Robust inference for ordinal response models[*]

Maria Iannario

Department of Political Sciences, University of Naples Federico II, Italy
e-mail: maria.iannario@unina.it

Anna Clara Monti

Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Italy
e-mail: acmonti@unisannio.it

Domenico Piccolo

Department of Political Sciences, University of Naples Federico II, Italy
e-mail: domenico.piccolo@unina.it

and

Elvezio Ronchetti

Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland
e-mail: Elvezio.Ronchetti@unige.ch

Abstract: The present paper deals with the robustness of estimators and tests for ordinal response models. In this context, gross-errors in the response variable, specific deviations due to some respondents' behavior, and outlying covariates can strongly affect the reliability of the maximum likelihood estimators and that of the related test procedures.

The paper highlights that the choice of the link function can affect the robustness of inferential methods, and presents a comparison among the most frequently used links. Subsequently robust M-estimators are proposed as an alternative to maximum likelihood estimators. Their asymptotic properties are derived analytically, while their performance in finite samples is investigated through extensive numerical experiments either at the model or when data contaminations occur. Wald and t-tests for comparing nested models, derived from M-estimators, are also proposed. M based inference is shown to outperform maximum likelihood inference, producing more reliable results when robustness is a concern.

1. Introduction

In recent years, the interest in ordinal data has been constantly growing since these data convey relevant information on several scientific and applied research areas, such as medicine, psychology, sociology, political sciences, economics, marketing, and so on. The motivation for this interest stems from the need to collect data on several aspects of current life of people and communities. Generally categorical data arise when items concerning opinions, preferences, judgements, evaluations, worries, etc., are expressed as ordered categories.

A large body of statistical literature has been devoted to models for ordinal data, as summarized in standard books such as Agresti (2010) and Tutz (2012) among others. The present paper focusses on models where the response variable is ordinal and depends on explanatory variables given by the subjects' covariates.

The most popular approach to ordered response models initially advocated by McCullagh (1980) is based on the assumption that a latent variable drives the response and the model is embedded within the Generalized Linear Model framework as formalized by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). A different perspective, more related to the psychological process of selection, leads to the CUB models (Piccolo, 2003; Iannario and Piccolo, 2016), so called because they parameterize the probability of a given response as a mixture of a shifted Binomial and a discrete Uniform random variable. This approach does not require the specification of a model for the latent variable and describes directly the effect of the covariates on the feeling and the uncertainty underlying the respondents' choices.

In spite of the huge body of literature on robustness both for continuous and discrete data in the past decades (see e.g. the books by Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel *et al.* (1986), Maronna *et al.* (2006)), the area of ordinal data has been somewhat neglected. Nevertheless, it has been recognized that respondents may deliberately or unconsciously choose a wrong category. This phenomenon, in addition to the occurrence of gross-errors (whose probability is never negligible) or to erratic behavior by a few respondents, produces a contamination of the assumed model distribution, which can alter the properties of estimators and tests.

Few papers in the robustness literature are related to this problem. An early reference is Hampel (1968), where in addition to the now classical infinitesimal approach to robustness, robust issues in the pure binomial case are discussed. This was followed up more recently by Ruckstuhl and Welsh (2001). In addition, Victoria-Feser and Ronchetti (1997) develop robust estimators for grouped data, which typically appear e.g. in income studies, by investigating the robustness properties of estimators with respect to the underlying continuous model. More recently, Iannario *et al.* (2016) deal with robustness for the class of CUB models,

and Croux *et al.* (2013) propose a robust estimator for ordinal response models with a logistic link function. In the literature about latent variable models, Moustaki and Victoria-Feser (2006) develop robust alternatives to LISREL and robust estimators for Generalized Linear Latent Variable Models. Finally, from a Bayesian perspective, some interest for detecting outliers with categorical and ordinal data has been advanced by Albert and Chib (1993; 1995), who compute Bayesian residuals for both binary and polychotomous response data. The approach has been extended also to sequential ordinal modelling (Albert and Chib, 2001) with the comparison of nonnested models. For a Bayesian modelling of categorical response data we refer to Chib (2005) who includes an overview of MCMC methods in this context.

In this paper we consider a rich class of ordinal response models based on a latent variable with covariates and different link functions. First, we study the impact of deviations from the underlying distribution on the reliability of the Maximum Likelihood (ML) estimator, and illustrate the results for the most common links: the probit, the logistic and the complementary log-log link. In particular, we highlight the role of the generalized residuals as diagnostic tools. Moreover, we compare the behavior of the most common links from a robustness point of view under different types of deviations. Then, as a robust alternative to the ML estimators (MLE), M-estimators are proposed, which yield reliable estimates of the value of the parameters, and can be used to derive robust testing procedures. They can be viewed as weighted MLEs, where the weights provide valuable diagnostic information on possible outliers and substructures in the data. Finally, in an extensive Monte Carlo study we investigate the finite sample behavior of our new robust estimators at the model and under various types of deviations that can appear in practice. In particular, we give recommendations on the choice of the tuning constant for robust estimators.

The paper is organized as follows: the next section provides a brief overview of the approach currently used in the analysis of ordinal response variables, and reviews classical ML inference in this context. Section 3 gives some insights on the robustness of current methods for ordinal response models with special emphasis on the behaviour of the generalized residuals for some popular links. Section 4 proposes a class of robust estimators for ordered response models, whose properties are numerically investigated in Section 5. Robust testing procedures are considered in Section 6. Some concluding remarks end the paper. In the Appendix we provide additional numerical evidence and some computational details. In particular, we investigate the choice of the tuning constant with respect to the efficiency and the robustness of the corresponding estimator.

2. Maximum likelihood inference for ordered response models

Let Y be an ordinal variable of interest which is linked to an underlying latent variable Y^* through the relationship

$$Y = j \quad \Longleftrightarrow \quad \alpha_{j-1} < Y^* \le \alpha_j, \quad j = 1, 2, \ldots, m, \tag{1}$$

where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = +\infty$ are the thresholds (cutpoints) of the continuous support of the latent variable, and m represents the given number of categories of Y.

The variable Y^*, in turn, depends on $p \geq 1$ covariates, so that for the i-th statistical unit we have the latent regression model

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p + \epsilon_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \ldots, n, \quad (2)$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and ϵ_i is a random variable whose distribution and density function are denoted by $G(\epsilon)$ and $g(\epsilon)$, respectively. Since Y^* is unobservable, a random sample is given by $(Y_i, \boldsymbol{X}_i)$, for $i = 1, 2, \ldots, n$.

Relationship (1) yields the following probability mass function for Y_i conditionally on $\boldsymbol{X}_i = \boldsymbol{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip})'$

$$Pr\left(Y_i = j \mid \boldsymbol{x}_i\right) = Pr\left(\alpha_{j-1} < Y_i^* \leq \alpha_j\right) = G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}),$$

for $j = 1, 2, \ldots, m$. In particular, since $\alpha_0 = -\infty$ and $\alpha_m = +\infty$, we have $G(\alpha_0 - \boldsymbol{x}_i'\boldsymbol{\beta}) = G(-\infty) = 0$ and $G(\alpha_m - \boldsymbol{x}_i'\boldsymbol{\beta}) = G(+\infty) = 1$, hence $Pr(Y_i = 1|\boldsymbol{x}_i) = G(\alpha_1 - \boldsymbol{x}_i'\boldsymbol{\beta})$ and $Pr(Y_i = m|\boldsymbol{x}_i) = 1 - G(\alpha_{m-1} - \boldsymbol{x}_i'\boldsymbol{\beta})$. Common specifications of $G(\epsilon)$ are the Gaussian, the logistic and the extreme value distributions (see Agresti, 2010, for an extensive review).

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ be the parameter vector, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m-1})'$ is the vector of the thresholds, then $\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta})$ where $\Omega(\boldsymbol{\theta})$ is an open subset of $\mathbb{R}^{p+m-1}$.

Given an observed random sample $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{X}$ be the matrix whose rows are given by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The log-likelihood function is

$$\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$$

with individual term

$$\begin{aligned} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) &= \sum_{j=1}^{m} I[y_i = j] \log Pr(Y_i = j|\boldsymbol{x}_i) \\ &= \sum_{j=1}^{m} I[y_i = j] \log\left[G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})\right], \end{aligned}$$

where $I[\omega]$ is an indicator function which takes the value 1 when ω holds and 0 otherwise. Notice that $\ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$ includes only one term, that is the j-th term of the sum if the i-th respondent chooses the j-th category.

The score function is $S(\boldsymbol{\theta}) = \sum_{i=1}^{n} s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$, where

$$s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} \qquad (3)$$

$$= \sum_{j=1}^{m} I[y_i = j] \left[\frac{1}{[G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})]} \frac{\partial Pr(Y_i = j | \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} \right],$$

and

$$\frac{\partial Pr(Y_i = j | \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial Pr(Y_i = j | \boldsymbol{x}_i)}{\partial \boldsymbol{\beta}'}, \frac{\partial Pr(Y_i = j | \boldsymbol{x}_i)}{\partial \boldsymbol{\alpha}'} \right)'.$$

Finally, the generic term of the information matrix $\mathcal{I}(\boldsymbol{\theta}, \boldsymbol{X})$ for a single observation, conditionally on $\boldsymbol{X} = \boldsymbol{x}$, is given by

$$
\begin{aligned}
\mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{x}) &= E_Y \left\{ -\frac{\partial^2 \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial \theta_r \, \partial \theta_s} \bigg| \boldsymbol{X} = \boldsymbol{x} \right\} \\
&= -\sum_{j=1}^{m} I[Y = j] \frac{\partial^2 \ell(\boldsymbol{\theta}, Y, \boldsymbol{x})}{\partial \theta_r \, \partial \theta_s} Pr(Y = j | \boldsymbol{x})
\end{aligned}
$$

for $(r, s) = 1, 2, \ldots, m+p-1$, and the elements of the unconditional information matrix $\mathcal{I}(\boldsymbol{\theta})$ are given by $\mathcal{I}_{rs}(\boldsymbol{\theta}) = E_{\boldsymbol{X}} \{ \mathcal{I}_{rs}(\boldsymbol{\theta}, \boldsymbol{X}) \}$.

For the thresholds we have $\partial Pr(Y_i = j | \boldsymbol{x}_i)/\partial \alpha_s = g(\alpha_s - \boldsymbol{x}_i'\boldsymbol{\beta}) \{ I[j = s] - I[j = s + 1] \}$, for $s = 1, 2, \ldots, m - 1$, where it is to be recalled that $g(\alpha_0 - \boldsymbol{x}_i'\boldsymbol{\beta}) = g(-\infty) = 0$ and $g(\alpha_m - \boldsymbol{x}_i'\boldsymbol{\beta}) = g(+\infty) = 0$. Thus, the elements of $s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ related to the thresholds are given by

$$s_{\alpha_s}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)}{\partial \alpha_s} = \frac{g(\alpha_s - \boldsymbol{x}_i'\boldsymbol{\beta}) \{ I[y_i = s] - I[y_i = s + 1] \}}{G(\alpha_{s+1} - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_s - \boldsymbol{x}_i'\boldsymbol{\beta})} \quad (4)$$

for $j = 1, \ldots, m - 1$. Notice that the score functions for the thresholds can be unbounded, when $\boldsymbol{x}_i'\boldsymbol{\beta} \to \pm\infty$, if the density at the numerator converges to zero slower than the denominator.

Since $\partial Pr(Y_i = j | \boldsymbol{x}_i)/\partial \beta_k = \{ g(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}) - g(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) \} x_{ik}$, the elements of $s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ related to the regression coefficients are given by

$$s_{\beta_k}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)}{\partial \beta_k} = \sum_{j=1}^{m} I[y_i = j] \frac{g(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}) - g(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta})}{G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})} x_{ik},$$

$$(5)$$

for $k = 1, 2, \ldots, p$.

Let us re-express (5) by means of the *generalized residuals* (Franses and Paap, 2004, p. 123)

$$e_{ij}(\boldsymbol{\theta}) = \frac{g(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - g(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}{G(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

$$(6)$$

where for $j = 1$ and $j = m$ (6) yields

$$e_{i1}(\boldsymbol{\theta}) = \frac{g(\alpha_1 - \boldsymbol{x}_i\beta)}{G(\alpha_1 - \boldsymbol{x}_i\beta)}, \qquad e_{im}(\boldsymbol{\theta}) = -\frac{g(\alpha_{m-1} - \boldsymbol{x}_i\beta)}{1 - G(\alpha_{m-1} - \boldsymbol{x}_i\beta)}.$$

Then, (5) becomes

$$s_{\beta_k}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = -\sum_{j=1}^{m} I[y_i = j]e_{ij}(\boldsymbol{\theta})\, x_{ik}. \qquad (7)$$

Notice again that, although the generalized residuals $e_{ij}(\boldsymbol{\theta})$ are defined for any $j \in \{1, 2, \ldots, m\}$, the individual score function for β_k, given by (7), contains only the j-th term such that $y_i = j$. The score function for the whole sample is given by

$$\sum_{i=1}^{n} s_{\beta_k}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} I[y_i = j]e_{ij}(\boldsymbol{\theta})\, x_{ik}, \qquad (8)$$

which has the same structure as the first-order condition of the Gaussian (ML) equation in the linear regression model, i.e. $\sum_{i=1}^{n} r_i \boldsymbol{x}_i = 0$, where the r_i's are the residuals. Equation (8) makes clear why the quantities $e_{ij}(\boldsymbol{\theta})$ are called generalized residuals: in ordinal response models they play a similar role to the r_i's in the linear model. Indeed their inspection is crucial to understanding the impact of anomalous observations on estimators.

