Robust and sparse logistic regression

Abstract

Logistic regression is one of the most popular statistical techniques for solving (binary) classification problems in various applications (e.g. credit scoring, cancer detection, ad click predictions and churn classification). Typically, the maximum likelihood estimator is used, which is very sensitive to outlying observations. In this paper, we propose a robust and sparse logistic regression estimator where robustness is achieved by means of the γ-divergence. An elastic net penalty ensures sparsity in the regression coefficients such that the model is more stable and interpretable. We show that the influence function is bounded and demonstrate its robustness properties in simulations. The good performance of the proposed estimator is also illustrated in an empirical application that deals with classifying the type of fuel used by cars.

Keywords: elastic net, γ-divergence, logistic regression, robustness, sparsity

1 Introduction

Logistic regression is a powerful way of describing a binary response variable depending on multiple explanatory variables. However, the classical maximum likelihood estimator breaks down when the data set contains contaminated observations (Carroll and Pederson, 1993; Bianco and Yohai, 1996) or when the number of possible explanatory variables is too large. To remedy the former, robust regression techniques have been proposed where the influence of each observation is bounded (see e.g. Morgenthaler, 1992; Croux and Haesbroeck, 2003; Bondell, 2005, 2008; Hosseinian and Morgenthaler, 2011; Hung et al, 2018). In order to regularize the method in high dimensions for variable selection, the elastic net penalty successfully shrinks the redundant variables to zero at the cost of a bias in the relevant regression coefficients (Zou and Hastie, 2005). While this solves the issue of variable selection, the elastic net estimator is not robust (Öllerer et al, 2015). Ideally, the regularization method which selects the relevant explanatory variables is robust to contamination in both the explanatory and response variables.

Hoffmann et al (2016) approach this problem by means of partial least squares and Kurnaz et al (2018) using M-estimation. Recently, Ponnet et al (2023) introduced a penalized robust double exponential estimator for generalized linear models with varying dispersion, which in absence of dispersion reduces to a penalized version of the robust estimator of Cantoni and Ronchetti (2001) studied by Avella-Medina and Ronchetti (2018). An alternative approach to robustness is by means of the γ-divergence for regression (Fujisawa and Eguchi, 2008), which generalizes the Kullback-Leibler divergence associated with maximum likelihood estimation. Kawashima and Fujisawa (2017) proposed a sparse robust linear regression technique via γ-divergence by adding an L_1-penalty term. Kawashima and Fujisawa (2019) introduced an algorithm for sparse generalized linear models based on stochastic optimization for data sets with more observations than explanatory variables. Kawashima and Fujisawa (2022) introduced an MM algorithm for

logistic γ-divergence regression based on data with more observations than explanatory variables.

In this paper, we propose a robust and sparse logistic regression estimator that selects the important variables by means of the elastic net penalty (Zou and Hastie, 2005). The proposed estimator is robust to outliers in both the predictor space and response variables as is shown by its bounded influence function. Simulations confirm the good properties and we demonstrate the usefulness in a practical application: predicting the type of fuel used by cars.

This paper is organised as follows. Section 2 introduces the estimator and proposes an efficient algorithm. In Section 3 we study its robustness by deriving the influence function. Section 4 contains the simulation study and Section 5 the empirical application. The conclusion in Section 6 ends the paper.

2 Robust and sparse logistic regression

In this section, we propose a robust and sparse logistic regression estimator based on the γ-divergence and provide an efficient algorithm to solve the optimization problem in practice.

2.1 Definition

We propose a robust and sparse logistic regression based on minimizing the γ-divergence for regression (Fujisawa and Eguchi, 2008) with an elastic net penalty term. This penalty term regularizes the regression problem and selects the important variables (Zou and Hastie, 2005).

For the uncontaminated data we assume that the conditional density function of the binary response y, with values 0 and 1, given the explanatory variables $\boldsymbol{x}$ is given by

$$f(y \mid \boldsymbol{x}) = \frac{\exp(y\boldsymbol{\theta}'\boldsymbol{x})}{1+\exp(\boldsymbol{\theta}'\boldsymbol{x})}. \tag{1}$$

If necessary, both $\boldsymbol{x}$ and $\boldsymbol{\theta}$ may include an intercept term. However, the actual observations may contain data errors and anomalies. This observed

distribution is represented by $g(\boldsymbol{x}, y)$ with conditional density

$$g(y \mid \boldsymbol{x}) = (1 - \varepsilon(\boldsymbol{x}))f(y \mid \boldsymbol{x}) + \varepsilon(\boldsymbol{x})\delta(y \mid \boldsymbol{x}), \tag{2}$$

where δ is the degenerate distribution and ε is a non-homogeneous contamination rate.

The γ-divergence for regression measures the distance between the observed conditional distribution g and the parametric distribution f.

Definition 2.1 (γ-divergence for logistic regression (Fujisawa and Eguchi (2008))**.** *For $\gamma > 0$, the γ-divergence for logistic regression is defined as*

$$\begin{aligned} D_\gamma(g(y \mid \boldsymbol{x}), f(y \mid \boldsymbol{x})) = \frac{1}{\gamma} \log \int \left(\int g(y \mid \boldsymbol{x})^{1+\gamma} dy \right)^{\frac{1}{1+\gamma}} g(\boldsymbol{x}) d\boldsymbol{x} \\ - \frac{1}{\gamma} \log \int \int \left(\frac{\exp(y(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})}{1 + \exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})} \right)^{\frac{\gamma}{1+\gamma}} g(\boldsymbol{x}, y) d\boldsymbol{x} dy. \end{aligned} \tag{3}$$

Fujisawa and Eguchi (2008) show several good properties of the γ-divergence under different types of contamination. The natural robustness to outliers makes this measure ideal for robust regression. The idea is to search for the parameters $\boldsymbol{\theta}$ that minimize the γ-divergence between the observed distribution $g(y \mid \boldsymbol{x})$ and the assumed parametric distribution $f(y \mid \boldsymbol{x})$. Since the first term in (3) does not depend on the parameters $\boldsymbol{\theta}$, this term can be ignored in the optimization. In addition to robustness, the proposed regression method behaves well when the number of explanatory variables exceeds the number of observations. This is achieved by adding to the γ-divergence an elastic net penalty, defined as

$$P_\alpha(\boldsymbol{\theta}_1) = (1 - \alpha)\frac{\|\boldsymbol{\theta}_1\|_2^2}{2} + \alpha\|\boldsymbol{\theta}_1\|_1, \tag{4}$$

with $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1)$ where $\boldsymbol{\theta}_1$ is the vector of regression coefficients without the intercept θ_0.

