Robust polytomous logistic regression

Miron, Julien; Poilane, Benjamin; Cantoni, Eva

How to cite

MIRON, Julien, POILANE, Benjamin, CANTONI, Eva. Robust polytomous logistic regression. In: Computational statistics & data analysis, 2022, p. 107564. doi: 10.1016/j.csda.2022.107564

This publication URL: https://archive-ouverte.unige.ch/unige:162777
Publication DOI: 10.1016/j.csda.2022.107564

Computational Statistics and Data Analysis 176 (2022) 107564

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Robust polytomous logistic regression

Julien Miron [1], Benjamin Poilane [*,1], Eva Cantoni

Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, 1211 Geneva, Switzerland

ARTICLE INFO

Article history:
Received 17 February 2021
Received in revised form 4 July 2022
Accepted 4 July 2022
Available online 13 July 2022

Keywords:
General linear models
M-estimators
Misclassification
Outliers
Polytomous regression
Robustness

ABSTRACT

In the context of polytomous regression, as with any generalized linear model, robustness issues are well documented. Existing robust estimators are designed to protect against misclassification, but do not protect against outlying covariates. It is shown that this can have a much bigger impact on estimation and testing than misclassification alone. To address this problem, two new estimators are introduced: a robust generalized linear model-type estimator and an optimal B-robust estimator, together with the corresponding Wald-type and score-type tests. Asymptotic distributions and variances of these estimators are provided as well as the asymptotic distributions of the test statistics under the null hypothesis. A complete comparison of the proposed new estimators and existing alternatives is presented. This is performed theoretically by studying the influence functions of the estimators, and empirically through simulations and applications to a medical dataset.

1. Introduction

Polytomous regression, or multinomial regression, is a classical tool of categorical data analysis, allowing the relationship between predictors and unordered categorical responses to be modeled (Agresti, 2012). It finds its applications in various disciplines, ranging from soil science (Kempen et al., 2009) to political studies (Mebane and Sekhon, 2004), to name a few. In particular, it is frequently used for inference on medical datasets where response variables can be medical procedures (Daniels and Gatsonis, 1997), behaviors of subjects (Blizzard and Hosmer, 2007), etc. Such datasets can contain outlying observations that, if not accounted for, can impact inference. Maximum likelihood estimation is not robust for the generalized linear models (GLMs) (Nelder and Wedderburn, 1972), which include polytomous regression. Robustness issues are documented by Pregibon (1981), Copas (1988) and Feng et al. (2014), among others, for binary regression (polytomous regression with two categories) and by Castilla et al. (2018) for polytomous regression.

One could be tempted to address the issue of outliers by removing them manually or by using a diagnostic tool such as Cook's distance or leverage measures; see Martín (2015) for the former and Lesaffre and Albert (1989) for the latter, both in the polytomous regression context. However, in addition to the reluctance to remove data from the analysis and the potential bias this could introduce, this solution could induce a masking effect where a few large outliers could mask others. The use of robust estimators is a better option.

Various approaches have been developed to obtain robust univariate GLM estimators. Optimal robust M-estimators are introduced in Stefanski et al. (1986), and Künsch et al. (1989) followed this work by restricting the optimality to the class

* Corresponding author at: Uni Mail – 5224, Bd du Pont-d'Arve 40, 1211 Geneva, Switzerland.
E-mail address: benjamin.poilane@unige.ch (B. Poilane).
[1] Contributed equally.

https://doi.org/10.1016/j.csda.2022.107564

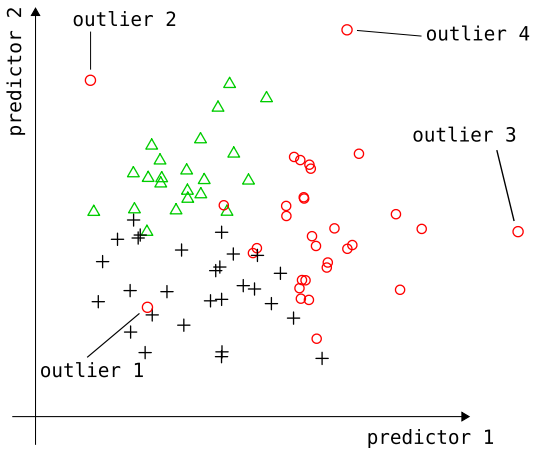

Fig. 1. Synthetic dataset, with two predictors and a response in three categories, denoted by different symbols, with four outliers. Outliers 1 and 2 have outlying responses (mislabelling) and outliers 2, 3 and 4 have outlying predictors.

of conditionally Fisher-consistent estimators. Cantoni and Ronchetti (2001) introduced quasi-likelihood-based estimators, implemented in the `robustbase` R package. More recently, Alqallaf and Agostinelli (2016) adapted the weighted maximum likelihood methodology developed by Markatou et al. (1998) to GLMs. One could also refer to Hung et al. (2018) for a minimum γ-divergence estimator. However, these methods apply to univariate GLMs and are not directly applicable to the multivariate setting of polytomous regression. Only a few adequate robust polytomous regression estimators are available: the robust generalized method of moments estimator by Wang (2014) and a minimum density power divergence (MDPD) estimator by Castilla et al. (2018). No reliable implementation of the generalized method of moments estimator of Wang (2014) is available and therefore it has not been assessed. On the other hand, the MDPD estimator of Castilla et al. (2018) is fast, efficient and offers protection against mislabelling.

However, as Fig. 1 illustrates, outliers can not only be observations with outlying responses (outliers 1 and 2 of Fig. 1) but also observations with extreme predictor values (outliers 2, 3, and 4 of Fig. 1). Although the minimum density power divergence estimator is more robust than the maximum likelihood estimator by protecting against mislabelling, it can still be highly affected by observations with outlying predictors, as shown in Sections 2.1 and 3.

The goal here is to document robustness issues in the polytomous regression context and to introduce two new robust estimators with their associated test statistics. The first is a B-robust weighted GLM estimator (RGLM), obtained by extending the robust GLM set-up of Cantoni and Ronchetti (2001, 2006) to the multinomial distribution, and the second is the optimal self-standardized B-robust (OBR) estimator, derived following Künsch et al. (1989). These two estimators improve on existing estimators by protecting not only against deviations in the responses, but also against outlying covariate values with a controlled efficiency loss. In addition, if the contamination scheme is unknown to the practitioner then the B-robust RGLM estimator and the associated test statistics should be favored because contamination in the predictor space can have a much stronger influence on inference than contamination in the responses, as illustrated in Section 3.

Section 2 details the maximum likelihood and the minimum density power divergence estimators and introduces the new robust GLM and the optimal B-robust estimators. Robustness properties of all of the aforementioned estimators are formally reviewed through the study of their influence functions; the corresponding Wald-type and score-type tests are also derived, as well as the asymptotic properties of both estimators and test statistics. Section 3 compares all of the estimators and the associated tests through extensive simulations. Section 4 applies the different methods to the *Vertebral column* dataset (Berthonnaud et al., 2005) and compares the estimators using cross-validation. Finally, Section 5 summarizes the findings and discusses extensions and further work.

2. Estimators for polytomous regression

For $i = 1, \ldots, n$, consider an independent sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, where $\boldsymbol{x}_i = \left(1, x_{i1}, ..., x_{ip}\right)^T$ denotes a vector of p predictors and an intercept, and $\boldsymbol{y}_i \in \{\boldsymbol{e}_1, ..., \boldsymbol{e}_k\}$ denotes a non-ordinal categorical response, where $\boldsymbol{e}_1, ..., \boldsymbol{e}_k$ is the canonical basis of $\mathbb{R}^k$. Assume that $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are independently drawn from a random variable $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X}$ follows an unspecified distribution F_X and $\boldsymbol{Y} \mid \boldsymbol{X}$ follows a multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)^T$, where π_i is the probability given $\boldsymbol{X}$ to belong to category i. For a vector $\boldsymbol{v}$, $\boldsymbol{v}^*$ denotes this vector without its last coordinate.

The polytomous regression model is a GLM (McCullagh and Nelder, 1983) defined through a matrix of parameters $\boldsymbol{\Gamma} \in \mathbb{R}^{(k-1)\times(p+1)}$ and a link function g. The probability vector $\boldsymbol{\pi}$ is made dependent on $k-1$ linear combinations of the coordinates of $\boldsymbol{x}$ through this link function g, which is a function from $\mathbb{R}^{k-1}$ to $[0,1]^k$ such that:

$$P\left(\boldsymbol{Y} = \boldsymbol{e}_j \mid \boldsymbol{X} = \boldsymbol{x}\right) = \pi_j = g^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{x}\right)_j, \quad \text{for } j = 1, \ldots, k. \tag{1}$$

A common choice for g is the logistic function. Noting $\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{x}$, it gives

$$\boldsymbol{\pi} = g^{-1}\left(\eta_1, \cdots, \eta_{k-1}\right) = \frac{1}{1+\sum_{j=1}^{k-1}\exp\left(\eta_j\right)}\left(\exp\left(\eta_1\right), .., \exp\left(\eta_{k-1}\right), 1\right)^T. \tag{2}$$

In what follows, π_y denotes $\boldsymbol{\pi}^{\boldsymbol{T}}\boldsymbol{y} = P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$.

2.1. Existing estimators

The maximum likelihood estimator $\hat{\boldsymbol{\Gamma}}_{ML}$ is the solution of the following estimating equations (McCullagh and Nelder, 1983, Chapter 5):

$$\sum_{i=1}^{n} \boldsymbol{\pi}^{*\prime T}_i \boldsymbol{V}^{*-1}_i \left(\boldsymbol{y}^*_i - \boldsymbol{\pi}^*_{\boldsymbol{i}}\right) \otimes \boldsymbol{x}_i = \boldsymbol{0}, \tag{3}$$

where $\boldsymbol{\pi}^{*\prime}_i \in \mathbb{R}^{(k-1)\times(k-1)}$ is the Jacobian matrix of $\boldsymbol{\pi}^*$ with respect to $\boldsymbol{\eta}$ evaluated at $\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{x}_i$, $\boldsymbol{V}^*_i = \mathrm{Var}[\boldsymbol{Y}^* \mid \boldsymbol{X} = \boldsymbol{x}_i] = \mathrm{diag}\left(\boldsymbol{\pi}^*_i\right) - \boldsymbol{\pi}^*_i\boldsymbol{\pi}^{*T}_i$ and $\otimes$ denotes the Kronecker product.

In the case of the polytomous regression with a logistic link defined by (2), $\boldsymbol{\pi}^{*\prime}_i = \boldsymbol{V}^*_i$ and Equation (3) simplifies to

$$\sum_{i=1}^{n} (\boldsymbol{y}^*_i - \boldsymbol{\pi}^*_i) \otimes \boldsymbol{x}_i = \boldsymbol{0}. \tag{4}$$

Define $\boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = (\boldsymbol{y}^* - \boldsymbol{\pi}^*) \otimes \boldsymbol{x}$, the estimating function of this estimator.

As detailed in Section 2.2, this estimator is not robust and can be highly biased by mislabelling and outlying $\boldsymbol{x}$ values. To overcome this issue, Castilla et al. (2018) extended the MDPD estimator from Ghosh and Basu (2016) to polytomous logistic regression. The resulting estimator $\hat{\boldsymbol{\Gamma}}_{MDPD}$ is the solution of the following estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{s}_{MDPD}\left(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\Gamma}\right) = \boldsymbol{0},$$

with

$$\boldsymbol{s}_{MDPD}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = w_\lambda\left(\pi_y\right)\boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) - \boldsymbol{\alpha}^{MDPD}(\boldsymbol{x}; \boldsymbol{\Gamma})$$

where $w_\lambda\left(\pi_y\right) = (\boldsymbol{\pi}^T\boldsymbol{y})^\lambda = \pi_y^\lambda$, $\boldsymbol{\alpha}^{MDPD}(\boldsymbol{x}; \boldsymbol{\Gamma}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}[w_\lambda(\pi_y)\boldsymbol{s}_{ML}(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})]$ and λ is a non-negative tuning parameter. The robustness of $\hat{\boldsymbol{\Gamma}}_{MDPD}$ increases with λ and if $\lambda = 0$, $\boldsymbol{s}_{MDPD} = \boldsymbol{s}_{ML}$ such that $\hat{\boldsymbol{\Gamma}}_{MDPD} = \hat{\boldsymbol{\Gamma}}_{ML}$.

