A New Robust Approach for Multinomial Logistic Regression With Complex Design Model

Elena Castilla and Pedro J. Chocano

Abstract—Robust estimators and Wald-type tests are developed for the multinomial logistic regression based on $\phi$-divergence measures. We compute the influence function of the proposed estimators and tests and discuss some consequences. Their robustness is illustrated by an extensive simulation study and two real examples.

Index Terms—Divergence measures, influence function, multinomial logistic regression model, robustness, Wald-type tests.

I. INTRODUCTION

MULTINOMIAL logistic regression model, also known as polytomous logistic regression model, is widely used in health and life sciences for analyzing nominal qualitative response variables and their relationship with respect to their corresponding explanatory variables or covariates. Reference [1] used hierarchical multinomial logistic regression models to examine how the rates of cardiac procedures depend on patient-level characteristics, including age, gender and race. Reference [2] also used multinomial logistic regression to detect uncommon risk factors related to oral cavity, larynx and lung cancers. Recently, [3] proposed a risk prediction model using semi-varying coefficient multinomial logistic regression to assess correct prediction rates when classifying the patients with early rheumatoid arthritis. Further examples of application of these methods can be found in [4], [5] and [6], among others. Although most of classical literature deals with the cases of simple random sampling scheme, the application of multinomial logistic regression model under complex survey setting (with stratification, clustering or unequal selection probabilities) can be found, for example, in [7]–[9] and [10].

Most of the results mentioned above are based on (pseudo) maximum likelihood estimators (PMLEs), which are well-known to be efficient, but also non-robust. Therefore, testing procedures based on MLEs may face serious robustness problems. Reference [11] developed density power divergence (DPD) based robust estimators (MDPDEs) and Wald-type tests for multinomial logistic regression model under simple random sampling. This approach was extended to complex design in [12]. In [13], pseudo minimum $\phi$-divergence estimators (PM$\phi$Es) as well as new estimators for the intra-cluster correlation coefficient were developed. Estimators in the Cressie Read subfamily with tuning parameter $\lambda > 0$ were shown to be an efficient alternative to classical PMLE ($\lambda = 0$) for small samples sizes.

It is well known that Cressie Read divergences with negative tuning parameter are generally robust against outliers, particularly, the Hellinger distance ($\lambda = -0.5$). Nevertheless, [13] did not consider estimators in the Cressie Read subfamily with negative tuning parameter and the robustness was not analyzed. Moreover, the theory developed in [13] is focused on estimation and does not consider hypothesis testing. In this paper we define PM$\phi$Es based Wald-type tests for testing composite null hypothesis studying their asymptotic distribution as well as their power. Regarding the robustness issue, the influence functions of the proposed estimators and Wald-type tests are studied and their explicit formulas are obtained. Finally, we empirically prove the robustness of PM$\phi$Es with $\lambda < 0$ through an extensive simulation study and two real data samples. In this paper we also provide the algorithms that are needed in order to compute some overdispersed multinomial distributions in the context of multinomial logistic regression models with complex design.

The organization of the paper is as follows. In Section II we present the multinomial logistic regression model and the framework necessary to define the PM$\phi$Es. Based on their asymptotic distribution, Wald-type tests are developed in Section III. In Section IV the study of the influence function of the proposed estimators and test statistics is detailed. An extensive simulation study and two numerical examples illustrate the robustness of the proposed estimators and Wald-type tests in Section V and Section VI, respectively. Finally, in the Appendix we present some extensions of the Monte Carlo simulation study.

II. MULTINOMIAL LOGISTIC REGRESSION MODEL WITH COMPLEX DESIGN

We consider a population $\Omega$ partitioned into $H$ strata. The data consist of $n_h$ clusters in stratum $h$. In the $i$-th cluster ($i = 1, \ldots, n_h$) within the $h$-th stratum ($h = 1, \ldots, H$) we have observed for the $j$-th unit ($j = 1, \ldots, m_{hi}$) the values of a categorical response variable $Y$ with $d + 1$ categories. Note
that we assume there are $H$ strata, $n_h$ clusters in stratum $h$ and $m_{hi}$ units in cluster $i$ of stratum $h$. The observed responses of the $(d+1)$-dimensional variable $Y$ are denoted by the $(d+1)$-dimensional classification vector.

$$y_{hij} = (y_{hij1}, \ldots, y_{hij,d+1})^T,$$

$h = 1, \ldots, H; i = 1, \ldots, n_h; j = 1, \ldots, m_{hi}$, with $y_{hijr} = 1$ if the $j$-th unit selected from the $i$-th cluster of the $h$-th stratum falls in the $r$-th category and $y_{hijr} = 0$ for $l \neq r$. It is very common when working with dummy or qualitative explanatory variables to consider that the $k + 1$ explanatory variables are common for all the individuals in the $i$-th cluster of the $h$-th stratum, being denoted as $x_{hi} = (x_{h0i}, x_{hi1}, \ldots, x_{hik})^T$, with the first one, $x_{h0i} = 1$, associated with the intercept.

Let us denote the sampling weight from the $i$-th cluster of the $h$-th stratum by $w_{hi}$. For each $i$, $h$ and $j$ the expectation of the $r$-th element of the random variable $Y_{hij} = (Y_{hij1}, \ldots, Y_{hij,d+1})^T$, corresponding to the realization $y_{hij}$, is determined by

$$\pi_{hir}(\beta) = E[Y_{hijr}|x_{hi}] = \Pr (Y_{hijr} = 1|x_{hi}) = \frac{\exp \{x_{hi}^T \beta_r\}}{1 + \sum_{l=1}^d \exp \{x_{hi}^T \beta_l\}}, \quad (1)$$

where $\beta_r = (\beta_{r0}, \beta_{r1}, \ldots, \beta_{rk})^T \in \mathbb{R}^{k+1}$, $r = 1, \ldots, d$ and the associated parameter space is given by $\Theta = \mathbb{R}^{d(k+1)}$. It is clear that

$$\pi_{hi,d+1}(\beta) = \frac{1}{1 + \sum_{l=1}^d \exp \{x_{hi}^T \beta_l\}}. \quad (2)$$

Let us assume that the expectation of $Y_{hij}$ does not depend on the unit number $j$. In what follows, we denote this property by "homogeneity". Note that this is not a strong assumption as we generally have random sampling with the clusters in each stratum (see [9], [12]). From now on, we denote by

$$Y_{hi} = \sum_{j=1}^{m_{hi}} Y_{hij} = \left(\sum_{j=1}^{m_{hi}} Y_{hij1}, \ldots, \sum_{j=1}^{m_{hi}} Y_{hij,d+1}\right)^T = (Y_{hi1}, \ldots, Y_{hi,d+1})^T$$

the random vector of counts in the $i$-th cluster of the $h$-th stratum and by $\pi_{hi}(\beta)$ the $(d + 1)$-dimensional probability vector with the elements given in (1), $\pi_{hi}(\beta) = (\pi_{h1i}(\beta), \ldots, \pi_{hi,d+1}(\beta))^T$.

Using this notation, the empirical and the theoretical probability vectors of the model are defined by

$$\hat{P} = \frac{1}{T}(w_{11}\tilde{y}_{11}^T, \ldots, w_{1n_1}\tilde{y}_{1n_1}^T, \ldots, w_{H1}\tilde{y}_{H1}^T, \ldots, w_{Hn_h}\tilde{y}_{Hn_h}^T)^T, \quad (3)$$

$$\pi(\beta) = \frac{1}{T}(w_{11}m_{11}\pi_{11}(\beta), \ldots, w_{1n_1}m_{1n_1}\pi_{1n_1}(\beta), \ldots, w_{Hn_h}m_{Hn_h}\pi_{Hn_h}(\beta))^T, \quad (4)$$

respectively, where $\tau = \sum_{h=1}^H \sum_{i=1}^{n_h} w_{hi}m_{hi}$. Probability vectors (3) and (4), both of dimension $(d + 1)\sum_{h=1}^H n_h$, are the key to constructing the definition of PMLEs.