Outlying values in the covariates induce both the numerator and the denominator of (6) to approach 0 and the final value of $e_{ij}(\boldsymbol{\theta})$, which can either diverge or vanish, depends on the speed of convergence of the two terms. However, if a large $\boldsymbol{x}_i$ is associated with a large residual, the corresponding statistical unit will have a dominating role in (8) when determining the estimate of the parameter. If instead the generalized residual tends to zero when $\boldsymbol{x}_i$ becomes large, then the impact of outlying covariates in the estimation process is limited. Consequently, it is recommended to choose the link function such that the generalized residuals, viewed as a function of $\boldsymbol{x}_i'\boldsymbol{\beta}$, are bounded. This point will be dealt with in greater detail in Section 3.

Notice however that the generalized residuals are affected also by anomalous data in the response Y. This variable has a limited support, $\{1, 2, \ldots, m\}$, hence unbounded outlying observations cannot occur. Nevertheless, a category inconsistent with the covariates may appear as a consequence of an incoherent selection by the respondent or of a collecting, reading or recording error. Such an event can produce an outlying residual which may have a non-negligible impact on (8), as illustrated in the following example.

EXAMPLE 1. Consider a categorical variable obtained through (1) where $Y^* = 1.5X + \epsilon$, X is a standard normal variable, $\epsilon \sim N(0,1)$ and $\boldsymbol{\alpha} = (-1.6, 0, 1.6)'$. An observed sample of $n = 30$ data has been collected, whose values are reported in Table 1.

Suppose that, by mistake, the first value of the response variable Y, which is 1, is erroneously reported as 4. Figure 1 shows the generalized residuals of the original and modified data. In the original data the residuals are displayed around 0 and no outlying values appear. The first generalized residual, which originally has the value 0.521, takes, in the modified dataset, the value -3.919 (this change is displayed by an arrow in the right panel). This residual turns out

TABLE 1
Dataset of Example 1

y	1	2	2	3	4	3	4	2	2	3
x	-1.39	0.04	-0.76	0.21	1.43	0.74	0.70	-0.23	0.20	1.21
y	3	2	2	4	3	3	2	1	2	2
x	0.32	-1.42	-0.41	1.00	0.96	0.92	-0.15	-1.22	-0.87	-1.04
y	1	3	2	4	2	2	3	1	1	2
x	-1.10	0.44	-0.20	1.68	-0.13	-0.20	0.05	-0.68	-0.73	-0.86

to be much lower than the other residuals and clearly outlying. Furthermore it is likely to magnify the role of x_1 in the estimation process.
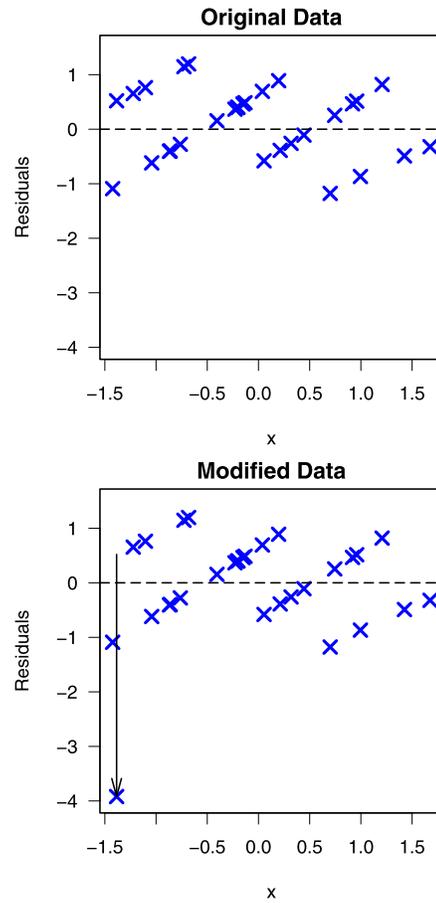

FIG 1. *Generalized residuals of the original and modified data of Example 1.*

3. Influence function, link functions and generalized residuals

In order to formalize the robustness properties of the MLE, we study its influence function (Hampel, 1974). In particular we focus on the boundness of this function, which is necessary for local robustness. The influence function is given by

$$IF(y, \boldsymbol{x}_i; ML, G) = \mathcal{I}^{-1}(\boldsymbol{\theta})s(y, \boldsymbol{x}_i; \boldsymbol{\theta}), \qquad y = 1, \ldots, m.$$

Since, for each parameter, the influence function is a linear combination of all the score functions, it depends on both the covariates $\boldsymbol{x}_i$ and the generalized residuals introduced in (6). While there is no control on the magnitude of the covariates, the generalized residuals can be bounded or unbounded according to the choice of the distribution function $G(\epsilon)$.

To investigate the behaviour of the generalized residuals, let $t = \boldsymbol{x}_i'\boldsymbol{\beta}$ and define the function $A_j(t)$ which, for given thresholds $(\alpha_1, \ldots, \alpha_{m-1})$ and any real t, is given by

$$A_j(t) = \frac{g(\alpha_j - t) - g(\alpha_{j-1} - t)}{G(\alpha_j - t) - G(\alpha_{j-1} - t)}, \qquad j = 1, 2, \ldots, m. \tag{9}$$

Hereafter the behaviour of $A_j(t)$ is examined when some of the most popular links used for ordinal response models, that is those derived from the Gaussian, the extreme value and the logistic distribution, are adopted.

- *Probit link*
  If ϵ has a normal distribution, (9) becomes

  $$A_j(t) = \frac{\phi(\alpha_j - t) - \phi(\alpha_{j-1} - t)}{\Phi(\alpha_j - t) - \Phi(\alpha_{j-1} - t)},$$

  where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution function, respectively. Figure 2 shows $A_j(t)$ for $j = 1, 2, \ldots, m$ when the thresholds are -2.5, -1, 0, 1 and 2.5. The functions $A_j(t)$ are unbounded as $t \to \pm\infty$ (except for $j = 1$ when $t \to -\infty$ and $j = m$ when $t \to +\infty$), hence the impact of large $\boldsymbol{x}_i'\boldsymbol{\beta}$ on the estimates can be amplified by unbounded generalized residuals.
  Furthermore, the score functions for the thresholds, defined by (4), are also unbounded. Consequently, when the probit link is adopted, a large $\boldsymbol{x}_i$ affects all the elements of $s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ and it is likely to have a non-negligible impact in the estimation process.
- *Complementary log-log link*
  Assume $G(\epsilon) = 1 - \exp\{-\exp(\epsilon)\}$, then (9) becomes

  $$A_j(t) = \frac{\{1 - G(\alpha_j - t)\}e^{\alpha_j - t} - \{1 - G(\alpha_{j-1} - t)\}e^{\alpha_{j-1} - t}}{G(\alpha_j - t) - G(\alpha_{j-1} - t)}.$$

  Figure 3 shows that the $A_j(t)$ functions for $j > 1$ are unbounded for negative values of t and converge to 1 as $t \to +\infty$ (notice the asymmetric
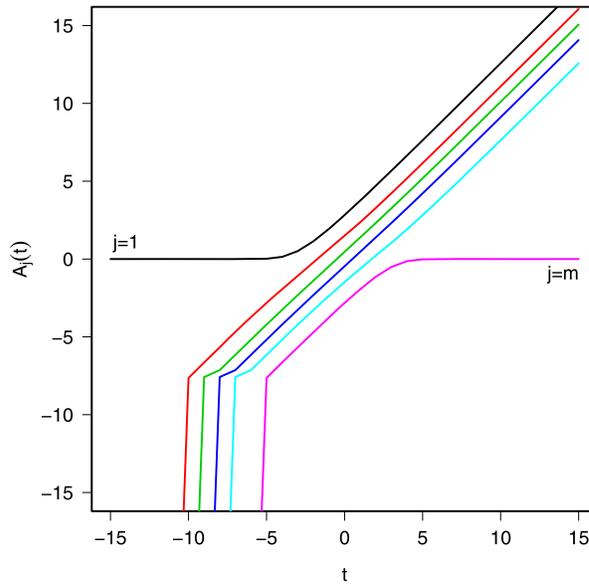
FIG 2. *$A_j(t)$ function for the probit link with thresholds $\boldsymbol{\alpha} = (-2.5, -1, 0, 1, 2.5)'$.*
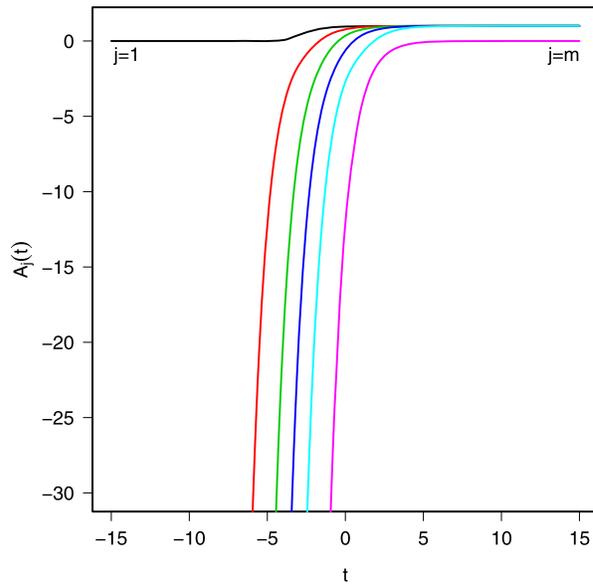

FIG 3. *$A_j(t)$ function for the complementary log-log link with thresholds $\alpha = (-2.5, -1, 0, 1, 2.5)'$.*

behaviour in the tails produced by this distribution). Moreover, for this link, analogously to the probit link, the score functions for the thresholds are also unbounded, so that the influence functions are unbounded as well.

- *Logistic link*
  If ϵ has a logistic distribution, (9) yields

$$A_j(t) = 1 - \left\{ \frac{1}{1 + e^{t-\alpha_j}} + \frac{1}{1 + e^{t-\alpha_{j-1}}} \right\}.$$

The functions $A_j(t)$ are illustrated in Figure 4. When a logistic link is adopted, the functions $A_j(t)$ are bounded everywhere, for any real t, and in particular $\lim_{t \to \pm\infty} A_j(t) = \pm 1$. Hence the generalized residuals are always bounded between -1 and 1, and they neither diverge nor amplify the impact on the estimators of outliers in the covariates.
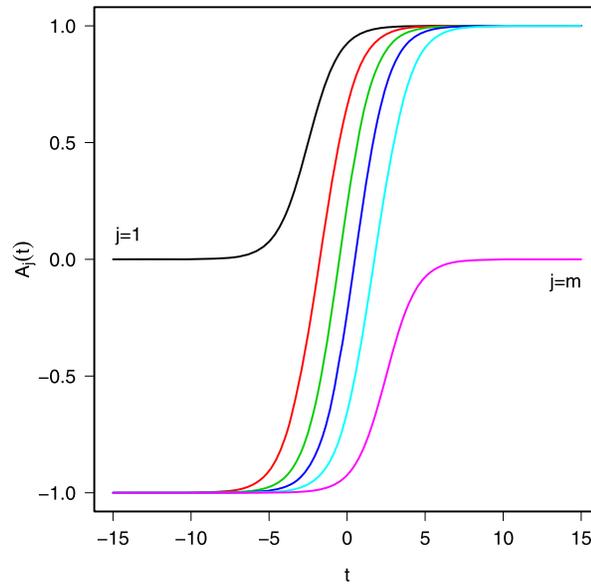
FIG 4. $A_j(t)$ *function for the logistic link with thresholds* $\alpha = (-2.5, -1, 0, 1, 2.5)'$.