2.2. Review of robustness concepts and properties

For a parameter $\boldsymbol{\Gamma}$, an M-estimator (Huber, 1964) $\hat{\boldsymbol{\Gamma}}_m$ is defined as the solution of

$$\sum_{i=1}^{n} \boldsymbol{s}_m\left(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{\Gamma}\right) = 0, \tag{5}$$

for a quite general function $\boldsymbol{s}_m$, called the estimating function of $\hat{\boldsymbol{\Gamma}}_m$. Denote by T_m the functional associated with the estimator $\hat{\boldsymbol{\Gamma}}_m$: for a distribution $F_{\boldsymbol{X},\boldsymbol{Y}}$ of the random variable $(\boldsymbol{X}, \boldsymbol{Y})$, $T_m(F_{\boldsymbol{X},\boldsymbol{Y}})$ is the solution in $\boldsymbol{\Gamma}$ of

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim F_{\boldsymbol{X},\boldsymbol{Y}}}\left[\boldsymbol{s}_m(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})\right] = 0.$$

If $\hat{F}_n$ denotes the empirical distribution function of the sample $\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)_{i=1,\ldots,n}$, then $\hat{\boldsymbol{\Gamma}}_m = T_m(\hat{F}_n)$. For $\varepsilon \in [0, 1]$, $\boldsymbol{x} \in \mathbb{R}^{p+1}$, and $\boldsymbol{y} \in \{\boldsymbol{e}_1, ..., \boldsymbol{e}_k\}$, define the contaminated distribution $F_\varepsilon = (1-\varepsilon)F_{\boldsymbol{X},\boldsymbol{Y}} + \varepsilon\Delta_{(\boldsymbol{x},\boldsymbol{y})}$, where $\Delta_{(\boldsymbol{x},\boldsymbol{y})}$ is a Dirac distribution at $(\boldsymbol{x}, \boldsymbol{y})$. If $\boldsymbol{\Gamma} = T_m(F_{\boldsymbol{X},\boldsymbol{Y}})$, the influence function of T_m at $\boldsymbol{\Gamma}$ evaluated at $(\boldsymbol{x}, \boldsymbol{y})$ is

$$IF_m(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = \lim_{\varepsilon\to 0}\frac{T_m(F_\varepsilon) - \boldsymbol{\Gamma}}{\varepsilon},$$

if this limit exists.

The influence function captures the asymptotic bias induced by a gross contamination by observations equal to $(\boldsymbol{x}, \boldsymbol{y})$, standardized by the mass of the contamination. High values of $IF_m(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma})$ indicate points with large influence on the estimation and thus weakness in the robustness of $\hat{\boldsymbol{\Gamma}}_m$. A bounded influence function is therefore an attractive feature of a robust estimator, and an estimator $\hat{\boldsymbol{\Gamma}}_m$ is said to be B-robust if its influence function is bounded with respect to $(\boldsymbol{x}, \boldsymbol{y})$.

B-robustness guarantees that a small contamination can only induce a limited bias on the estimate, whereas a non-B-robust estimator could be driven to arbitrarily large values by a single ill-placed observation.

For an M-estimator $\hat{\boldsymbol{\Gamma}}_m$ with an estimating function $\boldsymbol{s}_m$ differentiable with respect to $\boldsymbol{\Gamma}$, and such that the matrix $\boldsymbol{M}(\boldsymbol{\Gamma})$ defined below is non-singular, IF_m is proportional to the estimating function $\boldsymbol{s}_m$ (Hampel et al., 1986, Equation (4.2.9)):

$$IF_m(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = -\boldsymbol{M}(\boldsymbol{\Gamma})^{-1}\boldsymbol{s}_m(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}), \tag{6}$$

where $M(\boldsymbol{\Gamma}) = \mathbb{E}_{X,Y}\left[\partial \boldsymbol{s}_m(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})/\partial \boldsymbol{\Gamma}\right] = \mathbb{E}_{X,Y}\left[\boldsymbol{s}_m(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})\boldsymbol{s}_{ML}^T(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})\right]$.

Hence, the verification of the B-robustness of an M-estimator is reduced to the verification of the boundedness of its estimating function $\boldsymbol{s}_m$ with respect to $(\boldsymbol{x}, \boldsymbol{y})$. Noticing that $\hat{\boldsymbol{\Gamma}}_{ML}$ and $\hat{\boldsymbol{\Gamma}}_{MDPD}$ are M-estimators, their B-robustness is easily assessed through their estimating functions $\boldsymbol{s}_{ML}$ and $\boldsymbol{s}_{MDPD}$. Section A of the Supplementary material shows that if the space of the covariates is unbounded, $\hat{\boldsymbol{\Gamma}}_{ML}$ is not B-robust, and if $p \geq 2$, $\hat{\boldsymbol{\Gamma}}_{MDPD}$ is not B-robust either. The non-B-robustness of $\hat{\boldsymbol{\Gamma}}_{ML}$ and $\hat{\boldsymbol{\Gamma}}_{MDPD}$ motivated the introduction of the estimators $\hat{\boldsymbol{\Gamma}}_{RGLM}$ and $\hat{\boldsymbol{\Gamma}}_{OBR}$ proposed in Section 2.3.

B-robustness is necessary but not sufficient to guarantee stability of the estimators in practice, especially in the polytomous regression context. Indeed, as $\boldsymbol{y}$ can only take k distinct values, boundedness with respect to $(\boldsymbol{x}, \boldsymbol{y})$ is equivalent to boundedness with respect to $\boldsymbol{x}$. Hence, B-robustness only relates to robustness against outlying predictors (as presented by outliers 2, 3, and 4 in Fig. 1). However, a robust estimator should also be expected to decrease the influence of misclassified observations with non-outlying predictors, such as outlier 1 in Fig. 1. Robustness properties of estimators should be studied in more detail through the shape of their influence functions and their performance on contaminated datasets.

2.3. Robust GLM and optimal B-robust estimators

The robust GLM approach of Cantoni and Ronchetti (2001) for models with univariate responses can be extended to the multivariate setting of polytomous regression with a general link function.

Definition 1. The robust GLM estimator $\hat{\boldsymbol{\Gamma}}_{RGLM}$ is an M-estimator defined as the solution of

$$\sum_{i=1}^{n}\left[w_x(\boldsymbol{x}_i)\,\boldsymbol{\pi}_i^{*\prime T}\boldsymbol{V}_i^{*-1/2}\psi_{c_R}\{\boldsymbol{V}_i^{*-1/2}\left(\boldsymbol{y}_i^* - \boldsymbol{\pi}_i^*\right)\} \otimes \boldsymbol{x}_i - \boldsymbol{\alpha}^{RGLM}(\boldsymbol{x}_i; \boldsymbol{\Gamma})\right] = 0, \tag{7}$$

where $\boldsymbol{\alpha}^{RGLM}(\boldsymbol{x}_i; \boldsymbol{\Gamma}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_i}\left[w_x(\boldsymbol{x}_i)\,\psi_{c_R}\{\boldsymbol{V}_i^{*-1}\left(\boldsymbol{Y}^* - \boldsymbol{\pi}_i^*\right)\}\boldsymbol{\pi}_i^{*\prime} \otimes \boldsymbol{x}_i\right]$, w_x is a weighting function of the covariates to be defined, and, for a positive parameter c, ψ_c is the multivariate Huber function defined by

$$\psi_c(\boldsymbol{z}) = \begin{cases} \boldsymbol{z} & \text{if} \quad \|\boldsymbol{z}\|_2 \leq c \\ \dfrac{c}{\|\boldsymbol{z}\|_2}\boldsymbol{z} & \text{if} \quad \|\boldsymbol{z}\|_2 > c. \end{cases} \tag{8}$$

The choice of the weighting function w_x is discussed in Section 2.8.

Equation (7) can be interpreted as Equation (3) in which the standardized residuals $\boldsymbol{V}_i^{*-1/2}\left(\boldsymbol{y}_i^* - \boldsymbol{\pi}_i^*\right)$ have been Huberized to reduce the influence of misplaced points, weighted by $w_x(\boldsymbol{x_i})$, and to which the term $\boldsymbol{\alpha}^{RGLM}(\boldsymbol{x}_i; \boldsymbol{\Gamma})$ is added to ensure Fisher consistency. When using the logistic link function, $\boldsymbol{\pi}_i^{*\prime} = \boldsymbol{V}_i$ and $\|\boldsymbol{V}_i^{*-1/2}\left(\boldsymbol{y}_i^* - \boldsymbol{\pi}_i^*\right)\|_2 = \left(\pi_y^{-1} - 1\right)^{1/2}$ such that Equation (7) simplifies and yields the estimator $\hat{\boldsymbol{\Gamma}}_{RGLM}$ with estimating function:

$$\boldsymbol{s}_{RGLM}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = w_x(\boldsymbol{x})\,w_{c_R}\left(\pi_y\right)\boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) - \boldsymbol{\alpha}^{RGLM}(\boldsymbol{x}; \boldsymbol{\Gamma}), \tag{9}$$

where $w_{c_R}\left(\pi_y\right) = \min\left\{1, c_R\left(\pi_y^{-1} - 1\right)^{-1/2}\right\}$ (see Supplementary Section B for details on the derivation of this simplification). The constant $c_R > 0$ is a tuning parameter controlling robustness. Smaller values of c_R give more robust $\hat{\boldsymbol{\Gamma}}_{RGLM}$ (i.e., with smaller bounds on the influence function), and setting c_R to infinity makes $\hat{\boldsymbol{\Gamma}}_{RGLM}$ coincide with $\hat{\boldsymbol{\Gamma}}_{ML}$ (in this case ψ_{c_R} becomes the identity function).

Following Künsch et al. (1989), the optimal conditionally Fisher consistent B-robust estimator is defined in Definition 2.

Definition 2. The optimal conditionally Fisher-consistent B-robust estimator $\hat{\boldsymbol{\Gamma}}_{OBR}$ is the M-estimator defined by the estimating function:

$$\boldsymbol{s}_{OBR}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = \psi_{c_O}\{\boldsymbol{A}(\boldsymbol{\Gamma})\;\boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) - \boldsymbol{\alpha}^{OBR}(\boldsymbol{x}; \boldsymbol{\Gamma})\}, \tag{10}$$

where $\boldsymbol{A}(\boldsymbol{\Gamma}) \in \mathbb{R}^{(k-1)(p+1)\times(k-1)(p+1)}$ and $\boldsymbol{\alpha}^{OBR}(\boldsymbol{x}; \boldsymbol{\Gamma}) \in \mathbb{R}^{(k-1)(p+1)}$ are implicitly defined by:

$$\begin{cases} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}[\boldsymbol{s}_{OBR}(\boldsymbol{x}, \boldsymbol{Y}; \boldsymbol{\Gamma})] &= & 0 \\ \mathrm{Var}_{\boldsymbol{X},\boldsymbol{Y}}[\boldsymbol{s}_{OBR}(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\Gamma})] &= & \boldsymbol{I}_{(k-1)(p+1)} \end{cases} \tag{11}$$

and ψ_{c_O} is defined by Equation (8) with $c_O \geq \sqrt{(k-1)\times(p+1)}$.