**Definition 2.1:** Under homogeneity assumption within the clusters and taking into account the weights $w_{hi}$, the (weighted) pseudo-maximum likelihood estimator (PMLE) of $\beta$, denoted by $\hat{\beta}_P$, is obtained by maximizing

$$\mathcal{L}_P(\beta) = \sum_{h=1}^H \sum_{i=1}^{n_h} w_{hi} \log \pi_{hi}^T(\beta) \hat{y}_{hi}, \quad (5)$$

where $\log \pi_{hi}^T(\beta) = (\log \pi_{h1i}(\beta), \ldots, \log \pi_{hi,d+1}(\beta))$.

The pseudo-likelihood method (also denoted by quasi-likelihood in the literature) was originally defined by [14] for modeling data without a complete distributional knowledge. In the multinomial logistic regression model with complex survey, the distribution of $Y_{hi}$ might be unknown as their components jointly might be correlated, and the pseudo-likelihood considers an approximation of the likelihood under homogeneity assumption. See [9] and [15] for more details about the pseudo-likelihood.

**Remark 2.2:** If we consider that $Y_{hi}$ has a multinomial sampling scheme, which means that $Y_{hi,j}$, $j = 1, \ldots, m_{hi}$ are independent random variables with covariance matrix

$$\Sigma_{hi} = m_{hi}\Delta(\pi_{hi}(\beta)), \quad (6)$$

and $\Delta(\pi_{hi}(\beta)) = \text{diag}(\pi_{hi}(\beta)) - \pi_{hi}(\beta)\pi_{hi}^T(\beta)$, the term “pseudo” should be dropped because (5) is not an approximation. Here $\text{diag}(\pi_{hi}(\beta))$ denotes the matrix with the entries of $\pi_{hi}(\beta)$ along the diagonal. A weaker assumption is to consider that $Y_{hi}$ has a multinomial sampling scheme with an overdispersion parameter $\nu_{hi} = 1 + \rho_{hi}^2(m_{hi} - 1)$ and

$$\Sigma_{hi} = \nu_{hi}m_{hi}\Delta(\pi_{hi}(\beta)), \quad \text{but the distribution of } Y_{hi} \text{ is not in principle used for the estimators. Distributions such as Dirichlet Multinomial, Random Clumped and m-inflated belong to this family (see [10], [16] and [17] for details). In Appendix A-B we present the algorithms needed to compute these distributions in the context of multinomial logistic regression model with complex design.}

The PMLE can be obtained solving the following system of equations: $u(\beta) = 0_{d(k+1)}$, where $0_{d(k+1)}$ is the null vector of dimension $d(k+1)$,

$$u(\beta) = \sum_{h=1}^H \sum_{i=1}^{n_h} u_{hi}(\beta) \quad \text{and} \quad u_{hi}(\beta) = w_{hi}\pi_{hi}^T(\beta) \otimes x_{hi}. \quad (6)$$

Here $\otimes$ denotes the Kronecker product and $\pi_{hi}^T(\beta) = \tilde{y}_{hi}^* - m_{hi}\pi_{hi}(\beta)$, where the superscript * means the vector obtained by deleting the last component from the initial vector, i.e., $\tilde{y}_{hi}^* = (\tilde{y}_{h1i1}, \ldots, \tilde{y}_{hid})^T$ and $\pi_{hi}(\beta) = (\pi_{h1i}(\beta), \ldots, \pi_{hid}(\beta))^T$.

**Example 2.3 (Education in Malawi):** The 2010 Malawi Demographic and Health Survey (2010 MDHS, [18]) was implemented by the National Statistical Office (NSO) from June through November 2010, with a nationally representative sample of more than 27,000 households. The sample for the 2010 MDHS was designed to provide population and health indicator estimation at the national, regional, and district levels. Let us focus on Tables 2.3.1 and 2.3.2 of the
cited study, that present data on educational attainment for female and male household members age 6 and older, divided in five wealth quintile levels, which are considered as strata. We consider here a response variable with \(d+1 = 5\) categories: “no education”, “some primary”, “completed primary”, “some secondary”, “completed secondary or more”. For simplicity, the missing observations are not taken into account. Figure 1 shows the estimated probabilities by the PMLE of each one of the response categories for each gender. As expected, the proportion of women who have never attended any formal schooling is greater than the proportion of men. Moreover, the proportion of the population that has attained education declines with its level. In the ensuing work, we will present alternative estimators to the PMLE, which are seen to provide better performance in terms of robustness.

### A. PM\(\phi\)Es: Definition, Estimation and Asymptotic Distribution

**Definition 2.4:** Given the probability vectors (3) and (4), the family of phi-divergence measures between these probability vectors is given by

\[
d_\phi\left(\hat{\phi}, \pi(\beta)\right) = \frac{1}{\tau} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} m_{hi} \sum_{s=1}^{d+1} \pi_{his}(\beta) \phi\left(\frac{\hat{\pi}_{his}}{m_{hi} \pi_{his}(\beta)}\right),
\]

where \(\phi \in \Phi^* \) and \(\Phi^*\) denotes the class of all convex functions \(\phi : [0, \infty) \to \mathbb{R} \cup \{\infty\}\) such that \(\phi(1) = 0\), \(\phi''(1) > 0\) and we define \(0\phi(0/0) = 0\) and \(0\phi(p/0) = 0\) lim \(u/\phi(u)/u\).

Note that if \(\phi(x) = x \log x - x + 1\) in (7), then we have the so-called Kullback-Leibler divergence. For more details about phi-divergence measures see [19].

Let us now consider the PMLE in (5). It can be shown that the PMLE is related to the Kullback-Leibler divergence measure as follows:

\[
d_{KL}\left(\hat{\phi}, \pi(\beta)\right) = K - \frac{1}{\tau} L(\beta),
\]

with \(K\) being a constant that does not depend on \(\beta\) (see [13]). Hence, the maximization of \(L(\beta)\) is equivalent to the minimization of \(d_{KL}\left(\hat{\phi}, \pi(\beta)\right)\), i.e., the PMLE is the one which minimizes the Kullback-Leibler divergence between the empirical and theoretical probability vectors (3 and 4). Thus,

\[
\hat{\beta}_p = \arg \min_{\beta \in \Theta} d_{KL}\left(\hat{\phi}, \pi(\beta)\right).
\]

The definition of PM\(\phi\)E arises from the idea of generalizing definition (8) to other phi-divergence measures.

**Definition 2.5:** We consider the multinomial logistic regression model with complex survey defined in (1). The PM\(\phi\)E of \(\beta\) is defined by

\[
\hat{\beta}_{\phi,p} = \arg \min_{\beta \in \Theta} d_\phi\left(\hat{\phi}, \pi(\beta)\right).
\]

We shall now provide the equations to obtain the PM\(\phi\)Es. By equation (7) it is obvious that the PM\(\phi\)E of \(\beta\) is obtained by solving the following system of equations: \(u_{\phi}(\beta) = 0_{d(k+1)}\), where

\[
u_{\phi}(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} u_{\phi,hi}(\beta),
\]

with

\[
u_{\phi,hi}(\beta) = w_{hi} m_{hi} \frac{\partial \pi^T_{hi}(\beta)}{\partial \beta} f_{\phi,hi}(\hat{\gamma}_{m_{hi}}, \beta),
\]

\[
\frac{\partial \pi^T_{hi}(\beta)}{\partial \beta} = (I_{d \times d}, 0_{d \times 1}) \Delta(\pi_{hi}(\beta)) \otimes x_{hi},
\]

and

\[
\begin{align*}
\Delta(\pi_{hi}(\beta)) &= \left(f_{\phi,hi}(\hat{\gamma}_{m_{hi}}, \beta), \ldots, f_{\phi,hi(d+1)}(\hat{\gamma}_{m_{hi(d+1)}}, \beta)\right)^T, \\
f_{\phi,hi}(x, \beta) &= \frac{1}{\phi''(1)} \pi_{his}(\beta)^\phi(x) - \frac{1}{\phi''(1)} \phi(x) \pi_{his}(\beta).
\end{align*}
\]

Theorem 2.6 establishes the asymptotic distribution of the PM\(\phi\)Es, which will be the basis of the definition of the family of Wald-type tests in Section III.

