TEST (2016) 25:269–290
DOI 10.1007/s11749-015-0445-3

ORIGINAL PAPER

Robust estimation in generalized linear models: the density power divergence approach

Abhik Ghosh[1] · **Ayanendranath Basu**[1]

Received: 29 August 2014 / Accepted: 27 April 2015 / Published online: 28 May 2015

Abstract The generalized linear model is a very important tool for analyzing real data in several application domains where the relationship between the response and explanatory variables may not be linear or the distributions may not be normal in all the cases. Quite often such real data contain a significant number of outliers in relation to the standard parametric model used in the analysis; in such cases inference based on the maximum likelihood estimator could be unreliable. In this paper, we develop a robust estimation procedure for the generalized linear models that can generate robust estimators with little loss in efficiency. We will also explore two particular special cases in detail—Poisson regression for count data and logistic regression for binary data. We will also illustrate the performance of the proposed estimators through some real-life examples.

Keywords Density power divergence · Generalized linear model · Logistic regression · Poisson regression · Robustness

Mathematics Subject Classification 62F35 · 62J12

Electronic supplementary material The online version of this article (doi:10.1007/s11749-015-0445-3) contains supplementary material, which is available to authorized users.

This is part of the Ph.D. research work of A. Ghosh at the Indian Statistical Institute.

✉ Ayanendranath Basu
ayanbasu@isical.ac.in

Abhik Ghosh
abhianik@gmail.com

[1] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700 108, India

1 Introduction

Many real-life problems require suitable techniques to describe some response data through a set of related explanatory variables. Parametric regression helps the experimenter to model such scenarios by means of some pre-specified functional relationship between response and explanatory variables described through a set of real parameters. The most widely used regression model is linear regression for continuous responses. In practice, though, there are lots of different types of response data like count data, binary response data and others which arise frequently in real-life experiments. The generalized linear model is the general tool that can be used with all such types of response variables. It allows the experimenter to model the response variables by any distribution within a large family of distributions, namely the exponential family, and the expected response by any (suitably smooth) function of a linear combination of the explanatory variables. The ordinary linear regression is a special case of the above.

The classical estimation procedure in this context is the maximum likelihood estimation method which is asymptotically efficient but lacks robustness against outliers and model misspecification. In many real-life experiments, outliers show up as a matter or routine which influence the maximum likelihood estimators (MLEs) and often produce nonsensical results. So, there is a real need for developing robust estimation procedures for the generalized linear regression model. Although there is a crowded field of robust estimators in the ordinary linear regression problem, there exist only a few robust estimators for the generalized linear model. Cantoni and Ronchetti (2001) and Hosseinian (2009) present and discuss some such approaches that bound the Pearson residuals. There is another pathway in the literature which consists of bounding the unscaled deviance components in some special cases; see, eg., Bianco and Yohai (1996), Croux and Haesbroeck (2003) and Bianco et al. (2013). Aeberhard et al. (2014) provides a comparison between the two approaches in case of the negative binomial responses. However, most of these approaches consider the explanatory variables to be stochastic.

In this paper, we will develop an estimation procedure for the generalized linear model from a design perspective, where we will assume that the explanatory variables are fixed; each response is independent and follows the same distribution specified by the generalized linear model, but has different distributional parameters depending on the values of the corresponding explanatory variables. The idea is motivated by the work of Ghosh and Basu (2013) where a robust minimum divergence estimation procedure was developed under the general setup of independent but non-homogeneous observations using the density power divergence. This work considered the case of simple linear regression in detail. Here, we will follow a similar approach to develop the minimum density power divergence estimators of the parameters of the generalized linear model, which will be highly robust in presence of influential observations and also have comparable high efficiency. Like Cantoni and Ronchetti (2001) and Hosseinian (2009), our approach also bounds the Pearson residuals; hence the term "robustness" in this paper refers to bounded-influence robustness.

The rest of the paper is organized as follows. In Sect. 2 we briefly describe the model and develop the corresponding minimum density power divergence estimators. We also prove the asymptotic properties and present the influence function analysis of

the proposed estimator. We also present a short discussion on a data-driven choice of the tuning parameter α in Sect. 2.5. We will then explore the special case of Poisson regression for count data and logistic regression for binary data in Sects. 3 and 4, respectively. Section 5 contains the application of the proposed method to two real-life data sets. A brief comparison with some of the existing robust estimators is provided in Sect. 6. Finally the paper ends with some concluding remarks in Sect. 7. Some additional numerical examples are provided in the Supplementary Material.

2 The minimum density power divergence estimator in generalized linear models

2.1 The generalized linear model (GLM)

In generalized linear models, the response variables Y_i are independent and follow the general exponential family of distributions having density

$$f(y_i;\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right\}, \tag{1}$$

where the canonical parameter θ_i is a measure of location depending on the fixed predictor x_i and ϕ is the nuisance scale parameter. The mean μ_i of Y_i satisfies the relation $g(\mu_i)=\eta_i=x_i^T\beta$, where g is a monotone and differentiable link function and $\eta_i=x_i^T\beta$ is the linear predictor. Our main parameter of interest is the regression coefficient β and ϕ acts as the nuisance parameter which shows up only in the error variance. Clearly the generalized linear model allows us to choose several possible densities f from the exponential family and the link function g to form a wide variety of regression models.

By choosing f to be the normal density and g to be the identity link function the generalized linear model reduces to the usual normal linear regression model. Further, choosing f as the Poisson density and g as the log link $g(\mu)=\log(\mu)$, we get the Poisson regression case which is useful in modeling ordinal data and cases of overdispersion. For binomial f, choosing the Logit link function $g(\mu)=\log(\mu/(1-\mu))$ or the Probit link function $g(\mu)=\Phi^{-1}(\mu)$ generates the logistic and the Probit regression models, respectively, which are useful in modeling binary response variables.

2.2 The minimum density power divergence estimator (MDPDE)

We will define the minimum density power divergence estimators (MDPDEs) for the GLM with general density f and link function g so that we can estimate the regression coefficients for any regression model as a special case of it by substituting the form of f and g. In the later sections, we will consider some of these special cases in detail. The density power divergence (DPD) measure was developed by Basu et al. (1998) in terms of a tuning parameter $\alpha\geq 0$; the divergence between two densities h and f is given by

$$d_\alpha(h, f) = \int f^{1+\alpha} - \frac{1+\alpha}{\alpha} \int f^\alpha h + \frac{1}{\alpha} \int h^{1+\alpha}, \quad \text{if} \quad \alpha > 0,$$

and $d_0(h, f) = \lim_{\alpha \to 0} d_\alpha(h, f) = \int h \log(h/f)$. In practice, h represents the data density and f represents the model density (which depends on the unknown parameter). One then minimizes this divergence over the parameter space to get the minimum divergence estimate of the parameter. The situation is substantially simplified in case of the DPD because in this case the data distribution may be represented by ordinary empiricals, rather than its smoothed version.

