following since each GLM for categorical data can be specified using this triplet; see Peyhardi et al. (2013) for more details. 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 $\beta$. All the classical link functions $g = (g_1, \ldots, g_{J-1})$, rely on the same structure which we propose to write as

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

where $F$ is a continuous and strictly increasing cumulative distribution function (cdf) and $r = (r_1, \ldots, r_{J-1})^t$ is a diffeomorphism from $\mathcal{M}$ to an

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014. 

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 $\mathcal{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$. In the following we will consider four particular ratios. The *adjacent*, *sequential* and *cumulative* ratios respectively defined by $\pi_j/(\pi_j + \pi_{j+1})$, $\pi_j/(\pi_j + \ldots + \pi_J)$ and $\pi_1 + \ldots + \pi_j$ for $j = 1, \ldots, J-1$, assume order among categories but with different interpretations. We introduce the *reference* ratio, defined by $\pi_j/(\pi_j + \pi_J)$ for $j = 1, \ldots, J-1$, useful for nominal response variables.
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 only the first one depends on the $(r, F, Z)$ triplet.

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T \frac{\partial \mathcal{F}}{\partial \eta} \frac{\partial \pi}{\partial r}}_{(r,F,Z) \text{ dependent part}} \underbrace{\mathrm{Cov}(Y|X=x)^{-1}\,[y-\pi]}_{(r,F,Z) \text{ independent part}}. \tag{2}$$

We need only to evaluate the density function $\{f(\eta_j)\}_{j=1,\ldots,J-1}$ to compute the corresponding 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 Peyhardi (2013).

## 2 Partitioned conditional GLMs

The main idea consists in recursively partitioning the $J$ categories then specifying a conditional GLM at each step. This type of model is therefore referred to as partitioned conditional GLM. Such models have already been proposed, such as the nested logit model (McFadden, 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 specific explanatory variables and design matrices. We are also seeking to formally define partitioned conditional GLMs for an arbitrary number of levels in the hierarchy.

**PCGLM definition:** Let $J \geq 2$ and $1 \leq k \leq J-1$. A ***k*-partitioned conditional GLM** for categories $1, \ldots, J$ is defined by:

- A **partition tree $\boldsymbol{\mathcal{T}}$** of $\{1, \ldots, J\}$ with $\mathcal{V}^*$, the set of non-terminal vertices of cardinal $\boldsymbol{k}$. Let $\Omega_j^v$ be the children of vertex $v \in \mathcal{V}^*$.
- A **collection of models $\boldsymbol{\mathfrak{C} = \{(r^v, F^v, Z^v(x^v)) \mid v \in \mathcal{V}^*\}}$** 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 $v$.

**PCGLM estimation:** Using the partitioned conditional structure of model, the log-likelihood can be decomposed as $l = \sum_{v \in \mathcal{V}^*} l^v$, where $l^v$ represents the log-likelihood of GLM $(r^v, F^v, Z^v(x^v))$. Each component $l^v$ can be maximised individually, using (2), if all parameters $\{\beta^v\}_{v \in \mathcal{V}^*}$ are different.

**PCGLM selection:** The partition tree $\mathcal{T}$ and the collection of models $\mathcal{C}$ have to be selected using ordering assumption among categories.

- *Nominal data*: the partition tree $\mathcal{T}$ is built by aggregating similar categories - such as the nested logit model of McFadden (1978) - and $\mathcal{C}$ contains only reference models, appropriate for nominal data; see Peyhardi (2013).
- *Ordinal data*: we propose to adapt the Anderson's indistinguishability procedure (1984) for PCGLM selection.
- *Partially ordered data*: the partial ordering assumption among categories can be summarized by an Hasse diagram. Zhang and Ip (2012) defined an algorithm to build the partition tree $\mathcal{T}$ automatically from the Hasse diagram; see figure 1 with the pear tree dataset. It should be remarked that every partially-ordered variable $Y$ can be expressed in terms of elementary ordinal and nominal variables $\tilde{Y}_i$ (with at least one ordinal variable). We propose to build the partition tree $\mathcal{T}$ directly from these latent variables $\tilde{Y}_i$ to obtain a more interpretable structure. For these two methods of partition tree building, the main idea is to recursively partition the $J$ categories in order to use a simple (ordinal or nominal) GLM at each step.

## 3 Application to pear tree dataset

**Dataset description:** In winter 2001, the first annual shoot of 50 one-year-old trees was described by node. The presence of an immediate axillary shoot was noted at each successive node. Immediate shoots were classified into four categories according to their 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 $X_1$ (internode length).

**Results:** A higher likelihood and simpler interpretations were obtained using partial ordering information. The axillary production $Y$ of pear tree can be decomposed into two levels. Production first follows a sequential mechanism (ordinal model), giving latent bud, short shoot or long shoot

(first level of hierarchy; figure 1), which is strongly influenced by the internode length $X_1$ (the longer the internode, the longer the axillary shoot). The axillary shoot apex then differentiates or not into spine (second level of hierarchy; figure 1) depending on distance to growth unit end (second explanatory variable $X_2$ expressed in number of nodes).

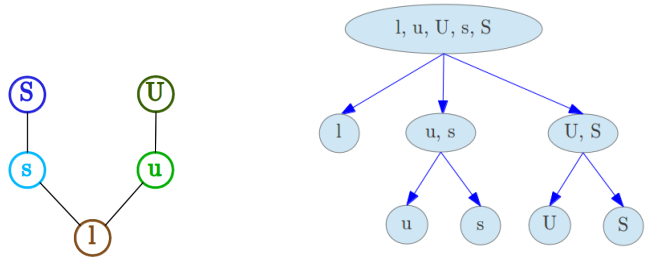


FIGURE 1. Hasse diagram and corresponding partition tree.

## References

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1 – 30.

McFadden, D. et al. (1978). Modelling the choice of residential location. *Institute of Transportation Studies, University of California.*

Peyhardi, J. (2013). A new GLM framework for analysing categorical data; application to plant structure and development. *PhD thesis.*

Peyhardi, J., Trottier, C. and Guédon, Y. (2013). A unifying framework for specifying generalized linear models for categorical data. *In 28th International Workshop on Statistical Modeling*, 331 – 335

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

Zhang, Q. and Ip, E.H. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine.*