**Theorem 2.6:** Let \(\hat{\beta}_{\phi,p}\) be the PM\(\phi\)E of parameter \(\beta\) for a multinomial logistic regression model with complex survey, let \(n\) be the total of clusters in all the strata of the sample and let \(\eta^*_h\) be an unknown proportion given by \(\lim_{n \to \infty} \frac{m_{hi}}{n} = \eta^*_h\), where \(h = 1, \ldots, H\). Then

\[\sqrt{n}(\hat{\beta}_{\phi,p} - \beta^0) \xrightarrow{L} N\left(0_{d(k+1)}, \mathbf{J}^{-1}(\beta^0) \mathbf{G}(\beta^0) \mathbf{J}^{-1}(\beta^0)\right),\]

where \(\beta^0\) is the true parameter value and

\[
\begin{align*}
\mathbf{J}(\beta) &= \lim_{n \to \infty} \mathbf{J}_n(\beta) = \frac{1}{n} \sum_{h=1}^{H} \eta^*_h \lim_{n \to \infty} \mathbf{J}_n^{(h)}(\beta), \\
\mathbf{G}(\beta) &= \lim_{n \to \infty} \mathbf{G}_n(\beta) = \frac{1}{n} \sum_{h=1}^{H} \eta^*_h \lim_{n \to \infty} \mathbf{G}_n^{(h)}(\beta),
\end{align*}
\]

with

\[
\begin{align*}
\mathbf{J}_n(\beta) &= \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} m_{hi} \Delta(\pi^*_m(\beta)) \otimes x_{hi} x_{hi}^T, \\
\mathbf{J}_n^{(h)}(\beta) &= \frac{1}{n} \sum_{i=1}^{n_h} w_{hi} m_{hi} \Delta(\pi^*_m(\beta)) \otimes x_{hi} x_{hi}^T, \\
\mathbf{G}_n(\beta) &= \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \mathbf{V} \left[U_{hi}(\beta)\right], \\
\mathbf{G}_n^{(h)}(\beta) &= \frac{1}{n} \sum_{i=1}^{n_h} \mathbf{V} \left[U_{hi}(\beta)\right], \\
\mathbf{V}[U_{hi}(\beta)] &= \frac{w^2_{hi}}{n} \mathbf{V}[\hat{\gamma}_{m_{hi}}] \otimes x_{hi} x_{hi}^T.
\end{align*}
\]

\(\mathbf{J}(\beta)\) is the Fisher information matrix. \(\mathbf{V}[\cdot]\) denotes the variance-covariance matrix of a random vector and \(U_{hi}(\beta)\) is the random variable generator of \(u_{hi}(\beta)\), given by (6).

**Proof:** For the convenience of the reader we give the relevant ideas from the proof of this theorem, which can be found in [13]. Following the same steps of the linearization method of [7], \(\mathbf{G}(\beta)\) and \(\mathbf{J}(\beta)\) are given by

\[
\begin{align*}
\mathbf{G}(\beta) &= \lim_{n \to \infty} \mathbf{V}\left[\frac{1}{\sqrt{n}} \mathbf{U}(\beta)\right], \\
\mathbf{J}(\beta) &= -\lim_{n \to \infty} \mathbf{n} \frac{\partial U^T_\phi(\beta)}{\partial \beta},
\end{align*}
\]
where $U_\phi(\beta)$ is the random variable generator of $u_\phi(\beta)$ given by (9). These are computed using a first Taylor expansion of $f_{\phi,h_1}(\theta_{m_{h_1}^i}, \beta)$ given in (10). Firstly,
\begin{equation}
    f_{\phi,h_1}(\theta_{m_{h_1}^i}, \beta) = \frac{\theta_{m_{h_1}^i} - \pi_{h_1}(\beta)}{\pi_{h_1}(\beta)} + o\left( \frac{\theta_{m_{h_1}^i} - \pi_{h_1}(\beta)}{\pi_{h_1}(\beta)} \right).
\end{equation}

Expanding it to the vectorial form
\begin{equation}
    f_{\phi,h_1}(\theta_{m_{h_1}^i}, \beta) = \text{diag}^{-1}(\pi_{h_1}(\beta))\left( \frac{\theta_{m_{h_1}^i} - \pi_{h_1}(\beta)}{\pi_{h_1}(\beta)} \right) + o\left( \frac{\theta_{m_{h_1}^i} - \pi_{h_1}(\beta)}{\pi_{h_1}(\beta)} \right) .
\end{equation}

The result follows substituting (13) in (9) and applying the Central Limit Theorem. Note that the effect of the $\phi$-divergence considered is lost when applying a first order approximation.

While the estimator $\hat{\beta}_{\phi,P}$ depends on the $\phi$ function, the asymptotic distribution does not. This is an important difference with respect to the (pseudo) MDPDE, whose asymptotic distribution depends on the corresponding DPD considered in the estimation. See Remark 2.9 for more details.

Remark 2.7: Matrices $J(\beta^\prime)$ and $G(\beta^\prime)$ of Theorem 2.6 can be consistently estimated as
\begin{equation}
    \hat{J}_n(\hat{\beta}_{\phi,P}) = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{h_{i}} m_{h_{i}} \Delta(\pi^*_{h_{i}}(\hat{\beta}_{\phi,P})) \otimes x_{h_{i}} x_{h_{i}}^T
\end{equation}
\begin{equation}
    \hat{G}_n(\hat{\beta}_{\phi,P}) = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left( u_{h_{i}}(\hat{\beta}_{\phi,P}) - \frac{1}{n} u(\hat{\beta}_{\phi,P}) \right) \otimes \left( u_{h_{i}}(\hat{\beta}_{\phi,P}) - \frac{1}{n} u(\hat{\beta}_{\phi,P}) \right)^T.
\end{equation}

An important family of phi-divergence measures is obtained by restricting $\phi$ from the family of convex functions $\Phi^*$ to the Cressie-Read subfamily, that is, $\phi$ is of the following form:
\begin{equation}
    \phi_\lambda(x) = \begin{cases}
        \frac{1}{\lambda(1+\lambda)} x^{\lambda+1} - x - \lambda(x-1), & \lambda \in \mathbb{R} \setminus \{-1,0\} \\
        \lim_{\lambda \to 0} \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda}, & \lambda \in \{-1,0\}.
    \end{cases}
\end{equation}

For the Cressie-Read subfamily it is established that for $\lambda \neq -1$
\begin{equation}
    u_{\phi_\lambda}(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} u_{\phi_\lambda,h_1}(\beta),
\end{equation}
where
\begin{equation}
    u_{\phi_\lambda,h_1}(\beta) = \frac{w_{h_{i}}}{(\lambda + 1) m_{h_{i}}} \Delta^*(\pi_{h_1}(\beta)) \otimes x_{h_{i}} x_{h_{i}}^T
\end{equation}
\begin{equation}
    \Delta^*(\pi_{h_1}(\beta)) = (I_d, 0_d) \Delta(\pi_{h_1}(\beta)).
\end{equation}

We can observe that for $\lambda = 0$ we have
\begin{equation}
    \phi_{\lambda=0}(x) = \lim_{\nu \to 0} \frac{1}{\nu(1+\nu)} [x^{\nu+1} - x - \nu(x-1)]
    = x \log x - x + 1
\end{equation}
and the associated phi-divergence coincides with the Kullback-Leibler divergence. Therefore, the PMLE of $\beta$ based on $\phi_{\lambda}(x)$ contains as special case the PMLE and $u_{\phi_1}(\beta)$ given in (6) matches with $u_{\phi_1,h_1}(\beta)$ given in (14). Other important divergences are obtained inside this family. For instance, for the values $\lambda = 1$, $\lambda = 2/3$ and $\lambda = -0.5$ it is obtained the chi-square divergence, the Cressie-Read divergence, and the Hellinger distance, respectively. Note that the Hellinger distance is well-known in statistical theory for its robustness ([20]). The study of the Influence Function of (Cressie-Read) PMLEs is presented in Section IV.

Remark 2.8: Throughout this paper we are referring to the case of complex sample survey. These all procedures can be easily simplified to the case of simple sample survey considering a single stratum and “observations” instead of clusters. Some work has been done within the phi-divergence measures and multinomial logistic regression (see [21], [22]) but, to the best of our knowledge, the robustness issue was not previously considered. In Section VI-B we provide an example to illustrate the application of the proposed methods also in this context.