Formally, we propose the robust sparse logistic (RoSLog) estimator as follows.

Definition 2.2 (RoSLog estimator)**.** *For $\gamma, \lambda > 0$, the parameters of the robust sparse logistic regression via γ-divergence are determined by*

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_\gamma &= \arg\min_{\boldsymbol{\theta}} \widehat{D}_\gamma(g(y \mid \boldsymbol{x}), f(y \mid \boldsymbol{x})) + \lambda P_\alpha(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} -\frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp(y_i(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}\right)^{\frac{\gamma}{1+\gamma}}\right) + \lambda P_\alpha(\boldsymbol{\theta}_1).\end{aligned} \tag{5}$$

We remark that for the case $\gamma = 0$, the RoSLog estimator coincides with the elastic net logistic regression estimator of Zou and Hastie (2005). In case $\gamma = \lambda = 0$, the proposed estimator reduces to the classical logistic regression estimator.

2.2 Parameter estimation

A solution to (5) can be obtained by means of a majorization-minimization (MM) algorithm (Ortega and Rheinboldt, 1970), where the majorization step in each iteration is based on the inequality

$$\begin{aligned}&-\frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp(y_i(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}\right)^{\frac{\gamma}{1+\gamma}}\right) \\ &\qquad \le -\frac{1}{\gamma}\sum_{i=1}^{n} w_i^{(m)}\log\left(\left(\frac{\exp(y_i(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x}_i)}\right)^{\frac{\gamma}{1+\gamma}}\right) + C,\end{aligned} \tag{6}$$

with C a constant. The weights $w_i^{(m)}$ are given by

$$w_i^{(m)} = \frac{1}{\sum_{j=1}^{m}\left(\frac{\exp(y_j(1+\gamma)\boldsymbol{\theta}'_{(m)}\boldsymbol{x}_j)}{1+\exp((1+\gamma)\boldsymbol{\theta}'_{(m)}\boldsymbol{x}_j)}\right)^{\frac{\gamma}{1+\gamma}}}\left(\frac{\exp(y_i(1+\gamma)\boldsymbol{\theta}'_{(m)}\boldsymbol{x}_i)}{1+\exp((1+\gamma)\boldsymbol{\theta}'_{(m)}\boldsymbol{x}_i)}\right)^{\frac{\gamma}{1+\gamma}}, \tag{7}$$

$i = 1, \ldots, n.$

These weights depend on the current regression estimate $\boldsymbol{\theta}_{(m)}$, obtained from the previous iteration.

Then, the minimization step in each iteration amounts to solving the following equation based on the current estimate $\boldsymbol{\theta}_{(m)}$:

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} -\sum_{i=1}^{n} w_i^{(m)}\log\left(\frac{\exp(y_i\boldsymbol{\beta}'\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right) + \lambda^\star P_{\alpha^\star}(\boldsymbol{\beta}_1). \tag{8}$$

Finally, the regression parameters are updated by $\boldsymbol{\theta}_{(m+1)} = \boldsymbol{\beta}/(1+\gamma)$. We remark that (8) contains the transformed penalty parameters $\lambda^\star$ and $\alpha^\star$ given by

$$\lambda^\star = \lambda \frac{1+\alpha\gamma}{1+\gamma} \qquad \text{and } \alpha^\star = \alpha \frac{1+\gamma}{1+\alpha\gamma}. \tag{9}$$

The MM step can be solved efficiently by existing algorithms as proposed in Friedman et al (2010). Pseudo-code for the RoSLog estimator is given in Algorithm 1.

Algorithm 1 MM algorithm for robust sparse logistic regression

1: Initial estimate $\boldsymbol{\theta}_{(0)}$
2: **while** not converged **do**
3: update weights $w_i^{(m)}, i = 1, \ldots, n$ (see (7))
4: solve $\boldsymbol{\beta}^\star = \arg\min_{\boldsymbol{\beta}} - \sum_{i=1}^n w_i^{(m)} \log\left(\frac{\exp(y_i \boldsymbol{\beta}' \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}\right) + \lambda^\star P_{\alpha^\star}(\boldsymbol{\beta}_1)$
5: set $\boldsymbol{\theta}_{(m+1)} = \frac{1}{1+\gamma}\boldsymbol{\beta}^\star$
6: **end while**
7: **end**

2.3 Initial estimate

The MM algorithm proposed above requires a good initial estimate. We follow Valdiviezo and Van Aelst (2019) and consider the following starting values for the procedure. Let $\boldsymbol{U}_1$ be the matrix with observations that are standardized marginally with a robust location and scale estimator. Moreover, consider the matrix $\boldsymbol{U}_2$ where the hyperbolic tangent is computed on each column of $\boldsymbol{U}_1$, after which the columns are again robustly standardized. Applying the traditional elastic net estimator on these two data sets as well as the original data set provides three deterministic starting values for Algorithm 1. In addition, the enetLTS estimate of Kurnaz et al (2018) is used as a starting value, yielding four starting values for the MM algorithm. Algorithm 1 is run with each of these four starting values separately and the solution with the smallest value of the objective function is returned as final estimates.

2.4 Selection of tuning parameters

The tuning parameters are selected by k-fold cross-validation, as commonly used in the literature. The data is split into k groups of approximately equal size and the RoSLog estimator is performed k times, each time leaving out one of the k groups. This is done for each combination of α and λ and the robust cross-validation metric equals

$$\mathrm{CV}(\alpha,\lambda) = -\frac{1}{\gamma}\log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\exp(y_i(1+\gamma)\widehat{\boldsymbol{\theta}}'_{[-i]}\boldsymbol{x}_i)}{1+\exp((1+\gamma)\widehat{\boldsymbol{\theta}}'_{[-i]}\boldsymbol{x}_i)}\right)^{\frac{\gamma}{1+\gamma}}\right), \tag{10}$$

where $\widehat{\boldsymbol{\theta}}_{[-i]}$ is the parameter estimate obtained when the i-th observation was in the left out group. Note that this cross-validation procedure uses the loss function in (5) to measure the quality of fit for each observation in the left out group. This guarantees that the cross-validation procedure selects the tuning constants with the same level of robustness as for the estimation of the regression parameters. Similar robust cross-validation procedures have been used by e.g. Khan et al (2010) and Bianco et al (2022). The values of α and λ that minimize the cross-validation metric are taken as the optimal choices.