The tuning parameter c_O acts in the same way as c_R for $\hat{\boldsymbol{\Gamma}}_{RGLM}$: smaller values of c_O give more robust $\hat{\boldsymbol{\Gamma}}_{OBR}$ and setting c_O to infinity makes $\hat{\boldsymbol{\Gamma}}_{OBR}$ coincide with $\hat{\boldsymbol{\Gamma}}_{ML}$. In practice, because no assumption is made on the distribution of $\boldsymbol{X}$, the second equation in (11) is approximated by:

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_i}[\boldsymbol{s}_{OBR}(\boldsymbol{x}_i,\boldsymbol{Y},\boldsymbol{\Gamma})]=\boldsymbol{I}_{(k-1)(p+1)}. \tag{12}$$

Using Equation (6), it is easy to check the B-robustness of both estimators:

Result 1. *B-robustness of $\hat{\boldsymbol{\Gamma}}_{RGLM}$ and $\hat{\boldsymbol{\Gamma}}_{OBR}$*

- *If $w_x(\boldsymbol{x})\cdot\boldsymbol{x}$ is bounded, the estimator $\hat{\boldsymbol{\Gamma}}_{RGLM}$ is B-robust.*
- *The estimator $\hat{\boldsymbol{\Gamma}}_{OBR}$ is B-robust if the matrix $\boldsymbol{A}$ is estimated from Equation (12) or is computed from the second Equation in (11) if the distribution of $\boldsymbol{X}$ is known.*

The proof of the first point follows directly from Equation (9) and the proof of the second point can be found in Künsch et al. (1989) under a more general setting. The estimator $\hat{\boldsymbol{\Gamma}}_{OBR}$ is optimal in the sense that it is admissible among the class of all conditionally Fisher-consistent M-estimators with a differentiable estimating function and with an influence function bounded by c_O for the metric associated with the inverse of their asymptotic variance matrix. For more details, see (Huber, 1981, Chapter 4.3).

2.4. Computational aspects

The R code accompanying this article provides functions to compute $\hat{\boldsymbol{\Gamma}}_{ML}$, $\hat{\boldsymbol{\Gamma}}_{RGLM}$ and $\hat{\boldsymbol{\Gamma}}_{OBR}$ as well as an implementation of $\hat{\boldsymbol{\Gamma}}_{MDPD}$, for comparison purposes in our simulation. The implementation of $\hat{\boldsymbol{\Gamma}}_{ML}$ and $\hat{\boldsymbol{\Gamma}}_{MDPD}$ uses Newton's method, and is guaranteed to converge for $\hat{\boldsymbol{\Gamma}}_{ML}$ from any starting point because the model likelihood is concave. Even though the density power divergence minimized by $\hat{\boldsymbol{\Gamma}}_{MDPD}$ is not convex, no numerical issues were encountered and Newton's method is adequate. The estimator $\hat{\boldsymbol{\Gamma}}_{RGLM}$ is estimated using Fisher's scoring algorithm (Lange, 2010, Chapter 14), which is fast and efficient with a well-chosen starting point. However, $\hat{\boldsymbol{\Gamma}}_{RGLM}$ corresponds to the maximum of a non-concave quasi-likelihood function. Thus, the choice of the starting point becomes important to avoid convergence issues, especially when the parameter dimension $(k-1)\times(p+1)$ is large. By default, the starting point is a modified maximum likelihood estimator computed only on observations with a w_x (details in Section 2.7 below) above a given threshold.

Due to the implicit definition of its estimating function, $\hat{\boldsymbol{\Gamma}}_{OBR}$ requires the use of the more complex IF algorithm detailed in Hampel et al. (1986). Indeed, the definition of the matrix $\boldsymbol{A}(\boldsymbol{\Gamma})$ and the vector $\boldsymbol{\alpha}^{OBR}(\boldsymbol{\Gamma})$ in Equations (11) or (12) is implicit. Hence, each estimating function evaluation requires an iterative procedure to compute $\boldsymbol{A}(\boldsymbol{\Gamma})$ and $\boldsymbol{\alpha}^{OBR}(\boldsymbol{\Gamma})$, leading to a more costly algorithm. Figure S1 in Supplementary Section C shows computational times of the estimators. Due to its iterative procedure, $\hat{\boldsymbol{\Gamma}}_{OBR}$ can take up to 50 times longer than $\hat{\boldsymbol{\Gamma}}_{RGLM}$, 75 times longer than $\hat{\boldsymbol{\Gamma}}_{MDPD}$, and 130 times longer than $\hat{\boldsymbol{\Gamma}}_{ML}$ when there are 18 parameters to estimate on 125 observations, which is not an extreme case (it corresponds to three categories and eight covariates).

2.5. Analytical comparisons

Fig. 2 shows the influence functions of $\hat{\boldsymbol{\Gamma}}_{ML}$ (plot a), $\hat{\boldsymbol{\Gamma}}_{MDPD}$ (plot b), $\hat{\boldsymbol{\Gamma}}_{RGLM}$ (plot c), and $\hat{\boldsymbol{\Gamma}}_{OBR}$ (plot d) with $p=2$ explanatory variables and $k=3$ categories. More precisely, it plots $\|IF_m(\boldsymbol{x},\boldsymbol{y}=\boldsymbol{e_1};\boldsymbol{\Gamma})\|_2$ as a function of $\boldsymbol{x}$, for m in {ML, MDPD, RGLM, OBR} and $\boldsymbol{\Gamma}=\begin{pmatrix}0 & 1.5 & \frac{\sqrt{3}}{2}\\ 0 & 0 & \sqrt{3}\end{pmatrix}$. This choice of $\boldsymbol{\Gamma}$, also used to generate the synthetic dataset of Fig. 1, is motivated by symmetry reasons: permuting the three categories corresponds to a one-third rotation of the covariate space. Hence, the influence functions evaluated at $\boldsymbol{y}=\boldsymbol{e}_2$ or $\boldsymbol{y}=\boldsymbol{e}_3$ are found by rotating the plots of Fig. 2 by $2\pi/3$ or $-2\pi/3$ respectively. The dotted red lines delimit regions where a category has a higher probability (computed with the true parameters) than others: the first category is more likely than others in the right region, the second category in the top left region, and the third one in the bottom left region. An informal interpretation of these plots would be that they indicate the bias induced by an extra observation with a response in the first category as a function of its position in the predictor space.

The B-robustness property can be checked graphically on Fig. 2. All influence functions go to zero when the covariates go to infinity in the right region of the plane. This implies that any observation in this region with a response in the first category sees its influence vanish as its covariates go to infinity. Thus, a well specified outlier such as outlier 3 is not influential for any estimator.

The influence functions of $\hat{\boldsymbol{\Gamma}}_{ML}$ and $\hat{\boldsymbol{\Gamma}}_{MDPD}$ are unbounded, whereas the influence functions of $\hat{\boldsymbol{\Gamma}}_{RGLM}$ and $\hat{\boldsymbol{\Gamma}}_{OBR}$ are bounded: the first two estimators are not B-robust whereas the last two are. An observation with a response in the first category but lying in the left half of the plane, such as outlier 1 or 2, can have an arbitrarily large influence on the

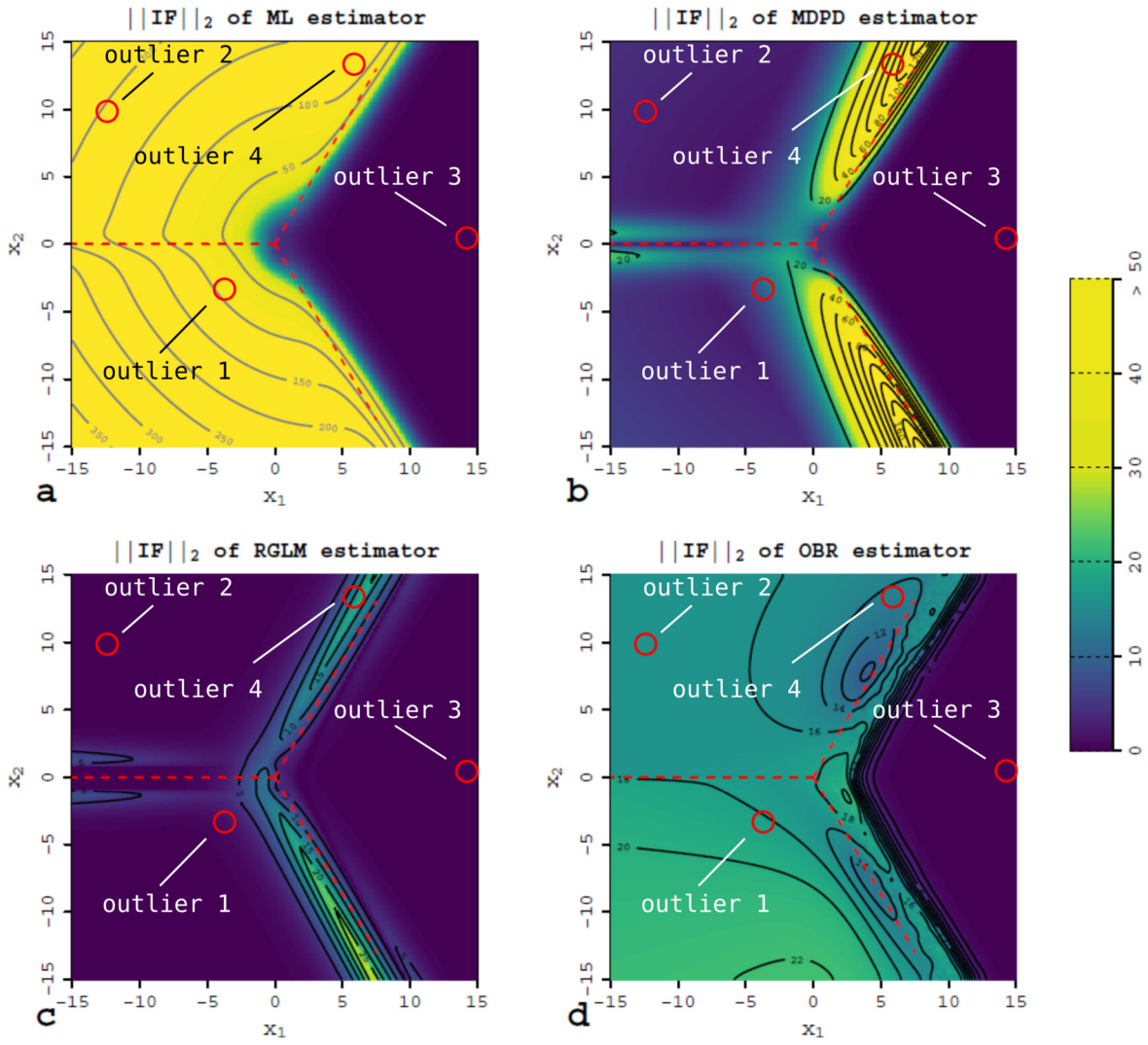

Fig. 2. Influence functions of $\hat{\mathbf{\Gamma}}_{ML}$ (a), $\hat{\mathbf{\Gamma}}_{MDPD}$ (b), $\hat{\mathbf{\Gamma}}_{RGLM}$ (c) and $\hat{\mathbf{\Gamma}}_{OBR}$ (d) estimators. Outliers of Fig. 1 have been reported on the plots to show their influence on the different estimators.

maximum likelihood estimator whereas all other estimators limit its influence. Although $\hat{\mathbf{\Gamma}}_{MDPD}$ is far more robust than $\hat{\mathbf{\Gamma}}_{ML}$, its weakness lies along the boundaries of the right region: if an observation with a response in first category falls in that region, its influence is proportional to $\|\boldsymbol{x}\|_2$. Outlier 4 can thus have a disproportionate influence on $\hat{\mathbf{\Gamma}}_{MDPD}$, but not on $\hat{\mathbf{\Gamma}}_{RGLM}$ or $\hat{\mathbf{\Gamma}}_{OBR}$.

A key difference between $\hat{\mathbf{\Gamma}}_{OBR}$ and the other robust estimators is that the influence functions of $\hat{\mathbf{\Gamma}}_{MDPD}$ and $\hat{\mathbf{\Gamma}}_{RGLM}$ go to zero with the norm of the predictor for grossly misclassified observations in the manner of redescending estimators, whereas the influence function of $\hat{\mathbf{\Gamma}}_{OBR}$ saturates: outlier 2 has a smaller influence than outlier 1 on $\hat{\mathbf{\Gamma}}_{MDPD}$ and $\hat{\mathbf{\Gamma}}_{RGLM}$, whereas these two outliers have roughly the same influence on $\hat{\mathbf{\Gamma}}_{OBR}$.

