

# Initial Robust Estimation in Generalized Linear Models

Claudio Agostinelli<sup>a</sup>, Marina Valdora.<sup>b,\*</sup>, Victor J. Yohai<sup>c</sup>

<sup>a</sup>*Department of Mathematics, University of Trento, Trento, Italy*

<sup>b</sup>*Departamento de Matematicas and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, University of Buenos Aires*

<sup>c</sup>*Departamento de Matematicas and Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, University of Buenos Aires, CONICET*

---

## Abstract

Generalized Linear Models are routinely used in data analysis. Classical estimators are based on the maximum likelihood principle and it is well known that the presence of outliers can have a large impact on them. Several robust procedures have been presented in the literature, being redescending M-estimators the most widely accepted. Based on non-convex loss functions, these estimators need a robust initial estimate, which is often obtained by subsampling techniques. However, as the number of unknown parameters increases, the number of subsamples needed in order for this method to be robust, soon makes it infeasible. Furthermore the subsampling procedure provides a non deterministic starting point. A new method for computing a robust initial estimator is proposed. This method is deterministic and demands a relatively short computational time, even for large numbers of covariates. The proposed method is applied to M-estimators based on transformations. In addition, an iteratively reweighted least squares algorithm is proposed for the computation of the final estimates. The new methods are studied by means of Monte Carlo experiments.

*Keywords:* Initial estimates, Outliers, Least squares estimators, M-estimators, Variance stabilizing transformations, Poisson regression

---

\*Corresponding Author

*URL:* [claudio.agostinelli@unitn.it](mailto:claudio.agostinelli@unitn.it) (Claudio Agostinelli), [mvaldora@gmail.com](mailto:mvaldora@gmail.com) (Marina Valdora.), [victoryohai@gmail.com](mailto:victoryohai@gmail.com) (Victor J. Yohai)

*Postal Address:* Marina Valdora, Departamento de Matematicas and Instituto de Cálculo. University of Buenos Aires. Intendente Güiraldes 2160, Ciudad Universitaria, C1428EGA, Buenos Aires, Argentina

## 1. Introduction

Robust estimators for generalized linear models (GLM) have been studied by many authors in recent years. Among them, we may cite Künsch et al. (1989), Cantoni and Ronchetti (2001), Bergesio and Yohai (2011), Bianco et al. (2013), Valdora and Yohai (2014) and Alqallaf and Agostinelli (2016). These proposals either lack robustness or require a robust initial estimator. We propose a method for computing an initial estimator which can be used to start an iterative algorithm, as needed by redescending estimators. We apply this method to the computation of M-estimators based on transformations (MT), proposed by Valdora and Yohai (2014). MT-estimators are a family of redescending M-estimators based on variance stabilizing transformations. A variance stabilizing transformation is a function, such that, if applied to a random variable with a distribution in a one-parameter family, the resulting random variable has an almost constant variance. For example in the case of the Poisson family, a function with this property is the square root (See, for example, Section 2.2 in Valdora and Yohai, 2014). Stabilizing the variance allows the correct scaling of the loss function used in the definition of M-estimators.

Consider a GLM in which  $y$  is the response and  $\mathbf{x}$  is a  $p$ -dimensional vector of explanatory variables. We assume that

$$g(\mu) = \beta_0^\top \mathbf{x}, \quad (1)$$

where  $\beta_0 \in \mathbb{R}^p$  is an unknown vector of parameters and  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a known link function. We further assume that

$$y|\mathbf{x} \sim F_\mu, \quad (2)$$

where  $F_\mu$  is a discrete or continuous distribution function in the exponential family of distributions in  $\mathbb{R}$  with a density of the form

$$f_\mu(y) = \exp(y\mu - b(\mu) + c(y)), \quad (3)$$

for given functions  $b$  and  $c$ .

Let  $t$  be a variance stabilizing transformation and  $\rho(u)$  be a function such that:

- (1)  $\rho(u)$  is a non-decreasing function of  $|u|$ ,
- (2)  $\rho(0) = 0$ ,
- (3) there exists  $k > 0$  such that  $\rho$  is strictly increasing in  $(0, k)$  and  $\rho$  is constant in  $(k, +\infty)$ .

MT-estimators are defined as

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} L(\hat{\boldsymbol{\beta}}), \quad (4)$$

where

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n \rho(t(y_i) - m(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}))),$$

and  $m$  is the function defined by

$$m(\mu) = \arg \min_{\gamma} \mathbb{E}_{\mu}(\rho(t(y) - \gamma)), \quad (5)$$

where  $\mathbb{E}_{\mu}(y)$  denotes the expectation of  $y$  when  $y \sim F_{\mu}$ . It is assumed that  $m$  is univocally defined, therefore (5) implies the Fisher consistency of  $\hat{\boldsymbol{\beta}}$ . Other assumptions necessary to have consistency and asymptotic normality of these estimators are listed in Valdora and Yohai (2014). The solution to (4) can be found by iterative methods which typically solve the corresponding system of estimating equations

$$\sum_{i=1}^n \psi(\mathbf{x}_i, y_i, \boldsymbol{\beta}) = 0. \quad (6)$$

where  $\psi(\mathbf{x}_i, y_i, \boldsymbol{\beta})$  is the derivative with respect to  $\boldsymbol{\beta}$  of  $\rho(t(y_i) - m(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})))$ . In Appendix A we provide an iteratively reweighted least squares (IRWLS) algorithm to find a solution of equation (6). The difficulty in the case of redescending M-estimators is that the objective function  $L(\boldsymbol{\beta})$  may have several local minima. As a consequence, the iterative procedure may converge to a solution to equation (6) that is not a solution to the optimization problem (4). To avoid this, one must begin the iterative algorithm at an initial estimator which is a very good approximation of the global minimum of  $L$ , i.e. the solution of (4). If  $p$  is small, this approximate solution may be obtained by the subsampling method (see, Valdora and Yohai, 2014). Based on the algorithm described in Rousseeuw and Leroy (1987) for linear models, this method consists in computing a finite set  $A$  of candidate solutions of (4) and then replace the minimization over  $\mathbb{R}^p$  by a minimization over  $A$ . The set

$A$  is obtained by randomly drawing subsamples of size  $p$  and then computing the maximum likelihood (ML) estimator based on the subsample. If the original sample contains a proportion  $\epsilon$  of outliers, then the probability that a given subsample is free of outliers is  $(1 - \epsilon)^p$  and the probability of having at least one subsample free of outliers is  $1 - (1 - (1 - \epsilon)^p)^N$ , where  $N$  is the number of subsamples drawn. If we want this probability to be greater than a given  $\alpha$ , we must draw a number of subsamples such that

$$1 - (1 - (1 - \epsilon)^p)^N > \alpha,$$

that is to say,

$$N > \frac{\log(\alpha)}{\log(1 - (1 - \epsilon)^p)} \simeq \left\lceil \frac{\log(\alpha)}{(1 - \epsilon)^p} \right\rceil.$$

This makes the algorithm infeasible for large  $p$ . Peña and Yohai (1999) studied this problem in the case of linear models, introducing an alternative method to compute the set of candidate solutions  $A$ . Their proposal succeeds in obtaining a set  $A$  which contains very good approximations of the actual solution and, on the other hand, requires the computation of a small number of subsamples, namely  $3p + 1$ . This makes the algorithm much faster and feasible even for very large values of  $p$ .

We modify the method introduced by Peña and Yohai (1999) in order to apply it to generalized linear models. We study its application to MT-estimators by means of an extensive Monte-Carlo study, which shows that the method is very fast and robust for large values of  $p$ .