Remarkably, the functions which appear on the right side of (4) in score functions for the thresholds are also bounded as illustrated in Figure 5. Hence the only source of unboundness in the score functions is given by the x_{ik}'s which are multiplied by the generalized residuals in (8).
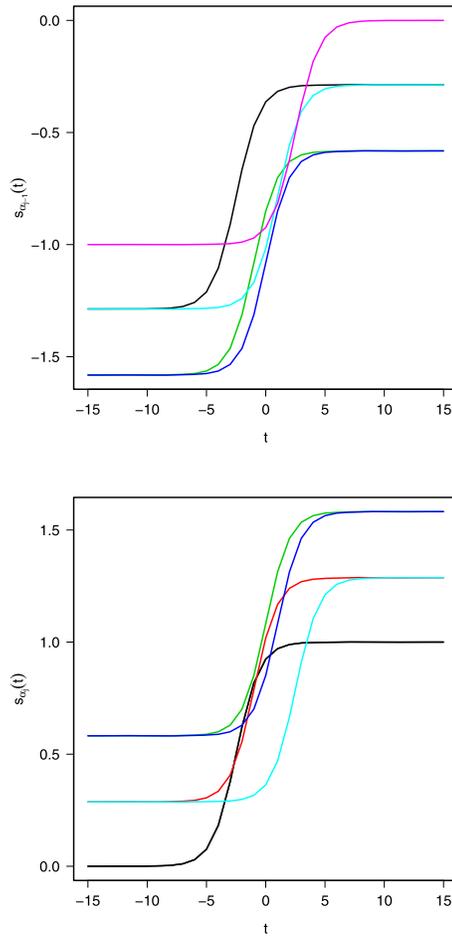
FIG 5. *Threshold score functions for the logistic link with thresholds $\alpha = (-2.5, -1, 0, 1, 2.5)'$.*

4. Robust estimation

A robust estimator for $\boldsymbol{\theta}$, with a bounded influence function, can be obtained by weighting the score functions so as to limit the influence of anomalous data. For this purpose an M-estimator (Hampel *et al.* 1986 p. 100, 230) is introduced, which is the implicit solution $\hat{\boldsymbol{\theta}}_M$ of

$$\sum_{i=1}^{n} \psi(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_M) = 0, \tag{10}$$

where

$$\psi(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \, w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) - a(\boldsymbol{\theta}), \tag{11}$$

and

$$a(\boldsymbol{\theta}) = E\big\{s(Y, \boldsymbol{X}; \boldsymbol{\theta})\, w(Y, \boldsymbol{X}; \boldsymbol{\theta})\big\}.$$

The weights $w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ in (11) are designed to downweight outlying observations in order to control their impact in the estimation. Of course, if $w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \equiv 1$ then $a(\boldsymbol{\theta}) \equiv 0$, $\psi(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ coincides with $s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$, and the solution of (10) is the MLE.

The term $a(\boldsymbol{\theta})$ in (11) is required to achieve Fisher consistency for $\hat{\boldsymbol{\theta}}_M$, and can be obtained as follows

$$a(\boldsymbol{\theta}) = E_X\bigg[E_Y\big\{s(Y, \boldsymbol{X}; \boldsymbol{\theta})w(Y, \boldsymbol{X}; \boldsymbol{\theta})|\boldsymbol{X}\big\}\bigg],$$

where by (3) it yields

$$
\begin{aligned}
E_Y\big\{s(Y, \boldsymbol{X}; \boldsymbol{\theta})w(Y, \boldsymbol{X}; \boldsymbol{\theta})|\boldsymbol{X} = \boldsymbol{x}_i\big\} &= \sum_{r=1}^{m} s(r, \boldsymbol{x}_i; \boldsymbol{\theta})w(r, \boldsymbol{x}_i; \boldsymbol{\theta})Pr(Y = r|\boldsymbol{x}_i) \\
&= \sum_{r=1}^{m} \frac{\partial Pr(Y = r|\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}\, w(r, \boldsymbol{x}_i, \boldsymbol{\theta}). \quad (12)
\end{aligned}
$$

Let $M(\boldsymbol{\theta}, \psi) = -E\bigg\{\dfrac{\partial}{\partial \boldsymbol{\theta}}\, \psi(Y, \boldsymbol{X}; \boldsymbol{\theta})\bigg\}$; the influence function of the M-estimator is given by

$$IF(y, \boldsymbol{x}_i; \psi, G) = M^{-1}(\boldsymbol{\theta}, \psi)\psi(y, \boldsymbol{x}_i; \boldsymbol{\theta}) \qquad \text{for } y = 1, 2, \ldots, m, \text{ and } \boldsymbol{x}_i \in \mathbb{R}^p.$$

In order to have a bounded influence function, the function $\psi(\cdot)$ needs to be bounded. The following two cases should be distinguished.

- *The covariates are outlier free.* This is either the case of qualitative explanatory variables, whose categories are represented by means of dichotomous regressors, or the case of finite explanatory variables without extreme values. Here the generalized residuals are the only source of unboundness of the score functions, and hence of the influence function. Consequently estimators with a bounded influence function can be obtained either by adopting the logistic link, whose generalized residuals have the limited range $(-1, 1)$, or by applying a downweighting of the residuals through an appropriate weight function when other links are adopted.
- *The covariates may contain outliers.* This case occurs when there are unlimited variables among the regressors. Here unboundness of the $\psi(\cdot)$ functions can derive either from outlying covariates or from unbounded generalized residuals, and the sequence of weights $w(\cdot)$ in (11) should limit the effect of both large $\boldsymbol{x}_i$ and large $e_{ij}(\boldsymbol{\theta})$.

When the regressors cannot contain outliers and the link is different from the logistic one, the only concern from the robustness viewpoint is given by large

generalized residuals. Hence, the following Huber weights (Hampel *et al.*, 1986, p. 104), which are functions of the magnitude of the residuals, can be used

$$
w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid < c \\ \dfrac{c}{\sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid}, & \text{if } \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \geq c, \end{cases}
$$

(13)

where c is a positive constant and the following relationship is exploited to evaluate the magnitude of the residuals

$$
\Big\| \sum_{j=1}^{m} I[y_i = j] \, e_{ij}(\boldsymbol{\theta}) \Big\| = \sqrt{\sum_{j=1}^{m} I[y_i = j] \, e_{ij}(\boldsymbol{\theta})^2} = \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid .
$$

The weights in (13) are decreasing functions of the magnitude of the residuals, when the latter exceed c in absolute value.

In the second case, when the regressors may contain outliers so that large $|e_{ij}(\boldsymbol{\theta})|$ and large $\boldsymbol{x}_i$ are a concern, the following weights, which are non-increasing functions of the magnitude of both the residuals and the covariates, can be applied

$$
w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\| < c \\ \dfrac{c}{\sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\|}, & \text{if } \sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\| \geq c. \end{cases}
$$

(14)

A Mahalanobis distance $\|\boldsymbol{x}_i\| = \left\{ (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X)' \hat{\boldsymbol{\Sigma}}_X^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X) \right\}^{1/2}$ can be used for the norm of the covariates in (14), which needs however to be based on a robust multivariate estimator of location $\hat{\boldsymbol{\mu}}_X$ and scatter $\hat{\boldsymbol{\Sigma}}_X$.

The threshold $\frac{c}{\|\boldsymbol{x}_i\|}$ defining the weights $w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ in (14) is no longer a constant, but depends on $\|\boldsymbol{x}_i\|$. More trimming is applied to a residual $\sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid$ associated with an outlying covariate, i.e. when $\|\boldsymbol{x}_i\|$ is large. This is analogous to the linear regression case; see Hampel *et al.* (1986), Fig.1, p. 322.

Notice also that if one adopts a logistic link, the generalized residuals are always bounded (see Section 3) and in principle only weights on the $\boldsymbol{x}_i$ may be needed. This is the robust estimator proposed by Croux *et al.* (2013). However the numerical experiments of Section 5 provide evidence that considering the interaction between regressors and generalized residuals can enhance the efficiency of the M-estimators, even with the logistic link.

Finally the value of the tuning constant c is to be chosen so that the loss of efficiency incurred by M-estimators with respect to the MLEs, does not exceed a given threshold (say 5% or 10%), when there is no contamination in the data; see Section 5.1.

Under general regularity conditions (Huber, 1981), the M-estimator $\hat{\boldsymbol{\theta}}_M$ is asymptotically normal, i.e.

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\right) \to N(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}, \psi)),$$

with asymptotic variance-covariance matrix

$$V(\boldsymbol{\theta}, \psi) = \boldsymbol{M}^{-1}(\boldsymbol{\theta}, \psi)\boldsymbol{Q}(\boldsymbol{\theta}, \psi)\boldsymbol{M}^{-1}(\boldsymbol{\theta}, \psi)$$

and $\boldsymbol{Q}(\boldsymbol{\theta}, \psi) = E\{\psi(Y, \boldsymbol{X}; \boldsymbol{\theta})\psi(Y, \boldsymbol{X}; \boldsymbol{\theta})'\}$.

$\boldsymbol{V}(\boldsymbol{\theta}, \psi)$ can be estimated by using sample estimates of $M(\boldsymbol{\theta}, \psi)$ and $Q(\boldsymbol{\theta}, \psi)$, i.e. $\hat{\boldsymbol{V}}_\psi = \hat{\boldsymbol{M}}^{-1}\hat{\boldsymbol{Q}}\hat{\boldsymbol{M}}^{-1}$, where

$$\hat{\boldsymbol{M}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_M}; \qquad \hat{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_M)\psi(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_M)'.$$

If we denote by $\boldsymbol{V}_{ML}$ and $\boldsymbol{V}_\psi$ the variance-covariance matrices of the ML and M-estimators, respectively, various relative efficiency measures (listed below) can be considered to evaluate the loss of efficiency at the model corresponding to a given c.

- *Trace criterion.* This is a global efficiency measure which compares the traces of the asymptotic variance-covariance matrices of the MLEs and the M-estimators,

$$\mathit{eff}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \frac{tr(\boldsymbol{V}_{ML})}{tr(\boldsymbol{V}_\psi)}.$$

- *Minimum variance ratio criterion* considers the minimum among the ratios between the asymptotic variances of the competing estimators,

$$\mathit{eff}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \min_i\left\{\frac{\boldsymbol{V}_{ML}^{ii}}{\boldsymbol{V}_\psi^{ii}}\right\}.$$

It evaluates the maximum loss of efficiency in the estimation of a single parameter, which can be encountered when the data are generated by the assumed model and an M-estimator is used.

- *Determinant criterion.* This is also a global measure of efficiency which compares the determinants of the asymptotic variance-covariance matrices,

$$\mathit{eff}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \frac{det(\boldsymbol{V}_{ML})}{det(\boldsymbol{V}_\psi)}.$$

All these criteria are functions of c. If the asymptotic variance-covariance matrices were known, given an efficiency target (say 0.95) for the M-estimator, the value of c could be chosen such that $eff(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = 1 - \nu$, where ν is the acceptable loss of efficiency. Unfortunately the asymptotic variance-covariance matrices are typically unknown, since they depend on various factors including the distribution of the covariates which is usually unavailable. Simulated versions of these criteria are therefore used in Section 5, where numerical investigations are carried out to determine suitable values of the tuning constant.

5. Numerical experiments

This Section investigates the finite sample behavior of our new robust estimators and discusses the choice of the tuning constant. Subsection 5.1 is devoted to identify which values of c guarantee that the loss of efficiency, when no contamination of the data occurs, is below the 5% threshold, whereas Subsection 5.2 explores the performance of M-estimators when the data are contaminated by gross errors or outliers. Both the probit and the logistic link are considered, since they are the most frequently used in practice. Due to space constraints, only the most relevant results are reported here, while an extended version of the numerical experiments is provided in the Appendix.

The analysis is carried out through numerical experiments on six models. The first two models have only qualitative covariates, whose categories are represented through dichotomous variables, so that the explanatory variables are outlier free, although anomalous data can still be encountered in the response variable. Hence for these two models robust estimation is performed only for the probit link. When the logistic link is adopted and $\boldsymbol{x}_i$ is made up of dichotomous variables, robust estimation is unnecessary since the generalized residuals belong to $(-1, 1)$ (see Sections 3 and 4). The following three models have 1, 2 and 3 continuous covariates respectively, which implies that outlying $\boldsymbol{x}_i$ are a concern. Finally, in the last model there are both dichotomous and continuous regressors. For models from 3 to 6 robust estimation is necessary with both the probit and the logistic link.