Suppose we have a data set (y_i, x_i); $i = 1, \ldots, n$ from the GLM with density f given by Eq. (1) and a general link function $g(\mu_i) = \eta_i = x_i^T \beta$. Further assume that the independent variables x_i are given and fixed so that we are indeed considering a fixed carrier generalized linear model. Then we have the setup of independent but non-homogeneous observations, where $y_1, \ldots, y_n$ are independent and y_i has density $f_i(.; (\beta, \phi)) = f(y_i; \theta_i, \phi)$ for all $i = 1, \ldots, n$. Hence, we can use the approach of Ghosh and Basu (2013), where the MDPDE for the independent but non-homogeneous observations was defined. Following this approach, the MDPDE of (β, ϕ) has to be obtained by minimizing

$$H_n(\beta, \phi) = \frac{1}{n} \sum_{i=1}^{n} V_i(Y_i; (\beta, \phi)),$$

$$\text{where} \quad V_i(Y_i; (\beta, \phi)) = \int f_i(y; (\beta, \phi))^{1+\alpha} \mathrm{d}y - \left(1 + \frac{1}{\alpha}\right) f_i(Y_i; (\beta, \phi))^\alpha.$$

Note that, in the usual GLM estimation, conventionally we use a robust estimate of scale parameter ϕ and then estimate the regression parameter β. One can perform simultaneous robust estimation of both the parameters, as in Huber's Proposal 2 (Huber 1983) in the linear case; however, such exceptions to the above convention are rare. Here, in the proposed minimum DPD estimation, we do simultaneously estimate β and ϕ robustly by just minimizing $H_n(\beta, \phi)$ with respect to both the parameters. The estimating equation of the parameters are then given by $\sum_{i=1}^{n} \nabla V_i(Y_i; (\beta, \phi)) = 0$, or,

$$\sum_{i=1}^{n} \left[\int u_i(y; (\beta, \phi) f_i(y; (\beta, \phi))^{1+\alpha} \mathrm{d}y - u_i(Y_i; (\beta, \phi) f_i(Y_i; (\beta, \phi))^\alpha \right] = 0.$$

where $u_i(y; (\beta, \phi)) = \nabla \log(f_i(y; (\beta, \phi))$; ∇ represents the derivative with respect to (β, ϕ), with ∇_β and ∇_ϕ denoting the indicated individual derivatives. Then, a simple calculation shows that

$$\nabla_\beta \log(f_i(y_i; (\beta, \phi)) = \frac{(y_i - \mu_i)}{\mathrm{Var}(y_i) g'(\mu_i)} x_i = K_{1i}(y_i; (\beta, \phi)) x_i,$$

$$\nabla_\phi \log(f_i(y_i; (\beta, \phi)) = -\frac{(y_i \theta_i - b(\theta_i))}{a^2(\phi)} a'(\phi) + \frac{\partial}{\partial \phi} c(y_i, \phi) = K_{2i}(y_i; (\beta, \phi)),$$

where K_{1i}, K_{2i} are the indicated functions. Thus, our estimating equations become

$$\sum_{i=1}^{n} x_i \left[\int K_{1i}(y;(\beta,\phi)) f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y - K_{1i}(y_i;(\beta,\phi)) f_i(y_i;(\beta,\phi))^{\alpha} \right] = 0, \tag{2}$$

$$\sum_{i=1}^{n} \left[\int K_{2i}(y;(\beta,\phi)) f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y - K_{2i}(y_i;(\beta,\phi)) f_i(y_i;(\beta,\phi))^{\alpha} \right] = 0. \tag{3}$$

However, if we want to ignore the nuisance parameter ϕ, as per the usual practice, and estimate β taking ϕ fixed (or, substituted suitably), it is enough to consider only estimating Eq. (2). Further, for $\alpha = 0$, we have

$$\int \frac{(y_i - \mu_i)}{\mathrm{Var}(y_i)} g'(\mu_i) x_i f_i(y_i;(\beta,\phi))^{1+\alpha} \mathrm{d}y = 0,$$

and hence the estimating equations for β (ignoring ϕ) simplify to

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\mathrm{Var}(Y_i) g'(\mu_i)} x_i = 0.$$

Note that this is just the maximum likelihood estimating equation and also is the same as the ordinary least squares (OLS) estimating equation for β assuming ϕ to be fixed. Thus, the MDPDE of β with $\alpha = 0$ equals the maximum likelihood estimator as well as the OLS estimator of β. That is the MDPDE proposed here is just a natural generalization of the MLE.

Also it is interesting to note that if our density f is such that $\int f(y;\theta_i,\phi)^{1+\alpha} \mathrm{d}y$ is independent of the location parameter θ_i, like the normal density, then we have $\int \frac{(y_i-\mu_i)}{\mathrm{Var}(y_i) g'(\mu_i)} x_i f_i(y_i;(\beta,\phi))^{1+\alpha} \mathrm{d}y = 0$ and hence the estimating Eq. (2) simplifies to

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\mathrm{Var}(Y_i) g'(\mu_i)} x_i f_i(Y_i;(\beta,\phi))^{\alpha} = 0. \tag{4}$$

2.3 Asymptotic properties

We will now derive the joint asymptotic distribution of the minimum density power divergence estimator $(\hat{\beta}, \hat{\phi})$ of the parameters (β, ϕ) obtained by solving the estimating Eqs. (2) and (3). For simplicity, we will assume that the true data-generating distribution also belongs to the model density with parameters (β^g, ϕ^g). Define, for $i = 1, \ldots, n$ and $j, k = 1, 2$,

$$\gamma_{ji} = \gamma_{ji}^{1+\alpha}(\beta,\phi) = \int K_{ji}(y;(\beta,\phi)) f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y,$$

$$\text{and } \gamma_{jki} = \gamma_{jki}^{1+\alpha}(\beta,\phi) = \int K_{ji}(y;(\beta,\phi)) K_{ki}(y;(\beta,\phi)) f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y,$$

$$\text{so that } N_i^{1+\alpha}(\beta,\phi) = \int u_i(y;(\beta,\phi) f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y = \begin{pmatrix} \gamma_{1i} x_i \\ \gamma_{2i} \end{pmatrix},$$

$$M_i^{1+\alpha}(\beta,\phi) = \int u_i(y;(\beta,\phi)) u_i(y;(\beta,\phi))^T f_i(y;(\beta,\phi))^{1+\alpha} \mathrm{d}y$$

$$= \begin{pmatrix} \gamma_{11i} x_i x_i^T & \gamma_{12i} x_i \\ \gamma_{12i} x_i^T & \gamma_{22i} \end{pmatrix}.$$

Now, put $\Gamma_j^{(\alpha)} = \mathrm{Diag}(\gamma_{ji})_{i=1,\ldots,n}$ and $\Gamma_{jk}^{(\alpha)} = \mathrm{Diag}(\gamma_{jki})_{i=1,\ldots,n}$ for $j,k = 1,2$ and $X^T = [x_1,\ldots,x_n]$. Then we have

$$\Psi_n(\beta,\phi) = \frac{1}{n}\sum_{i=1}^n M_i^{1+\alpha}(\beta,\phi) = \frac{1}{n}\begin{pmatrix} X^T \Gamma_{11}^{(\alpha)} X & X^T \Gamma_{12}^{(\alpha)} \mathbf{1} \\ \mathbf{1}^T \Gamma_{12}^{(\alpha)} X & \mathbf{1}^T \Gamma_{22}^{(\alpha)} \mathbf{1} \end{pmatrix}; \tag{5}$$

$$\Omega_n(\beta,\phi) = \frac{1}{n}\sum_{i=1}^n \left[M_i^{1+2\alpha}(\beta,\phi) - N_i^{1+\alpha}(\beta,\phi)(N_i^{1+\alpha}(\beta,\phi))^T \right] \tag{6}$$

$$= \frac{1}{n}\begin{pmatrix} X^T[\Gamma_{11}^{(2\alpha)} - \Gamma_1^{(\alpha)T}\Gamma_1^{(\alpha)}]X & X^T[\Gamma_{12}^{(2\alpha)} - \Gamma_1^{(\alpha)}\Gamma_2^{(\alpha)}]\mathbf{1} \\ \mathbf{1}^T[\Gamma_{12}^{(2\alpha)} - \Gamma_1^{(\alpha)}\Gamma_2^{(\alpha)}]X & \mathbf{1}^T[\Gamma_{22}^{(2\alpha)} - \Gamma_2^{(\alpha)T}\Gamma_2^{(\alpha)}]\mathbf{1} \end{pmatrix}. \tag{7}$$

Then, the asymptotic distribution of $(\hat{\beta},\hat{\phi})$ follows along the lines of Theorem 3.1 of Ghosh and Basu (2013), provided the Assumptions (A1)–(A7) hold in case of the generalized linear models. These assumptions are presented in the Supplementary material to this paper. Note that, Assumptions (A1)–(A3) hold directly from the properties of the exponential family of distributions.