3 Properties

In this section, we assume smoothness conditions on the penalty function $P_\alpha(\boldsymbol{\theta}_1)$, as in Öllerer et al (2015). Denote by $\nabla P_\alpha(\boldsymbol{\theta}_1)$ and $\nabla^2 P_\alpha(\boldsymbol{\theta}_1)$ the gradient and Hessian of P_α with respect to $\boldsymbol{\theta}_1$.

We measure robustness to outliers by means of the influence function, see for example Hampel et al (2011). The influence function describes the effect of infinitesimal, pointwise contamination at $(\boldsymbol{x}_0, y_0)$ on the estimator and is defined as

$$\mathrm{IF}((\boldsymbol{x}_0,y_0),\boldsymbol{\theta}_\gamma,F) = \frac{\partial}{\partial\varepsilon}\left[\boldsymbol{\theta}_\gamma\left(F_\varepsilon\right)\right]\Big|_{\varepsilon=0}, \tag{11}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{(\boldsymbol{x}_0,y_0)}$ is the contaminated distribution and F is the joint cumulative distribution function of the logistic model for $(\boldsymbol{X}, Y)$ in (1).

Here, we defined $\boldsymbol{\theta}_\gamma$ in its functional form for some distribution H as

$$\boldsymbol{\theta}_\gamma(H) = \arg\min_{\boldsymbol{\theta}} -\frac{1}{\gamma}\log\left(\mathbb{E}_H\left[\left(\frac{\exp(Y(1+\gamma)\boldsymbol{\theta}'\boldsymbol{X})}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{X})}\right)^{\frac{\gamma}{1+\gamma}}\right]\right) + \lambda P_\alpha(\boldsymbol{\theta}_1). \tag{12}$$

First, we introduce some notation to improve the exposition of the influence function. Denote by $v(\boldsymbol{x}, y)$ the weight function

$$v(\boldsymbol{x}, y) = \left(\frac{\exp(y(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})}\right)^{\frac{\gamma}{1+\gamma}} \tag{13}$$

and $\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \exp(\boldsymbol{\theta}'\boldsymbol{x})/(1+\exp(\boldsymbol{\theta}'\boldsymbol{x}))$.

The influence function of the RoSLog estimator is given below.

Proposition 3.1 (Influence function)**.** *The influence function under the logistic model F for $(\boldsymbol{X}, Y)$, given γ, λ and α, equals*

$$\begin{aligned} &IF((\boldsymbol{x}_0, y_0), \boldsymbol{\theta}_\gamma, F) \\ &\quad = \boldsymbol{A}^{-1} v(\boldsymbol{x}_0, y_0)\left[\boldsymbol{x}_0\left(y_0 - \pi(\boldsymbol{x}_0, (1+\gamma)\boldsymbol{\theta}_\gamma(F))\right) - \lambda\nabla P_\alpha(\boldsymbol{\theta}_\gamma(F))\right], \end{aligned} \tag{14}$$

where

$$\begin{aligned} \boldsymbol{A} = &\mathbb{E}_F\left[v(\boldsymbol{X}, Y)\left(\widetilde{\pi}(1-\widetilde{\pi})\boldsymbol{X}\boldsymbol{X}' + \lambda\nabla^2 P_\alpha(\boldsymbol{\theta}_\gamma(F))\right)\right] \\ &+ \gamma\,\mathbb{E}_F\left[v(\boldsymbol{X}, Y)\left(\widetilde{\pi}(1-\widetilde{\pi})\boldsymbol{X}\boldsymbol{X}' - (Y-\widetilde{\pi})^2\boldsymbol{X}\boldsymbol{X}'\right.\right. \\ &\left.\left.+\lambda(Y-\widetilde{\pi})\nabla P_\alpha(\boldsymbol{\theta}_\gamma(F))\boldsymbol{X}'\right)\right], \end{aligned} \tag{15}$$

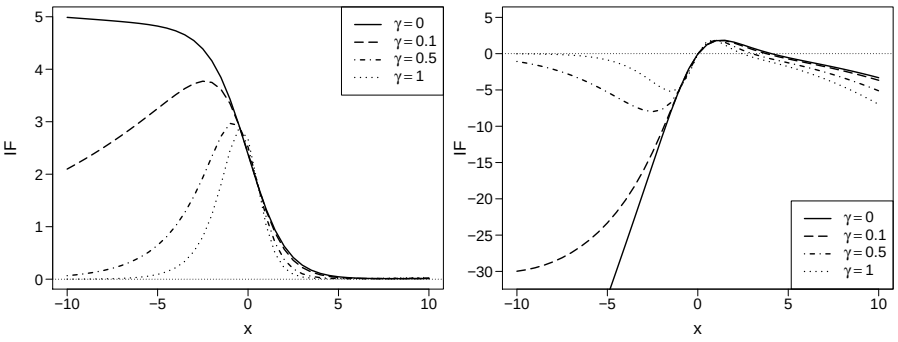
with $\widetilde{\pi} = \pi(\boldsymbol{X}, (1+\gamma)\boldsymbol{\theta}_\gamma(F))$.

When $\lambda = 0$, i.e. there is no sparsity penalty, then the second term in $\boldsymbol{A}$ vanishes because the bias on the estimator disappears. Moreover, if also $\gamma = 0$, then we recover the influence function of the maximum likelihood estimator for logistic regression.

Corollary 3.1. *The influence function of the RoSLog estimator is bounded. Moreover, for bad leverage points, the influence function redescends to zero, while for good leverage points the influence function is proportional to $\lambda\nabla P_\alpha(\boldsymbol{\theta}_\gamma(F))$.*

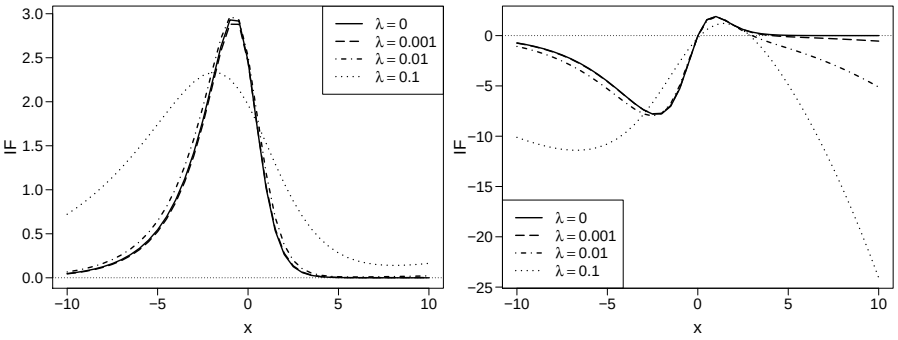
We illustrate Proposition 3.1 with an example. Consider the conditional density in (1) where there are 10 independent explanatory variables with standard Gaussian distribution. The true coefficients are $\theta_1 = 1$ and $\theta_j = 0$ for $j = 2, \ldots, 10$. We also include an intercept in the model whose true value θ_0 is zero.