Comparison of $\hat{\mathbf{\Gamma}}_{MDPD}$ and $\hat{\mathbf{\Gamma}}_{RGLM}$ can be pushed further by noticing that, setting $w_x = 1$ for $\hat{\mathbf{\Gamma}}_{RGLM}$, both estimators are weighted maximum likelihood estimators with estimating functions of the form:

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}; \mathbf{\Gamma}) = w(\pi_y)\, \boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \mathbf{\Gamma}) - \boldsymbol{\alpha}(\boldsymbol{x}; \mathbf{\Gamma}). \tag{13}$$

Weighting functions w_{c_R} and w_λ both downweight points with unlikely response under the model. However, Fig. 3 shows that w_{c_R} only downweights points with a probability less than $(1+c_R^2)^{-1}$, whereas w_λ (with $\lambda > 0$) downweights all points. The parameter c_R has the nice interpretation of controlling the probability under which an observation is downweighted by $\hat{\mathbf{\Gamma}}_{RGLM}$.

For some specific settings, $\hat{\mathbf{\Gamma}}_{RGLM}$ offers robustness properties similar to $\hat{\mathbf{\Gamma}}_{MDPD}$. One example is the *Mammography experience* data used in Castilla et al. (2018) (see Supplementary Section D). As this dataset only contains bounded covariates, most of which are binary, potential outliers can only influence estimation through their response (similar to outlier 1 in

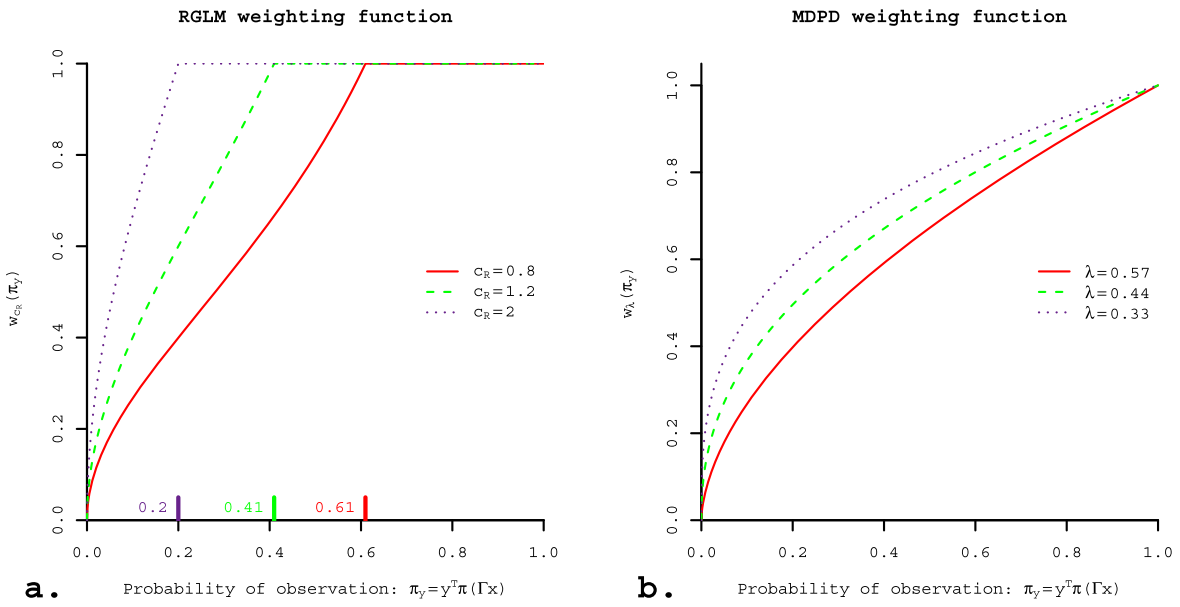

Fig. 3. Comparison of weights of $\hat{\mathbf{\Gamma}}_{RGLM}$ (a) and $\hat{\mathbf{\Gamma}}_{MDPD}$ (b) estimators as functions of the probability of an observation for different values of c_R and λ when $w_x = 1$. For each c_R value, the corresponding λ value is such that observations given a weight of 0.25 by $\hat{\mathbf{\Gamma}}_{RGLM}$ are also given a weight of 0.25 by $\hat{\mathbf{\Gamma}}_{MDPD}$. Such a representation is not possible for $\hat{\mathbf{\Gamma}}_{OBR}$ because it cannot simply be written as a weighted maximum likelihood estimator.

Fig. 1). Limiting the influence of covariates in the score function to achieve B-robustness, as the RGLM estimator does, is unnecessary in this case and does not yield further robustness towards such outliers. This is further seen in the results of Setting I in the simulations of Section 3.

2.6. Asymptotic properties, Wald-type and score-type tests

2.6.1. Asymptotic distributions

Due to the regularity of the polytomous logistic regression model, as all estimators considered here are M-estimators with differentiable score functions, their asymptotic distributions are normal with sandwich covariance matrices (Huber, 1981).

Result 2. *Let $\hat{\mathbf{\Gamma}}_m$ be an estimator of $\mathbf{\Gamma}$, for m in {ML, MDPD, RGLM, OBR}, and suppose that $\boldsymbol{Y}|\boldsymbol{X}$ follows a polytomous logistic model with parameter $\mathbf{\Gamma}$, then $\hat{\mathbf{\Gamma}}_m$ is conditionally Fisher-consistent and, if the matrix $\boldsymbol{M}_m$ defined below is non-singular,*

$$\sqrt{n}\left(\hat{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \mathbf{\Sigma}_m = \boldsymbol{M}_m^{-1}\boldsymbol{Q}_m\boldsymbol{M}_m^{-T}\right), \tag{14}$$

where $\hat{\boldsymbol{\gamma}}_m$ and $\boldsymbol{\gamma}$ denote the vectors of $\mathbb{R}^{(p+1)(k-1)}$ obtained by stacking the lines of $\hat{\mathbf{\Gamma}}_m$ and $\mathbf{\Gamma}$, respectively, $\boldsymbol{M}_m = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[\partial \boldsymbol{s}_m(\boldsymbol{X},\boldsymbol{Y},\mathbf{\Gamma})/\partial\boldsymbol{\gamma}]$, and $\boldsymbol{Q}_m = \mathrm{Var}_{\boldsymbol{X},\boldsymbol{Y}}[\boldsymbol{s}_m(\boldsymbol{X},\boldsymbol{Y};\mathbf{\Gamma})]$.

The matrices $\hat{\boldsymbol{M}}_m = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{Y}|X=\boldsymbol{x}_i}\left[\partial \boldsymbol{s}_m\left(\boldsymbol{x}_i, \boldsymbol{Y}, \hat{\mathbf{\Gamma}}_m\right)/\partial\boldsymbol{\gamma}\right]$ and $\hat{\boldsymbol{Q}}_m = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\boldsymbol{Y}|X=\boldsymbol{x}_i}\left[\boldsymbol{s}_m\left(\boldsymbol{x}_i, \boldsymbol{Y}, \hat{\mathbf{\Gamma}}_m\right)\right]$ are consistent estimators of $\boldsymbol{M}_m$ and $\boldsymbol{Q}_m$, respectively.

This result can be simplified in the case of $\hat{\mathbf{\Gamma}}_{ML}$. The second Bartlett identity implies that $\boldsymbol{M}_{ML} = -\boldsymbol{Q}_{ML}$, such that $\mathbf{\Sigma}_{ML}$ equals the Fisher information $\boldsymbol{Q}_{ML}^{-1}$. More detailed expressions of the asymptotic variances and their estimates are given in the Appendix.

2.6.2. Wald-type and score-type tests

To test a null hypothesis of the form $\mathcal{H}_0$: "$\boldsymbol{L}\boldsymbol{\gamma} = \boldsymbol{l}$", where $\boldsymbol{L}$ is a full rank matrix with $(p+1)(k-1)$ columns and $r \leq (p+1)(k-1)$ rows, one can use a Wald-type test, using the statistic $T_W = \boldsymbol{L}\hat{\boldsymbol{\gamma}}_m$.

Result 3. *Under the null hypothesis, if $\boldsymbol{L}\mathbf{\Sigma}_m\boldsymbol{L}^T$ is non-singular,*

$$n\left(\boldsymbol{L}\hat{\boldsymbol{\gamma}}_m - \boldsymbol{l}\right)^T\left(\boldsymbol{L}\mathbf{\Sigma}_m\boldsymbol{L}^T\right)^{-1}\left(\boldsymbol{L}\hat{\boldsymbol{\gamma}}_m - \boldsymbol{l}\right) \xrightarrow[n\to\infty]{D} \chi^2_r.$$

Alternatively, a score-type test can be constructed when the null hypothesis is of the form $\mathcal{H}_0$: "$\boldsymbol{\gamma}_{(2)} = \boldsymbol{l}$" where $\boldsymbol{\gamma}_{(2)}$ denotes a subvector of r coordinates of $\boldsymbol{\gamma}$. Such a hypothesis can be formulated as $\mathcal{H}_0$: "$\boldsymbol{L}\boldsymbol{\gamma} = \boldsymbol{l}$", where $\boldsymbol{L}$ is a full rank $r \times (p+1)(k-1)$ matrix of 0 and 1 whose rows sum to 1. Define $\boldsymbol{Z}_n = n^{-1}\sum_{i=1}^{n} \boldsymbol{L}\boldsymbol{s}_m\left(\boldsymbol{x}_i, \boldsymbol{y}_i; \hat{\boldsymbol{\Gamma}}_m^{\mathcal{H}_0}\right)$, where $\hat{\boldsymbol{\Gamma}}_m^{\mathcal{H}_0}$ denotes the estimate of $\boldsymbol{\Gamma}$ by estimator m under the constraint defined by $\mathcal{H}_0$, and the $r \times r$ matrices $\boldsymbol{\Sigma}_{m,\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{\Sigma}_m\boldsymbol{L}^T$ and $\boldsymbol{M}_{m,\boldsymbol{L}} = \left(\boldsymbol{L}\boldsymbol{M}_m^{-1}\boldsymbol{L}^T\right)^{-1}$. The score-type test uses the statistic ${R_n}^2 = \boldsymbol{Z}_n^T\left(\boldsymbol{M}_{m,\boldsymbol{L}}\boldsymbol{\Sigma}_{m,\boldsymbol{L}}\boldsymbol{M}_{m,\boldsymbol{L}}^T\right)^{-1}\boldsymbol{Z}_n$.

Result 4. *Under the null hypothesis, if $\boldsymbol{M}_{m,\boldsymbol{L}}\boldsymbol{\Sigma}_{m,\boldsymbol{L}}\boldsymbol{M}_{m,\boldsymbol{L}}^T$ is non-singular*

$$n{R_n}^2 \xrightarrow[n\to\infty]{D} \chi_r^2.$$

More general null hypotheses can be considered through linear transformations of the parameter $\boldsymbol{\gamma}$ giving more tedious expressions for $\boldsymbol{M}_{m,\boldsymbol{L}}$ and $\boldsymbol{\Sigma}_{m,\boldsymbol{L}}$. Results 3 and 4 follow directly from Proposition 2 of Heritier and Ronchetti (1994).