Theorem 1 *Under Assumptions (A1)–(A7) of* Ghosh and Basu (2013)*, there exists a consistent sequence* $(\hat{\beta}_n,\hat{\phi}_n)$ *of roots to the minimum DPD estimating Eqs.* (2) *and* (3)*. Also, the asymptotic distribution of* $\Omega_n^{-\frac{1}{2}}\Psi_n[\sqrt{n}((\hat{\beta}_n,\hat{\phi}_n) - (\beta^g,\phi^g))]$ *is* $(p+1)$*-dimensional normal with mean* 0 *and variance* I_{p+1}*, the identity matrix of dimension* $p+1$*, where* $\Psi_n = \Psi_n(\beta^g,\phi^g)$ *and* $\Omega_n = \Omega_n(\beta^g,\phi^g)$*.*

Note that the results of Theorem 1 would have been a direct consequence of standard M-estimation results provided the covariates are assumed to be stochastic. However, in this paper we are considering fixed design cases with non-stochastic covariates and the new Assumptions (A1)–(A7) are just the corresponding generalizations of the original assumptions of Huber (1964). See Ghosh and Basu (2013) for a more detailed discussion on these new assumptions. Similar generalizations in the context of the influence function in relation to the approach of Hampel et al. (1986) will be considered in the next subsection.

It follows from above theorem that the reciprocal of the matrix $\Psi_n^{-1}\Omega_n\Psi_n^{-1}$ gives a estimate of the asymptotic efficiency of the MDPDEs $(\hat{\beta}_n,\hat{\phi}_n)$. Though this depends

on the sample size n and the given covariates x_i, it will give a reasonable estimate of the asymptotic efficiency for large n.

Further, note that the asymptotic covariance of the estimators $\hat{\beta}_n$ and $\hat{\phi}_n$ is not in general 0 and hence these estimators are not asymptotically independent for all GLMs. However, for some particular cases including the normal linear regression case, they turn out to be independent. One possible set of sufficient conditions for their independence are $\gamma_{12i}^{1+2\alpha} = 0$ and $\gamma_{1i}^{1+\alpha}\gamma_{2i}^{1+\alpha} = 0$ for all i. These conditions hold for the normal linear regression case.

2.4 Influence function

To illustrate the robustness properties of the proposed estimation methodology for the generalized regression model, we will now consider the influence function of the MDPDE of the parameter $\theta = (\beta, \phi)$. For this we need to consider them in terms of a statistical functional at the true data-generating distribution $\underline{\mathbf{G}} = (G_1, \ldots, G_n)$. Let $T_\alpha^\beta(\underline{\mathbf{G}})$ and $T_\alpha^\phi(\underline{\mathbf{G}})$ denote the minimum DPD functionals for the parameters β and ϕ, respectively. Let $T_\alpha(\underline{\mathbf{G}}) = (T_\alpha^\beta(\underline{\mathbf{G}})^T, T_\alpha^\phi(\underline{\mathbf{G}}))^T$, which is defined by

$$\frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i(.), f_i(.; T_\alpha(\underline{\mathbf{G}}))) = \min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i(.), f_i(.;\theta)),$$

where g_i is the probability density function corresponding to G_i. We consider the contaminated density $g_{i,\epsilon} = (1-\epsilon)g_i + \epsilon\delta_{t_i}$ where t_i is the point of contamination and $G_{i,\epsilon}$ denotes the corresponding distribution function for all $i = 1, \ldots, n$. Let $\theta_\epsilon^{i_0} = T_\alpha(G_1, \cdot, G_{i_0-1}, G_{i_0,\epsilon}, \cdot, G_n)$ be the minimum DPD functional with contamination only in the i_0th direction. Then a fairly straightforward (albeit lengthy and tedious) calculation shows that the influence function of T_α for contamination at the direction i_0 will be

$$\begin{aligned}
\mathrm{IF}_{i_0}(t_{i_0}, T_\alpha, \underline{\mathbf{G}}) &= \Psi_n^{-1}\frac{1}{n}[f_{i_0}(t_{i_0};(\beta,\phi))^\alpha u_{i_0}(t_{i_0};(\beta,\phi)) - N_{i_0}^{1+\alpha}] \\
&= \Psi_n^{-1}\frac{1}{n}\begin{pmatrix} [f_{i_0}(t_{i_0};(\beta,\phi))^\alpha K_{1i_0}(t_{i_0};(\beta,\phi)) - \gamma_{1i_0}]x_i \\ \\ f_{i_0}(t_{i_0};(\beta,\phi))^\alpha K_{2i_0}(t_{i_0};(\beta,\phi)) - \gamma_{2i_0} \end{pmatrix}.
\end{aligned}$$

Note that for any fixed sample size n and any given (finite) values of X_is, if Ψ_n and γ_{ji_0}s are assumed to be bounded, the influence function of the MDPDE of (β, ϕ) will be bounded with respect to the contamination in any direction i_0 provided the terms $f_i(t_i;(\beta,\phi))^\alpha K_{ji}(t_i;(\beta,\phi))$ are bounded for all i and $j = 1, 2$. Under assumptions (A1)–(A7) of Ghosh and Basu (2013), Ψ_n and γ_{ji_0}s are necessarily bounded. This can be seen to hold for the majority of GLMs with $\alpha > 0$ because of the exponential nature of the density function and the polynomial nature of the functions $K_{ji}(t_i;(\beta,\phi))$. This demonstrates the robust nature of the MDPDE in most GLMs with $\alpha > 0$. However, for $\alpha = 0$ the term $f_i(t_i;(\beta,\phi))^\alpha K_{ji}(t_i;(\beta,\phi)) = K_{ji}(t_i;(\beta,\phi))$ is clearly unbounded implying the non-robust nature of the MLE in case of any GLM.

As discussed before, in the particular case when $\gamma_{12i}^{1+2\alpha} = 0$ and $\gamma_{1i}^{1+\alpha}\gamma_{2i}^{1+\alpha} = 0$ for all i (like the normal linear regression case), the minimum density power divergence estimator of β and ϕ become asymptotically independent and we can also separate out the influence function for the minimum density power divergence estimator of β and ϕ. Due to the special form of the matrix Ψ_n in this case, these two influence functions simplify, respectively, to

$$\text{IF}_{i_0}(t_{i_0}, T_\alpha^\beta, \underline{\mathbf{G}}) = (X^T \Gamma_{11}^{(\alpha)} X)^{-1} x_{i_0} [f_{i_0}(t_{i_0}; (\beta, \phi))^\alpha K_{1i_0}(t_{i_0}; (\beta, \phi)) - \gamma_{1i_0}],$$
$$\text{and} \quad \text{IF}_{i_0}(t_{i_0}, T_\alpha^\phi, \underline{\mathbf{G}}) = (\mathbf{1}^T \Gamma_{22}^{(\alpha)} \mathbf{1})^{-1} [f_{i_0}(t_{i_0}; (\beta, \phi))^\alpha K_{2i_0}(t_{i_0}; (\beta, \phi)) - \gamma_{2i_0}].$$

As in Ghosh and Basu (2013), in this context also we can define some measures of sensitivity based on the influence function, which is presented in the Supplementary material to this paper.

2.5 A data-driven choice of the tuning parameter α

The minimum DPD estimators depend on the choice of the tuning parameter $\alpha \geq 0$ defining the divergence. The properties of the MDPDE in the case of independent and identically distributed data have been extensively studied in the literature and it is well known that there is a trade-off between efficiency and robustness for varying α. Increasing α leads to greater robustness at the cost of efficiency. Ghosh and Basu (2013, 2015) also observed similar trade-offs for the linear regression case with fixed covariates. Here, we have observed the same phenomenon in the context of the proposed MDPDE for the Poisson and the logistic regression models (see Sects. 3 and 4 below). Therefore, it is necessary to carefully choose the tuning parameter α while using the MDPDE in any of the GLMs. In this section, we will try to present a possible approach to choose the optimum value of α based on the observed data at hand.