- MODEL 1. The response variable Y, generated by (1), assumes $m = 5$ categories. It depends on one qualitative explanatory variable with four categories, coded through 3 dichotomous $0 - 1$ variables X_1, X_2, and X_3, such that at most one of them can take the value 1. The error component in (2) is $N(0, 1)$, the regression coefficients are given by $\boldsymbol{\beta} = (2.5, 1.2, 0.5)'$, and the cutpoints are given by $\boldsymbol{\alpha} = (-0.7, 0, 1.5, 2.9)'$.
- MODEL 2. The response variable Y assumes 4 categories and depends on two qualitative variables W_i, for $i = 1, 2$. Each of them assumes three categories, coded by two dichotomous $0 - 1$ variables X_i^a and X_i^b such that $X_i^a + X_i^b \leq 1$. The latent variable is $Y^* = 2.5X_1^a + 1.0X_1^b + 3.6X_2^a + 1.8X_2^b + \epsilon$, where $\epsilon \sim N(0, 1)$ and the cutpoints are $\boldsymbol{\alpha} = (1.2, 2.8, 5)'$.
- MODEL 3. The response variable Y assumes 5 categories and depends on one regressor $X \sim N(0, 1)$. The latent variable is $Y^* = 1.5X + \epsilon$, and

the cutpoints are $\boldsymbol{\alpha} = (-1.7, -0.5, 0.5, 1.7)'$ when $\epsilon \sim N(0, 1)$, and $\boldsymbol{\alpha} = (-2.1, -0.6, 0.6, 2.1)'$ when ϵ has a logistic distribution with mean 0 and variance $\pi^2/3$. The cutpoints are defined so that they roughly correspond to the same percentiles of the latent variable Y^* for the probit and the logistic link.

- MODEL 4. The response variable Y assumes 4 categories, and depends on two regressors $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 4)$, with $\text{cov}(X_1, X_2) = 1.2$. The regression coefficients are given by $\boldsymbol{\beta} = (1.5, 0.7)'$, while the cutpoints are $\boldsymbol{\alpha} = (-2.3, 0, 2.3)'$ for the probit link and $\boldsymbol{\alpha} = (-2.6, 0, 2.6)'$ for the logistic one.

- MODEL 5. The response variable Y assumes 3 categories and the regressors are $X_1 \sim N(0, 1)$, $X_2 \sim N(0, 4)$ and $X_3 \sim N(0, 9)$, with $\text{cov}(X_1, X_2) = 1.5$, $\text{cov}(X_1, X_3) = 0.8$ and $\text{cov}(X_2, X_3) = 2.5$. The regression coefficients are given by $\boldsymbol{\beta} = (2.5, 1.2, 0.7)'$ while the cutpoints are $\boldsymbol{\alpha} = (-3.8, 3.8)'$ and $\boldsymbol{\alpha} = (-4, 4)'$ for the probit and the logistic link, respectively.

- MODEL 6. The response variable Y assumes 4 categories and depends on two explanatory variables $D \sim Bernoulli(0.5)$ and $X \sim N(0, 1)$. The latent regression model is $Y^* = 2.5D + 1.2X + 0.7XD + \epsilon$, and the cutpoints are $\boldsymbol{\alpha} = (-1, 1, 3)'$ and $\boldsymbol{\alpha} = (-1.4, 1.1, 3.4)'$ for the probit and the logistic link, respectively.

5.1. Tuning constant and loss of efficiency at the model

In order to measure the loss of efficiency of M-estimators, when the data are generated by the assumed model, 1000 samples of size $n = 200$ are drawn for each model, and both ML and M-estimators are computed. On the basis of experimental evidence, values between 0.5 and 4 are considered for the tuning constant c, on a grid with step 0.1.

When performing M-estimation, the weights are given by (13) for Models 1 and 2, and by (14) for the other models. Furthermore, when the logistic link is adopted, the question arises whether it is useful to include $|e_{ij}(\boldsymbol{\theta})|$ in the argument of the weights, in Models from 3 to 6, since the generalized residuals have a limited range. For this reason, two estimation procedures are considered for this link: the first one uses the weights defined in (14), while in the second procedure the weights are functions of $\|\boldsymbol{x}_i\|$ only, i.e.

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \min\{1, c/\|\boldsymbol{x}_i\|\}. \tag{15}$$

These weights are analogous to those used by Carroll and Pederson (1993) to obtain robust estimators for the logistic model with a dichotomous response variable.

In Model 3, since there is only one regressor, the norm $\|x_i\|$ for the i-th observation is computed as $|x_i - Med(X)|/MAD(X)$, where $Med(X)$ and $MAD(X)$ are the median and the normalized median absolute deviation of X. The same procedure is used in Model 6, where the norm is a function of the continuous variable only. In the other two models (4 and 5), the norm of the regressors

is given by the Mahalanobis distance, computed after estimating location and scale of $\boldsymbol{X}$ through the Stahel-Donoho procedure (Stahel, 1981; Donoho, 1982; see also Maronna and Yohai, 1995).

The three efficiency criteria introduced in Section 4 are applied to evaluate the loss of efficiency due to M-estimation at the model, with the asymptotic variance-covariance matrices of the estimators replaced by the Mean Square Error (MSE) matrices obtained from simulation.

The minimum values of c such that the loss of efficiency of M-estimators is not larger than 5%, are shown in Table 2.

TABLE 2

Minimum values of the tuning constant required to achieve an efficiency loss smaller than 5% when M-estimation is performed and there is no contamination in the data - $n = 200$.

Link	Efficiency Criterion	*Model* 1	2	3	4	5	6
Probit	Trace	1.1	1.2	1.1	1.7	1.7	1.1
	Min(MSE-ratio)	1.2	1.4	1.7	2.3	2.1	1.4
	Determinant	1.6	2.0	2.2	3.1	2.7	2.0
Logistic	Trace			0.6	0.9	0.9	0.7
	Min(MSE-ratio)			0.8	1.0	1.1	0.8
	Determinant			0.9	1.4	1.4	1.1
Logistic	Trace			0.8	1.1	1.0	0.8
with weights	Min(MSE-ratio)			1.1	1.4	1.3	0.9
(15)	Determinant			1.2	1.9	1.9	0.8

Some remarks follow.

- For a given loss of efficiency, the values of c are smaller for the logistic link than for the probit link. This evidence is explained by the circumstance that the logistic generalized residuals have a limited range and hence they can be expected to be smaller than the probit residuals. Therefore, the same value of c yields smaller weights with the probit link than with the logistic one, with a consequent larger loss of efficiency.
- For a fixed loss of efficiency, the Trace criterion requires smaller values of c, whereas the Determinant criterion requires larger values (i.e. a milder downweighting of the observations). On the whole, the value of c does not seem to be appreciably affected by the nature (dichotomous or continuous) and the number of regressors.
- When the probit link is adopted, in order to keep the loss of efficiency below 5%, c needs roughly to be between 1.1 and 1.7 according to the Trace criterion, between 1.2 and 2.3 according to the Minimum MSE ratio criterion, and between 1.6 and 3.1 according to the Determinant criterion. In short, when the target is a 5% loss at the model, c needs to be greater than 1, but a value of $c \geq 3$ might be too conservative.
- When the logistic link is adopted, $c < 1.5$ seems generally suitable to ensure that the loss of efficiency is below 5%, whatever criterion is used.

- When estimation with the logistic link is carried out with weights (15), the value of the tuning constant necessary to keep the loss of efficiency below 5% is usually slightly higher than that required by weights (14). In other words, for a given c, the loss of efficiency is larger when the residuals are neglected than when they are considered jointly with $\|\boldsymbol{x}_i\|$ in the weights. This provides evidence that it is important, when evaluating the magnitude of $\boldsymbol{x}_i$, to take the corresponding $|e_{ij}(\boldsymbol{\theta})|$ into account. In order to gain efficiency, two statistical units, such that $\|\boldsymbol{x}_i\|$ takes the same value, should be treated differently according to the size of the associated generalized residuals. For this reason, the investigation of this estimation method is not pursued further.

Figure 6 shows how the efficiency of the M-estimators varies with c, according to the three criteria, in case of Model 3 when the probit and the logistic link are adopted, respectively. When c decreases, the control on outlying observations increases and the ψ functions in (11) move away from the score functions, yielding a larger loss of efficiency. Consequently, the larger is the admissible loss of efficiency, the smaller is the value of the tuning constant. It is also to be remarked that the convergence of the three criteria to 1 is faster for the logistic link than for the probit one. Consequently, for a given c, the loss of efficiency at the model is smaller when the logistic link is adopted rather then when the probit one is used.

5.2. Tuning constant and robustness

After evaluating the loss of efficiency at the model due to M-estimation, it is interesting to investigate what is the gain in efficiency which can be achieved by robust estimation when data are contaminated. In what follows, some contaminations of the previous models are considered which highlight the performance of M-estimation when some observations deviate from the assumed model. In particular, three types of deviations are taken into account: gross errors in the response variable, outliers in continuous covariates and model misspecification.

In the comparison of alternative estimators, along with the Minimum MSE ratio, the maximum ratio between the MSE of the ML and M-estimators for each parameter,

$$\max_i \left\{ \frac{MSE_i^{ML}}{MSE_i^M} \right\},$$

is also considered. This additional criterion is useful when data contamination occurs, since under deviations from the model robust estimators are expected to be more efficient than MLEs. In particular, it points out what is the maximum gain in efficiency which can be achieved by M-estimation for a single parameter. For each numerical experiment 1˙000 samples of size $n = 200$ are generated.

CONTAMINATED MODEL 2 (*Shelter Effect*). Here we consider the case when five Y_i, which originally take value 1, 2 or 3, are changed into 4. This kind of contamination occurs when the selected category (in this case "four") can be

Probit Link

Tuning constant

Logistic Link

Tuning constant
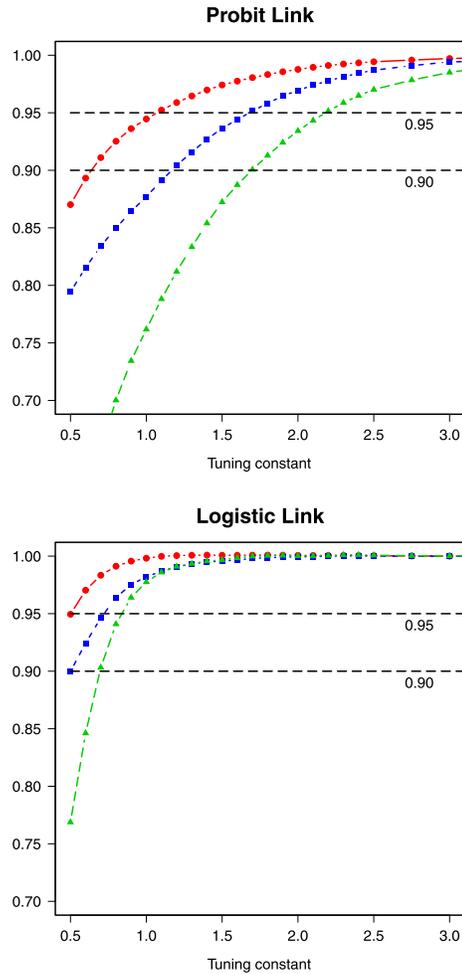
FIG 6. *Trace Efficiency Criterion (circles), Minimum MSE criterion (squares) and Determinant Criterion (triangles) in Model 3 - n = 200.*

regarded as a *shelter* category: a choice that the respondents feel comfortable with, although it appears incoherent with their profiles in terms of covariates (see Iannario, 2012, for a more extensive illustration of shelter choices).

The minimum and the maximum MSE-ratios useful for comparing the M-estimators with the MLEs are shown in Table 3. Values of c between 1 and 1.5 seem thoroughly appropriate to achieve robust estimation according to the two criteria. If say $c = 1.25$ is taken, the gain in efficiency in the estimation of a single parameter varies between 37.6% and roughly 140%, which is a quite remarkable achievement obtained by M-estimators. On the contrary, high values of c yield a poor performance of robust estimation.

TABLE 3

Efficiency criteria when M-estimation is performed in Model 2 with the probit link and a shelter effect occurs.

c	1	1.25	1.5	1.75	2	2.5	3	3.5
Min(MSE-ratio)	1.354	1.376	1.373	1.342	1.299	1.201	1.124	1.064
Max(MSE-ratio)	2.556	2.398	2.227	2.029	1.819	1.485	1.253	1.099

CONTAMINATED MODEL 3. Suppose that gross errors occur so that, for κ statistical units, the value of the regressor, which is $N(0,1)$, is erroneously recorded as 5. The MSE-ratio in Table 4 compares the MLE and the M-estimator (with $c = 1.5$) of the regression coefficient, when κ varies between 1 and 20, which - in a sample of size $n = 200$ - corresponds to an amount of contamination between 0.5% and 10%.

TABLE 4

MSE-ratio between the MLE and the M-estimator (with $c = 1.5$) of the regression coefficient in Model 3 when there are κ outliers equal to 5 in the explanatory variable.

κ	1	2	3	4	5	6	7
Probit link	6.963	15.912	23.146	26.697	28.012	27.046	25.253
Logistic link	1.283	2.212	3.401	4.377	4.930	5.315	5.368
κ		8	9	10	12	15	20
Probit link		23.072	20.840	18.419	14.576	10.541	6.600
Logistic link		5.098	4.905	4.538	3.853	2.821	1.788

Robust estimation produces a very large gain in efficiency, especially when the probit link is adopted. In fact, the efficiency as a function of the amount of contamination increases until a peak is reached, then decreases as a consequence of likely masking effects. Nevertheless, two remarks are to be made. The decrease in efficiency starts at a larger contamination for the logistic link with respect to the probit link; secondly, despite the decrease in efficiency, robust estimation remains still more efficient than ML estimation.