2.7. Choosing w_x

Multiple proposals exist in the literature regarding the choice of the weighting function w_x. Using a robust Mahalanobis distance $D(\tilde{\boldsymbol{x}}) = (\tilde{\boldsymbol{x}} - \hat{\mu}_x)^T\hat{\boldsymbol{\Sigma}}_x(\tilde{\boldsymbol{x}} - \hat{\mu}_x)$, where $\tilde{\boldsymbol{x}}$ is the vector $\boldsymbol{x}$ without its first coordinate, $\hat{\mu}_x$ is a robust estimator of center, and $\hat{\boldsymbol{\Sigma}}_x$ is a robust estimator of the covariance matrix, Croux et al. (2013) proposed to use

$$w_{x,1}(\boldsymbol{x}) = \frac{df}{df + D\left(\tilde{\boldsymbol{x}}\right)}, \tag{15}$$

where df is a tuning parameter. Alternatively, the implementation of the robust GLM estimator of Cantoni and Ronchetti (2001) in the `robustbase` R package offers the option

$$w_{x,2}(\boldsymbol{x}) = 1/(1 + 8\max\left\{0, (D\left(\tilde{\boldsymbol{x}}\right) - p)(2p)^{-1/2}\right\})^{1/2};$$

see also (Heritier et al., 2009, Chapter 5.3.1). $D\left(\tilde{\boldsymbol{x}}\right)$ can be computed using the minimum covariance determinant (MCD) method of Rousseeuw and Van Driessen (1999). Other weights can be derived from the linear model leverages $h_{ii} = \tilde{\boldsymbol{x}}_i^T\left(\sum_{j=1}^{n}\tilde{\boldsymbol{x}}_j\tilde{\boldsymbol{x}}_j^T\right)^{-1}\tilde{\boldsymbol{x}}_i$, for example $w_{x,3}(\boldsymbol{x}_i) = \sqrt{1-h_{ii}}$ or $w_{x,4}(\boldsymbol{x}_i) = (1-h_{ii})/\sqrt{h_{ii}}$ (Welsch, 1980). Weights $w_{x,1}$ and $w_{x,2}$ depend on $D\left(\tilde{\boldsymbol{x}}\right)$, which may be undefined on discrete covariates with few distinct outcomes if it is computed using MCD. Weights $w_{x,3}$ and $w_{x,4}$, on the other hand, are affected by strong outliers because of the definition of H. For the simulations in Section 3, $w_{x,1}$ is the most flexible option, as the parameter df can be tuned to give a desirable level of downweighting (see Section 2.8 for additional details on the tuning of df).

Given a weighting function w_x, one could naturally wonder about the properties of a weighted maximum likelihood estimator, $\hat{\boldsymbol{\Gamma}}_{wML}$, defined as the M-estimator associated with the estimating function $\boldsymbol{s}_{wML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma}) = w_x(\boldsymbol{x})\boldsymbol{s}_{ML}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\Gamma})$. This estimator would indeed be B-robust (if $w_x(\boldsymbol{x}) \cdot \boldsymbol{x}$ is bounded) but would offer poor protection against misclassification, as seen in Fig. 4, Section 3. However, this estimator can be helpful in the initial tuning of the constants used in w_x (see Section 2.8). Similarly, weighted versions of other estimators could be considered, for instance wMDPD, which is not considered in this work for fidelity reasons with the proposal by the authors.

2.8. Tuning constants

Apart from $\hat{\boldsymbol{\Gamma}}_{wML}$, every robust estimator mentioned in this section involves a tuning constant that acts as a robustness lever: as λ increases, and c_R and c_O decrease, the corresponding estimators gain robustness, but lose efficiency. A tuning method must be used to find the appropriate trade-off between robustness and efficiency. The method used in Section 4 consists of tuning these constants to reach a target asymptotic efficiency, τ, when comparing with $\hat{\boldsymbol{\Gamma}}_{ML}$, at the model.

The Fisher standardized efficiency, computed as the ratio of the Fisher standardized mean square error (MSE), defined by Definition 3 in Section 3.1, is used rather than the usual (non-standardized) efficiency. Indeed, the Fisher standardized MSE remains unchanged if a scaling is applied to a covariate and to the parameter $\boldsymbol{\Gamma}$ accordingly (e.g., switching from centimeters to inches if a covariate measures length). This is not the case for regular mean square error. However, any alternative notion of efficiency can be used.

If the true model parameters were known, the tuning could be carried out by searching for the robustness constant c_m such that τ equals $(k-1)(p+1)/\mathrm{tr}(\boldsymbol{Q}_{ML}\boldsymbol{\Sigma}_m)$, where $\boldsymbol{Q}_{ML}$ is the Fisher information matrix, $\boldsymbol{\Sigma}_m$ is the asymptotic variance of the estimator that depends on c_m, and $(k-1)(p+1)$ corresponds to the asymptotic Fisher standardized MSE of $\hat{\boldsymbol{\Gamma}}_{ML}$. In practice, because the true parameters are unknown, a robust estimate is used as a substitute in $\boldsymbol{Q}_{ML}$ and $\boldsymbol{\Sigma}_m$; see Section 3. For other tuning methods, see Castilla et al. (2018), Aeberhard et al. (2021), or Cantoni and Ronchetti (2001), among others.

The weighting function w_x used by $\hat{\mathbf{\Gamma}}_{RGLM}$ often requires the tuning of an additional constant (e.g. df if $w_{x,1}$ is used). On option to tune this constant to reach a target efficiency τ with respect to $\hat{\Gamma}_{ML}$ at the model is to perform it in two steps. First, choose df such that the resulting weighted maximum likelihood estimator $\hat{\mathbf{\Gamma}}_{wML}$ has an efficiency of τ^{δ}, with $\delta \in [0, 1]$, when compared to $\hat{\Gamma}_{ML}$. Second, tune the constant c_R such that $\hat{\mathbf{\Gamma}}_{RGLM}$ has an efficiency of $\tau^{1-\delta}$ when compared to $\hat{\mathbf{\Gamma}}_{wML}$. This ensures that $\hat{\mathbf{\Gamma}}_{RGLM}$ has a global efficiency of τ when compared to $\hat{\mathbf{\Gamma}}_{ML}$. Notice that using $\delta = 0$ yields $w_x = 1$, which does not protect against outlying predictors, and $\delta = 1$ yields $c_R = \infty$, which does not protect against misclassification. In Section 3, $\delta = 1/2$ is used to achieve the same loss of efficiency at each of the two tuning steps.

3. Simulation study

3.1. Estimation

In this section, the previously described estimators are compared through their estimates on clean and contaminated simulated datasets. The following setting is used to simulate the data: the response variable $\boldsymbol{Y}$ has $k = 3$ possible categories, and there are $p = 2$ covariates x_1, x_2, plus an intercept $x_0 = 1$. The parameter $\mathbf{\Gamma} \in \mathbb{R}^{2\times 3}$ is the same as that used to generate Figs. 1 and 2:

$$\mathbf{\Gamma} = \begin{pmatrix} 0 & 1.5 & \sqrt{3}/2 \\ 0 & 0 & \sqrt{3} \end{pmatrix}.$$

The explanatory variables x_1, x_2 are drawn as independent standard normal random variables and the responses $\boldsymbol{y}$ are simulated from a polytomous logistic regression model with parameter $\mathbf{\Gamma}$ according to Equations (1) and (2). This produces the clean datasets. The size of the datasets is set to $n = 500$. The clean datasets are modified with two contamination settings:

Setting I (contaminated responses) A percentage q of responses are replaced by categories drawn from a multinomial distribution with shuffled conditional probabilities $(\pi_3, \pi_1, \pi_2)^T$ instead of $(\pi_1, \pi_2, \pi_3)^T$. This setting can alternatively be interpreted as contaminating a clean dataset by drawing $\lfloor q \times n \rfloor$ of its responses from a polytomous logistic model with parameters $\begin{pmatrix} 0 & 0 & -\sqrt{3} \\ 0 & 1.5 & -\sqrt{3}/2 \end{pmatrix}$ instead of $\mathbf{\Gamma}$.

Setting II (contaminated covariates and responses) Setting I is extended by affecting $\boldsymbol{x}$ values. For the responses changed according to setting I, both covariate values of the corresponding datapoints are multiplied by 5, such that the contaminated datapoints correspond to the polytomous logistic model with parameters $\begin{pmatrix} 0 & 0 & -\sqrt{3} \\ 0 & 0.3 & -\sqrt{3}/2 \end{pmatrix}$ with explanatory variables drawn from a $\mathcal{N}(\mu = 0, \sigma = 5)$.

Contamination setting I only creates outlying responses and contamination setting II creates points with both outlying covariates and responses. For a given contamination setting s, ($s =$ I, II, see settings above), a given a contamination level q in $\{0\%, 1\%, \cdots, 20\%\}$, and a given estimator m in {ML, wML, MDPD, RGLM, OBR}, $R = 1'000$ replications are considered. The accuracy of the estimates $\left(\hat{\mathbf{\Gamma}}_m^{s,q,r}\right)_{r=1,\ldots,R}$ is summarized by their empirical Fisher-standardized mean squared error.

Definition 3. The Fisher-standardized mean squared error (FMSE) of an estimator $\hat{\mathbf{\Gamma}}$ is defined by:

$$FMSE(\hat{\mathbf{\Gamma}}, \mathbf{\Gamma}) = \mathbb{E}\left[\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right)^T \mathbf{Q}_{ML} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right)\right],$$

where $\boldsymbol{\gamma}$ (resp. $\hat{\boldsymbol{\gamma}}$) is the vector of $\mathbb{R}^{(p+1)(k-1)}$ obtained by stacking the lines of $\mathbf{\Gamma}$ (resp $\hat{\mathbf{\Gamma}}$) and $\mathbf{Q}_{ML}$ is the Fisher information matrix defined in Section 2.6.2.

The empirical FMSE of an estimator m through R simulations with a proportion q of observations issued from contamination setting s is

$$\text{eFMSE}(s, q, m) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\boldsymbol{\gamma}}_m^{s,q,r} - \boldsymbol{\gamma}\right)^T \mathbf{Q}_{ML}^r \left(\hat{\boldsymbol{\gamma}}_r^{s,q,m} - \boldsymbol{\gamma}\right), \tag{16}$$

where $\hat{\boldsymbol{\gamma}}_m^{s,q,r}$ (resp. $\boldsymbol{\gamma}$) is a vector of $\mathbb{R}^{(k-1)(p+1)}$ obtained by stacking the lines of $\hat{\mathbf{\Gamma}}_m^{s,q,r}$ (resp. $\mathbf{\Gamma}$) together and $\mathbf{Q}_{ML}^r$ is the Fisher information matrix computed on the r-th clean dataset by $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{V}_i(\boldsymbol{x}_i, \mathbf{\Gamma}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i^T$. The empirical efficiency (eEFF) of the estimators is computed as

$$\text{eEFF}(s, q, m) = \text{eFMSE}(s, 0, \text{ML}) / \text{eFMSE}(s, q, m). \tag{17}$$

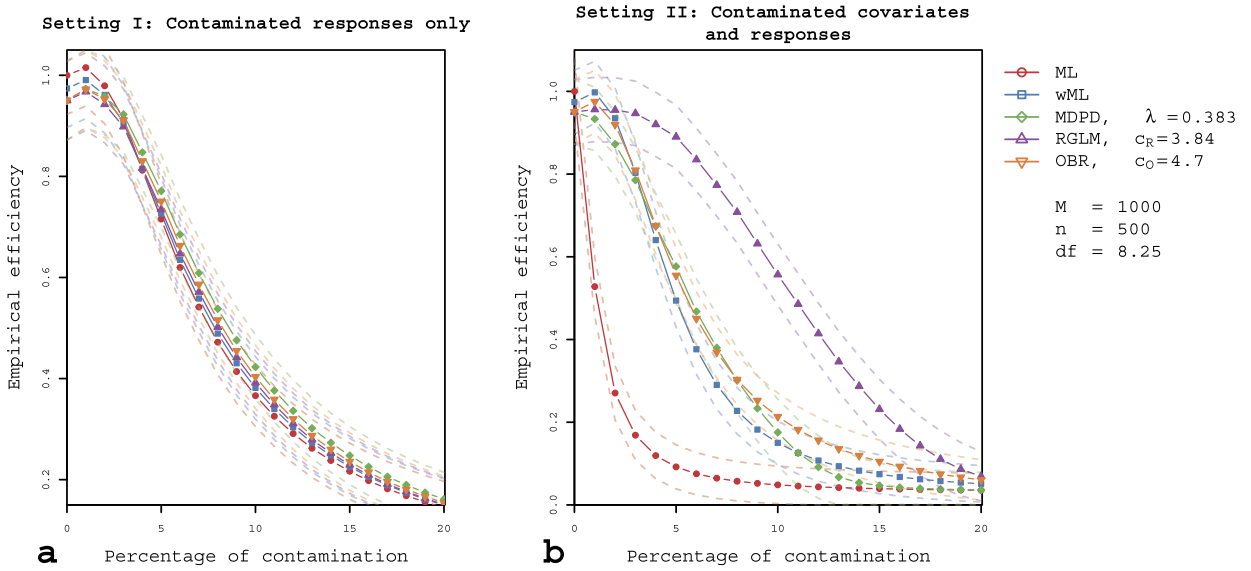

Fig. 4. Empirical efficiency of estimators on simulated contaminated datasets as a function of the percentage of contamination. Note that the y-axis scale is not the same for the two graphs.