In the context of the i.i.d. data problems, some data-driven choices for selecting the optimum tuning parameter in the minimum DPD estimation context have been proposed by Hong and Kim (2001) and Warwick and Jones (2005). Ghosh and Basu (2015) extended these approaches to the case of independent but non-homogeneous data and illustrated this approach for the case of linear regression through detailed simulation studies. In this present paper, we consider the GLM from its design perspective so that given the values of the explanatory variables x_i the response y_i is independent but not identically distributed. So, we can apply the results of Ghosh and Basu (2015) to choose a data-driven optimum choice of the tuning parameter α. Accordingly, we need to choose α by minimizing a consistent estimate of the mean square error (MSE) of the MDPDE $\hat{\theta} = (\hat{\beta}_\alpha, \hat{\phi}_\alpha)$ of the true parameter value $\theta^g = (\beta^g, \phi^g)$, defined as $E[(\hat{\theta}_\alpha - \theta^g)^T (\hat{\theta}_\alpha - \theta^g)]$, in the GLMs. The true parameter represents the larger component of a possible mixture distribution in the spirit of Warwick and Jones (2005). It follows from the asymptotic distribution of the MDPDE that, asymptotically

$$E[(\hat{\theta}_\alpha - \theta^g)^T (\hat{\theta}_\alpha - \theta^g)] = (\theta_\alpha - \theta^g)^T (\theta_\alpha - \theta^g) + \frac{1}{n} \text{Trace}[\Psi_n^{-1} \Omega_n \Psi_n^{-1}], \quad (8)$$

where Ψ_n and Ω_n are as defined in Sect. 2.3 and $\theta_\alpha = (\beta_\alpha, \phi_\alpha)$ is the parameter value minimizing the DPD measure between the true and model densities corresponding to tuning parameter α. Further, from the expressions of Ψ_n and Ω_n it is sufficient to find some consistent estimator of the quantities γ_{ji} and γ_{jki} for $j, k = 1, 2$ and $i = 1, \ldots, n$, which can be done by replacing the parameter value (β, ϕ) in their expressions by the corresponding MDPDEs $(\hat{\beta}_\alpha, \hat{\phi}_\alpha)$. Let us denote the resulting matrices by $\hat{\Psi}_n$ and $\hat{\Omega}_n$. To estimate the bias term, we will use $(\hat{\beta}_\alpha, \hat{\phi}_\alpha)$ as a consistent estimate of $(\beta_\alpha, \phi_\alpha)$. For estimating θ^g, we can use several "pilot" estimators which will in turn affect the final choice of the tuning parameter. Ghosh and Basu (2015) suggested, on the basis of an extensive simulation study, the choice of the MDPDE with $\alpha = 0.5$ as a reasonable pilot estimator. For any particular generalized model, we can find such a "good" pilot estimator through some simulation studies, and then use the observed data to choose the corresponding optimum tuning parameter value. Some examples illustrating this approach of choosing tuning parameter in case of the GLM is provided in the supplementary material.

Another perspective of the criterion (8) can be obtained by noting its similarity with the robust version of AIC (Heritier et al. 2009, p. 73, Eq. (3.31)) and its generalized version the GAIC (Heritier et al. 2009, p. 159). Although the trace term is different, the formulations are clearly in similar spirit which gives another interesting interpretation of this criterion.

Note that, the robustness of the proposed MDPDE, when the tuning parameter α is estimated from the data, also depends directly on the robustness of the estimation of α. Using the chain rule of derivatives, the robustness of the MDPDE with a data-driven α can be quantified by noting that its influence function is a multiple of the influence function of the fixed α estimator as obtained in Sect. 2.4 and the multiplier is nothing but the influence function of the estimator of optimum α itself. For the tuning parameter selection process described above, it can be verified empirically that the robustness of the optimum α estimator depends directly on that of the "pilot" estimator used; see Ghosh and Basu (2015) and Section 3 of the supplementary material of this paper for some numerical illustrations. Our experience in this regard indicates that the suggested choice of the MDPDE with $\alpha = 0.5$ as the "pilot" estimator is quite robust with respect to contamination in the data leading to robust selection of α. However, further research including more detailed empirical studies will improve our understanding of this complex issue; we hope to take up such research in the future.

Another important issue requires consideration in connection with the data-driven choice of the tuning parameter α. The asymptotic properties derived in Sect. 2.3 pertain to a fixed α. What about the asymptotic results under this data-driven choice? Clearly, in that case the final result will depend on the process of selecting the parameter α, and how good that process is. When the assumed model holds, and the data are pure, the classical method would work well and the chosen tuning parameter α should preferably remain close to 0. Large-scale simulation studies, not presented here, indicate that the adaptively chosen α is equal to or close to 0 in the overwhelming majority of the cases when the tuning parameter is adaptively selected using the Warwick and Jones (2005) approach and the data are pure; this phenomenon is observed for several different GLMs. While we do not have a general proof at this moment, our conjecture is that the estimator corresponding to the adaptively chosen tuning parameter will be

asymptotically equivalent to the maximum likelihood estimator under the model; at the least, the distribution of the estimator chosen through this adaptive routine will provide a good large sample approximation to that of the maximum likelihood estimator.

The description of the previous paragraph parallels the result of Theorem 1, where the asymptotic distribution for fixed α is provided under the model. However, when the model is misspecified or the data are contaminated, the description becomes more complicated. The theoretical optimal α then corresponds to the estimator which minimizes the sum of the square of the theoretical bias and the trace of the covariance matrix. We feel that whenever the data-driven estimate of α is consistent for the true (fixed) optimal value, the large sample asymptotic distribution of the fixed α estimator will provide a good approximation for the distribution of the adaptively chosen estimator. Clearly more research is needed on this topic.

3 Special Case I: Poisson regression for count data

The most useful regression tool for count data is the Poisson regression model where, given the values of explanatory variables, the response variables independently follow the Poisson distribution but with different mean parameters depending on the corresponding values of the explanatory variable. More precisely, let $(y_1, x_1), \ldots, (y_n, x_n)$ be the sample observations from the Poisson regression model. Assume that the values x_i of the explanatory variable are fixed. Then, in the Poisson regression model, the count variables y_i are assumed to be independent and have Poisson distributions with

$$E(y_i|x_i) = \mathrm{e}^{(x_i^T\beta)}$$

and we want to estimate the parameter β efficiently and robustly.

3.1 The MDPDE for Poisson regression

Poisson regression is indeed a special case of GLM with known shape parameter $\phi = 1$ and $\theta_i = \eta_i = x_i^T\beta$, $b(\theta_i) = \mathrm{e}^{\theta_i}$ and $c(y_i) = -\log(y_i!)$. Since here the mean is $\mu_i = \mathrm{e}^{(x_i^T\beta)} = \mathrm{e}^{\eta_i}$, the link function g is the natural logarithm function and the variance of y_i is also $\mathrm{e}^{(x_i^T\beta)}$. Thus, we can estimate the unknown parameter β using our minimum density power divergence estimation procedure as described earlier. Using the above notation and the form of the Poisson distribution, the minimum DPD estimating equation for $\alpha \geq 0$ becomes

$$\sum_{i=1}^{n}[\gamma_{1i}(\beta) - (y_i - \mathrm{e}^{(x_i^T\beta)})f_i(y;\beta)^\alpha]x_i = 0. \tag{9}$$

where $f_i(y;\beta)$ is the probability mass function of the Poisson distribution with mean $\mathrm{e}^{(x_i^T\beta)}$. In particular, for $\alpha = 0$, the above estimating equation simplifies to the maximum likelihood estimating equation given by

$$\sum_{i=1}^{n}(y_i - \mathrm{e}^{(x_i^T\beta)})x_i = 0. \tag{10}$$

However, for $\alpha > 0$, there is no simplified form for γ_{1i} and γ_{11i} so that we need to compute this quantities numerically and then numerically solve the estimating Eq. (9) with respect to β.