Figure 7 shows the MSE-ratio for the regression coefficient when c varies and $\kappa = 3$ for both the probit and the logistic link. Although the magnitude of the gain in efficiency is widely different for the two links, Figure 7 shows that c needs to be small, say $c \leq 2$, to achieve a more effective robust estimation.

CONTAMINATED MODEL 3 (*continued*). Table 5 shows the MSE-ratio between the MLE and the M-estimator of the regression coefficient in Model 3 when - because of a gross error - the value of x_i for a single statistical unit is replaced by an outlier, whose value varies between 3 and 10. For both the probit and logistic link the value of the tuning constant is $c = 1.5$. From the table it can be appreciated how the gain in efficiency of the M-estimators over the MLEs increases as the outlier moves away from the bulk of data. As in the previous experiment, the gain in efficiency is much larger for the probit link than for the logistic one, although robust estimation is very useful in both cases, as soon as the magnitude of the outlier becomes noticeable.

Probit Link

Tuning constant

Logistic Link

Tuning constant
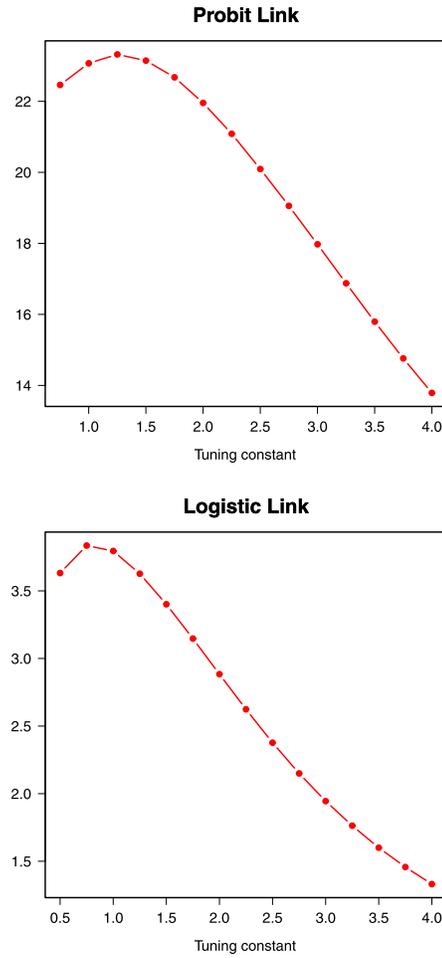
Fig 7. *MSE-ratio between the MLE and the M-estimator of the regression coefficient in Model 3 as c varies, when there are 3 outliers equal to 5 in the explanatory variable.*

TABLE 5
MSE-ratio between the MLE and the M-estimator (c = 1.5) of the regression coefficient in Model 3 when the outlier varies.

Outlier	3	4	5	6	7	8	9	10
Probit link	2.001	3.923	6.963	10.963	15.650	20.747	26.016	31.276
Logistic link	1.021	1.127	1.283	1.480	1.717	1.992	2.306	2.658

CONTAMINATED MODEL 4. The contamination considered for Model 4 is given by two statistical units whose regressors are jointly mis-reported as 5. Hence the amount of contamination is indeed very mild, that is 1%. Table 6 shows the Trace and the Determinant criterion, and the MSE-ratios between the MLEs

and the M-estimators of the regression coefficients when both the probit and the logistic link are adopted. Also this experiment highlights the considerable greater efficiency of M-estimators with respect to MLEs.

TABLE 6

Trace criterion, MSE-ratio for the regression coefficients and Determinant criterion in Model 4 under 1% contamination of the data ($x_1 = x_2 = 5$ for two statistical units).

	Trace	MSE-ratio		Determinant
c	Criterion	$\beta_1 = 1.5$	$\beta_2 = 0.7$	Criterion
		Probit link		
1.00	10.789	22.287	5.475	1.816
2.00	11.491	22.358	6.453	2.894
3.00	10.927	19.428	6.613	3.100
		Logistic link		
1.00	1.352	2.300	0.937	1.375
2.00	1.392	2.195	1.011	1.638
3.00	1.306	1.761	1.032	1.467

CONTAMINATED MODEL 5. Model 5 has been estimated by MLEs and M-estimators after inserting two extra unnecessary and independent regressors: $X_4 \sim N(0,1)$ and $X_5 \sim N(0,1)$. Hence the overparameterized model assumed for the latent variable in the estimation process is $Y^* = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, though $\beta_4 = \beta_5 = 0$. In addition the data have been contaminated by changing the value of X_1 into 5 for κ statistical units. Table 7 shows how the Trace criterion varies with κ when $c = 1.5$ for both the probit and the logistic link. The gain in efficiency provided by M-estimation is clearly evident.

TABLE 7

Trace criterion when Model 5 is misspecified by adding two unnecessary regressors, and there are κ anomalous values in X_1 ($c = 1.5$).

κ	1	2	3	4	5
Probit link	2.051	3.688	5.156	5.311	4.830
Logistic link	1.129	1.754	2.398	2.776	2.888

6. Robust testing procedures

Robust test procedures can be based on robust M-estimators by means of t-type statistics. As mentioned in Section 4, M-estimators are asymptotically normal, and the estimator $\hat{\boldsymbol{V}}_\psi$ of the asymptotic variance-covariance matrix is consistent for $V(\boldsymbol{\theta}, \psi)$. Hence, under the null hypothesis $H_0^k : \beta_k = 0$ for $k = 1, \ldots p$, we have

$$t_k = (\hat{\beta}_k^M - \beta_k) / \left\{ \hat{\boldsymbol{V}}_\beta^{kk} \right\}^{1/2} \overset{d}{\Rightarrow} N(0, 1),$$

where $\hat{\boldsymbol{V}}_\beta^{kk}$ is the k-th element on the diagonal of $\hat{\boldsymbol{V}}_\beta$, in turn $\hat{\boldsymbol{V}}_\beta$ is the submatrix of $\hat{\boldsymbol{V}}_\psi$ related to the regression coefficients and $\hat{\beta}_k^M$ is the k-th element of the M-estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$.

Wald type tests can also be carried out. Consider the hypothesis $H_0 : h(\boldsymbol{\theta}) = 0$, where $h(\boldsymbol{\theta})$ is a vector of d functions. Under H_0 we have

$$h(\hat{\boldsymbol{\theta}}_M)'\left\{H(\hat{\boldsymbol{\theta}}_M)'\hat{\boldsymbol{V}}_\psi H(\hat{\boldsymbol{\theta}}_M)\right\}^{-1}h(\hat{\boldsymbol{\theta}}_M) \stackrel{d}{\Rightarrow} \chi_d^2, \tag{16}$$

where $H(\hat{\boldsymbol{\theta}}_M)$ is the matrix of the derivatives of $h(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}_M$. Following Heritier and Ronchetti (1994), these tests inherit the robustness of the corresponding M-estimators and exhibit robustness of validity and robustness of efficiency.

MODEL 7. This model is similar to Model 3 but the response variable Y, which still depends on one regressor $X \sim N(0, 1)$, assumes 4 categories. The latent variable is $Y^* = 1.2X + \epsilon$, and the cutpoints are $\boldsymbol{\alpha} = (-1.5, 0, 1.5)'$ for the probit link, and $\boldsymbol{\alpha} = (-2, 0, 2)'$ for the logistic one. In the estimation process the model is misspecified by including a dichotomous regressor $D \sim Bernoulli(0.5)$ and its interaction with X. Hence the latent model assumed in the estimation is $Y^* = \beta_1 D + \beta_2 X + \beta_3 XD + \epsilon$. In addition there is a mild contamination (0.5%) of the data: the value of X is changed into 5 for a single statistical unit.

t-tests are carried out on the hypotheses $H_0^k : \beta_k = 0$ for $k = 1, 3$. The tests are based on both MLEs and M-estimators, and for the latter the tuning constant is set to $c = 1.5$. The tests are performed on $1\cdot000$ simulated samples of size $n = 200$, at the significance level 5% and 10%.

Table 8 shows the percentage of rejections of H_0^k for $k = 1, 3$. For the parameter β_1 the null hypothesis is true and gets generally rejected in a percentage of cases close to the nominal level. The simulated level of the test based on M-estimators, rather than on MLEs, is generally slightly closer to the nominal one. A remarkable difference in the performance of the tests occurs when testing the hypothesis H_0^3. Despite the parameter being zero, the percentage of rejections of the test based on the MLEs is high, especially when the probit link is adopted. In contrast, when the test is derived from M-estimators, the actual rate of rejection is satisfactorily close to the nominal level.

TABLE 8
Percentage of rejections in Model 7 of $H_0^k : \beta_k = 0, k = 1, 3$, under 0.5% of contamination in the data ($c = 1.5$).

	Probit link				Logistic link			
Significance level	5%		10%		5%		10%	
	β_1	β_3	β_1	β_3	β_1	β_3	β_1	β_3
MLE	4.2	41.4	9.2	54.1	4.5	8.6	10.3	16.5
M	4.5	5.6	9.7	9.6	4.4	4.8	10.2	9.4

Figure 8 shows the Receiver Operating Characteristics (ROC) curves for the t-tests based on the MLEs and on the M-estimators for the hypotheses H_0^1 and H_0^3, when the probit link is adopted. The true positive rates have been evaluated when the model for the latent variable is $Y^* = 0.7D + 1.2X + 0.5XD + \epsilon$. The curves make evident the better performance of the test based on M-estimators in terms of power as well as size.
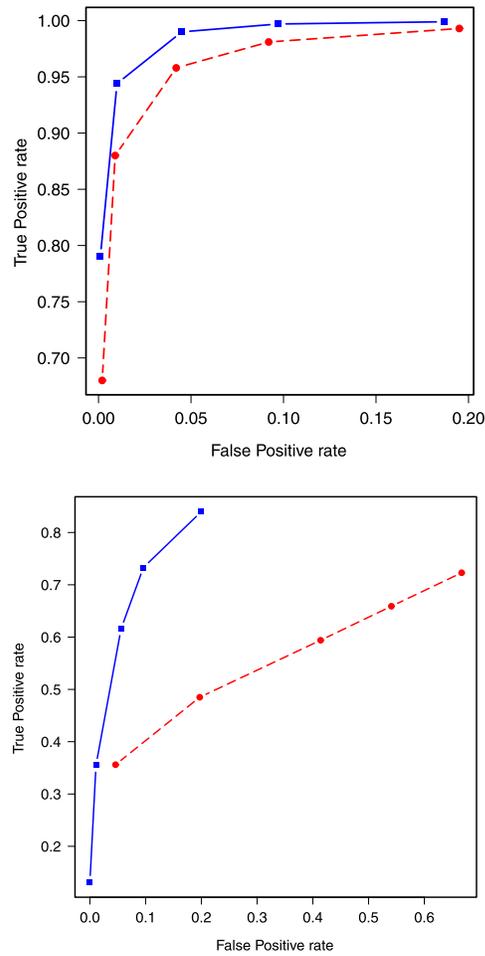
FIG 8. *ROC curves for the t-tests on β_1 (top panel) and β_3 (bottom panel) based on the MLEs (dashed line with circles) and on the M-estimators (solid line with squares) when the probit link is adopted - $n = 200$. The knots correspond to the nominal significance levels 0.1%, 1%, 5%, 10% and 20%.*

Finally, the hypothesis $H_0^D : \beta_1 = \beta_3 = 0$, to check whether D is a useful explanatory variable, is considered. When the model is estimated by MLEs, the tests are based on the Likelihood Ratio (LR) statistic, whereas when M-estimation is performed the tests are carried out through the Wald-type statistic introduced in (16).

Table 9 shows the percentage of rejections of the hypothesis H_0^D. The results are consistent with those of the t-tests. The percentage of rejections by the LR test are much higher than the nominal level, and this test becomes markedly too liberal when the probit link is adopted.

TABLE 9
Percentage of rejections in Model 7 of $H_0^D : \beta_1 = \beta_3 = 0$ under 0.5% of contamination in the data ($c = 1.5$).