As a particular case of the MT-estimator we define the Least Squares estimator based on Transformations (LST), which corresponds to  $\rho(u) = u^2$ , in the following way

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \left( t(y_i) - \mathbb{E}_{[g^{-1}(\mathbf{x}_i^T \beta)]} (t(y_i)) \right)^2. \quad (7)$$

This estimator can be seen as a natural generalization of the Least Squares estimator (LS) for linear models to the case of GLM. LST estimators are Fisher consistent, however since the corresponding  $\rho$  is not bounded, they are, in general, non robust.

The paper is organized as follows. In Section 2 we define the principal sensitivity components and explain how they can be used to detect outliers. In Section 3 we describe the proposed procedure in detail. In Section 4 we

describe the simulation study used to compare the proposed procedure with existing methods and give a summary of its result. In Section 5 we apply the proposed method to a real data example and compare the results to those obtained by other methods. In Section 6 we summarize our conclusions. In Appendix A we provide an iteratively reweighted least squares algorithm to find the solution to the optimization problems (4) and (7). We also provide supplementary material that contains a pseudo code of the proposed procedure, the complete results of the simulation study and the code necessary to reproduce the analysis of the real data example.

## 2. Detecting Outliers Using Principal Sensitivity Components

The classical statistic used to measure the influence of an observation is the Cook statistic introduced by Cook (1977) for linear models, which can be adapted for generalized linear models (see Chapter 12 of McCullagh and Nelder, 1989). This statistic is a measure of the distance between  $\hat{\beta}$ , the maximum likelihood estimator and  $\hat{\beta}_{(i)}$ , the maximum likelihood estimator computed without observation  $i$ . However, as it is well known (Maronna et al., 2006), this measure is non-robust and therefore, when there are several outliers, it may be completely unreliable.

The proposal of Peña and Yohai (1999) follows the same idea as the subsampling method, i.e., the candidate solutions  $A$  are obtained by computing the least squares estimates on subsamples. However, the subsamples are not chosen at random. Instead, they are chosen by deleting from the sample groups of outliers, which can potentially cause a masking effect. The set  $A$  will, in this way, contain candidates which are already quite robust estimates and therefore there is no need to have a large number of candidates as it is necessary for randomly chosen subsamples. In fact, the number of candidates in the set  $A$  is only  $3p + 1$ .

Let  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  be random vectors which follow a generalized linear model as defined by (2) and (1). Let  $\hat{\beta}$  be the LST estimator and let

$$\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_n)^\top = g^{-1}(\mathbf{X}\hat{\beta}),$$

be the vector of fitted values. Let  $\hat{\mu}_{i(j)}$  be the fitted value for observation  $i$  computed without using observation  $j$ , that is  $\hat{\mu}_{i(j)} = g^{-1}(\mathbf{x}_i^\top \hat{\beta}_{(j)})$ , where  $\hat{\beta}_{(j)}$  is the LST estimate based on the original sample without observation  $j$ . We define the  $i$ -th residual,  $e_i$ , as the difference between  $t_i = t(y_i)$  and

its predicted value  $\hat{t}_i = m(g^{-1}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}))$ , that is  $e_i = t_i - \hat{t}_i$ . Following the ideas introduced by Peña and Yohai (1999) for linear models, we define the sensitivity vectors as the vectors  $\mathbf{r}_i$  with entries

$$r_{ij} = \hat{t}_i - \hat{t}_{i(j)},$$

where  $\hat{t}_{i(j)} = m(\hat{\mu}_{i(j)})$  is the predicted value of  $t_i$  computed without using observation  $j$ . Then,  $r_{ij}$  is the sensitivity in forecasting  $t_i$  to the deletion of observation  $j$  and the sensitivity vectors are defined by

$$\mathbf{r}_i = (r_{i1}, \dots, r_{in}) \ , \quad 1 \leq i \leq n.$$

The sensitivity matrix  $\mathbf{R}$  is defined as the matrix whose rows are the vectors  $\mathbf{r}_i$ . The vector  $\mathbf{r}_i$  expresses how sensitive the prediction of  $y_i$  is to the deletion of each observation. The main idea in Peña and Yohai (1999) is to obtain candidates to start the algorithm to compute the robust estimator defined by (4), by eliminating those observations with large sensitivity. Since the sensitivity vectors have  $n$  componentes, which in general is a very large number, by similarity to what is proposed in Peña and Yohai (1999), we compute the first  $p$  principal components of  $\mathbf{R}$ . Let  $\mathbf{v}_1$  be defined by

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \sum_{i=1}^n (\mathbf{v}^\top \mathbf{r}_i)^2. \quad (8)$$

Note that  $\mathbf{v}_1$  is the direction in which the projections of the sensitivity vectors are the largest and

$$\mathbf{z}_1 = \mathbf{R}\mathbf{v}_1 \quad (9)$$

is the first principal component of the dataset whose elements are the rows of  $\mathbf{R}$ . Therefore, the largest entries in  $\mathbf{z}_1$  correspond to the largest terms in the sum in equation (8), which in turn correspond to the observations that have the largest projected sensitivity in the direction  $\mathbf{v}_1$ .

In the same way, we can define recursively  $\mathbf{v}_i$ ,  $2 \leq i \leq p$  as the solution to

$$\mathbf{v}_i = \arg \max_{\|\mathbf{v}\|=1} \sum_{i=1}^n (\mathbf{v}^\top \mathbf{r}_i)^2, \quad (10)$$

$$\text{subject to } \mathbf{v}_i \mathbf{v}_j = 0 \text{ for all } 1 \leq j < i. \quad (11)$$

The vectors  $\mathbf{v}_1, \dots, \mathbf{v}_p$  are the directions in which the projected sensitivity of the observations are the largest. The corresponding projections,

$$\mathbf{z}_i = \mathbf{R}\mathbf{v}_i, \quad (12)$$

are the principal components of  $\mathbf{R}$  which will be called principal sensitivity components. The entries of  $\mathbf{z}_i$  are the projections of the sensitivity vectors on the direction  $\mathbf{v}_i$ . Large entries correspond to observations whose projected sensitivity in the direction  $\mathbf{v}_i$  is large. Therefore, large entries are considered potential outliers.

In the case of linear models, Peña and Yohai (1999), showed that, if the sample is contaminated with less than  $(n - p + 1)/(2n - p + 1)$  high leverage outliers, then, either the least squares estimate is bounded or at least for one of the directions  $\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ , the absolute values of the entries corresponding to the outliers are larger than the median of these absolute values. For this reason, high-leverage outliers are expected to be extreme entries in at least one of the principal sensitivity components. The theorem implies that, if the sample is contaminated with less than  $(n - p + 1)/(2n - p + 1)$  high leverage outliers, the estimator will remain bounded.

### 3. Procedure for obtaining a robust initial estimate in generalized linear models

Consider a random sample following a generalized linear model as defined by (1), (2) and (3). We introduce a procedure to compute an approximation of  $\beta_0$  that will be used as an initial estimator in the IRWLS algorithm used to solve the estimating equation given in (6). The procedure has two stages: stage 1 aims at finding a highly robust but possibly inefficient estimate and stage 2 aims at increasing its efficiency.

*Stage 1.* In this stage, the idea is to find a robust, but possibly inefficient, estimate of  $\beta_0$  by an iterative procedure. In the  $k$ -th step of this iteration method, for  $k > 1$ , we set

$$\hat{\beta}^{(k)} = \arg \min_{\beta \in A_k} L(\beta). \quad (13)$$

In the first iteration ( $k = 1$ ) the set  $A_1$  is constructed as follows. We begin by computing the LST estimate with the complete sample and the principal sensitivity components. For each principal sensitivity component  $\mathbf{z}_i$  we compute three estimates by the LST method. The first estimate is computed

after eliminating the half of the observations corresponding to the smallest entries in  $\mathbf{z}_i$ , the second, after eliminating the half of the observations corresponding to the largest entries in  $\mathbf{z}_i$  and the third, after eliminating the half corresponding to the largest absolute values in  $\mathbf{z}_i$ . To these  $3p$  initial candidates we add the LST estimate computed using the complete sample, obtaining a set of  $3p + 1$  elements. Once we have  $A_1$  we obtain  $\hat{\beta}^{(1)}$  by minimizing  $L(\beta)$  over the elements of  $A_1$ .