3.2 Properties of the MDPDE

The asymptotic properties of the MDPDE of β under Poisson regression model follows directly from Theorem 1 (see Section 2 of the supplementary material for derivation).

Corollary 1 *Under Assumptions (A1)–(A7) of* Ghosh and Basu (2013)*, there exists a consistent sequence $\hat{\beta}_n = \hat{\beta}_n^{(\alpha)}$ of roots to the minimum DPD estimating Eqs. (9) for the tuning parameter α. Also, asymptotically,*

$$(X^T[\Gamma_{11}^{(2\alpha)}(\beta^g) - \Gamma_1^{(\alpha)2}(\beta^g)]X)^{-\frac{1}{2}}(X^T\Gamma_{11}^{(\alpha)}(\beta^g)X)(\hat{\beta}_n - \beta^g) \sim N_p(0, I_p).$$

Thus, the asymptotic efficiency of the different MDPDE $\hat{\beta}_n = \hat{\beta}_n^{(\alpha)}$ of β can be measured based on the asymptotic variance

$$\mathrm{AV}_\alpha(\beta^g) = (X^T\Gamma_{11}^{(\alpha)}(\beta^g)X)^{-1}(X^T[\Gamma_{11}^{(2\alpha)}(\beta^g) - \Gamma_1^{(\alpha)2}(\beta^g)]X)(X^T\Gamma_{11}^{(\alpha)}(\beta^g)X)^{-1},$$

which can be consistently estimated by replacing β^g with $\hat{\beta}_n$ in its expression, i.e., $\widehat{\mathrm{AV}}_\alpha = \mathrm{AV}_\alpha(\hat{\beta}_n)$. Thus, an estimate of the relative efficiency of the different MDPDEs of the ith component of the parameter vector β with respect to its MLE (or the OLS estimator) is given by

$$\widehat{\mathrm{RE}}_{i,\alpha} = \frac{i\text{th diagonal entry of } \widehat{\mathrm{AV}}_0}{i\text{th diagonal entry of } \widehat{\mathrm{AV}}_\alpha} \times 100.$$

Clearly, the above estimate of the relative efficiency depends on the sample size n and the choice of the given explanatory variables x_i. But it can be shown that the consistency of the estimator $\hat{\beta}_n$ implies that the above measure gives us a consistent estimator of the asymptotic relative efficiency if the x_is are chosen suitably. For example, X^TX must be bounded. We have presented the empirical value of this measure of relative efficiency for different sample sizes $n = 50$, respectively, under several different cases in Table 1; the same for $n = 100$ is provided in the supplementary material. We have reported six cases which are defined based on the true values of the regression coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and the given values of the explanatory variables x_i $(i = 1, \ldots, n)$ as follows: the parameter $p = 2$ in the first four cases; Cases I and II have $x_i = (1, \sqrt{i})$ while Cases III and IV have $x_i = \left(1, \frac{1}{i}\right)$; $\beta = (1, 1)$ for Cases I and III and $\beta = (1, 0.5)$ for Cases II and IV. The parameter $p = 3$ in Cases V and VI with common $x_i = \left(1, \sqrt{i}, \frac{1}{i^2}\right)$ and $\beta = (1, 1, 1)$, $\beta = (2, 1, 0.5)$,

Table 1 The estimated relative efficiencies of the MDPDE for various values of the tuning parameter α under different cases of Poisson regression with sample size $n = 50$

Case	Coefficients	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
I	β_0	100.0	100.0	98.3	91.2	80.9	73.5	58.1	37.9
	β_1	100.0	100.0	98.3	91.1	80.7	73.1	57.2	36.4
II	β_0	100.0	99.9	98.5	93.2	85.9	80.7	70.5	56.5
	β_1	100.0	99.8	98.4	93.0	85.7	80.5	70.0	55.5
III	β_0	100.0	100.0	98.8	94.5	88.9	85.1	77.6	67.7
	β_1	100.0	100.0	98.8	93.7	88.4	84.3	76.0	64.8
IV	β_0	100.0	100.0	98.7	94.4	88.9	85.1	77.8	68.0
	β_1	100.0	100.0	98.9	94.3	88.4	84.6	76.6	66.3
V	β_0	100.0	100.0	98.9	94.4	88.7	84.9	77.4	67.5
	β_1	100.0	100.0	98.9	94.3	88.1	84.3	76.1	65.7
	β_2	100.0	100.0	98.9	94.1	88.1	84.2	75.9	65.4
VI	β_0	100.0	100.0	98.7	94.2	88.1	84.0	76.0	65.5
	β_1	100.0	100.0	98.6	94.3	88.1	83.8	75.8	65.0
	β_2	100.0	100.0	98.6	94.3	88.1	83.7	75.7	64.8

respectively. All the simulations are done based on 1000 replications. It is clear from the tables that the loss of efficiency is quite negligible for the MDPDE with small positive α under each of the cases considered here. Even for large positive α near 0.5 we can get quite high efficiency if x_is are relatively small.

Next, to see the robustness of the MDPDE under the Poisson regression model, we will use the results from the Sect. 2.4. The influence function of the MDPDE in the direction i_0 simplifies to

$$\mathrm{IF}_{i_0}(t_{i_0}, T_\alpha^\beta, \underline{\mathbf{G}}) = (X^T \Gamma_{11}^{(\alpha)} X)^{-1} x_{i_0} \left[\frac{(t_{i_0} - \mathrm{e}^{(x_{i_0}^T \beta)})}{(t_{i_0}!)^\alpha} \mathrm{e}^{\alpha[t_{i_0}(x_{i_0}^T\beta) + \mathrm{e}^{(x_{i_0}^T\beta)}]} - \gamma_{1 i_0} \right].$$

Clearly, whenever the inverse of the first matrix exists, this influence function is bounded in t_{i_0} for any $\alpha > 0$ implying the robustness of the MDPDE with $\alpha > 0$. However, at $\alpha = 0$, $\gamma_{1 i_0} = 0$ and hence the influence function above further simplifies to

$$\mathrm{IF}_{i_0}(t_{i_0}, T_0^\beta, \underline{\mathbf{G}}) = (X^T \Gamma_{11}^{(0)} X)^{-1} x_{i_0} (t_{i_0} - \mathrm{e}^{(x_{i_0}^T \beta)}),$$

which is linear and hence unbounded in t_{i_0}. This indicates the non-robustness of the MLE and equivalently OLS of the regression parameter in case of the Poisson regression model. Figure 1 shows the influence function of the MDPDE for different α under several specific Poisson regression models and for sample size $n = 50$; the same for $n = 100$ is presented in the Supplementary material. The redescending nature of the influence function with increasing α is quite clear in all the figures.

Although these implications are visible in Table 1 and Fig. 1, it may be of importance to highlight them clearly in the text. There is a clear trade-off between robustness and efficiency over increasing α in this context. Small values of α provide a high degree of asymptotic efficiency; large values of α provide greater bounded-influence robustness as is evidenced by their highly stable influence functions.

4 Special Case II: logistic regression for binary data

Another important special case of the GLM is the logistic regression model which is used to model any categorical or binary dependent variable in terms of some explanatory variable. Given the value of the explanatory variable x_i, the binary outcome variable y_i (or the binary transform of the categorical variable) is assumed to follow a Bernoulli distribution with success probability π_i depending on the explanatory variable x_i (for each $i = 1, \ldots, n$). To ensure that the predicted values of π_i are in the interval $(0, 1)$, in the logistic model it is assumed that

$$\pi_i = \pi(x_i) = \frac{\mathrm{e}^{x_i^T \beta}}{1 + \mathrm{e}^{x_i^T \beta}}.$$

We will now assume that the x_is are fixed and consider the logistic regression model from its design perspective to estimate β efficiently and robustly.