		Probit link		*Logistic link*	
Significance level		5%	10%	5%	10%
MLE	(LR)	32.9	44.5	7.8	14.5
M	(Wald)	5.2	10.5	5.5	11.0

The ROC curves for the LR test and the Wald test based on M-estimators for H_0^D - when the probit link is adopted - are shown in Figure 9. In the comparison between the Wald robust test and the LR test, these curves confirm that the former test has an actual significance level closer to the nominal one and it more powerful. To summarize:
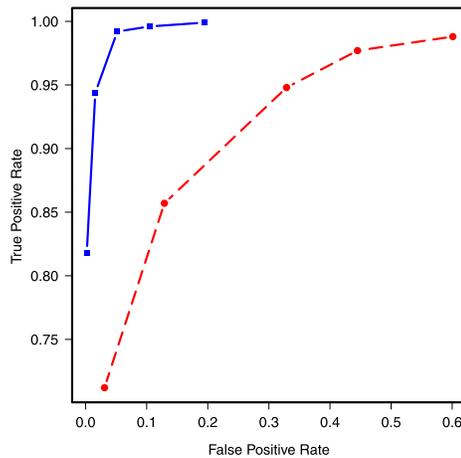
FIG 9. *ROC curves for LR test (dashed line with circles) and the robust Wald test (solid line with squares) on H_0^D when the probit link is adopted - $n = 200$. The knots correspond to the nominal significance levels 0.1%, 1%, 5%, 10% and 20%.*

- Both t and Wald type tests based on M-estimators are considerably more accurate (in terms of actual significance level) and more powerful than tests based on MLEs.
- If the MLEs are to be applied, the logistic link is to be recommended. It is more robust than the probit link, as it is reflected in the more reliable testing procedures.

Finally, the goodness of link test (Atkinson and Riani, 2000, p. 200-201) could be used as a complementary tool for guidance in the choice between alternative links.

7. Concluding remarks

The paper highlights that robustness is a relevant issue in modelling ordinal responses, and proposes robust estimation and testing procedures based on M-estimators. Such procedures provide reliable inference when data are contaminated and outperform procedures based on the MLEs.

Another relevant result developed in the paper concerns the impact of the link function. Actually, although there is a general agreement about the use and interpretation of estimated parameters with different link functions (logistic and probit link give practically equivalent results), our analyses point out that the behaviour of alternative links is completely different from the robustness viewpoint. In particular, analytical results as well as numerical experiments support the preference for the logistic link as a more robust link.

Finally the role of generalized residuals obtained by the estimated models with covariates is also investigated. The operational suggestion is to consider their plots as a fundamental diagnostic tool in the framework of ordinal response models.

8. Appendix

This appendix includes an extended version of the results of Section 5 related to numerical experiments intended to identify appropriate values of the tuning constant of M-estimators for ordered response models and to assess their robustness when data contamination occurs.

8.1. Computational details

The first step in the estimation process is the computation of the norm $\|\boldsymbol{x}_i\|$ of the covariates through a robust estimation of location and scale. In Model 3, since there is only one regressor, the norm $\|x_i\|$ is computed as $|x_i - Med(X)|/MAD(X)$, where $Med(X)$ and $MAD(X)$ are the median and the normalized median absolute deviation of X. The same procedure is used in Model 6, where the norm is a function of the continuous variable only. In the other two models (4 and 5), the norm of the regressors is given by the Mahalanobis distance $\|\boldsymbol{x}_i\| = \left\{ (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X)' \hat{\boldsymbol{\Sigma}}_X^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_X) \right\}^{1/2}$, after estimating location and scale of $\boldsymbol{X}$ through the Stahel-Dohono procedure (Stahel, 1981; Donoho, 1982; see also Maronna and Yohai, 1995). Given a direction $\boldsymbol{a} \in \mathbb{R}^p$ with $\|\boldsymbol{a}\| = 1$, denote by $\boldsymbol{a}'\boldsymbol{x}_i$ the projection of $\boldsymbol{x}_i$ along $\boldsymbol{a}$. The outlyingness of $\boldsymbol{x}_i$ with respect to the sample $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n)$ along $\boldsymbol{a}$ is defined by

$$t(\boldsymbol{x}_i, \boldsymbol{a}) = \frac{\boldsymbol{a}'\boldsymbol{x}_i - Med(\boldsymbol{a}'\boldsymbol{X})}{MAD(\boldsymbol{a}'\boldsymbol{X})}.$$

The outlyingness of $\boldsymbol{x}_i$ is given by $t(\boldsymbol{x}_i) = \max_a t(\boldsymbol{x}_i, \boldsymbol{a})$. The robust estimators of location and scatter are given by

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{X}} = \frac{1}{\displaystyle\sum_{i=1}^{n} W^X\{t(\boldsymbol{x}_i)\}} \sum_{i=1}^{n} \boldsymbol{x}_i W^X\{t(\boldsymbol{x}_i)\}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \frac{1}{\displaystyle\sum_{i=1}^{n} W^X\{t(\boldsymbol{x}_i)\}} \sum_{i=1}^{n} (x_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}})(x_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}})' W^X\{t(\boldsymbol{x}_i)\}$$

where $W^X(t)$ is a non-increasing function of t such as the Huber weights used by Maronna and Yohai (1995), that is $W^X(t) = \min\left\{1, (c^X/t)^2\right\}$ with tuning constant $c_X = \sqrt{\chi^2_{p,0.95}}$, where $\chi^2_{p,0.95}$ is the 95-th percentile of a χ^2_p distribution.

Then, for a given c, the solution of (10) is obtained by the Newton-Raphson algorithm.

Since

$$a(\boldsymbol{\theta}) = E_X\left[E_Y\left\{s(Y, \boldsymbol{X}; \boldsymbol{\theta})w(Y, \boldsymbol{X}; \boldsymbol{\theta})|\boldsymbol{X}\right\}\right],$$

where

$$E_Y\left\{s(Y, \boldsymbol{X}; \boldsymbol{\theta})w(Y, \boldsymbol{X}; \boldsymbol{\theta})|\boldsymbol{X} = \boldsymbol{x}_i\right\} = \sum_{r=1}^{m} \frac{\partial Pr(Y = r|\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} w(r, \boldsymbol{x}_i, \boldsymbol{\theta}), \quad (17)$$

at each iteration the value of $a(\boldsymbol{\theta})$ can be approximated as follows. For every $\boldsymbol{x}_i$ and the current value of the estimate $\hat{\boldsymbol{\theta}}_k$ at the k-th iteration, we approximate (17) by

$$\sum_{r=1}^{m} \frac{\partial Pr(Y = r|\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_k} w(r, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_k) \tag{18}$$

and then compute $a(\boldsymbol{\theta})$ by taking the sample average of (18) with respect to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

One important issue, which at the moment is still unsolved, is the choice of a good starting point for the Newton-Raphson algorithm. In the numerical experiments, the MLEs are used with satisfactory results. Indeed, the difficulties in finding alternative starting points provide a sound motivation for using the Huber weights, which correspond to a monotone ψ function. Alternative weights may be derived by considering redescending ψ functions, but it is well known that they would lead to multiple solutions of (10).

In order to evaluate the loss of efficiency produced by M-estimation at the model or the gain in efficiency achieved when data contamination occurs, simulated versions of the efficiency criteria are computed. Initially, the Mean Square

Error (MSE) matrix of the estimators is obtained from simulation; its generic element is

$$MSE_{ij} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_{ib} - \theta_i)(\hat{\theta}_{Jb} - \theta_j)$$

where B is the number of samples drawn, $\hat{\theta}_{ib}$ is the estimate obtained from the b-th sample and θ_i is the i-th parameter. Then the Trace criterion is computed as the ratio between the trace of the MSE matrix of the MLEs and that of the M-estimators,

$$\text{eff}_{Trace}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \frac{tr(MSE^{ML})}{tr(MSE^M)}.$$

The Minimum MSE ratio is obtained as the minimum among the ratios of the MSEs of the MLEs and those of the M-estimators for each parameter,

$$\text{eff}_{Min(MSE-ratio)}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \min_i \left\{ \frac{MSE_{ii}^{ML}}{MSE_{ii}^M} \right\}.$$

The Determinant criterion is obtained as the ratio between the determinant of the MSE matrices of the ML and M-estimators,

$$\text{eff}_{Det}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \frac{det(MSE^{ML})}{det(MSE^M)}.$$

Finally the Maximum MSE ratio is computed as

$$\text{eff}_{Max(MSE-ratio)}(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_{ML}; c) = \max_i \left\{ \frac{MSE_{ii}^{ML}}{MSE_{ii}^M} \right\}.$$

Briefly, in any criterion, the asymptotic variance-covariance matrices are replaced by the simulated MSE matrices.

8.2. Tuning constant and loss of efficiency at the model

In Subection 5.1, the minimum values of the tuning constant c such that the loss of efficiency at the model due to M-estimation does not exceed 5% are illustrated when the sample size is $n = 200$. Here some more extensive results are provided by considering also a smaller sample size $n = 100$ and by identifyng values of c suitable to keep the loss of efficiency below a larger threshold i.e. 10%. All the results are based on 1·000 simulated samples.

The minimum values of c such that the loss of efficiency of M-estimators is not larger than 5% or 10%, for $n = 100$ and $n = 200$, are shown in Tables 10, 11 and 12 for the probit link, the logistic one, and the logistic link with weights (15), respectively.

TABLE 10
Values of the tuning constant required to achieve a given efficiency of the M-estimator under no contamination in the data when the probit link is adopted.

	Trace		Min(MSE-ratio)		Determinant	
Model	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
Efficiency loss smaller than 5%						
1	1.2	1.1	1.3	1.2	1.7	1.6
2	1.6	1.2	1.9	1.4	2.4	2.0
3	1.3	1.1	2.2	1.7	2.3	2.2
4	1.8	1.7	2.3	2.3	2.6	3.1
5	1.1	1.7	1.3	2.1	1.8	2.7
6	1.5	1.1	1.9	1.4	2.2	2.0
Efficiency loss smaller than 10%						
1	0.7	0.7	0.9	0.8	1.5	1.4
2	1.0	0.7	1.1	0.8	2.0	1.7
3	0.8	0.7	1.6	1.2	1.8	1.7
4	1.3	1.2	1.7	1.7	2.3	2.5
5	0.9	1.2	1.0	1.6	1.7	2.4
6	1.1	0.9	1.4	1.1	1.9	1.7

TABLE 11
Values of the tuning constant required to achieve a given efficiency of the M-estimator under no contamination in the data when the logistic link is adopted.

	Trace		Min(MSE-ratio)		Determinant	
Model	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
Efficiency loss smaller than 5%						
3	0.7	0.6	1.1	0.8	1.3	0.9
4	1.0	0.9	1.3	1.0	1.5	1.4
5	1.2	0.9	1.4	1.1	2.0	1.4
6	0.8	0.7	1.0	0.8	1.3	1.1
Efficiency loss smaller than 10%						
3	< 0.5	< 0.5	0.8	0.6	1.0	0.7
4	0.7	0.6	0.9	0.7	1.3	1.2
5	0.9	0.7	1.0	0.8	1.7	1.3
6	0.6	0.5	0.6	0.7	1.0	0.9

Some remarks follow.

- By comparing the values of c when the sustainable loss of efficiency is 5% or 10%, we notice that the larger is the loss of efficiency, the smaller can be the value of c. Actually when c decreases the distance between the estimating function of the MLEs and that of the M-estimators increases, so that a smaller values of c implies a lower efficiency of the latter estimators at the model.
- For a given c, the loss of efficiency at the model measured by the Trace criterion is smaller than that measured by the Minimum MSE ratio, which in turn is smaller than that measured by the Determinant criterion. On the whole, the value of c does not seem to be largely affected by the nature (dichotomous or continuous) and the number of regressors.
- The values of c needed to keep the loss of efficiency under a given thresh-

TABLE 12

Values of the tuning constant required to achieve a given efficiency of the M- estimators under no contamination in the data when the logistic link with weights (15).

	Trace		Min(MSE-ratio)		Determinant	
Model	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
	Efficiency loss smaller than 5%					
3	1.0	0.8	1.2	1.1	1.6	1.2
4	1.2	1.1	1.4	1.4	1.8	1.9
5	1.3	1.0	1.5	1.3	2.2	1.9
6	< 0.5	0.8	0.6	0.9	< 0.5	0.8
	Efficiency loss smaller than 10%					
3	0.7	0.6	1.0	0.8	1.3	1.1
4	0.9	0.8	1.0	1.0	1.6	1.6
5	0.9	0.7	1.0	0.9	2.0	1.7
6	< 0.5	0.7	0.5	0.8	< 0.5	0.8

old are always smaller for the logistic link than for the probit link, since the generalized residuals produced by the former link tend to be smaller (having a limited range) and therefore yield a smaller argument for the Huber weights. As a consequence the convergence of the efficiency of the M-estimators to 1 (when c increases) is faster when the logistic link is adopted, as is also made evident by Figure 6.

- When the probit link is adopted (Table 10) and the Trace criterion is considered, c needs to be between 1.1 and 1.8 to ensure a loss of efficiency below 5% and between 0.7 and 1.3 to ensure a loss below 10%. When the Minimum MSE ratio is considered, suitable values of c are between 1.2 and 2.3 when the maximum admissible loss of efficiency is 5% and between 0.9 and 1.7 when it is 10%. The Determinant criterion can require $c \geq 3.1$ for a loss below 5% and $c \geq 2.5$ for a loss below 10%. In short, when the target is a 5% loss at the model, c needs to be greater than 1, but a value of $c \geq 3$ might be too conservative.