Suppose now that we are at iteration  $k > 1$ . Let  $0 < \alpha < 0.5$  be a quantile filtering constant; in all our simulations and examples we set  $\alpha = 0.05$ . For  $k > 1$ , we first delete the observations ( $i = 1, \dots, n$ ) such that  $y_i > F_{\hat{\mu}_i}^{-1}(1 - \alpha/2)$  or  $y_i < F_{\hat{\mu}_i}^{-1}(\alpha/2)$  where  $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^\top \hat{\beta}^{(k-1)})$ ; then with the remaining observations, we compute the LST estimator  $\hat{\beta}_{\text{LST}}^{(k)}$  and the corresponding principal sensitivity components. Let us remark that, for the computation of  $\hat{\beta}_{\text{LST}}^{(k)}$  we have deleted the observations that have large residuals, since  $\hat{\mu}_i$  is the fitted value obtained using  $\hat{\beta}^{(k-1)}$ . In this way, while candidates on the first step of the iteration are protected from high leverage outliers, candidate  $\hat{\beta}_{\text{LST}}^{(k)}$  is protected from low leverage outliers, which may not be extreme entries of the  $\mathbf{z}_i$ .

Now, the set  $A_k$  contains  $\hat{\beta}_{\text{LST}}^{(k)}$ ,  $\hat{\beta}^{(k-1)}$  and the  $3p$  LST estimates computed by deleting extreme values according to the new principal sensitivity components  $\mathbf{z}_i^{(k)}$  ( $i = 1, \dots, n$ ) as in the first iteration.  $\hat{\beta}^{(k)}$  is the element of  $A_k$  minimizing  $L(\beta)$ .

The iterations continue until  $\hat{\beta}^{(k)} \approx \hat{\beta}^{(k-1)}$ . Let  $\hat{\beta}_1$  be the final estimate obtained at this stage.

*Stage 2.* In this second stage we delete the observations  $y_i$  ( $i = 1, \dots, n$ ) such that  $y_i > F_{\hat{\mu}_i}^{-1}(1 - \alpha/2)$  or  $y_i < F_{\hat{\mu}_i}^{-1}(\alpha/2)$ , where  $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^\top \hat{\beta}_1)$  and compute the LST estimate  $\hat{\beta}^{(*)}$  with the reduced sample. Then, for each of the deleted observations we check whether  $y_i > F_{\hat{\mu}_i^{(*)}}^{-1}(1 - \alpha/2)$  or  $y_i < F_{\hat{\mu}_i^{(*)}}^{-1}(\alpha/2)$ , where  $\hat{\mu}_i^{(*)} = g^{-1}(\mathbf{x}_i^\top \hat{\beta}^{(*)})$ . Observations which are not within these bounds are finally eliminated; those which are within the bounds are restored to the sample. With the resulting set of observations we compute the LST estimate  $\hat{\beta}_2$  which is our proposed initial estimate.



#### 4. Monte Carlo Study

In this section we report the results of a Monte Carlo study in which we compare the MT-estimator computed with the proposed initial estimate (FMT), to the maximum likelihood estimator (ML), the robust quasi likelihood estimator (RQL) proposed by Cantoni and Ronchetti (2001), the Conditionally Unbiased Bounded Influence (CUBIF) estimator proposed by Künsch et al. (1989), and the MT-estimator beginning at an initial estimator computed by subsampling (SMT). For computing the RQL estimator, we used function `glmrob` from the R (R Core Team, 2018) package `robustbase` (Maechler et al., 2018), with method “Mql” and argument `weights.on.x` set to “robCov”, so that weights based on a robust Mahalanobis distance of the design matrix (intercept excluded) were used to downweight potential outliers in the  $\mathbf{x}$ -space. The CUBIF estimator was computed using function `cubinf`, available in the R package `robcbi` (Marazzi, 2018a). For the computation of the SMT estimator, the number of subsamples was set to 2500. Both FMT and SMT were computed using the iteratively reweighted least squares method described in the appendix and implemented in function `poissonMT` of the R package `poissonMT` (Agostinelli et al., 2018). They only differ in the starting point, obtained from functions `poissonMTinitial` and `poissonSSinitial` respectively. We study the case of Poisson regression and log link.

Let  $\mathbf{x} = (1, \mathbf{x}^*)$  be a random vector in  $\mathbb{R}^p$  such that  $\mathbf{x}^*$  is distributed as  $\mathcal{N}_{p-1}(\mathbf{0}, \mathbf{I})$  and let  $y$  be a random variable such that  $y|\mathbf{x} \sim \mathcal{P}(\exp(\boldsymbol{\beta}_0^\top \mathbf{x}))$ . Let  $\mathbf{e}_i$  be the vector of  $\mathbb{R}^p$  with all entries equal to zero except for the  $i$ -th entry which is equal to one. We considered five different models: in model 1, 4 and 5,  $\boldsymbol{\beta}_0 = \mathbf{e}_2$ ; in model 2,  $\boldsymbol{\beta}_0 = 2\mathbf{e}_1 + \mathbf{e}_2$ ; in model 3,  $\boldsymbol{\beta}_0 = 2\mathbf{e}_1 + 1\mathbf{e}_2 + 5\mathbf{e}_3$ . For each of these models we simulated the case in which the samples do not contain outliers and the case in which the samples have a proportion  $\epsilon$  of outliers. In models 1 to 4, all the outliers were placed at the point  $(\mathbf{x}_0, y_0)$ , with  $\mathbf{x}_0 = \mathbf{e}_1 + 3\mathbf{e}_2$  in models 1 to 3, while  $\mathbf{x}_0 = \mathbf{e}_1 + 3\mathbf{e}_2 + 4\mathbf{e}_3$  in model 4. The values of  $y_0$  belong to a grid ranging from  $\boldsymbol{\mu}_0 - K_1$  to  $\boldsymbol{\mu}_0 + K_2$ , where  $\boldsymbol{\mu}_0 = \exp(\boldsymbol{\beta}_0^\top \mathbf{x}) = \mathbb{E}_{\boldsymbol{\beta}_0}(y|\mathbf{x} = \mathbf{x}_0)$ . In model 5, the outliers were located at  $(\mathbf{x}_0, y_0)$ , with  $\mathbf{x}_0$  as in model 4 and  $y_0 \sim \mathcal{P}(\boldsymbol{\mu}_1)$ . The values of  $\boldsymbol{\mu}_1$  belong to a grid ranging from  $\boldsymbol{\mu}_0 - K_1$  to  $\boldsymbol{\mu}_0 + K_2$ , as before. The values of  $K_1$  and  $K_2$  and the grid step were chosen in a way such that the maximum mean squared error of our proposed estimator can be identified.

From our experience on linear models we know that point-mass contam-

ination is, in general, the worst type of contamination. Results on model 5 support this conjecture also for the Poisson regression model.

Given an estimator  $\hat{\beta}$ , we denote by MSE, the mean squared error defined by  $\mathbb{E}_{\beta_0}(\|\hat{\beta} - \beta_0\|^2)$ , where  $\|\cdot\|$  denotes the  $L_2$  norm. We estimate the MSE by

$$\hat{\text{MSE}} = \frac{1}{N} \sum_{j=1}^N \|\hat{\beta}_j - \beta_0\|^2,$$

where  $\hat{\beta}_j$  is the value of the estimator at the  $j$ -th replication and  $N$  is the number of replications which was chosen to equal 1000 for models 1 to 4 and 100 for model 5.