The tuning of the robustness constants associated with MDPD, RGLM and OBR estimators is made such that these estimators all have a 0.95 empirical efficiency on the clean datasets, compared to $\hat{\mathbf{\Gamma}}_{ML}$. The values of the resulting constants are given in the legend of Fig. 4. The weighting function w_x used in $\hat{\mathbf{\Gamma}}_{RGLM}$ is defined by Equation (15), where the tuning parameter df is tuned such that $\text{eEFF}(s, 0, \text{wML}) = \sqrt{0.95}$ (see Section 2.8 for more details). The estimator $\hat{\mathbf{\Gamma}}_{wML}$ was added for comparison purposes. The same value of the tuning parameter df of the downweighting function w_x was used for $\hat{\mathbf{\Gamma}}_{RGLM}$ and $\hat{\mathbf{\Gamma}}_{wML}$, which explains why the efficiency of $\hat{\mathbf{\Gamma}}_{wML}$ is different at $q = 0\%$.

Fig. 4 shows the empirical efficiencies, defined by Equation (17) of $\hat{\mathbf{\Gamma}}_{ML}$, $\hat{\mathbf{\Gamma}}_{wML}$, $\hat{\mathbf{\Gamma}}_{MDPD}$, $\hat{\mathbf{\Gamma}}_{RGLM}$ and $\hat{\mathbf{\Gamma}}_{OBR}$ as a function of the percentage q of contamination for settings I and II. Note that $q = 0$ corresponds to the clean datasets in both panels. In setting I (panel a), which only includes contaminated responses, the difference between the efficiencies of any two estimators is less than 0.05. Under large contamination, $\hat{\mathbf{\Gamma}}_{ML}$ has the lowest empirical efficiency, and $\hat{\mathbf{\Gamma}}_{MDPD}$ has the highest empirical efficiency because it does not unnecessarily downweight points as $\hat{\mathbf{\Gamma}}_{RGLM}$ does to protect against outlying covariates. However, as a 5% efficiency loss is usually comparable to the loss of efficiency traded for robustness at the model (no contamination), the gain is not sufficiently significant to motivate the use of robust estimators in setting I. However, the advantage of robust estimation is clear in the second setting (panel b) in which both the covariates and responses are contaminated. The empirical efficiency of $\hat{\mathbf{\Gamma}}_{ML}$ drops from 1 with no contamination, to 0.6 with 1% contamination, and to below 0.2 with more than 5% contamination. The non B-robust $\hat{\mathbf{\Gamma}}_{MDPD}$, while performing better than $\hat{\mathbf{\Gamma}}_{ML}$, performs significantly worse than the B-robust $\hat{\mathbf{\Gamma}}_{RGLM}$. The latter is clearly the best as its efficiency is barely affected by contamination of less than 5%. Note that this good performance is not only due to the protection against outlying covariates provided by the weighting function w_x because the weighted maximum likelihood estimator $\hat{\mathbf{\Gamma}}_{wML}$ performs significantly worse. Protection against both outlying covariates and mislabelling is necessary.

Boxplots of estimates of $\mathbf{\Gamma}_{13} = \sqrt{3}/2$ by the five considered estimators under setting II are shown in Fig. 5 for six contamination levels: 0%, 2%, 5%, 10%, 15%, and 20%. This contamination setting introduces a negative bias on all estimates. While the maximum likelihood estimators have a median bias of almost -0.5 with as little as 2% contamination, all robust estimators still cover the true parameter value in their interquartile range. For higher contamination, estimators $\hat{\mathbf{\Gamma}}_{wML}$, $\hat{\mathbf{\Gamma}}_{MDPD}$, and $\hat{\mathbf{\Gamma}}_{OBR}$ are significantly more biased than $\hat{\mathbf{\Gamma}}_{RGLM}$. Estimator $\hat{\mathbf{\Gamma}}_{MDPD}$ is slightly less biased than $\hat{\mathbf{\Gamma}}_{OBR}$ and $\hat{\mathbf{\Gamma}}_{wML}$ at first but its variance increases with contamination up to 10%, while the variances of $\hat{\mathbf{\Gamma}}_{OBR}$ and $\hat{\mathbf{\Gamma}}_{wML}$ remain stable, leading all three estimators to have similar empirical efficiencies. The boxplots for the other parameters estimates under setting II are given in Supplementary Section E.2 and show similar behavior. Similar boxplots for all the parameter estimates under setting I are also given in Supplementary Section E.1.

The efficiencies of all estimators seem to be higher with 1% or 2% contamination than with no contamination in setting I; this is because this contamination setting tends to shrink estimates towards 0. The FMSE of an estimator, as the usual MSE, can be decomposed as the sum of a bias component and a variance component. The shrinkage effect induced by contamination usually increases the bias component faster than it decreases the variance component. However, in this particular setting with a low proportion of contamination, the decrease of the variance components exceeds the increase in bias, leading to an increase of its standardized efficiency. This effect should not be expected to happen in general.

The behavior of $\hat{\mathbf{\Gamma}}_{OBR}$ could be surprising at first, but its optimality indicates it is the estimator with the smallest asymptotic variance under no contamination among estimators with a same bound on their influence functions. Hence, it

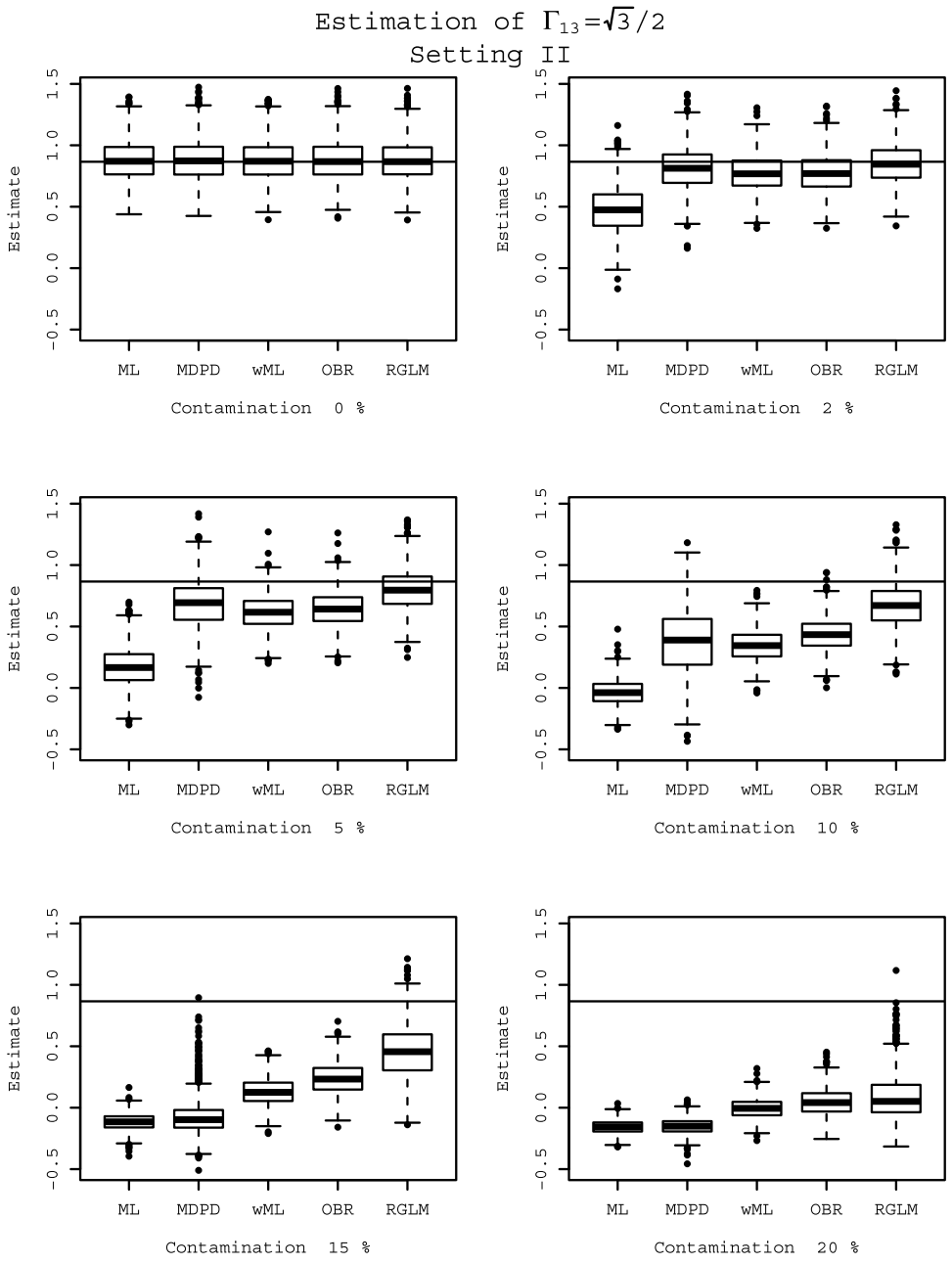

Fig. 5. Boxplots of the estimates of $\mathbf{\Gamma}_{13}$ by all five estimators for six different contamination levels in setting II. Solid lines indicate the true parameter value.

is expected that $\hat{\mathbf{\Gamma}}_{OBR}$ should be the least sensitive to a very small proportion of the worst case contamination. This is coherent with its behavior under contamination setting II with 1% contamination, where it has a higher efficiency than $\hat{\mathbf{\Gamma}}_{RGLM}$ and $\hat{\mathbf{\Gamma}}_{MDPD}$. It then performs poorly under contamination levels higher than 3%.

These simulations suggest that, by default, the B-robust estimator $\hat{\mathbf{\Gamma}}_{RGLM}$ should be favored when there is a possibility of outlying covariates, as it provides a significant increase of efficiency when the covariates are contaminated; when they are not, this estimator is only slightly less efficient than others.