Remark 2.9: The DPD, originally introduced in [23] and exhaustively studied in the monograph by [24], has been extensively studied in the last years for its balance between efficiency and robustness. Let us consider the multinomial logistic regression model with complex survey; following [11]
and [12], the (pseudo) MDPDE of $\beta$, denoted by $\hat{\beta}_{\lambda^*,p}$, is defined by
\[
\hat{\beta}_{\lambda^*,p} = \arg \min_{\beta} d_{\lambda^*}(\hat{p}, \pi(\beta)),
\]
where
\[
d_{\lambda^*}(\hat{p}, \pi(\beta)) = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} \times \left\{ \frac{1}{2} \sum_{i=1}^{n_h} m_{hi,\lambda^*}(\beta) \left( m_{hi,\lambda^*}(\beta) - 1 + \frac{\lambda^*}{\hat{g}_{ii}(\beta)} \right) \right\},
\]
for $\lambda^* > 0$, while for $\lambda^* = 0$ we have the PMLE. Its asymptotic distribution is given by
\[
\sqrt{n}(\hat{\beta}_{\lambda^*,p} - \beta^0) \xrightarrow{n \to \infty} N(0, \Omega_{\lambda^*,p}(\beta^0) \Omega_{\lambda^*,p}^{-1}(\beta^0)),
\]
where $\beta^0$ is the true parameter value and
\[
\Omega_{\lambda^*,p}(\beta) = \lim_{n \to \infty} \Omega_{n,\lambda^*,p}(\beta) = \lim_{n \to \infty} \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \Omega_{hi,\lambda^*,p}(\beta),
\]
\[
\Psi_{\lambda^*,p}(\beta) = \lim_{n \to \infty} \Psi_{n,\lambda^*,p}(\beta) = \lim_{n \to \infty} \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \Psi_{hi,\lambda^*,p}(\beta),
\]
with
\[
\Omega_{hi,\lambda^*,p}(\beta) = w_{hi}^2 \Delta^2(\pi_{hi}(\beta)) \text{diag} \lambda^*-1(\pi_{hi}(\beta)) \text{Var}[\hat{Y}_{hi}]
\times \text{diag} \lambda^*-1(\pi_{hi}(\beta)) \Delta^T(\pi_{hi}(\beta)) \otimes x_{hi} x_{hi}^T,
\]
\[
\Psi_{hi,\lambda^*,p}(\beta) = w_{hi} n^2 \Delta^2(\pi_{hi}(\beta)) \times \text{diag} \lambda^*-1(\pi_{hi}(\beta)) \Delta^T(\pi_{hi}(\beta)) \otimes x_{hi} x_{hi}^T.
\]

### III. WALD-TYPE TESTS

In the last years it has been very common in the statistical literature to consider Wald-type tests based on the minimum distance estimators instead of the MLE. The resulting tests have an excellent behaviour in relation to the robustness with a non-significant loss of efficiency, see for instance [25], [26].

In this section we introduce Wald-type test statistics based on the PMoEs as a generalization of the classical Wald test based on PMLE.

In this context, we are interested in testing
\[
H_0 : M^T \beta = m \quad \text{against} \quad M^T \beta \neq m,
\]
where $M$ is a $d(k+1) \times r$ full rank matrix with $r \leq d(k+1)$ and $m$ an $r$-vector.

**Definition 3.1:** Let $\hat{\beta}_{\phi,p}$ the PMoE of $\beta$ and denote
\[
\hat{V}_n(\hat{\beta}_{\phi,p}) = \frac{1}{n} \left( \hat{\beta}_{\phi,p} - \beta \right) G_n(\hat{\beta}_{\phi,p}) \left( \hat{\beta}_{\phi,p} - \beta \right) G_n^T(\hat{\beta}_{\phi,p}).
\]

Then the family of Wald-type test statistics for testing the null hypothesis given in (15) is defined as follows:
\[
W_n(\hat{\beta}_{\phi,p}) = n \left( M^T \hat{\beta}_{\phi,p} - m \right)^T M^T \hat{V}_n(\hat{\beta}_{\phi,p}) M \left( M^T \hat{\beta}_{\phi,p} - m \right),
\]
\[
\text{then it holds}
\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0)) \xrightarrow{n \to \infty} N(0, \sigma^2_W(\beta^0)),
\]
where
\[
\sigma^2_W(\beta^0) = 4 \left( M^T \beta^0 - m \right)^T M^T V(\beta^0) M^{-1} (M^T \beta^0 - m).
\]

**Theorem 3.2:** The asymptotic distribution of the Wald-type test statistics, $W_n(\hat{\beta}_{\phi,p})$, under the null hypothesis in (15), is a chi-square distribution with $r$ degrees of freedom $\chi^2_r$.

**Proof:** We have from (11) that
\[
\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0)) \xrightarrow{n \to \infty} N(0, V(\beta^0)),
\]
with $V(\beta^0) = J^{-1}(\beta^0) G(\beta^0) J^{-1}(\beta^0)$. Therefore,
\[
\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0)) \xrightarrow{n \to \infty} N(0, M^T V(\beta^0) M).
\]

Since rank($M$) = $r$, it follows that
\[
\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0)) \xrightarrow{n \to \infty} N(0, M^T V(\beta^0) M).
\]

Therefore, under $H_0$, we have that $W_n(\hat{\beta}_{\phi,p})$ defined in (16) converges in law to a chi-square distribution with $r$ degrees of freedom.

**Corollary 3.3:** Based on Theorem 3.2, the null hypothesis in (15) will be rejected if
\[
W_n(\hat{\beta}_{\phi,p}) > \chi^2_r \gamma_{1-a}
\]
being $\chi^2_r \gamma_{1-a}$ the upper $a$-th quantile of $\chi^2_r$.

The following theorem may be used to approximate the power function for the Wald-type test statistics given in (17).

**Theorem 3.4:** Let $\beta^0$ be the true value of the parameter and let us denote
\[
\ell^*(\hat{\beta}_{\phi,p},\beta^0) = \left( M^T \beta_1 - m \right)^T M^T V(\beta_2) M^{-1} (M^T \beta_1 - m).
\]

Then it holds
\[
\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0)) \xrightarrow{n \to \infty} N(0, \sigma^2_W(\beta^0)),
\]
where
\[
\sigma^2_W(\beta^0) = 4 \left( M^T \beta^0 - m \right)^T M^T V(\beta^0) M^{-1} (M^T \beta^0 - m).
\]

**Proof:** A first order Taylor expansion of $\ell^*(\hat{\beta}_{\phi,p},\beta^0)$ at $\hat{\beta}_{\phi,p}$ around $\beta^0$ gives
\[
\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) = \ell^*(\beta^0,\beta^0) + \frac{\partial \ell^*}{\partial \beta^T} \bigg|_{\beta=\beta^0} (\hat{\beta}_{\phi,p} - \beta^0) + o_p \left( \left\| \hat{\beta}_{\phi,p} - \beta^0 \right\| \right).
\]

The asymptotic distribution of $\sqrt{n}(\ell^*(\hat{\beta}_{\phi,p},\beta^0,\beta^0) - \ell^*(\beta^0,\beta^0))$ coincides with the asymptotic distribution of
\[
\sqrt{n} \left( \frac{\partial \ell^*}{\partial \beta^T} \bigg|_{\beta=\beta^0} (\hat{\beta}_{\phi,p} - \beta^0) \right),
\]
but
\[
\frac{\partial \ell^*}{\partial \beta^T} \bigg|_{\beta=\beta^0} = 2 (M^T \beta^0 - m)^T (M^T V(\beta^0) M)^{-1} M^T.
\]

From here, the result follows easily.
Theorem 3.5: Let $\beta^0 \in \Theta$ be the true value of the parameter such that $M^T \beta^0 \neq m$ and $\tilde{\beta}_{\phi,p} \xrightarrow{p \to} \beta^0$. The power function of the Wald-type test given in (17), is given by

$$\Pi(W_n(\beta_{\phi,p}) (\beta^0)) = 1 - \Phi_n \left( \frac{1}{\sigma_W (\beta^0)} \left( \chi_{r,\alpha}^2 - \sqrt{n} (\epsilon^p (\beta_{\phi,p}, \beta^0)) \right) \right),$$

where $\Phi_n (x)$ uniformly tends to the standard normal distribution as $n \to \infty$.