Figures 1a and 1b show the influence function of the RoSLog estimator for θ_0 and θ_1 at point $(\boldsymbol{x}_0, 1)$, where $\boldsymbol{x}_0 = (x, \ldots, x)$ for the case $\lambda = 0.01$ and $\alpha = 1$. When $\gamma = 0$, Figure 1b indicates that the influence function of the slope is unbounded. However, when γ increases, so does the robustness,



(a) Sensitivity of $\widehat{\theta}_0$ with respect to γ ($\lambda = 0.01$).

(b) Sensitivity of $\widehat{\theta}_1$ with respect to γ ($\lambda = 0.01$).



(c) Sensitivity of $\widehat{\theta}_0$ with respect to λ ($\gamma = 0.5$).

(d) Sensitivity of $\widehat{\theta}_1$ with respect to λ ($\gamma = 0.5$).

Fig. 1: The influence function of the RoSLog estimator for the interecept θ_0 and nonzero slope θ_1 at point $(\boldsymbol{x}_0, 1)$, where $\boldsymbol{x}_0 = (x, \ldots, x)$ for the case $\alpha = 1$.

and the influence function becomes redescending for bad leverage points. For the estimator of the intercept, both good and bad leverage points have zero influence when γ is large enough, as seen in Figure 1a.

In Figures 1c and 1d, γ is fixed to 0.5. Clearly, low to moderate λ have little effect on the influence function for the intercept and non-zero slope. However, for a large λ, the influence function is temporarily distorted before being equal to zero when all regression coefficients are shrunk to zero.

Following Theorem 4 of Avella-Medina and Ronchetti (2018), it can be shown that the proposed estimator using the lasso penalty is consistent for variable selection.

4 Simulations

The simulation set-up is similar to Kurnaz et al (2018). We consider two set-ups, a low-dimensional one where $n = 150$ and $p = 50$ and a high-dimensional one where $n = 100$ and $p = 200$. In each setting, the intercept equals one, 10% of the explanatory variables has a regression coefficient of one, while the other 90% have a coefficient of zero. The covariance matrix of the informative variables Σ_1 and of the noise variables Σ_0 are equal to $\Sigma_i = (\rho_i^{|j-k|})_{j,k=1,\ldots,p}$, with $\rho_0 = 0.9$ and $\rho_1 = 0.5$, respectively. The error ε_i is generated from a standard normal distribution and the corresponding responses are then given by

$$y_i = \begin{cases} 0 & \text{if } \boldsymbol{\theta}'\boldsymbol{x}_i + \varepsilon_i \leq 0, \\ 1 & \text{if } \boldsymbol{\theta}'\boldsymbol{x}_i + \varepsilon_i > 0. \end{cases} \tag{16}$$

Contamination is introduced by adding outliers in the informative variables only. 10% or 20% of the observations with label zero are randomly selected and their informative variables are replaced by independent draws from a Gaussian distribution $\mathcal{N}(20, 1)$. Moreover, these observations are assigned to the other group, so they get label one, such that bad leverage points are introduced.

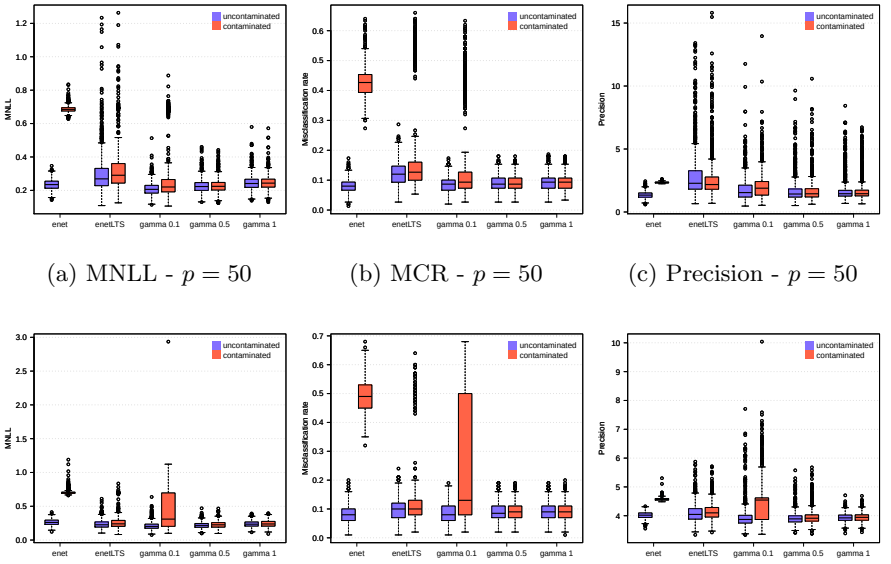
The estimates are evaluated on a clean dataset of the same dimension and size as the one used for estimation. We consider as performance measures the

mean of negative log-likelihood (MNLL), defined as

$$\text{MNLL}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i \log\left(\frac{\exp(\widehat{\boldsymbol{\theta}}'\boldsymbol{x})}{1+\exp(\widehat{\boldsymbol{\theta}}'\boldsymbol{x})}\right) -(1-y_i)\log\left(1-\frac{\exp(\widehat{\boldsymbol{\theta}}'\boldsymbol{x})}{1+\exp(\widehat{\boldsymbol{\theta}}'\boldsymbol{x})}\right)\right], \tag{17}$$

and the misclassification rate (MCR) for the classification rule which sets $\widehat{y}_i = 1$ if $f(y_i = 1 \mid \boldsymbol{x}, \widehat{\boldsymbol{\theta}}) > 0.5$.