3.2. Tests

In this section, the behavior of the Wald-type and score-type tests of the different estimators is examined through a simulation study similar to Section 3.1. An irrelevant explanatory variable, $\boldsymbol{x}_3$, with standard normal distribution, independent from the other covariates and from the response, is added to the previous simulation setting. The significance of this

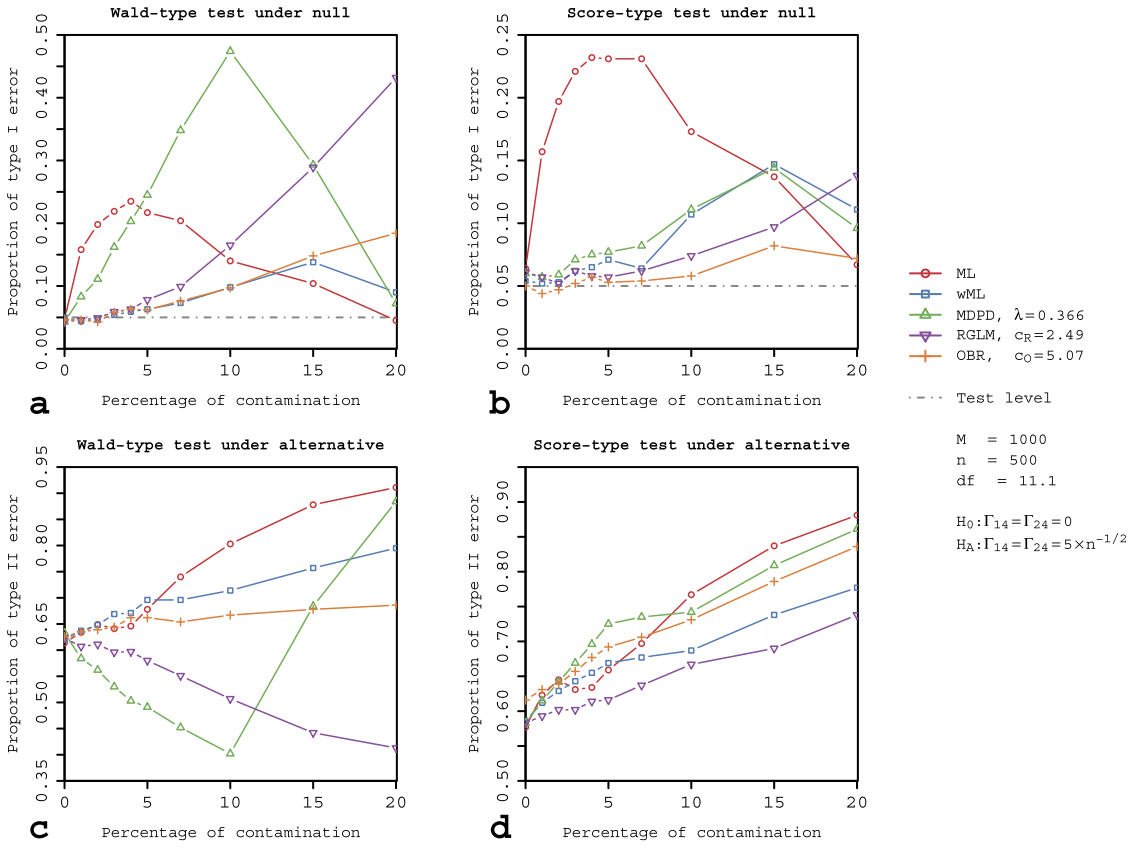

Fig. 6. Type I (first row) and type II (second row) errors of Wald-type (first column) and score-type (second column) tests under contamination setting II.

covariate is tested under different levels of contamination. Formally, there are $k = 3$ categories, $p = 3$ independent standard normal covariates, plus an intercept, and the model parameters matrix is

$$\mathbf{\Gamma} = \begin{pmatrix} 0 & 1.5 & \sqrt{3}/2 & 0 \\ 0 & 0 & \sqrt{3} & 0 \end{pmatrix}.$$

The null hypothesis $\mathcal{H}_0$: "$\Gamma_{14} = \Gamma_{24} = 0$" is tested against the alternative hypothesis $\mathcal{H}_A$: "$\Gamma_{14} = \Gamma_{24} = 5/\sqrt{n} = 0.224$". Setting II is used to contaminate the clean datasets with size $n = 500$. As before, $R = 1'000$ replications are made, for the following contamination levels: 0%, 1%, $\cdots$, 5%, 7%, 10%, 15%, and 20%. The tuning of the robustness constants is the same as in Section 3.1.

Fig. 6 shows the proportions of type I and type II errors for the Wald-type and the score-type tests at the 5% level. As expected, the Wald-type and score-type tests based on the maximum likelihood estimator are not robust. Their observed levels exceed 10% with 1% contamination and reach 25% with 5% contamination, whereas the nominal level is 5%, making these tests unreliable under contamination. For all other estimators, the score-type tests seem more robust than the Wald-type tests. The Wald-type test based on $\hat{\mathbf{\Gamma}}_{MDPD}$ is highly sensitive to contamination: although its power increases from 40% to 60% when reaching 10% contamination, its level climbs above 10% with 2% contamination and to 25% with 5% contamination, making this test unreliable. Level stability is of the utmost importance because an increase from the desired level is likely to lead to false discoveries: a gain in power does not compensate for an increased level. The Wald-type test associated with $\hat{\mathbf{\Gamma}}_{RGLM}$ remains reliable for contamination levels below 5% (its type I error remains under 7%), but it is strongly impacted by higher contamination levels, with a type I error exceeding 10% at a contamination level of 10% and exceeding 40% at a contamination level of 20%. The Wald-type tests associated with $\hat{\mathbf{\Gamma}}_{OBR}$ and $\hat{\mathbf{\Gamma}}_{wML}$ are more robust. Their levels stay quite close to the nominal 5% level when contamination is added, while their power only drops by 15% for $\hat{\mathbf{\Gamma}}_{wML}$ and 5% for $\hat{\mathbf{\Gamma}}_{OBR}$. The Score-type tests of the four robust estimators are more homogeneous but panels b and d of Fig. 6 tend to show superiority of the tests associated with $\hat{\mathbf{\Gamma}}_{RGLM}$ and $\hat{\mathbf{\Gamma}}_{OBR}$, whose type I error stay the closest to the nominal level, at least for contamination levels less than or equal to 15%.

The behavior of the Wald-type test associated with $\hat{\mathbf{\Gamma}}_{MDPD}$ may be surprising at first: its type I error suddenly drops when contamination levels exceed 10% while its type II error increases. This may be due to the influence of outliers on the estimation of the variance needed to compute the Wald-type test statistics: high contamination levels tend to shrink the variance estimate, yielding a smaller test statistic and thus lower rejection rate under both null or alternative hypotheses,

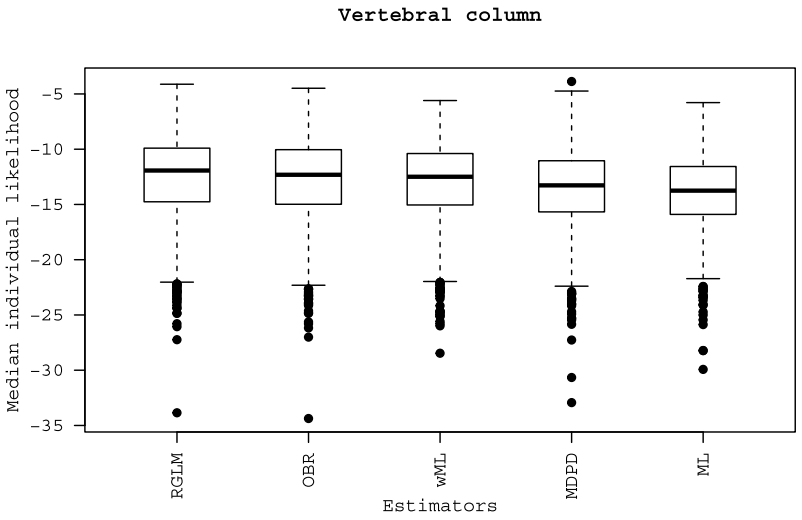

Fig. 7. Boxplots of 1'000 median individual likelihoods (MIL) in the *Vertebral column* dataset. The estimators are in descending order based on the median of the MIL. The tuning constants are $df = 6.040$, $\lambda = 0.487$, $c_R = 2.853$, $c_O = 4.736$.

giving lower type I error and larger type II error. The same effect is noted with the Wald-type test associated with $\hat{\mathbf{\Gamma}}_{ML}$. It is worth noticing that, despite its low efficiency for estimation, $\hat{\mathbf{\Gamma}}_{OBR}$ can be used for reliable and robust tests.

Also note that the contamination setting II preserves $\mathcal{H}_0$: the datasets generated under $\mathcal{H}_0$ and contaminated by this setting still fulfills $\mathcal{H}_0$. Indeed, the responses of the contaminated observations remain independent of covariate $\boldsymbol{x}_3$. Therefore, the large type I error rates shown in the first row of Fig. 6 are only due to non-robustness of the estimators and cannot be caused by deviations from $\mathcal{H}_0$.

4. Real data application: *Vertebral column* dataset

All estimators described in Section 2 are applied to the *Vertebral column* dataset (Berthonnaud et al., 2005), found on the *UCI Machine Learning Repository* (Dua and Graff, 2019). This dataset contains indices describing the shape and orientation of the different parts of the spines of 100 healthy subjects, 60 subjects suffering from disk hernia and 150 subjects suffering from spondylolisthesis. The resulting 310 observations contain six continuous covariates and a categorical response with three categories, corresponding to three diagnoses: healthy, disk hernia, and spondylolisthesis. The number of covariates has been reduced to three due to the high collinearity. The remaining variables are pelvic tilt, sacral slope, and pelvic radius. A polytomous regression model thus requires a matrix of $(3-1)\times(3+1)=8$ parameters. A basic data exploration reveals at least one observation that is clearly misclassified and multiple observations with *large* covariate values.

The estimators are compared through a 10-fold cross validation using a likelihood-based criterion. The method is as follows: removal of $n_{out} = \lfloor 0.1 \times n \rfloor$ observations, estimation of the parameters on the $(n - n_{out})$ remaining observations, and calculation of the individual log-likelihoods on the n_{out} removed points using the estimated parameters. The median of these n_{out} likelihoods is then extracted to provide a robust goodness-of-fit measure, not influenced by a few potential outlying observations. This process is repeated 1000 times. The results are given in Fig. 7. All estimators are tuned to the same asymptotic efficiency, 90%, according to the method described in Section 2.8. The asymptotic efficiencies needed for the tuning are computed by plugging in the value of $\hat{\mathbf{\Gamma}}_{OBR}$, estimated using $c_O = \sqrt{(k-1)\times(p+1)} = \sqrt{8}$, which is the most robust optimal B-robust estimator. Changes in the chosen estimator did not notably affect the tuning of the constants.

Fig. 7 shows a clear gain in using $\hat{\mathbf{\Gamma}}_{RGLM}$ over all other estimators. This dataset is a perfect example of a case for which the non-B-robust estimators $\hat{\mathbf{\Gamma}}_{MDPD}$ and $\hat{\mathbf{\Gamma}}_{ML}$ do not protect against outliers in the covariates, whereas $\hat{\mathbf{\Gamma}}_{wML}$ does not protect against misclassification. Similar to what was observed in Fig. 4 in Section 3, $\hat{\mathbf{\Gamma}}_{OBR}$ performs more poorly than $\hat{\mathbf{\Gamma}}_{RGLM}$ when outliers are present in the covariates.

5. Conclusion

Outlying covariates have been shown to have a much bigger impact on estimation and testing than misclassification alone. Two new estimators, the B-robust weighted GLM estimator and the optimal self-standardized B-robust estimator, guarantee robustness towards both misclassification and outlying covariates. Comparisons with existing alternatives on real and simulated datasets lead us to advocate for the use of $\hat{\mathbf{\Gamma}}_{RGLM}$. Using this estimator preventively (e.g., on a clean dataset) only leads to a small loss of efficiency, whereas using it on datasets with contamination prevents serious issues (biased estimates or misleading test conclusions).

An inferential point of view was adopted rather than a classification one. However, robust classification is a well-studied topic (see Bertsimas et al. (2019) and references therein) and polytomous regression is a popular classifier among the machine learning community. Robust polytomous regression classifiers can be found in Bootkrajang and Kabán (2012) and Yin et al. (2018) and two corresponding robust estimators could be derived as byproducts of these, although their statistical properties, such as asymptotic distribution or associated tests, have not been derived. Exploring in detail the close connections between robust estimation and robust classification with polytomous regression models would surely be enriching for both purposes.

An extension of the RGLM and OBR estimators to complex survey schemes would also be worth pursuing, as developed in Castilla et al. (2020) for the MDPD estimator. Tuning of the robustness constants would require further investigation. In particular, the comparison of an approach based on asymptotic variance minimization (Castilla et al., 2018), one based on a median downweighting criterion (Aeberhard et al., 2021), and the one used in this article, would be beneficial.

Construction of the weighting function w_x would also be worth investigating. Mahalanobis distance-based options are irrelevant when the distribution of the covariates is far from normal, and especially with discrete variables or a mix of continuous and discrete variables. A starting point could be a non-parametric method (see e.g., Dang and Serfling (2010)).

