Supplementary materials for
Partitioned conditional generalized linear models for categorical responses

Jean Peyhardi [1, 3], Catherine Trottier [2], and Yann Guédon [3]

[1] Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, 34095 Montpellier, France
[2] Institut Montpelliérain Alexander Grothendieck, Université Paul-Valéry Montpellier, 34199 Montpellier, France
[3] CIRAD, Amélioration Génétique et Adaptation des Plantes, and Inria, Virtual Plants, 34095 Montpellier, France

Address for correspondence: Jean Peyhardi, UFR Pharmacie 15 Avenue Charles Flahault, 34000 Montpellier, France.
E-mail: jean.peyhardi@umontpellier.fr.
Phone: (+33)4 11 75 96 81.

1 Examples of $(\boldsymbol{r}, \boldsymbol{F}, \boldsymbol{Z})$ specification

Table 1: $(\boldsymbol{r}, F, \boldsymbol{Z})$ specification of four generalized linear models for categorical responses

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

2 Proof of Proposition 1

The cardinal of vertex 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 subtree rooted at vertex v. Finally $|\mathcal{M}|$ denotes the number of independent regression equations of $\mathcal{M}$. Here we reason recursively on k, the cardinal of $\mathcal{V}^*$.

- **Initialisation:** For $k = 1$, the 1-PCGLM of any subset v of $\{1, \dots, J\}$ is a simple GLM for categorical responses with $|v| - 1$ regression equations and so the desired result.

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

  Let $\mathcal{M}$ be a $(k+1)$-PCGLM of $\{1, \dots, J\}$. Noting r the root vertex, 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 1-PCGLM 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

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

3 Indistinguishability procedure

3.1 Indistinguishability procedure with $(\boldsymbol{r}, F, \boldsymbol{Z})$ specification

Here we express the indistinguishability procedure in terms of canonical models by simply changing the design matrix. In fact, the hypothesis $H_{(3;r,s)}$ corresponds to the canonical (reference, logistic, $\boldsymbol{Z}_{r,s}$) model with

$$\boldsymbol{Z}_{r,s} = \left[\begin{array}{cccc|c|c} 1 & & & & \boldsymbol{x}^t & \\ & \ddots & & & \vdots & \\ \hline & & \ddots & & & \boldsymbol{x}^t \\ & & & \ddots & & \vdots \\ \hline & & & 1 & & \end{array}\right],$$

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

3.2 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, $\boldsymbol{Z}$) model with a block-structured design matrix $\boldsymbol{Z}$ 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 vertices. Let us describe this result in details using the block-structured design matrix $\boldsymbol{Z}_{r,s}$.

Proposition 1 *The canonical model (reference, logistic, $\boldsymbol{Z}_{r,s}$) is equivalent to the PCGLM specified in figure 1.*

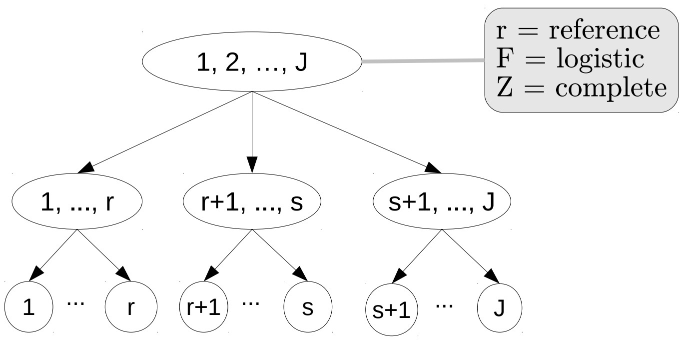

Figure 1: PCGLM specification of indistinguishability hypothesis $H_{(3,r,s)}$.

Proof: Assume that the distribution of Y is defined by the canonical (reference, logistic, $\boldsymbol{Z}_{r,s}$) model. We thus obtain

$$\frac{\pi_j}{\pi_J} = \begin{cases} \exp(\alpha_j + x^t\delta_1), & 1 \le j \le r, \\ \exp(\alpha_j + x^t\delta_2), & r < j \le s, \\ \exp(\alpha_j), & s < j \le J-1. \end{cases} \tag{3.1}$$

Let $\mathfrak{T}$ denote the partition tree of figure 1 and Ω_1, Ω_2 and Ω_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 (3.1), 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 $\boldsymbol{x}$ 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 follows exactly the expected PCGLM. As the parametrization is invertible, we obtain the equivalence.

Using this proposition, the canonical (reference, logistic, $\boldsymbol{Z}_{r,s}$) 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 Ω_1, Ω_2 and Ω_3).