We performed an extensive simulation study, considering  $p = 6, 10, 20, 30, 40, 50$  and  $100$ ,  $\epsilon = 0, 0.05, 0.1$  and  $0.15$  and  $n = 100, 400$  and  $1000$  (sample size 100 was not investigated for the case  $p = 100$ ). Complete results are reported in the Supplementary Material. In Figures 1 to 5 we plot the MSE as a function of  $y_0$  for samples of size  $n = 400$  with covariates of dimension  $p = 30$  and  $p = 100$ . The proportion of outliers is  $\epsilon = 0.10$ .

Figures 1 to 5 indicate that the proposed estimator has smaller MSE than all other proposals for almost all the contaminations considered in this setting. We study the MSE as a function of  $y_0$  and consider, as a measure of robustness, the maximum MSE for  $y_0 \in \mathbb{Z}_{\geq 0}$ . The proposed estimator has the smallest maximum MSE for all the models considered.

For smaller values of  $p$ , FMT performs generally better than SMT and better than the rest of the estimators as well. The difference in favour of FMT is larger when the outliers have high leverage and when  $p$  is large.

In Figure 6 we report the execution time for the different methods. This figure indicates that our proposed method is a great improvement over the subsampling method in these settings, as far as computational time is concerned.

## 5. Example: Right Heart Catheterization

This data set was used by Connors et al. (1996) to study the effect of right heart catheterization in critically ill patients. It contains data from 5735 patients from five medical centers in the USA between 1989 and 1994 on several variables. These variables include laboratory measurements taken on day one, dates of admission and discharge, category of the primary disease, and whether or not the right heart catheterization was performed, among

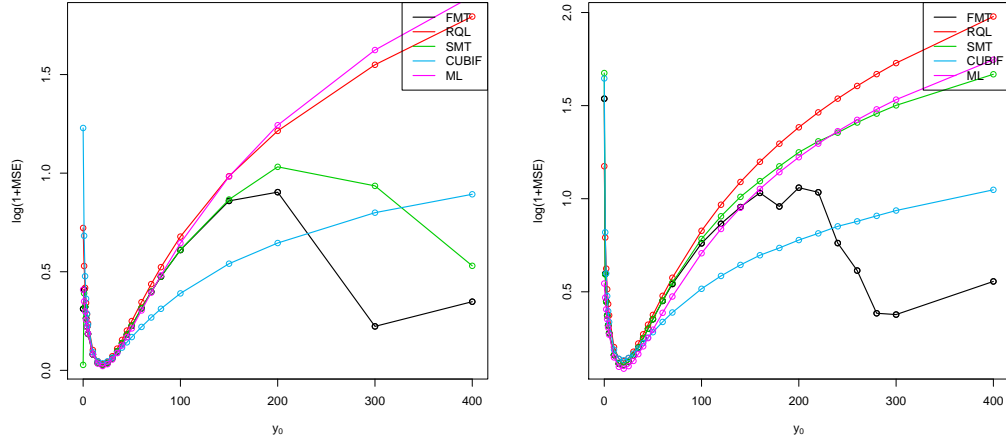


Figure 1: MSE for model 1,  $p = 30$  (left) and  $p = 100$  (right),  $n = 1000$  with 10% outliers at  $\mathbf{x}_0 = 1\mathbf{e}_1 + 3\mathbf{e}_2$ . Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

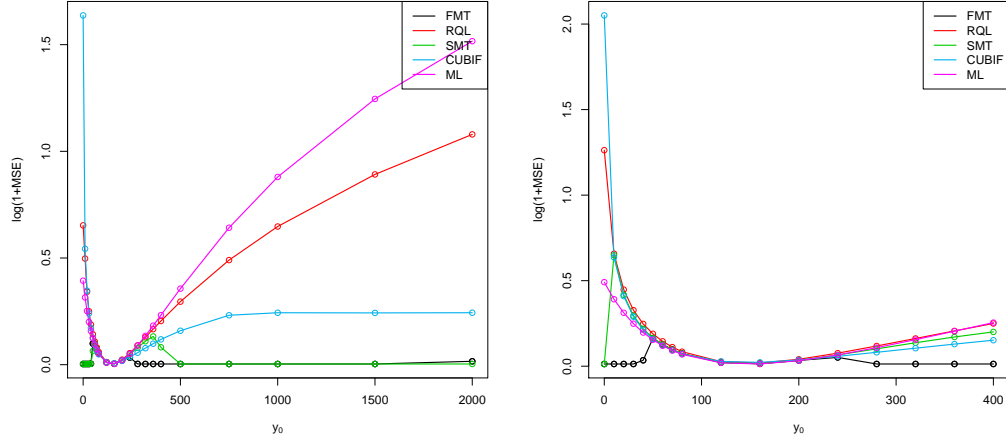


Figure 2: MSE for model 2,  $p = 30$  (left) and  $p = 100$  (right),  $n = 1000$  with 10% outliers at  $\mathbf{x}_0 = 1\mathbf{e}_1 + 3\mathbf{e}_2$ . Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

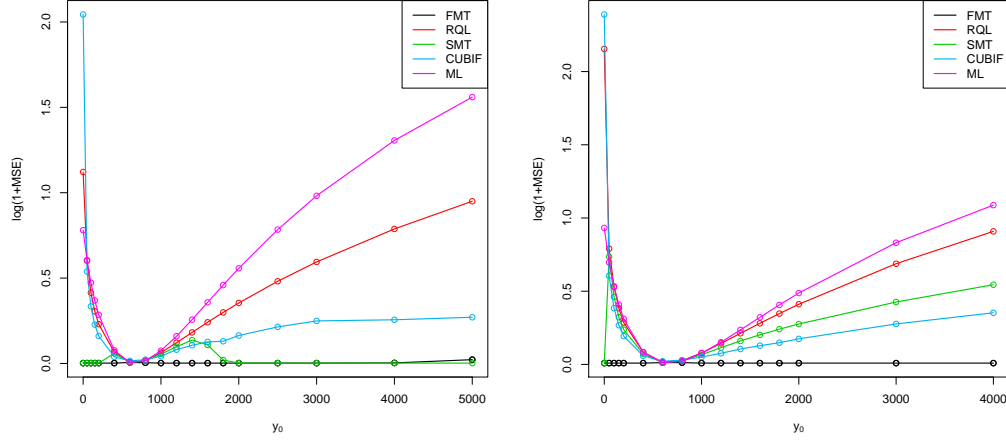


Figure 3: MSE for model 3,  $p = 30$  (left) and  $p = 100$  (right),  $n = 1000$  with 10% outliers at  $\mathbf{x}_0 = \mathbf{1e}_1 + 3\mathbf{e}_2$ . Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

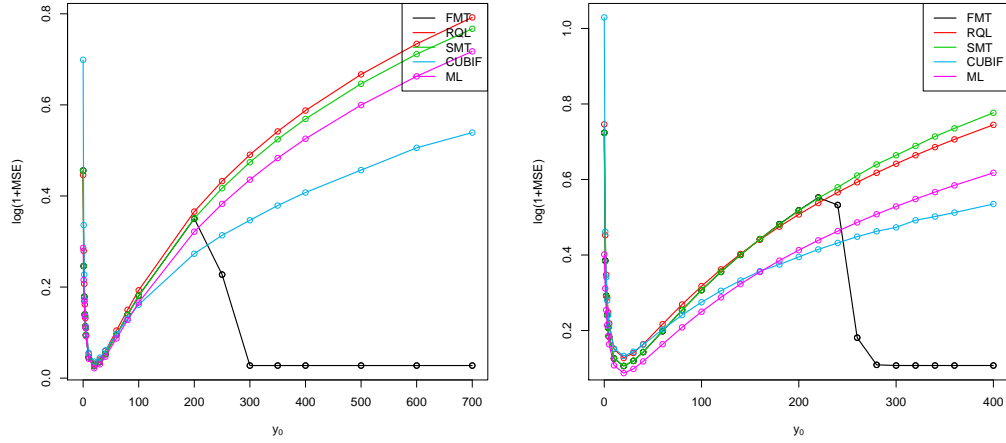


Figure 4: MSE for model 4,  $p = 30$  (left) and  $p = 100$  (right),  $n = 1000$  with 10% outliers at  $\mathbf{x}_0 = \mathbf{1e}_1 + 3\mathbf{e}_2 + 4\mathbf{e}_4$ . Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