- When the logistic link is adopted (Table 11) and the evaluation of the M-estimators is carried out through the Trace criterion, c needs to be between 0.6 and 1.2 if the maximum admissible loss of efficiency is 5% and smaller than 0.9 when it is 10%. The Minimum MSE criterion requires $0.8 \leq c \leq 1.4$ to keep the loss below 5% and $0.6 \leq c \leq 1$ to keep the loss below 10%. The Determinant criterion for the small sample size $n = 100$ can require also $c = 2$ to guarantee that the loss does not exceed 5% and $c = 1.7$ in order not to exceed 10%. $c = 2$ is the largest value of the tuning constant which ensures an efficiency loss not greater than 5% whatever criterion is used, but according to the first two criteria c can be even smaller, say roughly $c < 1.5$.

- When estimation with the logistic link is carried out with weights (15) (Table 12), the value of the tuning is generally slightly higher than what required by weights (14). As already pointed out in the paper, these results provide evidence in favour of weights (14) also for the logistic link despite the limited range of the generalized residuals. In order to reduce the loss of

efficiency, two statistical units, such that $\|\boldsymbol{x}_i\|$ takes the same value, should be treated differently according to the size of the associated generalized residuals.

8.3. Tuning constant and robustness

In order to investigate the gain in efficiency achieved by M-estimation in a neighborhood of the model, three types of deviations are taken into account: gross errors in the response variable, outliers in continuous covariates and model misspecification. The following examples provide the results of numerical experiment based on $1 \dot{.} 000$ samples of size $n = 200$.

CONTAMINATED MODEL 2. Here we consider the case, in Model 2, when 2 observations of the response Y, which should correspond to the fourth category, are erroneously recorded as belonging to the first one. The amount of contamination in the data due to gross errors is 1%.

Table 13 shows the values of the various criteria when M-estimation is performed with the probit link and c varies between 1 and 3.5. The gain in efficiency measured by the Trace criterion is remarkable and varies between 17.8% for a very large value of the tuning constant ($c = 3.5$) and over 100% when $c \leq 1.25$. The gain measured by the Determinant criterion varies between 2.9% when $c = 3.5$ and 42.9% when $c = 1.5$. According to the Minimum MSE criterion the minimum gain in efficiency varies between 2.4% and 7.5%. However the maximum gain in efficiency which can be obtained in the estimation of a single parameter varies between 21.8% and over 160%. The behaviour of the Minimum MSE ratio and that of the Determinant criterion are not monotone, indicating that c should not be too small (downweighting should not be excessive with respect to the amount of contamination). On the whole, suitable values of c appear those between 1 and 1.5, which are also consistent with a loss of efficiency smaller than 5% at the model according to the first two criteria.

TABLE 13
Efficiency criteria for M-estimation performed in Model 2 with the probit link when data are contaminated by 1% of gross errors in Y.

c	1	1.25	1.5	1.75	2	2.5	3	3.5
Trace	2.027	2.002	1.965	1.898	1.802	1.572	1.347	1.178
Determinant	1.269	1.402	1.429	1.389	1.318	1.190	1.079	1.029
Min(MSE-ratio)	1.050	1.065	1.075	1.072	1.068	1.067	1.045	1.024
Max(MSE-ratio)	2.606	2.538	2.438	2.303	2.139	1.767	1.442	1.218

CONTAMINATED MODEL 2 (*Shelter Effect*). This kind of contamination has already been considered in the paper, and deals with the case when - due to a shelter effect (Iannario, 2012) - five Y_i, which originally take values 1, 2 or 3, are changed into 4. For sake of completeness, in addition to the Minimum and Maximum MSE ratio criterion, here we report (in Table 14) also the value of

the Trace and Determinant Criterion. Whatever criterion is used, the gain in efficiency produced by M-estimation is clearly evident, especially for $1 \leq c \leq 1.5$.

TABLE 14
Efficiency criteria when M-estimation is performed in Model 2 with the probit link and a shelter effect occurs.

c	1	1.25	1.5	1.75	2	2.5	3	3.5
Trace	2.247	2.156	2.043	1.898	1.734	1.437	1.224	1.087
Determinant	1.572	1.585	1.505	1.390	1.255	1.114	1.036	0.999
Min(MSE-ratio)	1.354	1.376	1.373	1.342	1.299	1.201	1.124	1.064
Max(MSE-ratio)	2.556	2.398	2.227	2.029	1.819	1.485	1.253	1.099

CONTAMINATED MODEL 3. The contamination of Model 3 lies in the value of the regressor which is erroneously recorded as 5, for κ statistical units, where $\kappa = 1, \ldots, 5$. In the paper the efficiency of the M-estimator of the regression coefficient is investigated when $c = 1.5$. The efficiency of the M-estimators for a larger set of values of c is evaluated in Tables 15 and 16 for the probit and the logistic link respectively. Recall that, to keep the loss of efficiency at the model below 5%, c needs to be between 1.1 (for Trace criterion) and 2.2 (for the Determinant criterion) when the probit link is adopted, and between 0.6 and 0.9 when the logistic link is adopted.

Some remarks follow.

- M-estimation produces a gain in efficiency even for $c \geq 2.5$ when the probit link is adopted and for $c > 1$ when the logistic link is adopted, though such values are much higher than those necessary to keep the loss of efficiency below 5% at the model.
- According to the Trace and the Determinant criterion, M-estimation largely outperforms ML estimation whatever the amount of contamination, and the gain in efficiency increases remarkably with the number of outliers.
- Occasionaly M-estimation can produce a small loss of efficiency in the estimation of the cutpoints, which is however largely offset by the gain achieved in the estimation of the other parameters and especially of the regression coefficient. This is why the global efficiency measures (Trace and the Determinant criterion) provide outstanding evidence in favour of M-estimators.
- Although M-estimation is definitely more efficient when both the probit and the logistic link are adopted, in the former case the gain in efficiency is remarkably higher. This result is due to the intrinsic robustness of the logistic link, which - unlike the probit one - produces generalized residuals with a limited range. The unboundness of the generalized residuals of the probit link can generate huge loss of efficiency in ML estimation when data contamination occurs.

TABLE 15
Trace criterion, MSE ratio for each parameter and Determinant criterion in Model 3 with the probit link when there are κ outliers equal to 5 in the explanatory variable.

			MSE-ratio				
c	Trace	$\alpha_1 = -1.7$	$\alpha_2 = -0.5$	$\alpha_3 = 0.5$	$\alpha_4 = 1.7$	$\beta = 1.5$	Determinant
				$\kappa = 1$			
1.00	2.577	2.365	1.304	0.803	1.723	6.472	2.740
1.50	2.679	2.440	1.334	0.827	1.787	6.963	3.223
2.00	2.734	2.483	1.355	0.837	1.822	7.238	3.526
2.50	2.767	2.508	1.367	0.844	1.848	7.371	3.724
3.00	2.788	2.525	1.375	0.851	1.868	7.403	3.838
3.50	2.798	2.540	1.381	0.856	1.880	7.358	3.898
4.00	2.802	2.554	1.387	0.861	1.886	7.249	3.919
				$\kappa = 2$			
1.00	5.194	4.667	1.923	0.778	2.979	15.147	3.998
1.50	5.394	4.837	1.980	0.804	3.100	15.912	4.693
2.00	5.476	4.923	2.018	0.817	3.160	15.928	5.035
2.50	5.492	4.956	2.039	0.828	3.198	15.499	5.162
3.00	5.463	4.956	2.051	0.837	3.217	14.769	5.134
				$\kappa = 3$			
1.00	7.800	7.130	2.657	0.765	4.185	23.071	4.496
1.50	8.017	7.344	2.734	0.795	4.341	23.146	5.134
2.00	8.016	7.387	2.782	0.811	4.402	21.955	5.309
2.50	7.876	7.310	2.799	0.825	4.411	20.093	5.204
				$\kappa = 4$			
1.00	9.991	9.250	3.308	0.759	5.225	28.263	4.529
1.50	10.087	9.390	3.394	0.792	5.393	26.697	4.946
2.00	9.832	9.236	3.431	0.811	5.403	23.725	4.869
2.50	9.371	8.875	3.427	0.829	5.329	20.428	4.558
				$\kappa = 5$			
1.00	11.951	11.114	3.968	0.784	6.181	31.617	4.376
1.50	11.777	11.047	4.042	0.821	6.339	28.012	4.587
2.00	11.136	10.568	4.056	0.841	6.261	23.490	4.332
2.50	10.255	9.838	4.007	0.859	6.054	19.240	3.913

- The optimal value of the tuning constant seems to depend on the amount of contamination; when κ is small c can be as large as 3, nevertheless when the number of outliers increases low values of c are to be preferred.

In regard to the last point it is to be noticed that the behaviour of the efficiency measures, as in previous examples, is not monotone with respect to c. However some light on this issue is provided by Figure 10 which shows the Trace and the Maximum MSE ratio criterion when c varies for any value of κ. The plots make evident that when the amount of contamination in the data is mild, different values of c yield small changes in efficiency, whereas when the amount of contamination is large, the efficiency quickly decreases when c increases. Consequently c should be chosen as small as possible, say $c \leq 2$ for the probit link and $c \leq 1.5$ for the logistic one.

TABLE 16
*Trace criterion, MSE ratio for each parameter and Determinant criterion in Model 3 with
the logistic link when there are κ outliers equal to 5 in the explanatory variable.*

			MSE-ratio				
c	Trace	$\alpha_1 = -2.1$	$\alpha_2 = -0.6$	$\alpha_3 = 0.6$	$\alpha_4 = 2.1$	$\beta = 1.5$	Determinant
				$\kappa = 1$			
1.00	1.022	0.993	0.998	0.959	0.966	1.262	1.094
1.50	1.030	1.008	0.999	0.962	0.971	1.283	1.139
2.00	1.034	1.016	0.999	0.968	0.976	1.280	1.157
2.50	1.036	1.019	1.001	0.974	0.981	1.259	1.160
3.00	1.034	1.021	1.002	0.979	0.986	1.222	1.146
				$\kappa = 2$			
1.00	1.225	1.143	1.022	0.922	0.998	2.273	1.774
1.50	1.228	1.159	1.026	0.930	1.004	2.212	1.796
2.00	1.218	1.161	1.027	0.941	1.008	2.063	1.731
				$\kappa = 3$			
1.00	1.589	1.449	1.130	0.897	1.038	3.795	2.669
1.50	1.566	1.451	1.130	0.910	1.048	3.401	2.529
2.00	1.513	1.422	1.123	0.925	1.053	2.884	2.253
				$\kappa = 4$			
1.00	2.055	1.875	1.233	0.866	1.135	5.330	3.306
1.50	1.973	1.829	1.232	0.884	1.140	4.377	2.931
2.00	1.836	1.728	1.217	0.904	1.136	3.407	2.447
				$\kappa = 5$			
1.00	2.526	2.287	1.394	0.848	1.205	6.470	3.629
1.50	2.352	2.176	1.375	0.870	1.208	4.930	3.042
2.00	2.107	1.986	1.338	0.894	1.197	3.618	2.442

CONTAMINATED MODEL 3 (*continued*). The paper considers a further contamination of Model 3, produced by a gross error which replaces the value of x_i for one statistical unit by an outlier, whose value varies between 3 and 10. Figure 11 shows the various efficiency criteria when the value of the outlier varies (notice that the maximum gain in efficincy is achieved in the estimation of the regression coefficient). All criteria consistently point out an increasing gain in efficiency as the outlier moves away from the bulk of data.