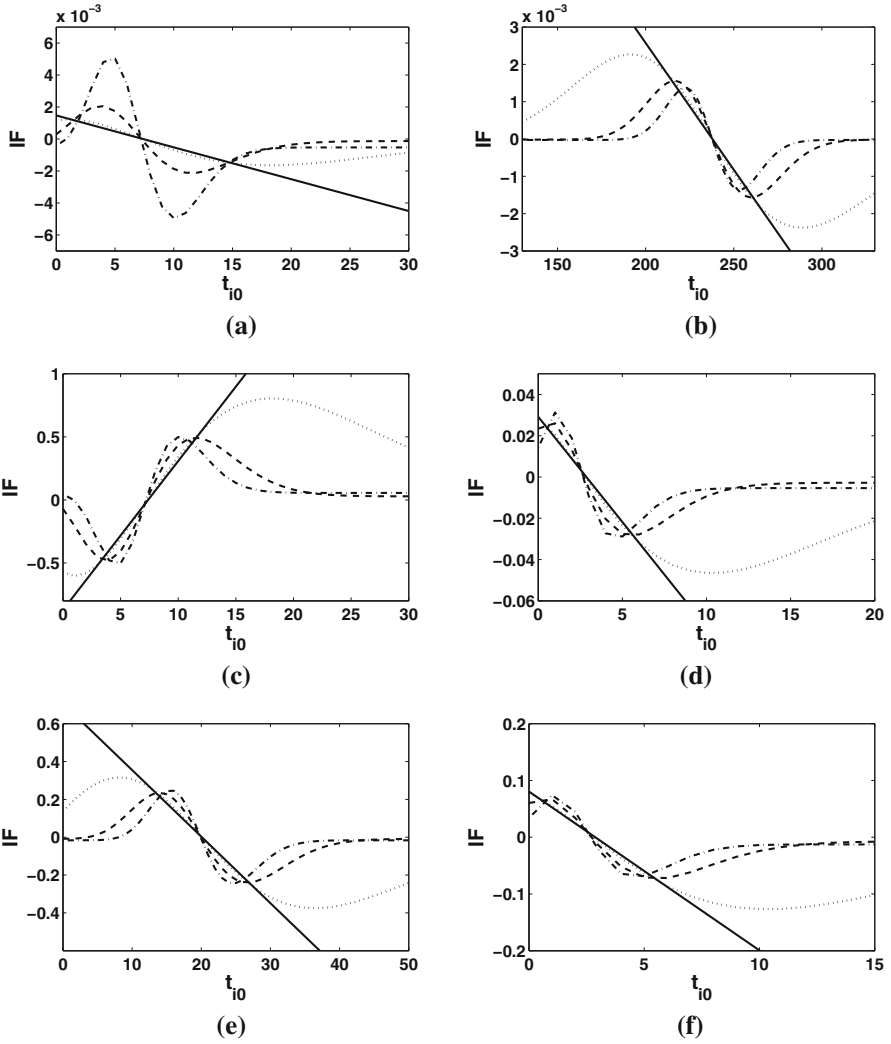

Fig. 1 Plot of the influence function of MDPDEs of the slope parameter β_1 for different α (*solid line* $\alpha = 0$, *dotted line* $\alpha = 0.1$, *dashed line* $\alpha = 0.5$ and *dashed-dotted line* $\alpha = 1$) and direction i_0 of contamination with $n = 50$. Here, Model (I)–(III) have $x_i = (1, \sqrt{i})^T$, $x_i = \left(1, \frac{1}{i}\right)^T$ and $x_i = \left(1, \frac{1}{i}, \frac{1}{i}\right)^T$, respectively, with $\beta_j = 1$ for all j. **a** Model I, $i_0 = 1$. **b** Model I, $i_0 = 20$. **c** Model II, $i_0 = 1$. **d** Model II, $i_0 = 20$. **e** Model III, $i_0 = 1$. **f** Model III, $i_0 = 20$

4.1 The MDPDE for logistic regression

We can treat the logistic regression model as a particular case of the GLM with known shape parameter $\phi = 1$ and $\theta_i = \eta_i = x_i^T \beta$, $c(y_i) = 0$. The distribution of y_i is Bernoulli with mean $\mu_i = \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, and $\text{var}(y_i) = \pi_i(1 - \pi_i) = \frac{e^{\eta_i}}{(1+e^{\eta_i})^2}$. Thus, the link function g is the logit function and so we can use the minimum DPD estimation

procedure discussed in Sect. 2 to estimate β robustly. Using the above notations and the form of the Bernoulli distribution, the minimum DPD estimating equation for $\alpha \geq 0$ is given by,

$$\sum_{i=1}^{n}\left[\frac{e^{x_i^T\beta}(e^{\alpha(x_i^T\beta)}-1)}{(1+e^{x_i^T\beta})^{2+\alpha}}-\left(y_i-\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)\frac{e^{\alpha(x_i^T\beta)y_i}}{(1+e^{x_i^T\beta})^{\alpha}}\right]x_i=0, \tag{11}$$

which can be further simplified to

$$\sum_{i=1}^{n}(1-2y_i)e^{(x_i^T\beta)(1-y_i)}\,\frac{(e^{\alpha(x_i^T\beta)}+e^{x_i^T\beta})}{(1+e^{x_i^T\beta})^{2+\alpha}}x_i=0. \tag{12}$$

We can easily solve the above with respect to β to compute the MDPDE for any $\alpha \geq 0$. In particular, for $\alpha = 0$, Eq. (11) simplifies to

$$\sum_{i=1}^{n}\left(y_i-\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)x_i=0, \tag{13}$$

which is the maximum likelihood estimating equation. Once again the minimum DPD estimating equation is just a generalization of the maximum likelihood estimating equation.

4.2 Properties of the MDPDE

We will now present the asymptotic distribution of the MDPDE of β in the logistic regression case as it follows from Theorem 1. In this special case, we have the following result.

Corollary 2 *Under Assumptions (A1)–(A7) of* Ghosh and Basu (2013)*, there exists a consistent sequence $\hat{\beta}_n = \hat{\beta}_n^{(\alpha)}$ of roots to the minimum DPD estimating Eq. (12) at the tuning parameter α. Also, the asymptotic distribution of*

$$\left(\sum_{i=1}^{n}e^{x_i^T\beta^g}\frac{(e^{\alpha(x_i^T\beta^g)}+e^{x_i^T\beta^g})^2}{(1+e^{x_i^T\beta^g})^{4+2\alpha}}(x_ix_i^T)\right)^{-\frac{1}{2}}$$
$$\times\left(\sum_{i=1}^{n}e^{x_i^T\beta^g}\frac{(e^{\alpha(x_i^T\beta^g)}+e^{x_i^T\beta^g})}{(1+e^{x_i^T\beta^g})^{3+\alpha}}(x_ix_i^T)\right)(\hat{\beta}_n-\beta^g)$$

is p-dimensional normal with mean 0 and variance I_p.

As argued in Sect. 3.2 for the Poisson regression, the asymptotic efficiency of the different MDPDE $\hat{\beta}_n = \hat{\beta}_n^{(\alpha)}$ of β for the logistic regression can also be measured in terms of its asymptotic variance $\mathrm{AV}_\alpha(\beta^g)$, which can be again estimated consistently by $\widehat{\mathrm{AV}}_\alpha = \mathrm{AV}_\alpha(\hat{\beta}_n)$.