Proof:

$$\Pi(W_n(\beta_{\phi,p}) (\beta^0)) \geq P_{\beta^0} (W_n(\beta_{\phi,p}) > \chi_{r,\alpha}^2)$$

$$= P_{\beta^0} (n^{1/2} (\beta_{\phi,p} - \beta^0) > \chi_{r,\alpha}^2)$$

$$= P_{\beta^0} \left( \sqrt{n} \left( \epsilon^p (\beta_{\phi,p}, \beta^0) - \epsilon^p (0, \beta^0) \right) \right)$$

$$> \chi_{r,\alpha}^2 \sqrt{n} - \sqrt{n} (\epsilon^p (0, \beta^0))$$

$$= 1 - P_{\beta^0} \left( \frac{1}{\sigma_W (\beta^0)} \left( \chi_{r,\alpha}^2 n - \sqrt{n} (\epsilon^p (0, \beta^0)) \right) \right),$$

$$1 - \Phi_n \left( \frac{1}{\sigma_W (\beta^0)} \left( \chi_{r,\alpha}^2 n - \sqrt{n} (\epsilon^p (0, \beta^0)) \right) \right),$$

where $\Phi_n (x)$ uniformly tends to the standard normal distribution $\phi (x)$ as $n \to \infty$.

Corollary 3.6: It is clear that

$$\lim_{n \to \infty} \Pi(W_n(\beta_{\phi,p}) (\beta^0)) = 1$$

for all $\alpha \in (0, 1)$. Therefore, the Wald-type tests are consistent in the sense of [27].

Remark 3.7: Theorem 3.5 can be applied to obtain the necessary sample size so that the Wald-type tests have a determinate fix power, i.e., $\Pi(W_n(\beta_{\phi,p}) (\beta^0)) \equiv \pi^0$. It is given by

$$n = \left[ A + B + \sqrt{A(A + 2B)} \right] + 1,$$

where $[x]$ denotes the largest integer less than or equal to $x$, $A = \sigma_W^2 (\beta^0) (\Phi^- (1 - \pi^0))^2$ and $B = 2 \epsilon^p (\beta^0, \beta^0) \chi_{r,\alpha}^2$.

We may also find approximations of the power function of the Wald-type tests given in (16) at an alternative hypothesis close to the null hypothesis. Let $\beta_n \in \Theta - \Theta_0$ be a given alternative, and let $\beta_0 \in \Theta_0$ (null hypothesis) be the element closest to $\beta_n$ in terms of the Euclidean distance. We may introduce contiguous alternative hypotheses by considering a fixed $d \in \mathbb{R}^{(k+1)}$ and to permit $\beta_n$ moving towards $\beta_0$ as $n$ increases through the relation

$$H_{1,n} : \beta_n = \beta_0 + n^{-1/2} d.$$

Let us now relax the condition $M^T \beta^0 = m$ that defines the null hypothesis. Let $\delta \in \mathbb{R}^r$ and consider the following sequence, $\beta_n$, of parameters moving towards $\beta_0$ according to

$$H_{1,n} : M^T \beta_n = m - n^{-1/2} \delta.$$

Theorem 3.8: The asymptotic distribution of $W_n(\beta_{\phi,p})$ is given by:

(a) Under $H_{1,n}$, $W_n(\beta_{\phi,p}) \xrightarrow{n \to \infty} \chi_r^2 (\Delta)$, where $\Delta$ is the parameter of non-centrality given by

$$\Delta = d^T M \left[ M^T V(\beta_0) M \right]^{-1} d.$$

Proof: It is easy to see that

$$M^T \beta_{\phi,p} = m - n^{-1/2} M^T d + M^T \tilde{\beta}_{\phi,p} - \beta_n.$$

We know, under $H_{1,n}$, that $\sqrt{n}(\tilde{\beta}_{\phi,p} - \beta_n) \xrightarrow{n \to \infty} N(0, \Sigma)$ and $\beta_n \xrightarrow{n \to \infty} \beta_0$. Therefore, since $M^T \beta_0 = m$,

$$\sqrt{n}(M^T \tilde{\beta}_{\phi,p} - m) \xrightarrow{n \to \infty} N \left( \left( M^T d, M^T V(\beta_0) M \right) \right).$$

But it is known that if $Z \in N(\mu, \Sigma)$, $\Sigma$ is a symmetric projection of rank $k$ and $\Sigma \mu \neq 0$, then $Z^T Z$ is a chi-square distribution with $k$ degrees of freedom and non-centrality parameter $\mu^T \mu$. So considering the quadratic form

$$W_n(\beta_{\phi,p}) = Z^T Z,$$

with $Z = \sqrt{n} \left[ M^T V(\beta_{\phi,p})^{-1} M \right]^{-1/2} (M^T \tilde{\beta}_{\phi,p} - m)$

$$\xrightarrow{n \to \infty} N \left( \left[ M^T V(\beta_0) M \right]^{-1/2} M^T d, I_{r \times r} \right),$$

the result is immediate with the non-centrality parameter being

$$d^T M \left[ M^T V(\beta_0) M \right]^{-1} M d.$$}

The point (b) is straightforward taking into account that the equivalence between the hypotheses (19) and (20) is given by $M^T d = \delta$.

IV. INFLUENCE FUNCTION ANALYSIS

An important concept in robustness theory is the influence function (IF) defined by

$$IF(t, U, F) = \lim_{\varepsilon \to 0} \frac{U(F_\varepsilon) - U(F)}{\varepsilon} = \frac{\partial U(F)}{\partial \varepsilon} \bigg|_{\varepsilon = 0^+},$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon \Delta$, with $\varepsilon$ being the contamination proportion and $\Delta$ being the degenerate distribution at the contamination point $t$. Thus, the (first-order) IF, as a function of
estimators. Here, we extend their definition for the case of non-homogeneous set-up. In fact, the observations within a cluster of a stratum are iid but the observations in different cluster and stratum are independent non-homogeneous. So we need to modify the definition of the IF accordingly. Recently, [29]–[31] have discussed the extended definition of the IF for the independent but non-homogeneous observations; however, all these works have been developed for DPD-based estimators. Here, we extend their definition for the case of multinomial logistic regression and PMDEs. Note that, in the cited paper of [13], the IF of the proposed estimators was not computed. The goal of this section is to compute the IF of the proposed estimators and Wald-type tests and discuss its implication in the robustness of the proposed statistics.

We first need to define the statistical functional $U_\phi(\beta)$ corresponding to the PMDEs as the minimizer of the phi-divergence between the true and the model densities. This is defined as the minimizer of

$$H_\phi(\beta) = \frac{1}{T} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{s=1}^{d+1} \pi_{his}(\beta) \phi \left( \frac{g_{his}}{\pi_{his}(\beta)} \right).$$

Under appropriate differentiability conditions, it is given by the solution of the estimating equations

$$\frac{\partial H_\phi(\beta)}{\partial \beta} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} m_{hi} \frac{\partial \pi_{hi}^T(\beta)}{\partial \beta} f_{\phi,hi}(g_{hi},\beta) = 0.$$

Particularly, for the Cressie-Read subfamily

$$\frac{\partial H_{\lambda}(\beta)}{\partial \beta} = \frac{1}{T} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \lambda + 1 \frac{\partial \pi_{hi}^T(\beta)}{\partial \beta} \text{diag}^{-1}(\lambda + 1)(\pi_{hi}(\beta)) g_{hi}^{\lambda + 1} = 0.$$

For simplicity, let us first assume that the contamination is only in one cluster probability $g_{h_{0}i_{0}}$ for some $h_{0}$ and $i_{0}$. Consider the contaminated probability vector

$$g_{h_{0}i_{0},e} = (1 - e) \pi_{h_{0}i_{0}}(\beta^0) + e \delta_t,$$

where $\delta_t$ is the contamination proportion and $\delta_t$ is the degenerate probability at the outlier point $t = (t_1,\ldots,t_{d+1}) \in \{0,1\}^{d+1}$ with $s_{d+1} t_s = 1$ and

$$g_{hi} = \begin{cases} \pi_{hi}(\beta^0) & \text{if } (i,h) \neq (i_0,h_0), \\ g_{h_{0}i_{0},e} & \text{if } (i,h) = (i_0,h_0). \end{cases}$$

Let $g_{e}$ denote the corresponding contaminated full probability vector, which is the same as $g$ except $g_{h_{0}i_{0}}$ being replaced by $g_{h_{0}i_{0},e}$. Let $G_e$ be the corresponding contaminated distribution vector. We replace $\beta$ in (22) by $\beta_{e}^0 = U_\phi(G_e)$. Then, we have Equation (23), shown at the bottom of the next page.