CONTAMINATED MODEL 4. The contamination of Model 4 (already considered in the paper) is given by two statistical units whose regressors are jointly mis-reported as 5. Figure 12 shows the efficiency of M-estimators measured by the four criteria. A loss of efficiency can be incurred in the estimation of the cutpoints (this is why the Minimum MSE criterion is below 1), which is nevertheless largely offset by the gain in efficiency obtainable in the estimation of the regression coefficients. Values of c between 1.5 and 2 seem to be optimal to achieve a robust estimation according to the Trace and the Maximum MSE criterion for both links, and also for the Determinant criterion in the case of the logistic link.
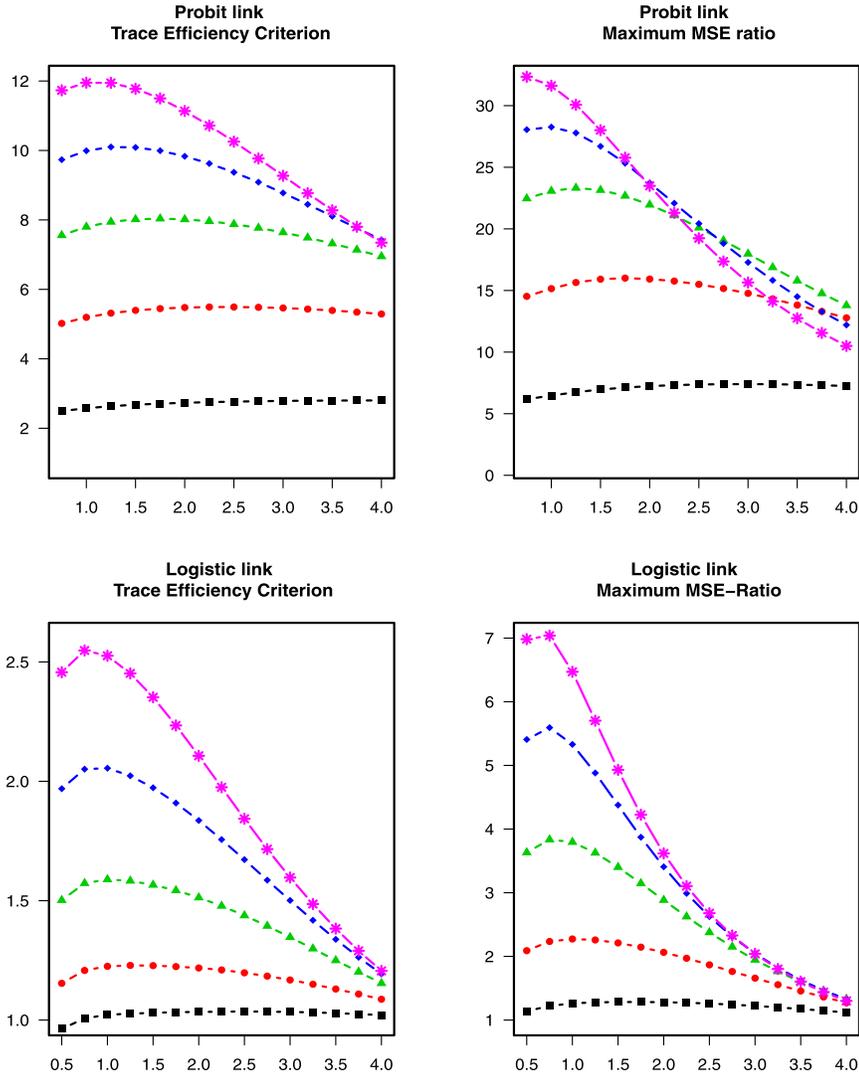
FIG 10. *Efficiency Criteria in Model 3 when c varies for* $\kappa = 1$ *(squares)*, $\kappa = 2$ *(circles)*, $\kappa = 3$ *(triangles)*, $\kappa = 4$ *(diamonds) and* $\kappa = 5$ *(stars)*.

CONTAMINATED MODEL 5. To investigate the robustness properties of M-estimation in Model 5, the paper considers both a misspecification of the model, obtained by inserting two unnecessary regressors in the estimation process, and a mild contamination produced by changing the value of X_1 into 5 for κ statistical units. Table 17 shows how the Trace and the Determinant criterion vary with κ when $c = 1.5$ for both the probit and the logistic link. The gain in efficiency provided by M-estimation is clearly evident.
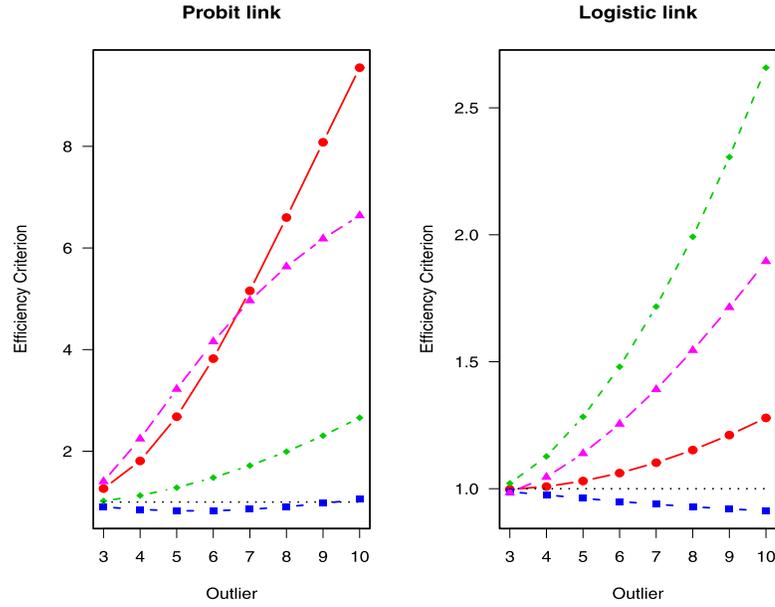
FIG 11. *Trace Criterion (circles), Minimum MSE-ratio criterion (squares), Maximum MSE ratio criterion (diamonds) and Determinant criterion (triangles) in Model 3 contaminated by an outlier with value between 3 and 10.*
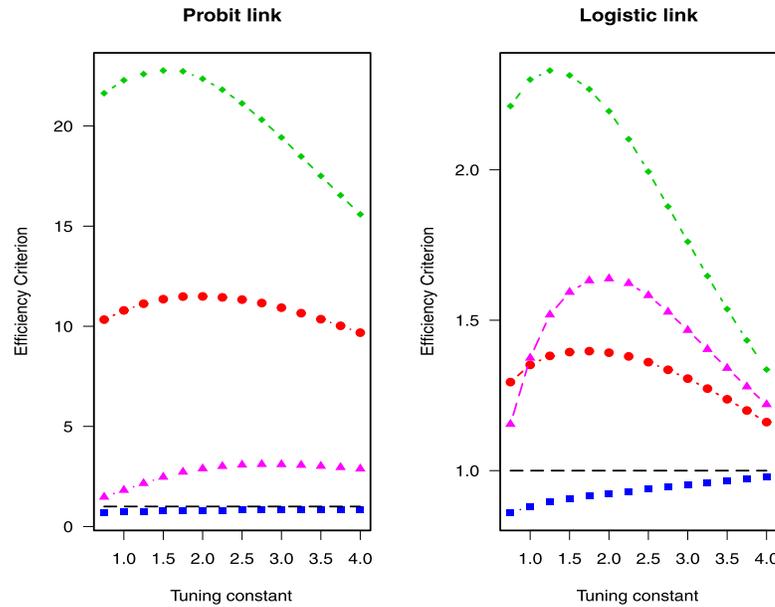

FIG 12. *Trace Criterion (circles), Minimum MSE-ratio criterion (squares), Maximum MSE ratio criterion (diamonds) and Determinant criterion (triangles) in Model 4 under 1% contamination of the data ($x_1 = x_2 = 5$ for two statistical units).*

Trace and Determinant criterion when Model 5 is misspecified by adding two unnecessary regressors, and there are κ anomalous values in X_1 ($c = 1.5$).

	Trace				
κ	1	2	3	4	5
Probit link	2.051	3.688	5.156	5.311	4.830
Logistic link	1.129	1.754	2.398	2.776	2.888
	Determinant				
Probit link	0.993	1.256	1.301	1.335	1.237
Logistic link	1.294	2.094	2.665	2.603	2.453

In addition Figure 13 shows the MSEs of the M-estimators versus those of the MLEs when κ varies. As the amount of contamination increases, the MLEs lose progressively more efficiency, whereas the MSEs of the M-estimators keep pretty stable.
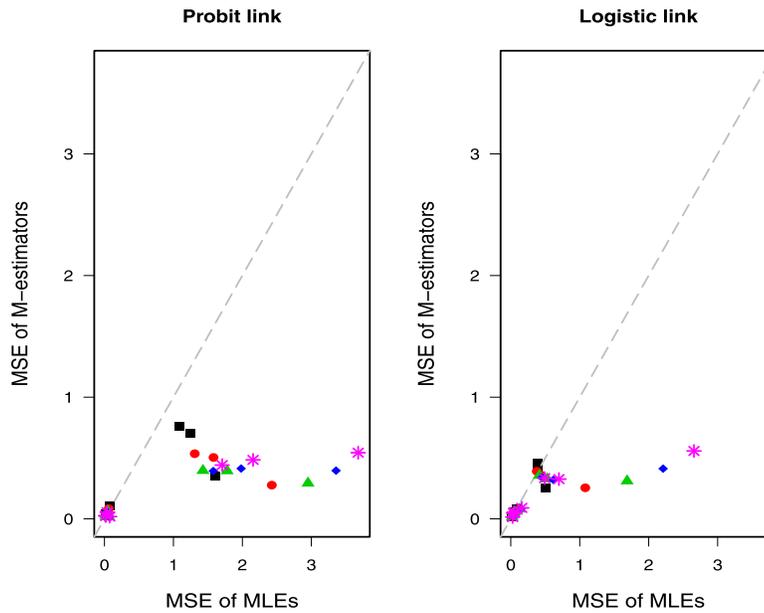

FIG 13. *MSEs of MLEs and M-estimators in Model 5, misspecified by adding two unnecessary regressors, and there are $\kappa = 1$ (squares), $\kappa = 2$ (bullet), $\kappa = 3$ (triangles), $\kappa = 4$ (diamonds) and $\kappa = 5$ (stars) anomalous values in X_1 ($c = 1.5$).*

References

Agresti A. (2010). *Analysis of Ordinal Categorical Data*, 2nd edition, J.Wiley & Sons, Hoboken. MR2742515

Albert J.H., Chib S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669–679. MR1224394

Albert J.H., Chib S. (1995). Bayesian residual analysis for binary response regression, *Biometrika*, **82**, 747–748. MR1380812

Albert J.H., Chib S. (2001). Sequential ordinal modeling with applications to survival data, *Biometrics*, **57**, 829–836. MR1863450

Atkinson A., Riani M. (2000). *Robust Diagnostic Regression Analysis*, Springer, New York. MR1884997

Carroll R. J., Pederson S. (1993). On robustness in the logistic regression model, *Journal of the Royal Statistical Society, Series B*, **55**, 693–706. MR1223937

Chib S. (2005). Modeling and analysis for categorical response data, in: D. Dey and C.R. Rao eds., *Handbook of Statistics*, **25**, 835–867. MR2490549

Croux C., Haesbroeck G., Ruwet C. (2013). Robust estimation for ordinal regression, *Journal of Statistical Planning and Inference*, **143**, 1486–1499. MR3070253

Donoho D. L. (1982). *Breakdown properties of multivariate location estimators*, Ph.D. Qualifying paper, Harvard University.

Franses P.H., Paap R. (2004). *Quantitative Models in Marketing Research*, Cambridge University Press, Cambridge.

Hampel F.R. (1968). *Contribution to the Theory of Robust Estimation*, Ph.D Thesis, University of California, Berkeley. MR2617979

Hampel F. R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**, 383–393. MR0362657

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, J. Wiley & Sons, New York. MR0829458

Heritier S., Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models, *Journal of the American Statistical Association*, **89**, 897–904. MR1294733

Huber P.J. (1981). *Robust Statistics*, J. Wiley & Sons, New York. MR0606374

Huber P.J., Ronchetti E. (2009). *Robust Statistics*, 2nd ed., J. Wiley & Sons, New York. MR2488795

Iannario M. (2012). Modelling *shelter* choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, **21**, 1–22. MR2881627

Iannario M., Piccolo D. (2016). A comprehensive framework of regression models for ordinal data, *METRON*, **74**, 233–252. MR3534537

Iannario M., Monti A. C., Piccolo D. (2016). Robustness issues for CUB models, *TEST*, **25**, 731–750. MR3554413

Maronna R.A., Martin R.D., Yohai V. J. (2006). *Robust Statistics: Theory and Methods*, J. Wiley & Sons, New York. MR2238141

Maronna R. A., Yohai V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, **90**, 330–341. MR1325140

McCullagh P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B*, **42**, 109–142. MR0583347

McCullagh P., Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition, Chapman and Hall, London. MR3223057

Moustaki I., Victoria-Feser M. P. (2006). Bounded-influence robust estimation in generalized linear latent variable models, *Journal of the American Statistical Association*, **101**, 644–653. MR2256179

Nelder J.A., Wedderburn R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A*, **135**, 370–384. MR0375592

Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, **5**, 85–104.

Ruckstuhl A.F., Welsh A.H. (2001), Robust fitting of the binomial model, *The Annals of Statistics*, **29**, 1117–1136. MR1869243

Stahel W. A. (1981). *Breakdown of covariance estimators*, Research Report 31, Fachgruppe für Statistik, ETH, Zurich.

Tutz G (2012). *Regression for Categorical Data*, Cambridge University Press, Cambridge. MR2856629

Victoria-Feser M.P., Ronchetti E. (1997). Robust estimation for grouped data, *Journal of the American Statistical Association*, **92**, 333–340. MR1436121