As in the Poisson regression case, here also we can compute the values of relative efficiencies of the MDPDEs of the coefficients of the logistic regression model based on $\widehat{\mathrm{AV}}_\alpha$. This measure of relative efficiency clearly depends on the value of β and X_is. We present the empirical estimate of the relative efficiencies of the MDPDE in case of the logistic regression model in Table 2 for sample size $n = 50$ and the same for $n = 100$ is presented in the Supplementary material. These are calculated based on a simulation study with 1000 replications under several different cases of logistic regressions. These cases are defined based on the given values of the explanatory variables x_i $(i = 1, \ldots, n)$ as in the case of Poisson regression, but now with the true regression coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ being $(0.1, 0.1)$, $(0.001, 0.0001)$, $(1, 1)$, $(0.1, 0.1)$, $(0.1, 0.1, 0.1)$ and $(0.01, 0.001, 0.0001)$, respectively, for Cases I–VI. It is clearly seen from the tables that for any value of the parameter and the explanatory variables, the loss of efficiency is negligible for small $\alpha > 0$. Further, if the values of $x_i^T \beta$ is small, then we can get quite high efficiency even for large positive α near 0.5.

5 Real data examples

In this section, we will explore the performance of the proposed MDPDEs in Poisson and logistic regression models by applying it on two interesting real data sets. Application to several other real data sets are presented in the supplementary material. For all the applications, the estimators are computed by minimizing the corresponding objective function through the software "R"; the minimization is performed using the "optim" function of R under suitable convergence criteria. The "R" code used is available from the authors.

5.1 Epilepsy data

First we consider an interesting data set consisting of 59 epilepsy patients from Thall and Vail (1990). The data were obtained from a clinical trial carried out by Leppik et al. (1985) where the patients were treated by the anti-epileptic drug "progabide" or a placebo with randomized assignment. Then the total number of epilepsy attacks was noted which we model by an appropriate set of explanatory variables through a Poisson regression model (Hosseinian 2009). The variables considered in this regard are "Base", the eight-week baseline seizure rate prior to randomization in multiples of 4, "Age", the patient's age in multiple of 10 years, and "Trt", a binary indicator for the treatment–control group. Also, the interaction between treatment and baseline seizure rate is important in this case, because it represents either higher or lower seizure rate for the treatment group compared to the placebo group depending on the baseline count. In fact, the drug decreases the epilepsy only if the baseline count becomes sufficiently large in number with respect to some critical threshold.

The data were also analyzed by Hosseinian (2009) who compared the maximum likelihood estimator with the robust methodologies proposed by herself in the same paper and those by Cantoni and Ronchetti (2001). There it was observed that the data contain some outlying observations due to which the interaction effect between treatment and baseline seizure rate turns out to be insignificant based on the maximum

Table 2 The estimated relative efficiencies of the MDPDE for various values of the tuning parameter α under different cases of logistic regression with sample size $n = 50$

Case	Coefficients	$\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
I	β_0	100.0	99.0	90.7	74.6	67.6	61.3	50.4	37.5
	β_1	100.0	99.2	92.7	79.6	73.8	68.4	58.7	46.7
II	β_0	100.0	99.3	93.3	81.2	75.8	70.7	61.5	50.0
	β_1	100.0	99.3	93.3	81.2	75.8	70.7	61.5	50.0
III	β_0	100.0	98.6	86.7	65.2	56.5	49.0	36.8	23.9
	β_1	100.0	98.1	82.8	56.9	47.2	39.2	27.1	15.6
IV	β_0	100.0	99.3	92.8	79.8	74.0	68.7	59.1	47.2
	β_1	100.0	99.2	92.4	79.0	73.0	67.5	57.7	45.6
V	β_0	100.0	99.3	92.7	79.8	74.0	68.6	59.0	47.1
	β_1	100.0	99.2	92.6	79.4	73.5	68.1	58.4	46.3
	β_2	100.0	99.2	92.4	78.9	72.9	67.4	57.5	45.4
VI	β_0	100.0	99.3	93.3	81.1	75.6	70.5	61.3	49.7
	β_1	100.0	99.3	93.3	81.1	75.6	70.5	61.3	49.7
	β_2	100.0	99.3	93.3	81.1	75.6	70.5	61.3	49.7

Table 3 The minimum density power divergence estimates, their standard errors and p values for the epilepsy data

	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
Intercept						
Estimate	1.9888	2.1089	1.9106	1.9691	2.0060	1.9653
SE ($\times 100$)	13.6518	15.2509	12.6869	13.7081	14.9185	17.0043
p value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Trt						
Estimate	−0.2375	−0.3169	−0.3871	−0.3893	−0.3516	−0.3186
SE ($\times 100$)	7.6816	8.6812	7.9139	8.4566	9.1111	10.1787
p value	0.0030	0.0006	0.0000	0.0000	0.0003	0.0027
Base						
Estimate	0.0858	0.0866	0.1689	0.1631	0.1622	0.1562
SE ($\times 100$)	0.3698	0.4101	0.2778	0.3055	0.3359	0.3959
p value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Age						
Estimate	0.2308	0.1153	0.0408	0.0362	0.0242	0.0559
SE ($\times 100$)	4.1498	4.7242	3.9374	4.2416	4.6138	5.2119
p value	0.0000	0.0177	0.3045	0.3972	0.6017	0.2878
Trt $\times$ Base						
Estimate	0.0069	0.0107	0.0156	0.0165	0.0131	0.0098
SE ($\times 100$)	0.4443	0.4893	0.3230	0.3537	0.3888	0.4600
p value	0.1283	0.0323	0.0000	0.0000	0.0013	0.0373

likelihood estimator whereas the robust estimators show this interaction to be significant. Here, we will apply our proposed robust minimum density power divergence estimators for this epilepsy data set and try to see if our proposed estimators are also robust enough to differentiate with maximum likelihood estimator for the interaction effect.

Table 3 presents the parameter estimates, their asymptotic standard errors and corresponding p values based on the minimum density power divergence estimator with different α. Clearly the estimators corresponding to $\alpha \geq 0.3$ are quite different from the maximum likelihood estimator and for these estimators the interaction effect is also significant under the Poisson regression model. Indeed, these estimators are quite similar to the robust estimators considered in Hosseinian (2009) but, as we have described earlier, have superior asymptotic properties.

5.2 Damaged carrots data

As an interesting data example leading to the logistic regression model, we consider the damaged carrots dataset of Phelps (1982). The data set was obtained from a soil experiment trial containing the proportion of insect-damaged carrots with three blocks

Table 4 The minimum density power divergence estimates, their standard errors and p values for the damaged carrots data

	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$
Intercept						
Estimate	1.4805	1.4880	1.4974	1.5157	1.5310	1.5569
SE	0.6562	0.6648	0.6859	0.7118	0.7406	0.7868
p value	0.0339	0.0352	0.0395	0.0441	0.0501	0.0599
Logdose						
Estimate	−1.8175	−1.8163	−1.8102	−1.8102	−1.8102	−1.8152
SE	0.3439	0.3484	0.3601	0.3749	0.3917	0.4183
p value	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
Block1						
Estimate	0.5421	0.5330	0.5149	0.4969	0.4824	0.4654
SE	0.2318	0.2338	0.2392	0.2462	0.2542	0.2668
p value	0.0284	0.0322	0.0421	0.0554	0.0704	0.0945
Block2						
Estimate	0.8430	0.8284	0.7973	0.7710	0.7483	0.7240
SE	0.2260	0.2283	0.2344	0.2422	0.2510	0.2649
p value	0.0011	0.0014	0.0025	0.0041	0.0067	0.0119

and eight dose levels of insecticide in the experiments and was discussed by Williams (1987). McCullagh and Nelder (1989) used these data to illustrate the identification methods for isolated departures from the model through an outlier in the y-space present in the data (14th observation; dose level 6 and block 2). Later Cantoni and Ronchetti (2001) modeled these data by a binomial logistic model to illustrate the performance of their proposed robust estimators. However, it can be checked easily that the observation 14 is only an outlier in the y-space and not a leverage point.