Now, we are going to get the derivative of (23) with respect to $\varepsilon$, which is as in Equation (24), shown at the bottom of the next page.

Now, evaluating in $\varepsilon = 0$ and simplifying, we have

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} m_{hi} \frac{\partial \pi_{hi}^T(\beta)}{\partial \beta} \text{diag}^{-1}(\lambda + 1)(\pi_{hi}(\beta^0)) \frac{\partial \pi_{hi}^T(\beta)}{\partial \beta} |_{\beta = \beta^0} \times \pi_{hi}^{\lambda + 1}(\beta^0) I(\theta_{h_0i_0},U_\phi(\beta),F)$$

$$- \left\{ \sum_{h=1}^{H} \sum_{i=1}^{n_h} \omega_{hi} m_{hi} \frac{\partial \pi_{hi}^T(\beta)}{\partial \beta} \pi_{hi}^{\lambda + 1}(\beta^0) \text{diag}^{-1}(\lambda + 1)(\pi_{hi}(\beta^0)) \right\} = 0.$$

From this, Theorem 4.1 follows straightforward.

**Theorem 4.1:** Let us consider the multinomial logistic regression model under complex design given in (1). The IF of the PMDEs with respect to the $i_0$ cluster in the $h_0$ stratum is given by

$$IF(t_{h_0i_0},U_\phi(\beta),F) = \Psi_{\beta}^{-1}(\beta) u_{\beta_{h_0i_0}}^*(\beta),$$

where $u_{\beta_{h_0i_0}}^*(\beta)$ and $\Psi_{\beta}(\beta)$ are given in Equations (26) and (27), shown at the bottom of the next page, respectively.

Let us now study the IF of proposed Wald-type tests statistics defined in Section III. In our context, the functional associated with the Wald-type test evaluated at $U_\beta(G)$ is given by

$$W_n(U_\beta(G)) = n \left( M^T U_\beta(G) - m \right)^T \left( M^T V_n(U_\beta(G)) M \right)^{-1} \times \left( M^T U_\beta(G) - m \right).$$

The IF of general Wald-type tests under such non-homogeneous set-up has been extensively studied in [26] for the case of DPD estimators. Here, we can follow that the first-order IF of $W_n$, defined as the first order derivative of its value at the contaminated distribution with respect to $\varepsilon$ at $\varepsilon = 0$, becomes null at the null distribution. Therefore, the first order IF is not informative in this case of Wald-type tests, and we need to investigate the second-order IF, denoted by $IF^{(2)}$, of $W_n$. Through some computations we obtain the following proposition.

**Theorem 4.2:** Let us consider the multinomial logistic regression model under complex design given in (1). The second-order IF of the functional associated with the Wald-type tests with respect to the $i_0$ cluster in the $h_0$ stratum is given by

$$IF^{(2)}(t_{h_0i_0},W_n,F) = 2IF^{(2)}(t_{h_0i_0},U_\phi(\beta),F) M \left( M^T V_n(U_\beta(G)) M \right)^{-1} \times M^T IF(t_{h_0i_0},U_\phi(\beta),F),$$

where $IF(t_{h_0i_0},U_\phi(\beta),F)$ is the IF of the PMDEs given in (25).

Note that similar results for the IF of both the PMDEs and Wald-type test statistics are given in the case there is contamination in some of the clusters within some stratum.
In [20] it was established that the estimators based on the minimum $\phi$-divergences, including the class of Cressie-Read divergences, have the same unbounded IF. Namely, it was established that the IF was the inverse of the Fisher information matrix multiplied by the (contaminated) score vector. However, these results were developed in the context of a iid sample. In our setting, it was necessary to compute the IF in a complex design structure. Result (25) is in concordance with Proposition 1 in [20]. Moreover, the IF is identical to the IF of the PMLE. This result suggests that the MP$\phi$Es are asymptotically fully efficient at the model, so the method provides an efficient estimator of the model parameters when the model is true. On the other hand, it also indicates that the IF of the MP$\phi$Es is potentially unbounded. If we study the IF of the Wald-type tests statistics, then the same conclusion is obtained because its second-order IF (28) is a quadratic function of the corresponding IF of the PM$\phi$Es. However, several authors ([20], [32], [33]) have shown the limitation of the IF approach for this case since some minimum $\phi$-divergences do exhibit strong robustness properties. For instance, Cressie-Read divergences with negative tuning parameter $\lambda$, with the well known Hellinger distance as a particular case ($\lambda = -0.5$).

A similar idea as in Theorem 11, where it was established that all the MP$\phi$Es have the same asymptotic variance-

\begin{equation}
\frac{\partial H_{\phi_0}(\beta)}{\partial \beta} \bigg|_{\beta = \beta_{0,0}^0} = -\lambda + 1 \sum_{h=1}^{H} \sum_{i=1}^{n} \left\{ \frac{\omega_{hi} m_{hi}}{\lambda + 1} \frac{\partial \pi_{hi}(\beta)}{\partial \beta} \bigg|_{\beta = \beta_{0,0}^0} \text{diag}^{-1}(\pi_{hi}(\beta_{0,0}^0)) \right\} -\lambda + 2 \sum_{h=1}^{H} \sum_{i=1}^{n} \left\{ \frac{\omega_{hi} m_{hi}}{\lambda + 1} \frac{\partial^2 \pi_{hi}(\beta)}{\partial \beta^2} \bigg|_{\beta = \beta_{0,0}^0} \text{diag}^{-1}(\pi_{hi}(\beta_{0,0}^0)) \right\} + \lambda + 1 \sum_{h=1}^{H} \sum_{i=1}^{n} \left\{ \frac{\omega_{hi} m_{hi}}{\lambda + 1} \frac{\partial \pi_{hi}(\beta)}{\partial \beta} \bigg|_{\beta = \beta_{0,0}^0} \text{diag}^{-1}(\pi_{hi}(\beta_{0,0}^0)) \right\} -\lambda + 1 \sum_{h=1}^{H} \sum_{i=1}^{n} \left\{ \frac{\omega_{hi} m_{hi}}{\lambda + 1} \frac{\partial^2 \pi_{hi}(\beta)}{\partial \beta^2} \bigg|_{\beta = \beta_{0,0}^0} \text{diag}^{-1}(\pi_{hi}(\beta_{0,0}^0)) \right\} \right.

\end{equation}

where $\omega_{hi}$ is the weight associated with the $i$th observation in the $h$th group, $m_{hi}$ is the maximum likelihood estimate of the $i$th parameter in the $h$th group, and $\lambda$ is the tuning parameter. The result is analogous to the one obtained in [20] for the Wald-type tests statistics, but it is valid for all $\phi$-divergences, including the Cressie-Read divergences.
covariance matrix, leads us to understand why the IF is also invariant to the choice of \( \phi \). Let us denote

\[
p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) = \frac{\hat{y}_{hi}}{m_{hi}} - \pi_{hi}(\beta),
\]

and the Pearson residual functions associated to the \( i \)-th cluster in the \( h \)-th stratum. These are measures of discrepancy between the observed and the fitted model. Now, \( f_{\phi,hi}(x, \beta) \) and \( f_{\phi,hi}(\hat{y}_{hi}, \beta) \) defined in (10) can be seen as a function of (29), that is,

\[
f_{\phi,hi}(x, \beta) = A_{\phi} \left( p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) \right),
\]

\[
f_{\phi,hi}(\hat{y}_{hi}, \beta) = A_{\phi} \left( p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) \right)
\]

\[
= \left( A_{\phi} \left( p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) \right), \ldots, A_{\phi} \left( p \left( \frac{\hat{y}_{hi,(d+1)}}{m_{hi}}, \beta \right) \right) \right)^{T}.
\]

Therefore, the estimating equations given in (9) can be expressed by

\[
\sum_{h=1}^{H} \sum_{i=1}^{n_{h}} w_{hi} m_{hi} \frac{\partial \pi_{hi}(\beta)}{\partial \beta} A_{\phi} \left( p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) \right) = 0. \tag{30}
\]

The function \( A_{\phi}(z) \) is called the residual adjustment function (RAF, [20]). Note that the PMLE is obtained by taking in (30) the linear RAF \( A(z) = z \). From here and without loss of generality, we will focus on the class of Cressie-Read divergences, for which

\[
A_{\lambda}(z) = \frac{(1 + z)^{1+\lambda} - 1}{1 + \lambda}.
\]

Reference [20] showed that the RAF may have a critical effect on the robustness of the corresponding estimators. As large outliers correspond to large values of \( p \left( \frac{\hat{y}_{hi}}{m_{hi}}, \beta \right) \), the RAF of the more robust MP\( \phi \)Es should shrink the effect of such residuals relative to the classical PMLE. In Figure 2 we show the RAF for different values of the tuning parameter \( \lambda \). It is shown that for negative values of \( \lambda \) there is a down-weighting effect on large outliers (see also [32], Figure 1 for the generalized Hellinger family or [20], Figure 3 for some particular divergences). Note that in our multinomial model, this behavior is extended to \( d + 1 \) categories.