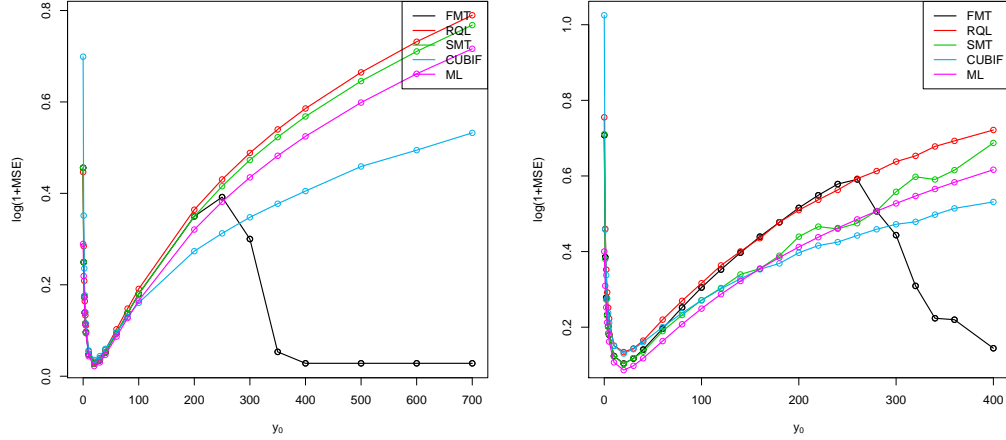


Figure 5: MSE for model 5,  $p = 30$  (left) and  $p = 100$  (right),  $n = 1000$  with 10% outliers at  $\mathbf{x}_0 = 1\mathbf{e}_1 + 3\mathbf{e}_2 + 4\mathbf{e}_4$ . Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

other features. A detailed description of the covariates can be found in Connors et al. (1996). The data were downloaded from the repository at Vanderbilt University, specifically from

<http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/rhc.csv>

We concentrate on the data corresponding to patients with congestive heart failure (CHF) as primary disease category. This leaves us with 456 observations, for which we want to use the available variables to explain the length of hospital stay. Since the study only involves patients that have been in hospital for 2 or more days, we define the response variable as  $y = \text{length of hospital stay} - 2$ , computed as discharge date minus admission date minus 2. The matrix  $\mathbf{x}$  of covariates contains information on 57 variables for each of the 456 patients. We assume that  $y|\mathbf{x}$  follows a Poisson distribution with mean  $\mu = \exp(\boldsymbol{\beta}^\top \mathbf{x})$  and we seek to estimate  $\boldsymbol{\beta}$  and to study its usefulness to explain and predict the length of hospital stay. We compute all the estimates using the complete sample of patients from CHF category.

After computing the estimators, we compute and draw boxplots of the deviance residuals for each fit. These boxplots are given in Figure 7. We also give the medians of the deviance residuals for each fit in Table 1.

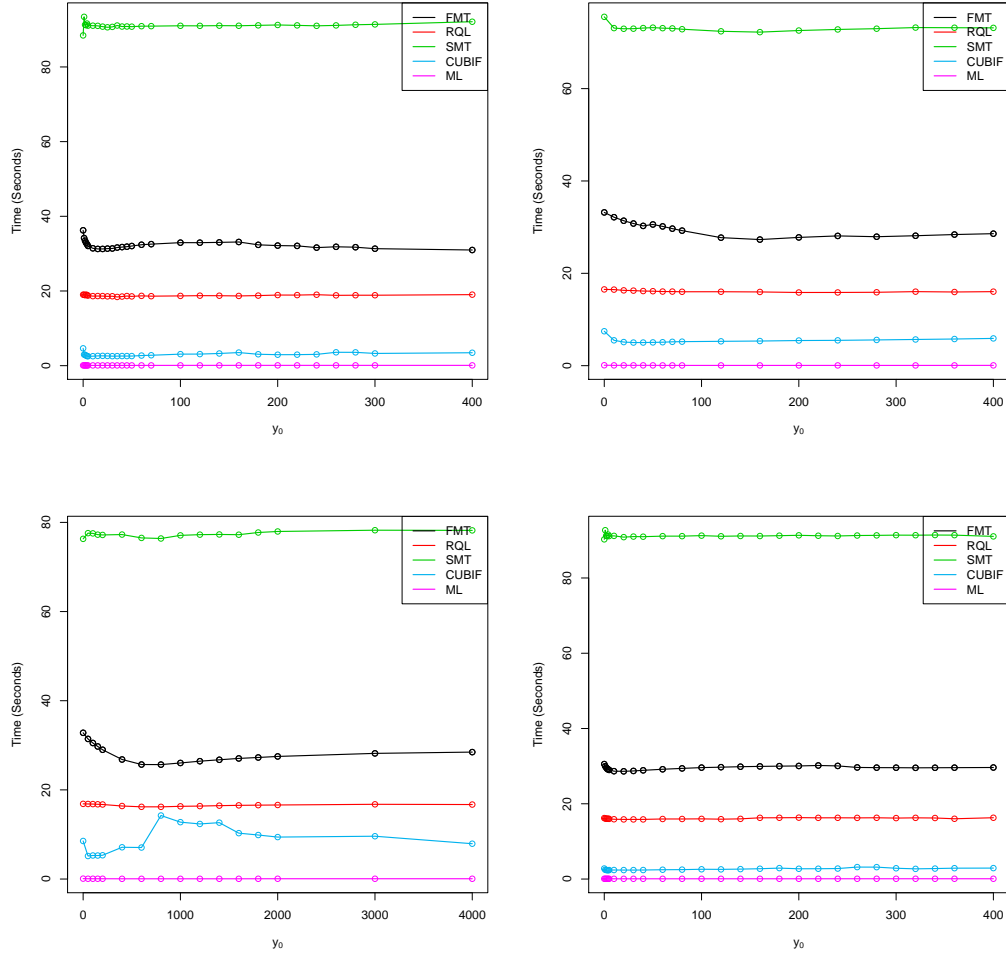


Figure 6: Execution time, in seconds, for  $p = 100$ ,  $n = 1000$  and 10% outliers. First row: models 1 and 2, second row: models 3 and 4. Black: FMT, red: RQL, green: SMT and light blue: CUBIF.