We now apply the minimum density power divergence estimation method for several different α to explore the performance of the proposed method in case of the presence of outlier only in the y-space. Table 4 presents the parameter estimates, their asymptotic standard errors and corresponding p values for different tuning parameters α. The estimates corresponding to $\alpha \geq 0.3$ again turn out to be highly robust and also similar to the robust estimator obtained by Cantoni and Ronchetti (2001). Also, for these estimators the indicator of Block 1 turns out to be insignificant which became significant in case of the maximum likelihood estimator (corresponding to $\alpha = 0$) due to the presence of the outlying observation.

6 Comparison with existing robust estimators in GLM

Here, we briefly consider a comparison of our proposed estimators with some existing robust estimators. As noted previously, there are few robust inference procedures in the literature of GLM; only the Poisson, logistic and negative binomial regression models

with stochastic covariates have got some attention. On the contrary, our proposal considers non-stochastic covariates and, therefore, is not theoretically comparable to the existing methods. However, from a practical point of view they can be adapted to solve real-life problems with fixed covariates and hence numerical comparisons can be of some interest.

Two existing methods appear to be close to our proposal in the sense of bounding the Pearson residual. One is the approach of Hosseinian (2009) who has proposed weighted likelihood-type robust estimators by following the L_q quasi-likelihood approach of Morgenthaler (1992); the other is by Cantoni and Ronchetti (2001) who have considered a class of Mallows-type M-estimators as a special case of the generalized estimating equation of Preisser and Qaqish (1999). The second work is itself a special case of Cantoni (2004).

Hosseinian (2009) only proposed robust estimators for the Poisson and logistic regression cases and provided no general form for all GLMs. Further, the proposed estimating equations in Hosseinian (2009) are not asymptotically unbiased implying an inconsistent estimator. Our proposal does not have this theoretical flaw and, in addition, is completely general. Accordingly, further comparison with the Hosseinian (2009) work does not appear to be useful.

On the contrary, the goal of the Cantoni and Ronchetti (2001) work was not to just introduce a new robust estimator for GLM; rather it aims to develop a comprehensive robust analysis (estimation, testing and model selection through the analysis of deviance) that would complement the classical analysis. Their estimators (and those proposed in the current paper) have unbiased estimating equations at the model; hence it is easy to establish the theoretical consistency results of these estimators unlike the case of Hosseinian (2009).

On the robustness issue, the estimators proposed in Cantoni and Ronchetti (2001) have bounded-influence functions which is also the case with our proposed MDPDEs. Our estimators appear to have competitive or better robustness properties compared to Cantoni and Ronchetti (2001). For illustration, let us consider the Epilepsy Data example modeled by Poisson regression. The analysis based on the proposed MDPDE has been presented in Table 3, which shows that the MDPDE with $\alpha \in [0.3, 0.7]$ can successfully ignore the outliers in the data and generate robust insights. In this example, the effect of the outliers is actually on the significance of coefficients of the variables "Age" and "Trt $\times$ Base". Our analysis shows that while the "Age" variable is significant for $\alpha = 0$, this false significance is quickly turned around by moderate values of α. Similarly the true significance of the coefficient of "Trt $\times$ Base" is masked at $\alpha = 0$, but clearly observed at larger values of α. The MDPDE for the coefficient of "Age" for $\alpha \in [0.3, 0.7]$ vary from 0.04 to 0.02 with p values of the order of 0.3, and those for "Trt $\times$ Base" ranges from 0.016 to 0.013 with p values less than 10^{-4}. The coefficients of "Age" and "Trt $\times$ Base" obtained by the techniques of Cantoni and Ronchetti (2001) are 0.16 and 0.012 with p value of 0.0008 and 0.02, respectively (Hosseinian 2009, Table 12, p. 125). Therefore, the proposed MDPDEs with moderate α seem to produce more robust/competitive estimators compared to the Cantoni and Ronchetti (2001) estimators. Similar results can be observed for the other real data examples.

We hope to conduct an extensive simulation study in the future to get a more comprehensive idea about the comparisons of the proposed estimators with all the estimators mentioned in the Sect. 1 over the twin goals of efficiency and robustness.

7 Conclusions

In this paper, we have proposed a new general methodology for robust estimation in case of generalized linear models and considered two prominent special cases—Poisson regression and logistic regression. We have established the robustness properties of the proposed method in terms of the influence function analysis and applied it to several real data sets having different types of outliers. Our method appears to perform competitively in comparison with existing techniques in terms of its robustness properties and capability of generalization to all GLMs. Our method is also a bona fide optimization procedure; selecting the correct solution is, therefore, easier than the estimating equation-based competitors. On the whole, we expect that the proposed estimators will help the researchers in several application domains to estimate the model parameters in any generalized linear model efficiently and robustly.

Acknowledgments The authors thank the editor, the associate editor and three anonymous referees for several useful suggestions that led to an improved version of the manuscript.

References

Aeberhard WH, Cantoni E, Heritier S (2014) Robust inference in the negative binomial regression model with an application to falls data. Biometrics 70(4):920–931. doi:10.1111/biom.12212

Basu A, Harris IR, Hjort NL, Jones MC (1998) Robust and efficient estimation by minimizing a density power divergence. Biometrika 85(3):549–559. doi:10.1093/biomet/85.3.549

Bianco AM, Boente G, Rodrigues IM (2013) Resistant estimators in Poisson and Gamma models with missing responses and an application to outlier detection. J Multivar Anal 114:209–226. doi:10.1016/j.jmva.2012.08.008

Bianco AM, Yohai VJ (1996) Robust estimation in the logistic regression model. In: Rieder H (ed) Robust statistics, data analysis, and computer intensive methods. Springer, New York, pp 17–34

Cantoni E (2004) A robust approach to longitudinal data analysis. Can J Stat 32(2):169–180. doi:10.2307/3315940

Cantoni E, Ronchetti E (2001) Robust inference for generalised linear models. J Am Stat Assoc 96(455):1022–1030. doi:10.1198/016214501753209004

Croux C, Haesbroeck G (2003) Implementing the Bianco and Yohai estimator for logistic regression. Comput Stat Data Anal 44(1–2):273–295. doi:10.1016/S0167-9473(03)00042-2

Ghosh A, Basu A (2013) Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. Electron J Stat 7:2420–2456. doi:10.1214/13-EJS847

Ghosh A, Basu A (2015) Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach. J Appl Stat. doi:10.1080/02664763.2015.1016901

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Heritier S, Cantoni E, Copt S, Victoria-Feser MP (2009) Robust methods in biostatistics. Wiley, New York

Hong C, Kim Y (2001) Automatic selection of the tuning parameter in the minimum density power divergence estimation. J Korean Stat Soc 30:453–465

Hosseinian S (2009) Robust inference for generalized linear models: binary and poisson regression. Thesis, Ecole Polytechnique Federal de Lausanne

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35(1):73–101. doi:10.1214/aoms/1177703732

Huber PJ (1983) Minimax aspects of bounded-influence regression (with discussion). J Am Stat Assoc 78:66–80. doi:10.1080/01621459.1983.10477928

Leppik IE et al (1985) A double-blind crossover evaluation of progabide in partial seizures. Neurology 35:285

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Morgenthaler S (1992) Least-absolute-deviations fits for generalized linear models. Biometrika 79(4):747–754. doi:10.1093/biomet/79.4.747

Phelps K (1982) Use of the complementary log-log function to describe dose–response relationships in insecticide evaluation field trials. In: Gilchrist R (ed) Lecture notes in statistics, vol. 14. GLIM.82: Proceedings of the international conference on generalized linear models. Springer, New York

Preisser JS, Qaqish BF (1999) Robust regression for clustered data with applications to binary regression. Biometrics 55:574–579. doi:10.1111/j.0006-341X.1999.00574.x

Thall PF, Vail SC (1990) Some covariance models for longitudinal count data with overdispersion. Biometrics 46(3):657–671

Warwick J, Jones MC (2005) Choosing a robustness tuning parameter. J Stat Comput Simul 75:581–588. doi:10.1080/00949650412331299120

Williams DA (1987) Generalised linear model diagnostics using the deviance and single case deletions. Appl Stat 36:181–191