Moreover, the behavior of \( A_{\lambda}(z) \) for \( z \) near 0 determines the behavior of the IF (21) for \( \epsilon \) near 0. All the MP\( \phi \)Es have the same IF because all have the same first order approximation

\[
A_{\lambda}(z) \approx z,
\]

which is, in fact, equal to the value of the RAF for the PMLE. The second order approximation is given by

\[
A_{\lambda}(z) \approx z + \frac{\lambda^{2}}{2},
\]

and the tuning parameter \( \lambda \) plays an important role in measuring the trade-off between robustness and efficiency. As noted in [20] (Section 4), negative values of \( \lambda \) often guarantee

the second order approximation of the bias function of the estimators under contamination to be smaller than the first order approximation, so we can expect the behavior of the PM\( \phi \)Es with \( \lambda < 0 \) to be more robust.

In the following two sections we empirically illustrate the robustness of PM\( \phi \)Es with negative tuning parameter through an extensive simulation study and some real data examples.

V. MONTE CARLO SIMULATION STUDY

In this section we develop a simulation study in order to illustrate the robustness of the proposed estimators and Wald-type tests based on them. Following the simulation studies proposed in [11] and [12], we consider \( H = 4 \) strata with \( n_{h} \) clusters of the same size \( m \), with \( m = 20 \) and \( n_{h} \in \{10, 20, \ldots, 60\} \) for \( h = 1, \ldots, H \). We consider \( d + 1 = 3 \) categories on the response variable, depending on \( k = 2 \) explanatory variables. The response variable \( Y_{hi} \), described as

\[
E[Y_{hi}] = m \pi_{hi} (\beta^{0}) \quad \text{and} \quad V[Y_{hi}] = \nu_{m} m \Delta (\pi_{hi} (\beta^{0})),
\]

\[
\nu_{m} = 1 + \rho^{2}(m - 1), \quad i = 1, \ldots, n_{h}, \quad h = 1, \ldots, H,
\]

is considered to follow the m-Inflated multinomial distribution (see Remark 2.2), with parameters \( \rho^{2} = 0.5 \) and \( \pi_{hi} (\beta^{0}) \), given by the logistic relationship (1) with

\[
\beta^{0} = (\beta_{01}, \beta_{11}, \beta_{21}, \beta_{02}, \beta_{12}, \beta_{22})^{T}
\]

\[
= (0, -0.9, 0.1, 0.6, -1.2, 0.8)^{T}
\]

and \( x_{hi} \sim N(0, I) \) for all \( i = 1, \ldots, n_{h}, \ h = 1, \ldots, H \). In order to study the robustness issue, these simulations are repeated under contaminated data having 10% outliers. These outliers are generated by permuting the elements of the outcome variable, such that categories 1, 2, 3 are classified as categories 3, 1, 2 for the outlying observations.
Note that this view of considering outliers as classification errors in the multinomial logistic regression model is, in fact, in line with the general literature on robust analysis of categorical data ([34], [35]) and is covered with the theory developed in Section IV, where our “outlier producing measure” indeed provides classification error if the outlier...
In this scenario, the root of mean square error (RMSE) for the Cressie-Read PM$_\phi$Es of $\beta$ with $\lambda \in \{-0.5, -0.3, 0, 2/3\}$ is studied, both for the contaminated and not-contaminated cases (see top of Figure 3). To compute the accuracy in terms of contrast, we consider the testing problem

$$H_0: \beta_{11} = -0.9 \text{ vs. } H_1: \beta_{11} \neq -0.9.$$ 

For computing the empirical test level, we measured the proportion of Wald-type test statistics exceeding the corresponding chi-square critical value. The simulated test powers were also obtained under $H_1$ in a similar manner (here we consider $\beta_{11} = -1.5$). We used a nominal level of 0.05. Both levels and powers are presented in the middle and bottom of Figure 3.

It is observed that PMLE ($\lambda = 0$) presents the best behavior in terms of efficiency for the non-contaminated setting. In addition, PM$_\phi$E with $\lambda = 0.66$ presents a RMSE lower than PM$_\phi$E with negative values of $\lambda$. On the other hand, when it is considered the contaminated setting, better empirical levels are observed for negative values of the tuning parameter $\lambda$, particularly, for $\lambda = -0.5$. In terms of powers, negative values of $\lambda$ present better behavior in both settings, non-contaminated and contaminated. Positive values of $\lambda$ are presented as a good alternative only in terms of efficiency for small sample sizes, in concordance with [13]. Other alternative scenarios are considered in Appendix A-A.

### VI. Numerical Examples

#### A. Education in Malawi (Continuation)

Let us continue with the 2010 MDHS presented in Example 2.3. As pointed out there, the 2010 MDHS presents data on educational attainment for female and male by its wealth quintile level. In this section we make a comparison of the behavior between different PM$_\phi$Es, when estimating the probabilities of the response categories. For this purpose, and after estimating the PM$_\phi$Es in a grid of tuning parameters $\lambda \in \{-0.5, 0.7\}$ in the Cressie-Read subfamily, we measure a weighted standardized mean absolute error (SMAE) of the estimated probabilities against the observed probabilities. This is done by distinguishing the strata (wealth quintiles), the clusters (female and male) and jointly, as it can be seen in Figure 4. The lowest SMAEs are obtained for negative values of $\lambda$. Then, they seem to offer a better behavior than the classical PMLE.
B. Mammography Experience Data

As noted in Remark 2.8, multinomial logistic regression model under complex sample design is an extension of the classical one, evaluated under a simple sample design. Therefore, the tools developed in this paper can be also applied to these cases, in which the data design may be much simpler. In this section we study the Mammography

---

Fig. 5. Mammography example: predicted category probabilities of the response variable for the MLE ($\lambda = 0$) and M$\phi$E$\lambda$s with $\lambda = 2/3$ and $\lambda = -0.5$. 
experience data, a subset of a study by the University of Massachusetts Medical School, introduced in [36] and recently studied by [37] and [11]. This study, which assess factors associated with women’s knowledge, attitude and behavior towards mammography, involves 412 individuals, grouped in 125 distinct covariates values (which, somehow correspond to the “clusters” in a more complex survey) and 8 explanatory variables, detailed in the cited bibliography. The response variable ME (Mammography experience) is a categorical factor with three levels: “Never”, “Within a Year” and “Over a Year”. As suggested by [37], the groups of observations associated with covariate values $x_i$ for $i \in \{1, 3, 17, 35, 75, 81, 102\}$ can be treated as outliers. So this data set is a perfect candidate to show the robustness performance of the proposed estimators.

We compute the minimum $\phi$-divergence estimators (M$\phi$Es) of $\beta$, where $\lambda \in \{-0.5, 0, 2/3\}$, for the full dataset and also for the outliers deleted dataset. Moreover, we plot the corresponding (estimated) category probabilities for each available distinct covariate values. The left panel of Figure 5 presents these category probabilities for the first category, while the right panel presents these category probabilities for the second category. Results clearly indicate the significant variation of the MLE and M$\phi$E with $\lambda = 2/3$ in the presence or absence of the outliers (red circles and blue triangles, respectively). However, the M$\phi$E with $\lambda = -0.5$ is shown to be much more stable, which is in concordance with its theoretical robustness.

In [11] this dataset was also analyzed in order to illustrate the robustness of other family of estimators, those based on DPD divergences, the MDPDEs. This family is also parameterized by a tuning parameter, let say, $\lambda^* \geq 0$, and contains the MLE as a particular case for $\lambda^* = 0$ (see Remark 2.9 for more details). The question that could arise here is the point on using PM$\phi$Es instead of MDPDEs. In this regard, the efficiency of both family of estimators is compared in the following way: for each pair of tuning parameters $(\lambda, \lambda^*)$ in a grid on $[-0.5, 0] \times [0, 1]$, the M$\phi$Es and MDPDEs are computed and the estimated probabilities for each of the categories of the response variable are obtained for each of the $I = 125$ observations. Then, we count the number of
times, in these 125 observations, that the $M\phi E$ presents a lower error (the estimated probability is closer to the observed probability) than the MDPDE. The higher this value is, the better is the $M\phi E$ with respect to the MDPDE. If the value is under $[I/2] = 63$, then the MDPDE is preferable. These results are illustrated in the two heat plots (for the first and
Fig. 9. RMSEs (top), empirical levels (middle) and empirical powers (bottom). Non-contaminated and contaminated settings (left and right, respectively). Dirichlet Multinomial distribution.

second category, the third is omitted since it is similar) on Figure 6. We can observe how $M\phi\text{Es}$ with a low value of $\lambda$ improves any MDPDE, while MDPDEs with a large value of $\lambda^*$ only improves PM$\phi$Es with tuning parameters near to 0. The efficiency of the MLE is not comparable to any other option.
Now, we want to evaluate the robustness of the proposed Wald-type tests. We consider the problem of testing
\[ H_0 : \beta_{SYMPT_1} = 0, \]
\[ H_0 : \beta_{SYMPT_2} = \beta_{SYMPT_3}, \]
for the variable SYMPT (“You do not need a mammogram unless you develop symptoms: 1, strongly agree; 2, agree; 3, disagree; 4, strongly disagree”). The p-values obtained based on the proposed test are plotted over \( \lambda \) in Figure 7 for both the full and the outlier deleted data. Clearly, the test decision at the significance level \( \alpha = 0.1 \) changes completely in the presence of outliers for \( \lambda \) near to 0.

VII. CONCLUDING REMARKS

In this paper, we present robust estimators (PMoE) and Wald-type tests based on them for the multinomial logistic regression under complex survey, by means of \( \phi \)-divergence measures. Particularly, our study focuses on the Cressie-Read subfamily of divergences, which are modeled by a tuning parameter \( \lambda \). It is theoretically discussed and empirically illustrated how PMoEs and Wald-type tests with \( \lambda < 0 \) are more robust than the classical PMLE and Wald-test, presenting an interesting alternative to them. We believe that this method may be of special interest for analyzing demographic and health surveys, such as the ones presented in Section VI-B, as well as overall complex surveys for developing countries.

One of the problems that arise here is the following. Given any data set, the choice of the tuning parameter \( \lambda \). Robustness is usually accompanied with a loss of efficiency and other factors such as sample size. These issues can be determinant in this decision. One possible way to make this choice is as follows: in a grid of possible tuning parameters, apply a measure of discrepancy to the data. Then, the tuning parameter that leads to the minimum discrepancy-statistic can be chosen as the “optimal” one. This is, somehow, the idea followed in previous examples (see Figure 4 and 6). Another alternative is the one proposed by [38], which consists on minimizing the estimated mean square error, computed as the sum of the squared (estimated) bias and variance. One of the main drawbacks of this method is the fact that it depends on an pilot estimator to estimate the bias. This problem was also highlighted recently in [39], where an “iterative Warwick and Jones algorithm” (IWJ algorithm) is proposed. Application of these methods and a development of new ones will be a challenging and interesting problem for further consideration.

APPENDIX A

A. Some Extensions of the Simulation Study

We extend the simulation study presented in Section V to other scenarios. Particularly, we first study the same scheme as in Section V but considering two different overdistributed distributions for the response variable: the Random-Clumped and the Dirichlet Multinomial distributions. Results are presented in Figure 8 and Figure 9.

Same conclusions as in Section V are obtained, illustrating again the robustness of proposed estimators and Wald-type tests against classical PMLE.

B. Algorithms for m-Inflated, Random Clumped and Dirichlet Multinomial Distributions in the Context of Multinomial Logistic Regression Models With Complex Design

We present the algorithms that are needed to compute the Random-Clumped ([10]), Dirichlet-multinomial ([40]) and m-Inflated ([41]) multinomial distributions in the context of multinomial logistic regression models with complex design. We consider without loss of generality that the intra-cluster correlation parameter is equal in all the clusters and strata \( (\rho_{hi} = \rho, h = 1, \ldots, H, i = 1, \ldots, n_h) \).

1) m-Inflated Distribution:

**Goal:** Generation of response variable with m-Inflated distribution of parameters \( \rho \) and \( \pi(\beta) \), in a scenario with \( H \) strata and \( n_h \) clusters in the stratum \( h, h = 1, \ldots, H \).

1: for \( h = 1, \ldots, H \) do
2: for \( i = 1, \ldots, n_h \) do
3: \( k_1 \leftarrow Ber(\rho^2) \)
4: if \( k_1 = 0 \) then
5: \( y_{hi} \leftarrow M(m_{hi}, \pi_h(\beta)) \)
6: else
7: \( y_{hi} \leftarrow m_{hi} \times M(1, \pi_h(\beta)) \)
8: end if
9: end for
10: end for
11: return \( y = (y_{11}, \ldots, y_{Hn_H})^T \)

2) Random-Clumped Distribution:

**Goal:** Generation of response variable with Random Clumped distribution of parameters \( \rho \) and \( \pi(\beta) \), in a scenario with \( H \) strata and \( n_h \) clusters in the stratum \( h, h = 1, \ldots, H \).

1: for \( h = 1, \ldots, H \) do
2: for \( i = 1, \ldots, n_h \) do
3: \( y^{(0)} \leftarrow M(1, \pi_h(\beta)) \)
4: \( k_1 \leftarrow Bin(m_{hi}, \rho) \)
5: \( y^{(1)} = M(m_{hi} - k_1, \pi_h(\beta)) \)
6: \( y_{hi} = y^{(0)} \times k_1 + y^{(1)} \)
7: end for
8: end for
9: return \( y = (y_{11}, \ldots, y_{Hn_H})^T \)

3) Dirichlet-Multinomial Distribution:

**Goal:** Generation of response variable with Dirichlet Multinomial distribution of parameters \( \rho \) and \( \pi(\beta) \), in a scenario with \( H \) strata and \( n_h \) clusters in the stratum \( h, h = 1, \ldots, H \).

1: for \( h = 1, \ldots, H \) do
2: for \( i = 1, \ldots, n_h \) do
3: \( \alpha_1 \leftarrow \frac{1 - \rho^2}{\rho^2} \pi_{hi1}(\beta) \)
4: \( \alpha_2 \leftarrow \frac{1 - \rho^2}{\rho^2}(1 - \pi_{hi1}(\beta)) \)
5: \( y_{hi1} \leftarrow Bin(m_{hi1} Beta(\alpha_1, \alpha_2)) \)
6: for \( r = 1, \ldots, d \) do
7: \( \alpha_1 \leftarrow \frac{1 - \rho^2}{\rho^2}\pi_{hi}(\beta) \)
8: \( \alpha_2 \leftarrow \frac{1 - \rho^2}{\rho^2}(1 - \sum_{i=1}^{r-1} \pi_{hi}(\beta)) \)
9: \( y_{hir} \leftarrow Bin(m_{hi} - \sum_{i=1}^{r-1} y_{hir}, Beta(\alpha_1, \alpha_2)) \)
10: end for
11: \( y_{hi,d+1} = m_{hi} - \sum_{r=1}^{d} y_{hir} \)
12: \( y_{hi} = (y_{hi1}, \ldots, y_{hi,d+1}) \)
13: end for
14: end for
15: return \( y = (y_{11}, \ldots, y_{Hn_H})^T \)

ACKNOWLEDGMENT

The authors thank the referees and corresponding Editor for their helpful comments and suggestions.