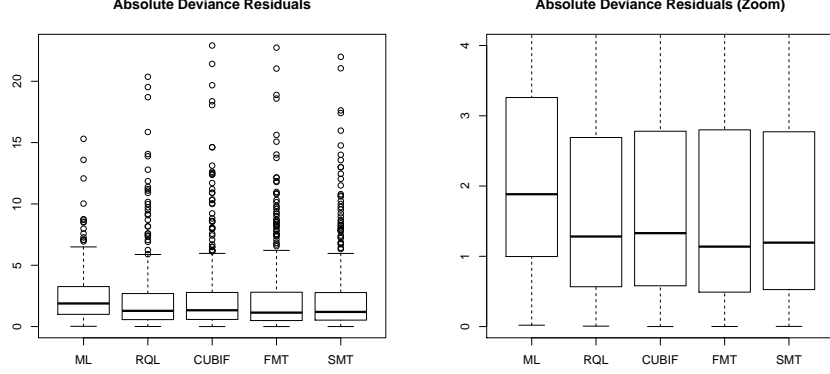


Figure 7: Absolute deviance residuals for RHC data

ML	RQL	CUBIF	FMT	SMT
1.88	1.28	1.50	1.14	1.19

Table 1: Median of absolute deviance residuals

The median of the deviance residuals of the FMT fit is the smallest. On the other hand the computational time of SMT is approximately 66 seconds, while the computational time of FMT is only 6 seconds. So we not only succeed in greatly decreasing the computational time but it seems that we also find a better solution in the sense that the deviance residuals of FMT are smaller than those of SMT at least for half the observations. We observe that FMT is slightly better at finding the minimum of the objective function (4) than SMT, since  $L(\hat{\beta}_{FMT}) = 170.4383$  while  $L(\hat{\beta}_{SMT}) = 170.5213$ , where  $\hat{\beta}_{FMT}$  and  $\hat{\beta}_{SMT}$  are the estimates obtained by the FMT and the SMT methods respectively.

In Figure 8 we compare other quantiles of the residuals, we plot the  $q$ -quantiles of the absolute deviance residuals versus  $q$  for each of the methods. We see that quantiles of the absolute deviance residuals corresponding to FMT and SMT are smaller than those of the other estimators for approximately 75% of the observations.

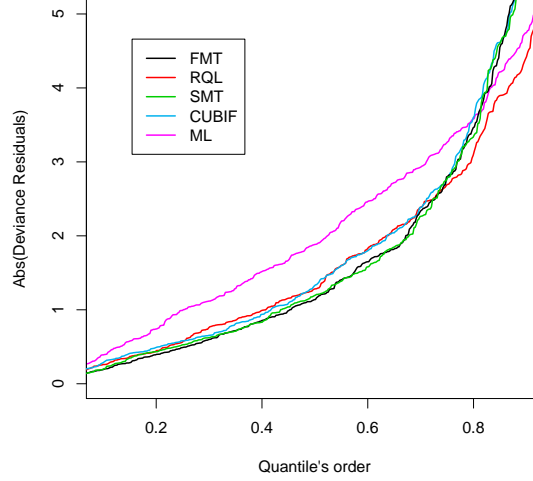


Figure 8: Quantiles of deviance residuals for RHC data

RQL	CUBIF	FMT	SMT
84	105	100	97

Table 2: Number of outliers detected by each method

In order to decide which observations are outliers we use the following bootstrap procedure. We choose a random observation and, using the vector of covariates  $\mathbf{x}$ , we generate its response according to the assumed model and the estimated parameter, that is to say, we simulate the sample from  $y \sim \mathcal{P}(\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}))$ . Then, we compute the deviance residual corresponding to  $(\mathbf{x}, y)$  and  $\hat{\boldsymbol{\beta}}$ . We repeat this procedure 5000 times and, in this way, we generate a sample of deviance residuals that follows the nominal model. Observations that are larger than the 0.9995 quantile or smaller the 0.0005 quantile of this sample are considered outliers. The number of outliers detected by each of the robust estimators considered is given in Table 2.

In Figure 9 we draw scatter plots of the deviance residuals generated using methods RQL, CUBIF and SMT vs the deviance residuals generated using



method FMT. In each plot, red points represent the observations that are considered outliers by both methods that are being considered, green points are considered outliers only by FMT and blue points are considered outliers only by the other method (RQL, CUBIF and SMT respectively). We see that, while FMT disagrees several points with RQL and CUBIF, it only disagrees in very few points with SMT. This is expected, since both SMT and FMT methods minimize the same objective function. However, these few differences may account for the small improvement in the median of the deviance residuals and in the minimum attained.

Finally, we compute the robust weights based on each of the methods and compare them using the scatter plots in Figure 10. In these plots the colours are chosen according to the FMT and SMT methods. This means that red points represent the observations that are considered outliers by both FMT and SMT, green points are considered outliers only by FMT and blue points are considered outliers only by SMT. This figure again shows the agreement between SMT and FMT and their disagreement with RQL and CUBIF. It also shows that both FMT and SMT give zero weights to the outliers, while RQL and CUBIF give positive weights to all observations. This partly explains their lower robustness.

## 6. Conclusion

We introduce a deterministic robust initial estimator for generalized linear models, which is used as a starting point for an iteratively reweighted least squares algorithm to obtain an MT-estimator. We illustrate the procedure for the Poisson model. Monte Carlo experiments show that MT-estimators computed with the proposed initial estimator have a small bounded mean squared error exhibiting a redescending behaviour. This is not the case for other proposals such as ML, RQL, CUBIF estimators. Finally, MT-estimators computed with an initial estimator based on subsampling, not only have a larger mean squared error than MT-estimators computed with the proposed initial estimator, but their computational time is much longer as well.

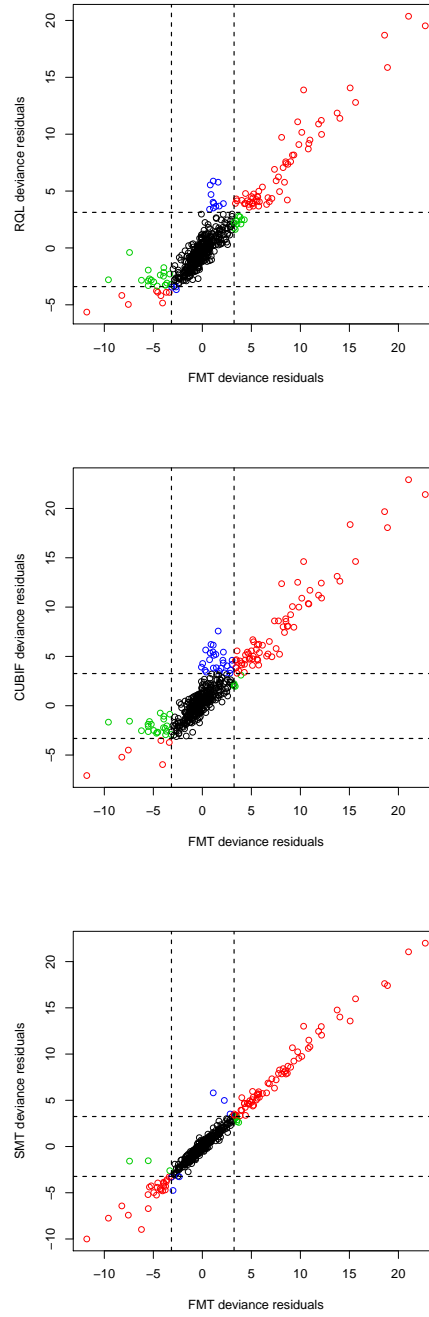


Figure 9: Deviance residuals generated by different methods vs deviance residuals generated by FMT method for RHC data. Dotted lines indicate 0.0005 and 0.9995 quantiles of the bootstrapped deviance residuals. Red points represent the observations that are considered outliers by both methods, green points are considered outliers only by FMT and blue points are considered outliers only by the other method considered in the plot (RQL, CUBIF and SMT respectively).