Nevertheless, the main challenge of the methods considered here are the numerical aspects. The dimensions of the parameter of the polytomous regression model easily become large and these methods can only be applied to datasets of reasonable dimensions and for which responses overlap. Having a dataset where observations with different categories are linearly separable (or close) in the predictor space yields numerical instability. This separability issue gets more and more problematic when the number of covariates or the number of categories is large. Introducing an adequate penalization to shrink the parameters could force convergence of the methods, and allow their use on bigger datasets, even when categories are linearly separated in the space of covariates. In particular, extending the procedure of Kosmidis and Lunardon (2020) to robust polytomous regression sounds quite promising as it could guarantee convergence while reducing the bias.

Acknowledgements

All three authors are employed by the University of Geneva. This project was pursued without any further funding. We thank the reviewers and the associate editor for their careful review and comments that have led to a much-improved version of the manuscript.

Appendix A. Asymptotic variances of estimators

We give here the expressions of $\boldsymbol{M}_m$ and $\boldsymbol{Q}_m$ in Equation (14). The covariates are assumed to be random variables. In case of a fix design, $\mathbb{E}_X[f(X)]$ should be replaced by $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ for any measurable function f. In practice, matrices $\boldsymbol{M}_m$ and $\boldsymbol{Q}_m$ can be estimated by replacing $\boldsymbol{\Gamma}$ by $\hat{\boldsymbol{\Gamma}}_m$ and by replacing the expectation on X ($\mathbb{E}_X[\ldots]$) by the average over all $\boldsymbol{x}_i$ values ($\frac{1}{n}\sum_{i=1}^{n}[\ldots]$).

- ML estimator:

$$\boldsymbol{\Sigma}_{ML} = \boldsymbol{Q}_{ML} = -\boldsymbol{M}_{ML}^{-1} = \mathbb{E}_X\left[\boldsymbol{V}^* \otimes XX^T\right]^{-1}$$

- MDPD estimator: $\boldsymbol{\Sigma}_{MDPD} = \boldsymbol{M}_{MDPD}^{-1}\boldsymbol{Q}_{MDPD}\boldsymbol{M}_{MDPD}^{-T}$ with:

$$\boldsymbol{M}_{MDPD} = \mathbb{E}_X\left[\left(\boldsymbol{V}^*\mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda-1}\boldsymbol{V}^* + \left(\boldsymbol{\pi}^T\boldsymbol{e}_k\right)^{\lambda+1}\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}\right)\otimes XX^T\right]$$

and

$$\boldsymbol{Q}_{MDPD} = \mathbb{E}_X\left[\left(\boldsymbol{V}^*\mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda-1}\boldsymbol{V}^*\mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda-1}\boldsymbol{V}^* + \boldsymbol{A} + \boldsymbol{B}\right)\otimes XX^T\right]$$

where

$$\boldsymbol{A} = \left(\boldsymbol{\pi}^T\boldsymbol{e}_k\right)^{2\lambda+1}\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T} - \left(\boldsymbol{\pi}^T\boldsymbol{e}_k\right)^{2\lambda+2}\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}$$

and

$$\boldsymbol{B} = \left(\boldsymbol{\pi}^T\boldsymbol{e}_k\right)^{\lambda+1}\left(\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}\mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda} + \mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda}\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T} - 2\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}\mathrm{diag}\left(\boldsymbol{\pi}^*\right)^{\lambda-1}\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}\right)$$

These expressions are different from the ones in Castilla et al. (2018), which we found to be incorrect.

- RGLM estimator: $\boldsymbol{\Sigma}_{RGLM} = \boldsymbol{M}_{RGLM}^{-1}\boldsymbol{Q}_{RGLM}\boldsymbol{M}_{RGLM}^{-T}$ with:

$$\begin{aligned}\boldsymbol{M}_{RGLM} &= \mathbb{E}_X\left[w_x(X)\left[\boldsymbol{V}^*\boldsymbol{W}^* + (\zeta\boldsymbol{I}^* - \boldsymbol{W}^*)\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}\right]\otimes XX^T\right]\\ \boldsymbol{Q}_{RGLM} &= \mathbb{E}_X\left[w_x^2(X)\left[\boldsymbol{V}^*\boldsymbol{W}^* + (\zeta_2\boldsymbol{I}^* - \boldsymbol{W}^*)\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T} - \boldsymbol{C}\right]\otimes XX^T\right]\end{aligned}$$

where

$\boldsymbol{I}^*$ is the identity matrix of $\mathbb{R}^{(k-1)\times(k-1)}$,
$\zeta = \mathbb{E}_{Y|X}[w_c(Y^T\boldsymbol{\pi})] = \sum_{j=1}^{k} \pi_j w_c(\pi_j)$,
$\zeta_2 = \mathbb{E}_{Y|X}[w_c(Y^T\boldsymbol{\pi})^2] = \sum_{j=1}^{k} \pi_j w_c(\pi_j)^2$,
$\boldsymbol{W}^*$ is a diagonal matrix of $\mathbb{R}^{(k-1)\times(k-1)}$ whose j-th diagonal element is $w_c(\pi_j)$

and

$$\boldsymbol{C} = (\zeta \boldsymbol{I}^* - \boldsymbol{W}^*)\boldsymbol{\pi}^*\boldsymbol{\pi}^{*T}(\zeta \boldsymbol{I}^* - \boldsymbol{W}^*)\,.$$

- OBR estimator: as $\boldsymbol{Q}_{OBR} = \boldsymbol{I}$ by property of the estimator, $\boldsymbol{\Sigma}_{OBR} = \boldsymbol{M}_{OBR}^{-1}\boldsymbol{M}_{OBR}^{-T}$ with

$$\boldsymbol{M}_{OBR} = \mathbb{E}_X\left[\sum_{j=1}^{k} \boldsymbol{\pi}_j \psi_{c_O}\left\{A(\boldsymbol{\Gamma})(\boldsymbol{e}_j^* - \boldsymbol{\pi}^*) \otimes X - \boldsymbol{\alpha}^{OBR}(X, \boldsymbol{\Gamma})\right\}(\boldsymbol{e}_j^* - \boldsymbol{\pi}^*) \otimes X\right]$$

The computation of the $\hat{\boldsymbol{\Gamma}}_{OBR}$ estimator using the IF algorithm gives, as a byproduct, an estimate $\hat{\boldsymbol{A}}(\hat{\boldsymbol{\Gamma}}_{OBR})$ of the matrix $\boldsymbol{A}(\boldsymbol{\Gamma})$ above which can be plugged in the above formula to obtain an estimate of the asymptotic variance $\boldsymbol{\Sigma}_{OBR}$.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107564.

References

Aeberhard, W.H., Cantoni, E., Marra, G., Radice, R., 2021. Robust fitting for generalized additive models for location, scale and shape. Stat. Comput. 31 (11), 1–16.
Agresti, A., 2012. Categorical Data Analysis. Wiley.
Alqallaf, F., Agostinelli, C., 2016. Robust inference in generalized linear models. Commun. Stat., Simul. Comput. 45 (9), 3053–3073.
Berthonnaud, E., Dimnet, J., Roussouly, P., Labelle, H., 2005. Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. J. Spinal Disord. Tech. 1, 40–47.
Bertsimas, D., Dunn, J., Pawlowski, C., Zhuo, Y.D., 2019. Robust classification. INFORMS J. Optim. 1 (1), 2–34.
Blizzard, L., Hosmer, D.W., 2007. The log multinomial regression model for nominal outcomes with more than two attributes. Biom. J. 49 (6), 889–902.
Bootkrajang, J., Kabán, A., 2012. Label-noise robust logistic regression and its applications. In: Flach, P.A., De Bie, T., Cristianini, N. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 143–158.
Cantoni, E., Ronchetti, E., 2001. Robust inference for generalized linear models. J. Am. Stat. Assoc. 96 (455), 1022–1030.
Cantoni, E., Ronchetti, E., 2006. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. J. Health Econ. 25 (2), 198–213.
Castilla, E., Ghosh, A., Martin, N., Pardo, L., 2018. New robust statistical procedures for the polytomous logistic regression models. Biometrics 74 (4), 1282–1291.
Castilla, E., Ghosh, A., Martin, N., Pardo, L., 2020. Robust semiparametric inference for polytomous logistic regression with complex survey design. Adv. Data Anal. Classif., 1–34.
Copas, J.B., 1988. Binary regression models for contaminated data. J. R. Stat. Soc. B 50 (2), 225–253.
Croux, C., Haesbroeck, G., Ruwet, C., 2013. Robust estimation for ordinal regression. J. Stat. Plan. Inference 143 (9), 1486–1499.
Dang, X., Serfling, R., 2010. Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. J. Stat. Plan. Inference 140 (1), 198–213.
Daniels, M.J., Gatsonis, C., 1997. Hierarchical polytomous regression models with applications to health services research. Stat. Med. 16 (20), 2311–2325.
Dua, D., Graff, C., 2019. UCI machine learning repository. http://archive.ics.uci.edu/ml.
Feng, J., Xu, H., Mannor, S., Yan, S., 2014. Robust logistic regression and classification. In: Advances in Neural Information Processing Systems 27'. Curran Associates, Inc., pp. 253–261.
Ghosh, A., Basu, A., 2016. Robust estimation in generalized linear models: the density power divergence approach. Test 25 (2), 269–290.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons.
Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.-P., 2009. Robust Methods in Biostatistics, vol. 825. John Wiley & Sons.
Heritier, S., Ronchetti, E., 1994. Robust bounded-influence tests in general parametric models. J. Am. Stat. Assoc. 89 (427), 897–904.
Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35, 73–101.
Huber, P.J., 1981. Robust Statistics. Wiley.
Hung, H., Jou, Z.-Y., Huang, S.-Y., 2018. Robust mislabel logistic regression without modeling mislabel probabilities. Biometrics 74 (1), 145–154.
Kempen, B., Brus, D.J., Heuvelink, G.B., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma 151 (3), 311–326.
Künsch, H.R., Stefanski, L.A., Carroll, R.J., 1989. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. J. Am. Stat. Assoc. 84 (406), 460–466.
Kosmidis, I., Lunardon, N., 2020. Empirical bias-reducing adjustments to estimating functions. ArXiv preprint. arXiv:2001.03786.
Lange, K., 2010. Numerical Analysis for Statisticians, 2nd edition. Springer.
Lesaffre, E., Albert, A., 1989. Multiple-group logistic regression diagnostics. J. R. Stat. Soc., Ser. C, Appl. Stat. 38 (3), 425–440.
Markatou, M., Basu, A., Lindsay, B.G., 1998. Weighted likelihood equations with bootstrap root search. J. Am. Stat. Assoc. 93 (442), 740–750.
Martín, N., 2015. Using Cook's distance in polytomous logistic regression. Br. J. Math. Stat. Psychol. 68 (1), 84–115.
McCullagh, P., Nelder, J., 1983. Generalized Linear Models. Chapman & Hall/CRC.
Mebane Jr., W.R., Sekhon, J.S., 2004. Robust estimation and outlier detection for overdispersed multinomial models of count data. Am. J. Polit. Sci. 48 (2), 392–411.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. R. Stat. Soc. A 135 (3), 370–384.
Pregibon, D., 1981. Logistic regression diagnostics. Ann. Stat. 9 (4), 705–724.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41 (3), 212–223.
Stefanski, L.A., Carroll, R.J., Ruppert, D., 1986. Optimally bounded score functions for generalized linear models with application to logistic regression. Biometrika 73, 413–424.
Wang, X., 2014. Modified generalized method of moments for a robust estimation of polytomous logistic model. PeerJ 2. https://doi.org/10.7717/peerj.467.
Welsch, R.E., 1980. Regression sensitivity analysis and bounded-influence estimation. In: Kmenta, J., Ramsey, J.B. (Eds.), Evaluation of Econometric Models. Academic Press, pp. 153–167.
Yin, M., Zeng, D., Gao, J., Wu, Z., Xie, S., 2018. Robust multinomial logistic regression based on RPCA. IEEE J. Sel. Top. Signal Process. 12 (6), 1144–1154.