Figure 2 and Figure 3 display the results for 1000 replications when $\alpha = 1$ for all estimators with respectively 10% or 20% of contaminated observations. Clearly, the classical elastic net logistic regression is heavily influenced by the outliers. The LTS elastic net is more robust although it is also highly affected by the outliers in a small fraction of the replications. This can be seen by comparing the MNLL and MCR statistics between the clean and contaminated



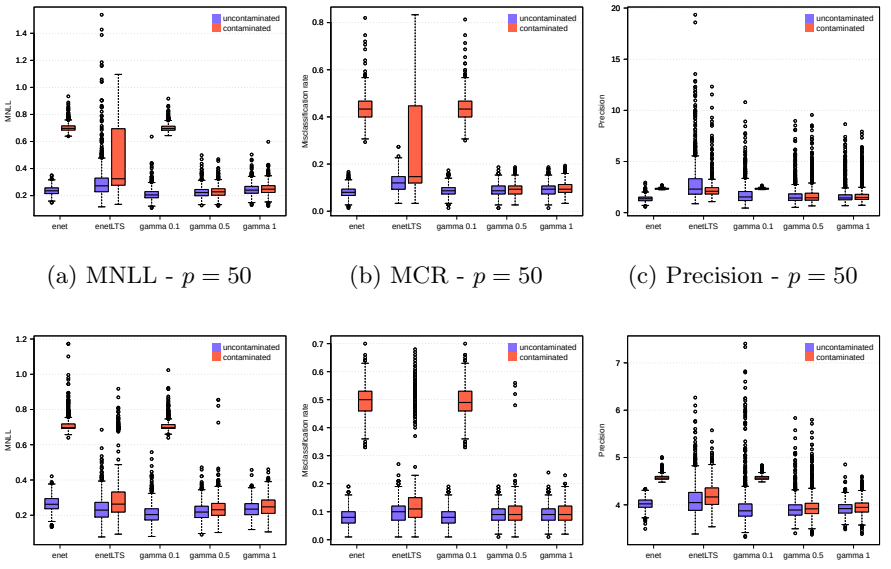
(a) MNLL - $p = 50$ (b) MCR - $p = 50$ (c) Precision - $p = 50$

(d) MNLL - $p = 200$ (e) MCR - $p = 200$ (f) Precision - $p = 200$

Fig. 2: Estimation accuracy for uncontaminated data and data with 10% of bad leverage points. All estimators use $\alpha = 1$.

datasets. We observe that the proposed sparse γ-logistic regression is robust when γ is large enough. Especially, the lower variance in MNLL compared to the LTS procedure is remarkable.

Figure 4 and Figure 5 show that the false positive rate of the robust methods is comparable with and without contamination. This is in strong contrast to the false positive rate of the elastic net estimator that is significantly lower, indicating that the non-robust elastic net estimator has the tendency to select no variables when outliers are present. The false negative rate is low for all robust methods, for the proposed sparse γ-logistic regression definitely when γ is large enough, and close to one in the contaminated case for the non-robust elastic net estimator.



(a) MNLL - $p = 50$ (b) MCR - $p = 50$ (c) Precision - $p = 50$

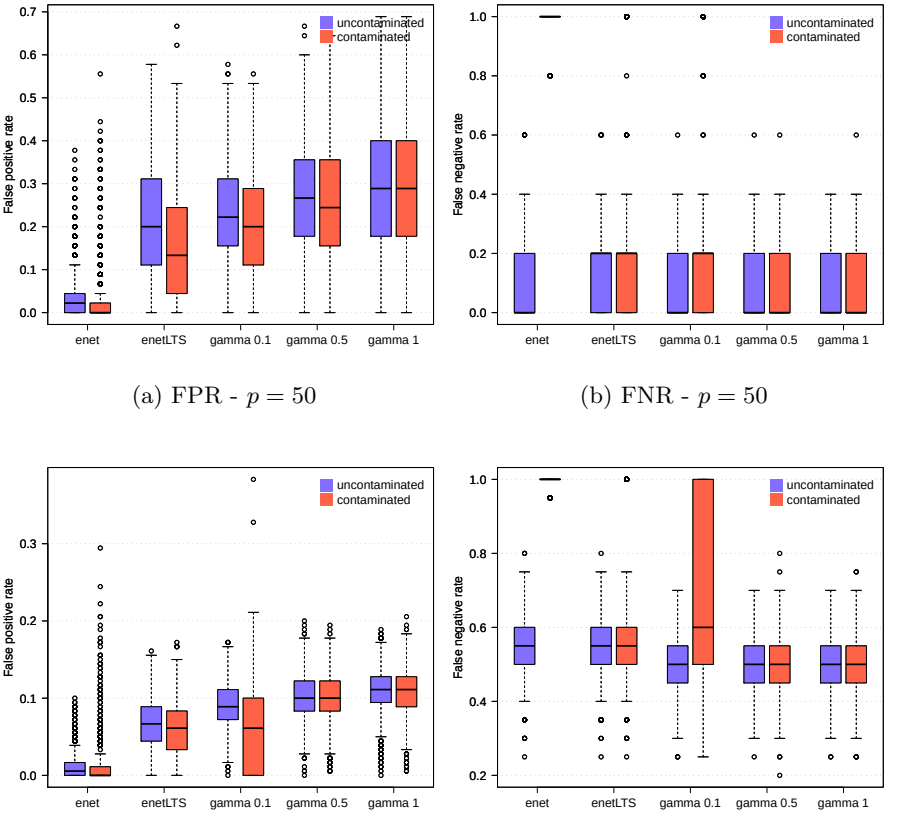
(d) MNLL - $p = 200$ (e) MCR - $p = 200$ (f) Precision - $p = 200$

Fig. 3: Estimation accuracy for uncontaminated data and data with 20% of bad leverage points. All estimators use $\alpha = 1$.

5 Real data illustration

We now compare the performance of RoSLog and classical elastic net logistic regression on a dataset that was scraped from the website of the popular British television show Top Gear[1] by Alfons et al (2016). The Top Gear dataset contains 29 numerical and categorical variables of $n = 242$ cars, after omitting observations with missing values. The dataset is included in the R package

[1] http://www.topgear.com/uk/



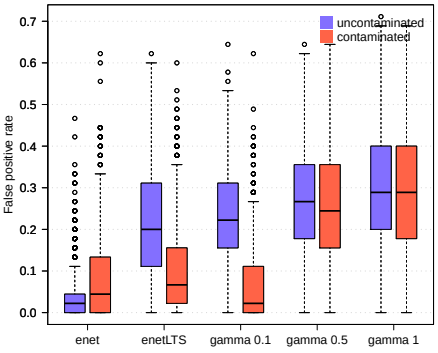
(a) FPR - $p = 50$ (b) FNR - $p = 50$

(c) FPR - $p = 200$ (d) FNR - $p = 200$

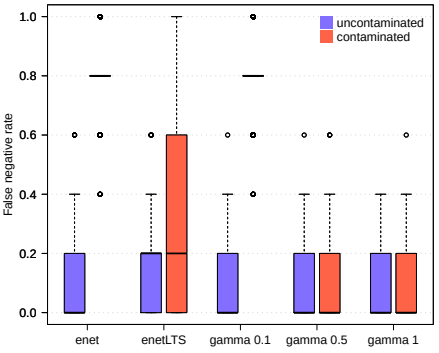
Fig. 4: Variable selection performance for uncontaminated data and data with 10% of bad leverage points. All estimators use $\alpha = 1$.

`robustHD` (Alfons, 2021) and we refer to this package for an explanation of the variables.

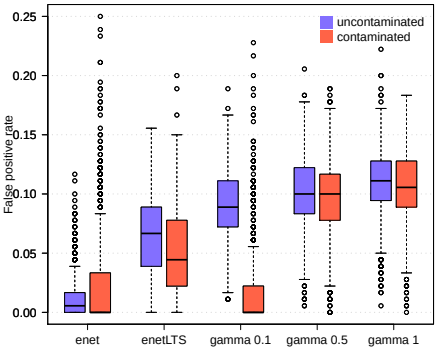
In this study, we aim to predict the type of fuel based on the 11 available continuous variables. These explanatory variables are robustly standardized. To allow for possible non-linearities in the model, we add all first order interaction of these variables. This yields 77 predictor variables. For the penalty we use $\alpha = 1$.



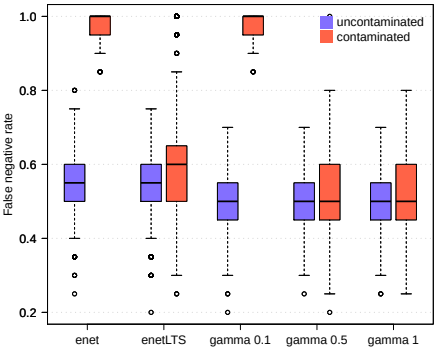
(a) FPR - $p = 50$



(b) FNR - $p = 50$



(c) FPR - $p = 200$



(d) FNR - $p = 200$

Fig. 5: Variable selection performance for uncontaminated data and data with 20% of bad leverage points. All estimators use $\alpha = 1$.

To select the value of γ for the RoSLog estimator, we follow the ideas of Riani et al (2020) by monitoring the performance of the RoSLog estimator on a grid of γ values between 0.05 and 1 with a step size of 0.05. To measure the prediction accuracy of the resulting fits we have first identified outliers according to the most robust solution, obtained with $\gamma = 1$. This yields 7 outliers in this example. We then split the complete data set in a training part containing 70% of the observations and a test set with the remaining 30%, where we do not include any of the identified outliers in the test set. We considered 10 such random splits of the data and calculated the average misclassification rate for each value of γ. Figure 6 shows the resulting averaged misclassification rates. It can be seen that the misclassicifaction rate drops until γ reaches the value 0.45 and then increases again. Hence, we consider the RoSLog with $\gamma = 0.45$ as the optimal solution.

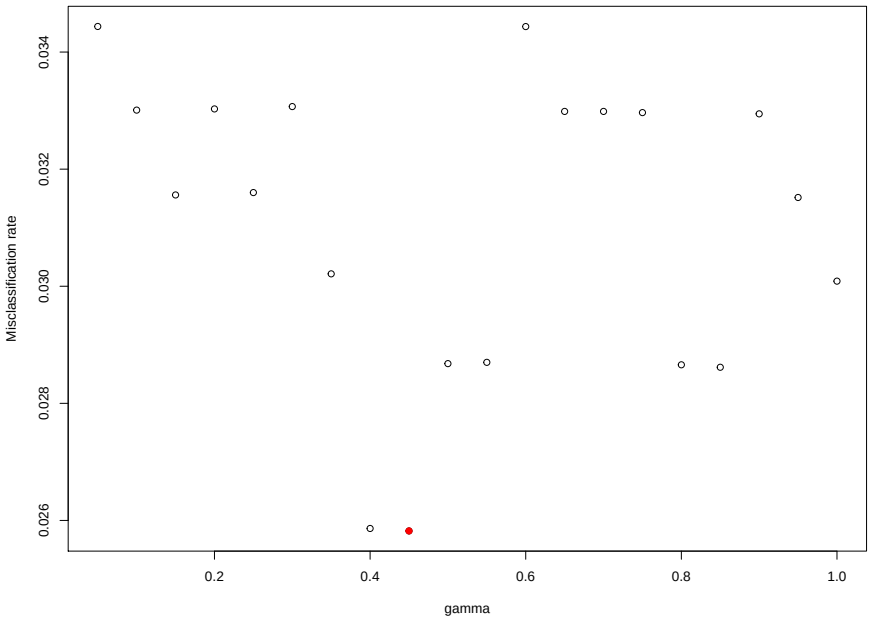


Fig. 6: The misclassification rate averaged over 10 datasets, for 20 values of γ. In red $\gamma = 0.45$, the value which yields the smallest misclassification rate.

We obtain an average misclassification rate of 0.0258 for the RoSLog and 0.0913 for enet, respectively. Hence, the RoSLog clearly outperforms enet. We conclude that the RoSLog is less influenced by the outliers in the training data than enet.

From Figure 7 we can see the seven cars are flagged by the RoSLog with $\gamma = 0.45$, namely Chevrolet Captiva, Hyundai i30, Jaguar XF Sportbrake, Mercedes-Benz C-Class, Mini Convertible, Nissan Juke and Peugeot 308. The Mini Convertible runs on diesel, which is exceptional for such a small car. All the other outliers use petrol. The Chevrolet Captiva stands out because it is a big car for a petrol engine. The Hyundai i30, Nissan Juke and Peugeot 308 have a low price, low BHP and high MPG, what is more likely to be expected from a diesel engine. Moreover, from our investigations on the internet it seems that some of these cars (e.g. Hyundai i30) seem to be mainly (or even only) available with a diesel engine as opposed to a petrol engine as stated in the dataset.

Both the RoSLog and enet estimated models yield a lot of zero coefficients, respectively 52 and 56 out of 77 variables in total. Therefore they both provide sparse model estimates, which increases stability and interpretability. Table 1 gives the coefficients of the selected variables by enet and the RoSLog with $\gamma = 0.45$. For several variables, it intuitively makes sense that they are related with the response. For example, BHP (engine power) has a positive coefficient in both models meaning that cars with a high BHP tend to run on petrol. The opposite can be said about torque, cars with a high torque value are more likely to have a diesel engine.

6 Conclusion

We proposed a robust and sparse logistic regression estimator based on γ-divergence. Sparsity is achieved by the elastic net penalty, regularizing the estimator in high dimensions. Robustness of the method is shown by its bounded influence function. In particular, the influence function equals zero for

bad leverage points. Simulations confirm the good performance of the estimator and the practical usefulness is demonstrated on a real data set, classifying cars by their type of fuel.

The proposed estimator focuses on identifying and down-weighting outlying observations which are also called casewise or rowwise outliers, i.e. rows of the data matrix. However, downweighting complete cases becomes inefficient for modern big data which contain relatively few cases in comparison to the number of variables. Often only a few cells in outlying cases are actually contaminated and the remaining cells still contain valuable information. Such contaminated cells are called cellwise outliers, elementwise outliers or entrywise outliers. On the other hand, a small fraction of cellwise outliers can affect a very large fraction of cases in high dimensional data, easily exceeding 50% of

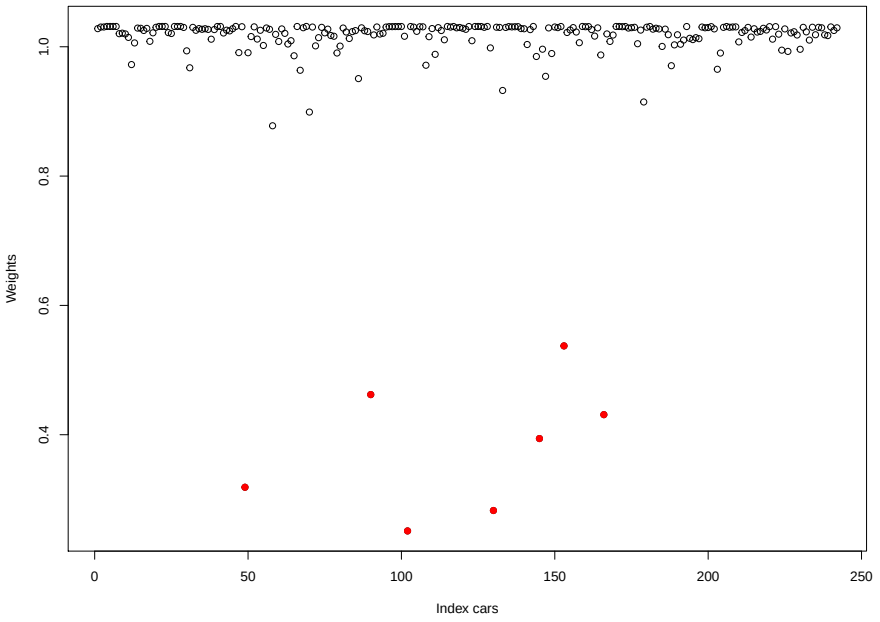


Fig. 7: The outlying weights obtained with RoSLog for $\gamma = 0.45$, indicating the seven outliers.

the observations, such that casewise robust methods become unreliable (Alqallaf et al, 2009). Therefore, an interesting topic for future research is to adapt divergence based estimators for robustness against cellwise outliers.

	enet	$\gamma = 0.45$
Intercept	-0.2065	-0.4161
Displacement	-1.9338	-0.3159
BHP	3.7034	2.5059
Torque	-5.4272	-4.4442
Acceleration	-1.2835	-0.9697
MPG	-1.5977	-1.0162
Weight	-0.4190	-0.6274
Width	-0.0309	
Height	-0.4233	-0.3220
Displacement × Torque		0.0587
Displacement × Acceleration	-0.1480	
Displacement × MPG	-1.1777	
Displacement × Length	0.1672	
Displacement × Width		0.1019
BHP × MPG	-0.3704	-0.1956
Torque × Torque	0.4429	0.1285
Torque × Acceleration		-0.1053
Torque × MPG	0.8757	
Torque × Weight		0.3647
Acceleration × Acceleration	-0.4492	
Acceleration × TopSpeed	-0.03162	
Acceleration × MPG	-0.4664	-0.6762
Acceleration × Width	-0.3122	
Acceleration × Height		0.0597
TopSpeed × MPG	-0.9466	-0.8414
TopSpeed × Length	0.2024	0.0779
MPG × Weight	0.2406	0.2236
MPG × Length	0.8920	
MPG × Height	-0.6236	-0.3849
Weight × Length	0.5449	0.0698
Width × Height	-0.4361	-0.1833

Table 1: The coefficients of the selected variables in the Top Gear data set by enet and RoSLog with $\gamma = 0.45$.

References

Alfons A (2021) Robusthd: an r package for robust regression with high-dimensional data. Journal of Open Source Software 6(67):3786

Alfons A, Croux C, Gelper S (2016) Robust groupwise least angle regression. Computational Statistics & Data Analysis 93:421–435

Alqallaf F, Van Aelst S, Yohai VJ, et al (2009) Propagation of outliers in multivariate data. The Annals of Statistics pp 311–331

Avella-Medina M, Ronchetti E (2018) Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika 105(1):31–44

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

Bianco AM, Boente G, Chebi G (2022) Penalized robust estimators in sparse logistic regression. TEST 31:563–594. https://doi.org/10.1007/s11749-021-00792-w

Bondell HD (2005) Minimum distance estimation for the logistic regression model. Biometrika 92(3):724–731

Bondell HD (2008) A characteristic function approach to the biased sampling model, with application to robust logistic regression. Journal of Statistical Planning and Inference 138(3):742 – 755

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. Journal of the American Statistical Association 96:1022–1030. https://doi.org/10.2307/2670248

Carroll RJ, Pederson S (1993) On robustness in the logistic regression model. Journal of the Royal Statistical Society: Series B (Methodological) 55(3):693–706

Croux C, Haesbroeck G (2003) Implementing the bianco and yohai estimator for logistic regression. Computational statistics & data analysis 44(1-2):273–295

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22

Fujisawa H, Eguchi S (2008) Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis 99(9):2053–2081

Hampel FR, Ronchetti EM, Rousseeuw PJ, et al (2011) Robust Statistics: the Approach based on Influence Functions. John Wiley and Sons, New York

Hoffmann I, Filzmoser P, Serneels S, et al (2016) Sparse and robust PLS for binary classification. Journal of Chemometrics 30(4):153–162

Hosseinian S, Morgenthaler S (2011) Robust binary regression. Journal of Statistical Planning and Inference 141(4):1497 – 1509. https://doi.org/http://dx.doi.org/10.1016/j.jspi.2010.11.015, URL http://www.sciencedirect.com/science/article/pii/S037837581000515X

Hung H, Jou ZY, Huang SY (2018) Robust mislabel logistic regression without modeling mislabel probabilities. Biometrics 74(1):145–154

Kawashima T, Fujisawa H (2017) Robust and sparse regression via γ-divergence. Entropy 19(11):608

Kawashima T, Fujisawa H (2019) Robust and sparse regression in generalized linear model by stochastic optimization. Japanese Journal of Statistics and Data Science 2:465–489. https://doi.org/https://doi.org/10.1007/s42081-019-00049-9

Kawashima T, Fujisawa H (2022) Robust regression against heavy heterogeneous contamination. Metrika pp 1–22

Khan JA, Van Aelst S, Zamar RH (2010) Fast robust estimation of prediction error based on resampling. Computational Statistics & Data Analysis 54(12):3121–3130. https://doi.org/https://doi.org/10.1016/j.csda.2010.01.031, URL https://www.sciencedirect.com/science/article/pii/S0167947310000460

Kurnaz FS, Hoffmann I, Filzmoser P (2018) Robust and sparse estimation methods for high-dimensional linear and logistic regression. Chemometrics and Intelligent Laboratory Systems 172:211–222

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

Öllerer V, Croux C, Alfons A (2015) The influence function of penalized regression estimators. Statistics 49(4):741–765

Ortega JM, Rheinboldt WC (1970) Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, Orlando

Ponnet J, Segaert P, Van Aelst S, et al (2023) Robust inference and modeling of mean and dispersion for generalized linear models. Journal of the American Statistical Association to appear. https://doi.org/10.1080/01621459.2022.2140054

Riani M, Atkinson A, Corbellini A, et al (2020) Robust regression with density power divergence: Theory, comparisons, and data analysis. Entropy 22:399. https://doi.org/10.3390/e22040399

Valdiviezo HC, Van Aelst S (2019) Fast computation of robust subspace estimators. Computational Statistics & Data Analysis 134:171–185

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2):301–320

A Proof of Proposition 3.1

In the proof below, we define v explicitly as a function of $\boldsymbol{\theta}$,

$$v(\boldsymbol{x}, y, \boldsymbol{\theta}) = \left(\frac{\exp(y(1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})}{1+\exp((1+\gamma)\boldsymbol{\theta}'\boldsymbol{x})} \right)^{\frac{\gamma}{1+\gamma}}. \tag{18}$$

The first order conditions on (12) for the contaminated distribution F_ε are

$$\begin{aligned} 0 = \mathbb{E}_{F_\varepsilon}\left[\boldsymbol{X} v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F_\varepsilon))\left[Y - \pi(\boldsymbol{X}, (1+\gamma)\boldsymbol{\theta}_\gamma(F_\varepsilon))\right]\right] \\ - \lambda\, \mathbb{E}_{F_\varepsilon}\left[v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F_\varepsilon))\right] \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F_\varepsilon)), \end{aligned} \tag{19}$$

By definition of F_ε it follows that

$$\begin{aligned} 0 = (1-\varepsilon)\, \mathbb{E}_F\left[\boldsymbol{X} v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F_\varepsilon))\left[Y - \pi(X, (1+\gamma)\boldsymbol{\theta}_\gamma(F_\varepsilon))\right]\right] \\ + \varepsilon \boldsymbol{x}_0 v(\boldsymbol{x}_0, y_0, \boldsymbol{\theta}_\gamma(F_\varepsilon))\left(y_0 - \pi(\boldsymbol{x}_0, (1+\gamma)\boldsymbol{\theta}_\gamma(F_\varepsilon))\right) \\ - (1-\varepsilon)\, \mathbb{E}_F\left[v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F_\varepsilon))\right] \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F_\varepsilon)) \\ - \varepsilon\lambda v(\boldsymbol{x}_0, y_0, \boldsymbol{\theta}_\gamma(F_\varepsilon)) \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F_\varepsilon)). \end{aligned} \tag{20}$$

Taking the derivative with respect to ε and evaluating at $\varepsilon = 0$ yields

$$\begin{aligned} 0 = &- \mathbb{E}_F\left[\boldsymbol{X} v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F))\left[Y - \pi(\boldsymbol{X}, (1+\gamma)\boldsymbol{\theta}_\gamma(F))\right]\right] \\ &+ \lambda\, \mathbb{E}_F\left[v(\boldsymbol{X}, Y, \boldsymbol{\theta}_\gamma(F))\right] \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F)) \\ &+ \boldsymbol{x}_0 v(\boldsymbol{x}_0, y_0, \boldsymbol{\theta}_\gamma(F))\left(y_0 - \pi(\boldsymbol{x}_0, (1+\gamma)\boldsymbol{\theta}_\gamma(F))\right) \\ &- \lambda v(\boldsymbol{x}_0, y_0, \boldsymbol{\theta}_\gamma(F)) \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F)) \\ &- \Bigg(\mathbb{E}_F\left[v(\boldsymbol{X}, Y)\left(\widetilde{\pi}(1-\widetilde{\pi})\boldsymbol{X}\boldsymbol{X}' + \lambda \nabla^2 P_\alpha(\boldsymbol{\theta}_\gamma(F))\right)\right] \\ &\quad + \gamma\, \mathbb{E}_F\Big[v(\boldsymbol{X}, Y)\big(\widetilde{\pi}(1-\widetilde{\pi})\boldsymbol{X}\boldsymbol{X}' - (Y-\widetilde{\pi})^2 \boldsymbol{X}\boldsymbol{X}' \\ &\qquad + \lambda(Y-\widetilde{\pi}) \nabla P_\alpha(\boldsymbol{\theta}_\gamma(F))\boldsymbol{X}'\big)\Big] \Bigg) \frac{\partial}{\partial \varepsilon}\left[\boldsymbol{\theta}_\gamma\left(F_\varepsilon\right)\right]\Big|_{\varepsilon=0}, \end{aligned} \tag{21}$$

where $\widetilde{\pi} = \pi(\boldsymbol{X}, (1+\gamma)\boldsymbol{\theta}_\gamma(F))$. The first two terms are the estimating equations of $\boldsymbol{\theta}_\gamma(F)$ and are thus equal to zero. This proves Theorem 3.1.