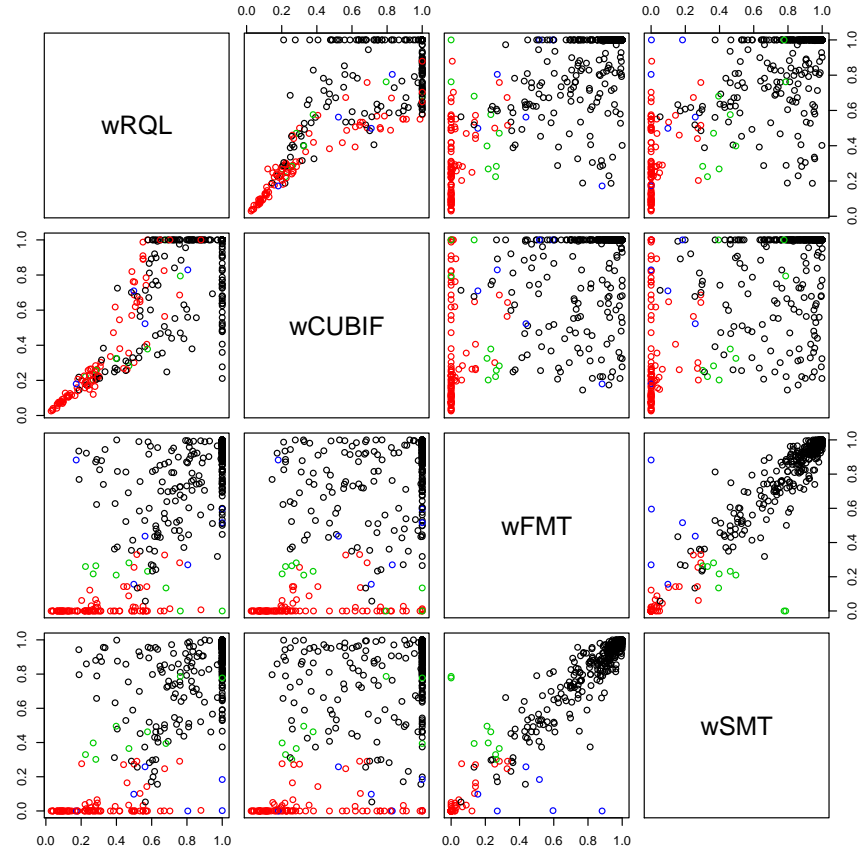


Figure 10: Scatterplots comparing the robust weights for each of the methods. Red points represent the observations that are considered outliers by both FMT and SMT, green points are considered outliers only by FMT and blue points are considered outliers only by SMT.

## Acknowledgments

Victor J. Yohai was partially supported by Grants PICT 2014-0351 from ANPCYT and 20020130100279BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

Marina Valdora was partially supported by Grant 20020130100279BA from the Universidad de Buenos Aires at Buenos Aires, Argentina.

This research was partially supported by the Italian-Argentinian project “Robust procedures to predict cost and duration of hospital stays” funded by the collaboration program MINCYT-MAE (IT1306-AR14MO6).

We would like to thank the authors of the R packages `checkmate` (Lang, 2017), `corrplot` (Wei and Simko, 2017), `MASS` (Venables and Ripley, 2002), `Rmpi` (Yu, 2002), `robcbi` (Marazzi, 2018a), `robeth` (Marazzi, 2018b), `robust` (Wang et al., 2017), `robustbase` (Maechler et al., 2018), `snow` (Tierney et al., 2016) and `xtbale` (Dahl, 2016) and of course many other packages and functions from R (R Core Team, 2018) since their works were of great help in the development of our R package and the running of the Monte Carlo simulation study and the examples.

## Appendix A. Computational details and algorithms

In this Appendix we describe the iteratively reweighted least squares algorithms that were used to compute the LST and the MT estimators. Suppose that we have an initial estimator  $\beta_0$  and call  $s(t) = m(g^{-1}(t))$ . Then, using a Taylor expansion of order one we can approximate  $m(g^{-1}(\mathbf{x}_i^\top \beta)) = s(\mathbf{x}_i^\top \beta)$  by

$$s(\mathbf{x}_i^\top \beta_0) + s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i^\top (\beta - \beta_0). \quad (\text{A.1})$$

Then, an approximate value to the LST estimator can be found as the value  $\beta_1$  that minimizes

$$\sum_{i=1}^n (t(y_i) - s(\mathbf{x}_i^\top \beta_0) - s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i^\top (\beta - \beta_0))^2.$$

Therefore,  $\beta_1 - \beta_0$  is the LS estimator for a linear model with responses  $t(y_1), \dots, t(y_n)$  and regressor vectors  $s'(\mathbf{x}_1^\top \beta_0) \mathbf{x}_1, \dots, s'(\mathbf{x}_n^\top \beta_0) \mathbf{x}_n$  and consequently

$$\beta_1 = \beta_0 + (\mathbf{X}^\top \mathbf{W} (\mathbf{X} \beta_0)^2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{X} \beta_0) (\mathbf{T} - s(\mathbf{X} \beta_0)), \quad (\text{A.2})$$

where  $\mathbf{X}$  is the  $n \times p$  matrix whose  $i$ -th row is  $\mathbf{x}_i^\top$ ,  $\mathbf{s}(\mathbf{X}\boldsymbol{\beta}) = (s(\mathbf{x}_1^\top \boldsymbol{\beta}), \dots, s(\mathbf{x}_n^\top \boldsymbol{\beta}))^\top$ ,  $\mathbf{W}(\mathbf{X}\boldsymbol{\beta})$  is the diagonal matrix with diagonal elements  $s'(\mathbf{x}_1^\top \boldsymbol{\beta}), \dots, s'(\mathbf{x}_n^\top \boldsymbol{\beta})$  and  $\mathbf{T} = (t(y_1), \dots, t(y_n))^\top$ .

An iterative procedure to compute the LST estimator can be obtained by

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + (\mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}_k)^2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}_k)(\mathbf{T} - \mathbf{s}(\mathbf{X}\boldsymbol{\beta}_k)). \quad (\text{A.3})$$

Iterations will continue until  $\|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\| / \|\boldsymbol{\beta}_k\| \leq \delta$ , where  $\delta$  is the error tolerance.

Suppose that  $\boldsymbol{\beta}_k$  converges to  $\boldsymbol{\beta}^*$ ; then this value should satisfy the LST estimating equation. In fact, taking limit in both sides of (A.3) we get

$$(\mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}^*)^2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}^*)(\mathbf{T} - \mathbf{s}(\boldsymbol{\beta}^*)) = 0,$$

which is equivalent to

$$\mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}^*)(\mathbf{T} - \mathbf{s}(\boldsymbol{\beta}^*)) = 0.$$

Then,  $\boldsymbol{\beta}^*$  satisfies the estimating equation of the LST estimator.

To start the algorithm, it will be convenient to write equation (A.2) in the following, slightly different, way

$$\boldsymbol{\beta}_1 = (\mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}_0)^2 \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}_0)^2 \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{W}(\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{T} - \mathbf{s}(\mathbf{X}\boldsymbol{\beta}_0))). \quad (\text{A.4})$$

Observe that, according to (A.4), to compute  $\boldsymbol{\beta}_1$  we only need to give  $\boldsymbol{\eta}_0 = \mathbf{X}\boldsymbol{\beta}_0$ . Then, since for Poisson regression and log link,  $\mathbf{x}_i^\top \boldsymbol{\beta} = \log(\mathbb{E}(y_i))$ , it seems reasonable to take  $\boldsymbol{\eta}_0 = (\log(y_1 + 0.1), \dots, \log(y_n + 0.1))^\top$ . The value 0.1 is added to avoid numerical problem when  $y_i = 0$ . To compute the LST estimators in our procedure, only one iteration is performed. The reason is that, for these auxiliary estimators, the accuracy is not as important as the speed at which they can be computed. Our experiments show that there is no noticeable loss in the precision of the final estimate by doing this but, on the other hand, the computation times decrease significantly.

We describe now an analogous iterative algorithm for computing the MT estimator. Suppose that we have an initial robust estimator  $\boldsymbol{\beta}_0$ . We compute a new value using two approximations. As in the case of the LST estimator, replacing, in (4),  $m(g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}))$  by (A.1) we consider the approximate loss function

$$\sum_{i=1}^n \rho(t(y_i) - s(\mathbf{x}_i^\top \boldsymbol{\beta}_0) - s'(\mathbf{x}_i^\top \boldsymbol{\beta}_0) \mathbf{x}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}_0)).$$

Differentiating with respect to  $\beta$  we obtain the estimating equation

$$\sum_{i=1}^n \psi(t(y_i) - s(\mathbf{x}_i^\top \beta_0) - s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i^\top (\beta - \beta_0)) s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i = 0, \quad (\text{A.5})$$

where  $\psi = \rho'$ . Note that this equation can be written as

$$\sum_{i=1}^n (t(y_i) - s(\mathbf{x}_i^\top \beta_0) - s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i^\top (\beta - \beta_0)) w(\mathbf{x}_i^\top \beta, \mathbf{x}_i^\top \beta_0) s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i, \quad (\text{A.6})$$

where

$$w(u, v) = \frac{\psi(t(y_i) - s(v) - s'(v)(u - v))}{t(y_i) - s(v) - s'(v)(u - v)}.$$

Since  $\beta$  should be close to  $\beta_0$ , the second approximation is to replace, in (A.6),  $w(\mathbf{x}_i^\top \beta, \mathbf{x}_i^\top \beta_0)$  by  $w^*(\mathbf{x}_i^\top \beta_0) = w(\mathbf{x}_i^\top \beta_0, \mathbf{x}_i^\top \beta_0)$ . Then  $\beta_1$  is defined as the solution to the approximate estimating equation:

$$\sum_{i=1}^n (t(y_i) - s(\mathbf{x}_i^\top \beta_0) - s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i^\top (\beta - \beta_0)) w^*(\mathbf{x}_i^\top \beta_0) s'(\mathbf{x}_i^\top \beta_0) \mathbf{x}_i,$$

and is given by

$$\beta_1 = \beta_0 + (\mathbf{X}^\top \mathbf{W}^2(\mathbf{X}\beta_0) \mathbf{W}^*(\mathbf{X}\beta_0) \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{X}\beta_0) \mathbf{W}^*(\mathbf{X}\beta_0) (\mathbf{T} - \mathbf{s}(\mathbf{X}\beta_0)),$$

where  $\mathbf{W}^*(\mathbf{X}\beta)$  is the  $n \times n$  diagonal matrix with diagonal elements  $w^*(\mathbf{x}_1^\top \beta), \dots, w^*(\mathbf{x}_n^\top \beta)$ .

Then, the iterative procedure to compute the MT estimator is given by

$$\beta_{k+1} = \beta_k + (\mathbf{X}^\top \mathbf{W}^2(\mathbf{X}\beta_k)^\top \mathbf{W}^*(\mathbf{X}\beta_k) \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}(\mathbf{X}\beta_k) \mathbf{W}^*(\mathbf{X}\beta_k) (\mathbf{T} - \mathbf{s}(\mathbf{X}\beta_k)). \quad (\text{A.7})$$

Suppose that  $\beta_k \rightarrow \beta^*$ , then taking limits in both sides of (A.7), we get

$$\mathbf{X}^\top \mathbf{W}(\mathbf{X}\beta^*) \mathbf{W}^*(\mathbf{X}\beta^*) (\mathbf{T} - \mathbf{s}(\mathbf{X}\beta^*)) = \mathbf{0},$$

and this is equivalent to

$$\mathbf{X}^\top \mathbf{W}(\mathbf{X}\beta^*) \Psi(\mathbf{X}\beta^*) = \mathbf{0}, \quad (\text{A.8})$$

where  $\Psi(\mathbf{X}\beta) = (\psi(t(y_1) - s(\mathbf{x}_1^\top \beta)), \dots, \psi(t(y_n) - s(\mathbf{x}_n^\top \beta)))^\top$ . Then  $\beta^*$  satisfies the estimating equation of the MT estimator.

- Agostinelli, C., Valdora, M., Yohai, V., 2018. Robust M-Estimators Based on Transformations for Poisson Model. R package version 0.3-5.  
URL <https://cran.r-project.org/web/packages/poissonMT/index.html>
- Alqallaf, F., Agostinelli, C., 2016. Robust inference in generalized linear models. *Communications in Statistics - Simulation and Computation* 45 (9), 3053–3073.
- Bergesio, A., Yohai, V., 2011. Projection estimators for generalized linear models. *Journal of the American Statistical Association* 106, 661–671.
- Bianco, A., Boente, G., Rodrigues, I., 2013. Resistant estimators in Poisson and gamma models with missing responses and an application to outlier detection. *Journal of Multivariate Analysis* 114, 209–226.
- Cantoni, E., Ronchetti, E., 2001. Robust inference for generalized linear models. *Journal of the American Statistical Association* 96, 1022–1030.
- Connors, A., Speroff, T., Dawson, N., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., Knaus, W., 1996. The effectiveness of right heart catheterization in the initial care of critically ill patients. *The Journal of the American Medical Association* 276 (11), 889–897.
- Cook, R., 1977. Detection of influential observation in linear regression. *Technometrics* 19 (1), 15–18.
- Dahl, D. B., 2016. xtable: Export Tables to LaTeX or HTML. R package version 1.8-2.  
URL <https://CRAN.R-project.org/package=xtable>
- Künsch, H., Stefanski, L., Carroll, R., 1989. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association* 84, 460–466.
- Lang, M., 2017. checkmate: Fast argument checks for defensive r programming. *The R Journal* 9 (1), 437–445.  
URL <https://journal.r-project.org/archive/2017/RJ-2017-028/index.html>

- Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E., Anna di Palma, M., 2018. robustbase: Basic Robust Statistics. R package version 0.93-1.  
URL <http://robustbase.r-forge.r-project.org/>
- Marazzi, A., 2018a. robcbi: Conditionally Unbiased Bounded Influence Estimates. R package version 1.1-2.  
URL <https://CRAN.R-project.org/package=robcbi>
- Marazzi, A., 2018b. robeth: R Functions for Robust Statistics. R package version 2.7-2.  
URL <https://CRAN.R-project.org/package=robeth>
- Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics. Theorey and Methods. Wiley.
- McCullagh, P., Nelder, J., 1989. Generalized Linear Models, 2nd Edition. Chapman and Hall/CRC.
- Peña, D., Yohai, V., 1999. A fast procedure for outlier diagnostics in large regression problems. Journal of the American Statistical Association 94, 434–445.
- R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.  
URL <https://www.R-project.org/>
- Rousseeuw, P., Leroy, A., 1987. Robust regression and outlier detection. Wiley and Sons.
- Tierney, L., Rossini, A. J., Li, N., Sevcikova, H., 2016. snow: Simple Network of Workstations. R package version 0.4-2.  
URL <https://CRAN.R-project.org/package=snow>
- Valdora, M., Yohai, V., 2014. Robust estimators for generalized linear models. Journal of Statistical Planning and Inference 146, 31–48.
- Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York, iSBN 0-387-95457-0.  
URL <http://www.stats.ox.ac.uk/pub/MASS4>



- Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin, D., Maechler, M., Konis., K., 2017. robust: Port of the S+ "Robust Library". R package version 0.4-18.  
URL <https://CRAN.R-project.org/package=robust>
- Wei, T., Simko, V., 2017. R package "corrplot": Visualization of a Correlation Matrix. (Version 0.84).  
URL <https://github.com/taiyun/corrplot>
- Yu, H., 2002. Rmpi: Parallel statistical computing in r. R News 2 (2), 10–14.  
URL [https://cran.r-project.org/doc/Rnews/Rnews\\_2002-2.pdf](https://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf)