Robust inference and modeling of mean and dispersion for generalized linear models

Robust Inference and Modeling of Mean and Dispersion for Generalized Linear Models

Abstract

Generalized Linear Models (GLMs) are a popular class of regression models when the responses follow a distribution in the exponential family. In real data the variability often deviates from the relation imposed by the exponential family distribution, which results in over- or underdispersion. Dispersion effects may even vary in the data. Such data sets do not follow the traditional GLM distributional assumptions, leading to unreliable inference. Therefore, the family of double exponential distributions has been proposed, which models both the mean and the dispersion as a function of covariates in the GLM framework. Since standard maximum likelihood inference is highly susceptible to the possible presence of outliers, we propose the robust double exponential (RDE) estimator. Asymptotic properties and robustness of the RDE estimator are discussed. A generalized robust quasi-deviance measure is introduced which constitutes the basis for a stable robust test. Simulations for binomial and Poisson models show the excellent performance of the RDE estimator and corresponding robust tests. Penalized versions of the RDE estimator are developed for sparse estimation with high-dimensional data and for flexible estimation via generalized additive models (GAMs). Real data applications illustrate the relevance of robust inference for dispersion effects in GLMs and GAMs.

Keywords: Double exponential family, Likelihood ratio test, M-estimator, Influence function, Penalization

1 Introduction

Generalized Linear Models (GLMs) form a unified way of modeling the mean response when the responses follow an exponential family distribution (see e.g. McCullagh and Nelder, 1989). In practice, real data often display a larger or smaller variability than expected under a standard GLM. In these cases the data are said to be over- or underdispersed, respectively. Such data sets typically invalidate the standard GLM distributional assumptions. Moreover, dispersion effects may be different for subgroups in the data or depend on a set of covariates. It is critical to account for dispersion for several reasons. First, correct inference, e.g. confidence intervals for the mean response, depends on the dispersion (Smyth, 1989; Cai et al., 2008). Secondly, neglecting dispersion may result in a loss of efficiency and a bias in the estimation of the regression coefficients in the mean model (Smyth and Verbyla, 1999; Antoniadis et al., 2016). Thirdly, the dispersion model itself may be the main focus of interest (Lian et al., 2015).

To model the dispersion in a GLM framework, Efron (1986) proposed the family of double exponential distributions. It generalizes the single parameter exponential family by including an additional parameter to model the dispersion. More formally, suppose that the variable Y follows a one-parameter exponential family with parameter μ and density $e_Y(y;\mu)$, denoted by $Y \sim \mathrm{EF}(\mu)$. The variance of Y, which may depend on the parameter μ, is denoted by $V(\mu)$. Then, the corresponding double exponential family with parameters μ and $\theta > 0$ is defined as

$$\overline{f}(y;\mu,\theta) = c(\mu,\theta)\theta^{1/2} e_Y(y;\mu)^{\theta} e_Y(y;y)^{1-\theta}.$$

A variable Y with a distribution belonging to the double exponential family is denoted by $Y \sim \mathrm{DEF}(\mu,\theta)$. Efron (1986) showed that the normalizing constant $c(\mu,\theta)$ which ensures that $\overline{f}(y;\mu,\theta)$ is a density, is approximately equal to 1. In practice, one may thus approximate $\overline{f}(y;\mu,\theta)$ by $f(y;\mu,\theta)$ which is obtained by setting $c(\mu,\theta) = 1$. Efron (1986) also showed that $\mathrm{E}\left[Y\right] \approx \mu$ and $\mathrm{Var}\left[Y\right] \approx \frac{V(\mu)}{\theta}$. Hence, the parameter θ represents underdispersion when $\theta > 1$ and overdispersion when $\theta < 1$. Note that when $\theta = 1$, the density $\overline{f}(y;\mu,\theta)$ reduces to $e_Y(y;\mu)$. Therefore, the single parameter exponential family is obtained as a special case.

In a regression context, we assume that $Y\,|\,\boldsymbol{x},\boldsymbol{z} \sim \mathrm{DEF}(\mu,\theta)$. The parameters μ and θ thus depend on predictor variables $\boldsymbol{x} \in \mathbb{R}^{p_1}$ and $\boldsymbol{z} \in \mathbb{R}^{p_2}$, respectively. This leads to the following combined regression model

$$\mu = h(\boldsymbol{x}^t\boldsymbol{\beta}) \quad \text{and} \quad \theta = g(\boldsymbol{z}^t\boldsymbol{\gamma}), \tag{1}$$

with h and g monotone functions and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors of length p_1 and p_2, respectively. Note that we consider functions h and g that are invertible. Moreover, g should be positive since $\theta > 0$. A natural choice for the link function h is to take the same choice as for the standard GLM in absence of dispersion. For the dispersion model such a natural choice for the link function is not available (Efron, 1986). A common choice for g is an exponential function. Other possibilities are the inverse $g(t) = 1/(1+t)$ (Lee and Nelder, 2000) or the logistic-like function $g(t) = 1.25/(1+\exp(-t))$ of Efron (1986).

Based on a random sample of n observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, the maximum likelihood estimates (MLE) for the parameters of the above model are obtained by solving the classical score equations:

$$\sum_{i=1}^{n} \boldsymbol{U}\left(y_i,\mu_i,\theta_i\right) = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{U}_{\boldsymbol{\beta}}\left(y_i,\mu_i,\theta_i\right) \\ \boldsymbol{U}_{\boldsymbol{\gamma}}\left(y_i,\mu_i,\theta_i\right) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{U}_{\mu_i}\,\mu_i' \\ \boldsymbol{U}_{\theta_i}\,\theta_i' \end{pmatrix} = \boldsymbol{0}. \tag{2}$$

Here, $\mu_i' = \frac{\partial}{\partial\boldsymbol{\beta}}\mu_i$, $\theta_i' = \frac{\partial}{\partial\boldsymbol{\gamma}}\theta_i$, and $\boldsymbol{U}_{\mu_i} = \frac{\partial}{\partial\mu_i}\boldsymbol{\mathcal{L}}\left(y_i,\mu_i,\theta_i\right)$, $\boldsymbol{U}_{\theta_i} = \frac{\partial}{\partial\theta_i}\boldsymbol{\mathcal{L}}\left(y_i,\mu_i,\theta_i\right)$ with $\boldsymbol{\mathcal{L}}\left(y,\mu,\theta\right)$ the log likelihood function corresponding to the double exponential generalized linear model in (1). Hence, $\boldsymbol{U}(y,\mu,\theta)$ is the score function corresponding to the MLE. However, it is well-known that the maximum likelihood estimator (CDE) and associated inference is extremely sensitive to model deviations and outliers in the data.

Other approaches to model the mean response and the dispersion simultaneously have been proposed in the literature. Nelder and Pregibon (1987) proposed the extended quasi-likelihood framework where the deviance is modeled to account for the dispersion. Lee and Nelder (1998) proposed a restricted Extended Quasi-Likelihood estimator (CEQL) which uses adjusted deviances to reduce the bias when estimating the dispersion parameters with a relatively large number of mean parameters. Lee and Nelder (2000) showed that extended quasi-likelihood inference and maximum likelihood inference in double exponential models lead to identical results. Other proposals in the statistical literature are pseudo likelihood (Davidian and Carroll, 1987), double generalized linear models (Smyth, 1989) and dispersion models (Jørgensen, 1987; Jørgensen, 1997). Comparisons can be found in Nelder and Lee (1992) and Davidian and Carroll (1988).

The non-robustness of maximum likelihood inference implies that outliers may bias the parameter estimates and confidence intervals and also hypothesis tests may become unreliable and/or uninformative. Therefore, various robust alternatives have been proposed in the context of GLMs, such as Cantoni and Ronchetti (2001), Bergesio and Yohai (2011), Valdora and Yohai (2014) and Ghosh and Basu (2016). Several contributions focus on a specific GLM. Robust logistic regression has been studied by Künsch et al. (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996), Croux and Haesbroeck (2003), Bondell (2005, 2008) and Hosseinian and Morgenthaler (2011), whereas the negative binomial case has been studied by Aeberhard et al. (2014, 2017) and Amiguet et al. (2017). For the Gamma model, robust estimators were proposed by Bianco et al. (2005) and Cantoni and Ronchetti (2006). To our knowledge, only two robust GLM methods have focused on modeling the dispersion. Croux et al. (2012) and Neykov et al. (2012) proposed robustified versions of the extended quasi-likelihood approach. Croux et al. (2012) introduced robust M-estimators for Generalized Additive Models (GAMs), including GLMs as a special case. Neykov et al. (2012) exploited the idea of trimming to obtain robust estimators. However, robust inference for GLMs and robust dispersion tests have not been considered for these proposals.

In this paper, we present a robust estimator for double exponential family GLMs and we develop associated tests for robust inference based on these flexible models. The robust inference allows in particular to test for the presence of dispersion. The proposed method allows to model both the mean μ and dispersion θ in (1) based on a (possibly different) set of predictors $\boldsymbol{x}$ and $\boldsymbol{z}$ and is valid for any double exponential distribution.

The remainder of this paper is structured as follows. Section 2 presents the robust double exponential estimator in general. We discuss the popular Poisson and binomial models as particular cases. In Section 3, we construct robust inference for double exponential GLMs based on robust likelihood ratio techniques. The finite-sample performance of the robust inference is investigated by means of simulations in Section 4. In Section 5 we illustrate the methodology on some real data examples. In Section 6 we develop penalized RDE estimators. We consider both penalties to obtain sparsity in high-dimensional settings and regularization penalties in the context of flexible smooth estimation via GAMs. Section 7 concludes with a final discussion. Derivations of theoretical results are given in the Appendix and Supplementary Material which also contains additional results.

2 The Robust Double Exponential (RDE) Estimator

2.1 General double exponential GLMs

The MLE for double exponential GLMs corresponding to the estimating equations (2) is very sensitive to outlier(s) in both the response and the explanatory variables. Therefore, we consider a general class of M-estimators of Mallows' type as a robust alternative. Our robust double exponential (RDE) estimator is defined as the solution of the following estimating equations:

$$\sum_{i=1}^{n} \Psi\left(y_i, \mu_i, \theta_i\right) = \sum_{i=1}^{n} \begin{pmatrix} \Psi_{\boldsymbol{\beta}}\left(y_i, \mu_i, \theta_i\right) \\ \Psi_{\boldsymbol{\gamma}}\left(y_i, \mu_i, \theta_i\right) \end{pmatrix} = \mathbf{0}, \tag{3}$$

where

$$\begin{aligned} \Psi_{\boldsymbol{\beta}}\left(y_i, \mu_i, \theta_i\right) &= \nu_1(y_i, \mu_i, \theta_i) w_1(\boldsymbol{x}_i, \boldsymbol{z}_i) \mu_i' - a_1, \qquad (4) \\ \Psi_{\boldsymbol{\gamma}}\left(y_i, \mu_i, \theta_i\right) &= \nu_2(y_i, \mu_i, \theta_i) w_2(\boldsymbol{x}_i, \boldsymbol{z}_i) \theta_i' - a_2. \end{aligned}$$

The constants a_1 and a_2 make the estimator Fisher consistent and are given by

$$a_1 = \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}\left[\nu_1(Y_j, \mu_j, \theta_j)\right] w_1(\boldsymbol{x}_j, \boldsymbol{z}_j) \mu_j' \quad \text{and} \quad a_2 = \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}\left[\nu_2(Y_j, \mu_j, \theta_j)\right] w_2(\boldsymbol{x}_j, \boldsymbol{z}_j) \theta_j',$$

where the expectations are with respect to the conditional distributions $Y_j \,|\, \boldsymbol{x}, \boldsymbol{z}$ which follow the double exponential GLM in (1) with mean μ_j and dispersion parameter θ_j. The RDE estimates can be calculated by using Fisher scoring and alternating between the $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ updates as outlined in Appendix 8.2.

The RDE estimator in (3) is an M-estimator with score function $\Psi\left(y, \mu, \theta\right)$. An important measure of the robustness of the RDE estimator is its influence function (Huber, 1981; Hampel et al., 1986). Intuitively, the influence function measures the change in the estimator when the model is perturbed by an infinitesimal small amount of contamination at location $(y, \boldsymbol{x}, \boldsymbol{z})$. Estimators with an unbounded influence function are extremely sensitive to perturbations in the data because a small amount of contamination can already have an arbitrarily large effect on the estimator. Therefore, estimators with bounded influence function are preferred. We now derive the influence function of the RDE estimator and investigate under which assumptions boundedness is obtained.

The influence function of an M-estimator is given by $\mathrm{IF}(y, \Psi, F) = M(\Psi, F)^{-1}\Psi(y, \mu, \theta)$, with $M(\Psi, F) = -\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\eta}}\Psi(Y, \mu, \theta)\right]$ where $\boldsymbol{\eta} = \left(\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t\right)^t$ is the parameter vector containing all the model parameters. For the RDE estimator an expression for $M(\Psi, F)$ is derived in the supplementary material. Since the influence function of an M-estimator is proportional to its score function $\Psi(y, \mu, \theta)$, choosing a bounded score function leads to a robust RDE estimator with bounded influence function. Note that the MLE in (2) is an M-estimator with score function $\Psi(y, \mu, \theta) = \boldsymbol{U}(y, \mu, \theta)$ which generally is unbounded, confirming the non-robustness of the MLE. To guarantee the boundedness of the score functions $\Psi_{\boldsymbol{\beta}}$ and $\Psi_{\boldsymbol{\gamma}}$, bounded functions $\nu_1(y, \mu, \theta)$ and $\nu_2(y, \mu, \theta)$ are needed to control large deviations in the response, while the weight functions $w_1(\boldsymbol{x}, \boldsymbol{z})$ and $w_2(\boldsymbol{x}, \boldsymbol{z})$ in (4) are needed to downweight the effect of leverage points in the $\boldsymbol{x}$ and/or $\boldsymbol{z}$-space.

An intuitively appealing choice for $\nu_1(y, \mu, \theta)$ and $\nu_2(y, \mu, \theta)$ is

$$\begin{aligned}\nu_1(y, \mu, \theta) &= v_1(r)\boldsymbol{U}_\mu, \\ \nu_2(y, \mu, \theta) &= v_2(r)\boldsymbol{U}_\theta, \qquad (5)\end{aligned}$$

with $r = (y - \mu)/\sqrt{V(\mu)/\theta}$, the scaled Pearson residual of an observation. Here, $v_1(r)$ and

$v_2(r)$ are weight functions that should downweight the contribution of outlying responses to the standard score functions $\boldsymbol{U}_\mu$ and $\boldsymbol{U}_\theta$ in (2). As the form of the score functions $\boldsymbol{U}_\mu$ and $\boldsymbol{U}_\theta$ is determined by the likelihood of the specific double exponential model, these weight functions should be carefully chosen such that they are able to reduce the effect of potential outliers in the response sufficiently, resulting in an estimator with bounded influence function, see the examples in Section 2.2. A common choice for these weight functions is $v_j(r) = \psi(r)/r;\ j = 1, 2$ where $\psi(r)$ diminishes the effect of large residuals. A popular choice is the Huber function defined as $\psi_{\mathrm{H}}(r, c) = \max(-c, \min(c, r))$ with c a tuning constant providing a trade-off between efficiency and robustness. Alternatively, the redescending Tukey bisquare function $\psi_{\mathrm{T}}(r, c) = \left((r/c)^2 - 1\right)^2 r\, I(|r| \leqslant c)$ can be used. For more information, we refer the reader to Rousseeuw and Leroy (2005).

The functions $w_1(\boldsymbol{x}, \boldsymbol{z})$ and $w_2(\boldsymbol{x}, \boldsymbol{z})$ are used to downweight potential leverage points and may be chosen to factor over the arguments. That is, $w_1(\boldsymbol{x}, \boldsymbol{z}) = w_X(\boldsymbol{x})w_Z(\boldsymbol{z}) = w_2(\boldsymbol{x}, \boldsymbol{z})$ for example, where $w_X(\boldsymbol{x})$ and $w_Z(\boldsymbol{z})$ are often taken to be the inverse of a robustly estimated Mahalanobis distance. For a p-dimensional variable U this weight function is given by $w_U(\boldsymbol{u}) = d(\boldsymbol{u}, \hat{\boldsymbol{\mu}}_U, \hat{\Sigma}_U)^{-1/2}$ with $d(\boldsymbol{u}, \hat{\boldsymbol{\mu}}_U, \hat{\Sigma}_U) = (\boldsymbol{u} - \hat{\boldsymbol{\mu}}_U)^t \hat{\Sigma}_U^{-1} (\boldsymbol{u} - \hat{\boldsymbol{\mu}}_U)$ where $\hat{\boldsymbol{\mu}}_U$ and $\hat{\Sigma}_U$ are robust location and scatter matrix estimates of U, respectively. These estimates can be obtained by high-breakdown estimators of location and scatter such as the minimum covariance determinant (MCD) estimator (Rousseeuw, 1984), S-estimators (Lopuhaä, 1989) or MM-estimators (Tatsuoka and Tyler, 2000), for instance. Alternatively, a hard cutoff rule may be used. In this case, all observations whose robustly estimated Mahalanobis distance exceeds a cutoff, e.g. $\chi^2_{p,0.975}$ which denotes the 97.5% quantile from a χ^2_p-distribution, are given weight zero while the remaining observations receive weight 1, i.e. $w_U(\boldsymbol{u}) = I(d(\boldsymbol{u}, \hat{\boldsymbol{\mu}}_U, \hat{\Sigma}_U) \leq \chi^2_{p,0.975})$ with $I(\cdot)$ the indicator function.

Note that by taking $w_1(\boldsymbol{x}, \boldsymbol{z}) = w_2(\boldsymbol{x}, \boldsymbol{z}) = 1$ and $v_1(r) = v_2(r) = 1$ in (5), we recover the standard MLE. Moreover, when there is no dispersion, i.e. all $\theta_i = 1$, the RDE estimator simplifies to the robust estimator of Cantoni and Ronchetti (2001), hence our proposal can be seen as a generalization of their robust estimator for GLMs.

M-estimators are consistent and asymptotically normally distributed under suitable conditions. We assume conditions (A1)-(A9) stated in the supplementary material which correspond to those in Cantoni and Ronchetti (2001) and have previously been studied by Huber (1981), Clarke (1986) and Bednarski (1993) among others. Let $F_{\boldsymbol{\eta}}$ denote the model distribution corresponding to the double exponential GLM in (1) with parameter $\boldsymbol{\eta} = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)^t$. Then, the asymptotic variance of M-estimators at $F_{\boldsymbol{\eta}}$ is given by

$$\Omega = M(\Psi, F_{\boldsymbol{\eta}})^{-1} Q(\Psi, F_{\boldsymbol{\eta}}) M(\Psi, F_{\boldsymbol{\eta}})^{-t},$$

with $Q(\Psi, F_{\boldsymbol{\eta}}) = \mathrm{E}\left[\Psi(Y, \mu, \theta)\Psi(Y, \mu, \theta)^t\right]$. For the RDE estimator an expression for the matrix $Q(\Psi, F_{\boldsymbol{\eta}})$ is derived in the supplementary material.

The ratio of the trace of the asymptotic variances of the RDE estimator and the MLE at the double exponential generalized linear model $F_{\boldsymbol{\eta}}$ yields the asymptotic mean squared error (AMSE) (Heritier et al., 2009) of the RDE estimator $\hat{\boldsymbol{\eta}}$

$$\mathrm{AMSE}(\hat{\boldsymbol{\eta}}, F_{\boldsymbol{\eta}}) = \frac{\mathrm{tr}\left(\mathrm{E}\left[\boldsymbol{U}(Y, \mu, \theta)\boldsymbol{U}(Y, \mu, \theta)^t\right]^{-1}\right)}{\mathrm{tr}\left(\Omega\right)}. \tag{6}$$

This AMSE measures the loss of efficiency of the RDE estimator with respect to the MLE at the model distribution $F_{\boldsymbol{\eta}}$. In practice, the AMSE can be estimated by replacing $F_{\boldsymbol{\eta}}$ by its empirical counterpart. The AMSE in (6) measures the relative efficiency for estimation of the complete parameter vector $\boldsymbol{\eta}$. When the focus is mainly on inference for the mean model regression parameters $\boldsymbol{\beta}$ (given the vector $\boldsymbol{\gamma}$), then its relative efficiency can be determined by replacing $\boldsymbol{U}$ and Ψ by $\boldsymbol{U}_{\boldsymbol{\beta}}$ and $\Psi_{\boldsymbol{\beta}}$, respectively, in (6). The AMSE can be used to tune the weight functions in the RDE estimator, i.e. to determine values of the tuning constants, such that a predetermined efficiency is obtained.

2.2 The RDE estimator for Poisson and binomial models

The double exponential version of two highly popular GLMs, the Poisson and binomial model, will now be discussed in more detail. Similar arguments hold for other double exponential GLMs.

We first consider the case where Y follows a double Poisson distribution, denoted as $Y \sim \mathrm{DEP}(\mu, \theta)$. For the corresponding double Poisson GLM, i.e. $Y|\boldsymbol{x}, \boldsymbol{z} \sim \mathrm{DEP}(\mu, \theta)$, the exponential function is chosen for both link functions h and g in (1). Note that the exponential function is the natural choice for h and it also fulfills the conditions on g. The score functions for the double Poisson GLM can now easily be derived and are given by

$$\boldsymbol{U}_\mu = \frac{y-\mu}{\mu/\theta} \quad \text{and} \quad \boldsymbol{U}_\theta = \frac{1}{2\theta} - \mu + y \ln\left(\frac{\mu \exp(1)}{y}\right) I(y > 0).$$

As discussed in the previous section, the weight functions $v_1(r)$ and $v_2(r)$ in (5) should be chosen carefully to downweight the effect of outliers. As expected, the choice $v_1(r) = \psi_H(r)/r$ suffices to obtain a bounded function $\nu_1(y, \mu, \theta)$ for the mean model. However, it can be seen that the same choice $v_2(r) = \psi_H(r)/r$ does not suffice to obtain a bounded function $\nu_2(y, \mu, \theta)$ for the dispersion model. Therefore, it is needed to include a faster decreasing function to make $\nu_2(y, \mu, \theta)$ bounded. We propose $v_2(r) = (\psi_H(r, c)/r)^2$ and this weight function will be used in the remainder of the manuscript.

Secondly, we focus on the double binomial distribution, denoted as $Y \sim \mathrm{DEB}(\mu, \theta)$. The corresponding one parameter exponential family is $\mathrm{Bin}(m, \mu)/m$ such that y is an element of $\{0,\ 1/m,\ 2/m,\ \ldots,\ 1\}$. Note that for a sample $y_i, i \in \{1, \ldots, n\}$, from the double binomial distribution, it is possible for m to depend on i as well. However, to simplify notation we drop this dependence in the remainder of the paper without loss of generality. For the double binomial GLM, i.e. $Y|\boldsymbol{x}, \boldsymbol{z} \sim \mathrm{DEB}(\mu, \theta)$ we take the natural logit function for the link function h, while we keep the exponential function for g. The score functions for this double binomial GLM then become

$$\boldsymbol{U}_\mu = \frac{(y-\mu)}{\mu(1-\mu)/(m\theta)} \quad \text{and} \quad \boldsymbol{U}_\theta = \frac{1}{2\theta} + my \ln\left(\frac{\mu}{y}\right) I(y \neq 0) + m(1-y) \ln\left(\frac{1-\mu}{1-y}\right) I(y \neq 1).$$

Similarly as in the Poisson case, we need a fast decreasing function such as $v_2(r) = (\psi_H(r, c)/r)^2$ to bound the function $\nu_2(y, \mu, \theta)$ for the dispersion model.

3 Robust Inference

The notion of deviance is a popular concept to perform inference and model selection in GLMs. To develop robust inference for double exponential GLMs, we introduce a robust generalized quasi-deviance to measure the quality of a fit. The generalized quasi-deviance is defined as

$$D_{QM}(\boldsymbol{y}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\theta}}, \boldsymbol{\mu}, \boldsymbol{\theta}) = -2 \sum_{i=1}^{n} Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \mu_i, \theta_i), \tag{7}$$

where $Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \mu_i, \theta_i)$ is given by

$$\begin{aligned} Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \mu_i, \theta_i) = &\int_{\tilde{s}_{1i}}^{\tilde{\mu}_i} \nu_1(y_i, s, \theta_i) w_1(\boldsymbol{x}_i, \boldsymbol{z}_i)\, ds - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{s}_{2j}}^{\tilde{\mu}_j} \mathrm{E}\left[\nu_1(Y_j, s, \theta_j)\right] w_1(\boldsymbol{x}_j, \boldsymbol{z}_j)\, ds \\ &+ \int_{\tilde{t}_{1i}}^{\tilde{\theta}_i} \nu_2(y_i, \mu_i, t) w_2(\boldsymbol{x}_i, \boldsymbol{z}_i)\, dt - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}_{2j}}^{\tilde{\theta}_j} \mathrm{E}\left[\nu_2(Y_j, \mu_j, t)\right] w_2(\boldsymbol{x}_j, \boldsymbol{z}_j)\, dt. \end{aligned} \tag{8}$$

Here, the values $\tilde{s}_{1i}$, $\tilde{s}_{2j}$, $\tilde{t}_{1i}$ and $\tilde{t}_{2j}$ are determined such that $\nu_1(y_i, \tilde{s}_{1i}, \theta_i) = 0$, $\mathrm{E}\left[\nu_1(Y_j, \tilde{s}_{2j}, \theta_j)\right] = 0$, $\nu_2(y_i, \mu_i, \tilde{t}_{1i}) = 0$ and $\mathrm{E}\left[\nu_2(Y_j, \mu_j, \tilde{t}_{2j})\right] = 0$, respectively. Hence, they are independent of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\theta}}$. The robust generalized quasi-deviance in (7) takes the quality of the fit in both the mean and dispersion model into account. Indeed, the first two terms in (8) measure the goodness of fit of the regression model for the mean, i.e. $\mu_i = h(\boldsymbol{x}_i^t \boldsymbol{\beta})$ while the last two terms measure the goodness of fit of the regression model for the dispersion, i.e. $\theta_i = g(\boldsymbol{z}_i^t \boldsymbol{\gamma})$. Note that in absence of dispersion, i.e. $\theta = 1$, the last two terms in (8) become zero and the robust generalized quasi-deviance in (7) reduces to the robust quasi-deviance proposed by Cantoni and Ronchetti (2001).

The generalized quasi-deviance provides a useful basis for robust inference and model selection. We focus on the comparison of two nested models $M_{p-q} \subset M_p$ with $p-q$ and p parameters, respectively. In particular, let us partition the vector $\boldsymbol{\eta}^t = (\boldsymbol{\eta}_1^t, \boldsymbol{\eta}_2^t)$ into $(p-q)$ components for $\boldsymbol{\eta}_1^t$ and q components for $\boldsymbol{\eta}_2^t$, then we consider testing the null hypothesis $H_0 : \boldsymbol{\eta}_2 = \boldsymbol{0}$ without loss of generality (after re-arranging the components of $\boldsymbol{\eta}$ if necessary). Let $\hat{\boldsymbol{\eta}}$ denote the RDE estimator of $\boldsymbol{\eta}$ in the full model, obtained by solving (3). Similarly,

$\hat{\boldsymbol{\eta}}_1^{(0)}$ is the RDE estimator of $\boldsymbol{\eta}_1$ in the reduced model under the null hypothesis. Let $\hat{\mu}_i$, $\hat{\theta}_i$ and $\mu_i^{(0)}$, $\theta_i^{(0)}$ denote the corresponding quantities in the full and reduced model, respectively. Based on the robust generalized quasi-deviance in (7), a natural measure for the discrepancy between the two nested models is given by

$$\begin{aligned}\Lambda_{QM} =& D_{QM}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) - D_{QM}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) \\ =& 2\sum_{i=1}^{n} \Big[Q_M(y_i, \hat{\mu}_i, \hat{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) - Q_M(y_i, \hat{\mu}_i^{(0)}, \hat{\theta}_i^{(0)}, \hat{\mu}_i, \hat{\theta}_i) \Big].\end{aligned} \tag{9}$$

Note that Λ_{QM} is independent of $\tilde{s}_{1i}$, $\tilde{s}_{2j}$, $\tilde{t}_{1i}$ and $\tilde{t}_{2j}$. The asymptotic distribution of the test statistic Λ_{QM} is given by the following proposition which is proven in Appendix 8.1.

Theorem 1. *Assume conditions (A1)-(A9) (see the supplementary material) for distribution* $F_{\boldsymbol{\eta}}$ *under* $H_0 : \boldsymbol{\eta}_2 = \mathbf{0}$ *and that* $M\left(\Psi_{\boldsymbol{\beta}}, F_{\boldsymbol{\eta}}\right)$ *and* $M\left(\Psi_{\boldsymbol{\gamma}}, F_{\boldsymbol{\eta}}\right)$ *are symmetric positive definite. Let* q_1 *and* q_2 *respectively denote the number of components of* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$ *assumed to be zero under the null hypothesis.*

1. *Under* H_0, Λ_{QM} *is asymptotically distributed as* $\sum_{i=1}^{q_1} \lambda_i^{\boldsymbol{\beta}} N_i^2 + \sum_{j=1}^{q_2} \lambda_j^{\boldsymbol{\gamma}} N_j^2$, *where the* N_i *and* N_j *are independent standard normal variables. The values* $\lambda_1^{\boldsymbol{\beta}} \geqslant \lambda_2^{\boldsymbol{\beta}} \geqslant \ldots \geqslant \lambda_{q_1}^{\boldsymbol{\beta}} > 0$ *correspond to the* q_1 *positive eigenvalues of the matrix* $Q\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}}\right) \cdot \left(M^{-1}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}}\right) - \widetilde{M}^{*^+}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}}\right)\right)$ *where* $\widetilde{M}^{*^+}$ *is such that* $\widetilde{M}^{*^+}_{11} = M^{-1}_{(11)}$, $\widetilde{M}^{*^+}_{12} = \widetilde{M}^{*^+}_{21} = \widetilde{M}^{*^+}_{22} = 0$, *using the notation* $M^{1}_{(11)}$ *to denote the* M*-matrix of* $\Psi_{\boldsymbol{\beta}}$ *restricted to the first* $p_1 - q_1$ *components. Similarly the values* $\lambda_1^{\boldsymbol{\gamma}} \geqslant \lambda_2^{\boldsymbol{\gamma}} \geqslant \ldots \geqslant \lambda_{q_2}^{\boldsymbol{\gamma}} > 0$ *correspond to the* q_2 *positive eigenvalues of the matrix* $Q\left(\Psi^{\boldsymbol{\gamma}}, F_{\boldsymbol{\gamma}}\right)\left(M^{-1}\left(\Psi^{\boldsymbol{\gamma}}, F_{\boldsymbol{\gamma}}\right) - \widetilde{M}^{*^+}\left(\Psi^{\boldsymbol{\gamma}}, F_{\boldsymbol{\gamma}}\right)\right)$.

2. *Consider the sequence of contiguous alternatives* $H_{1,n} : \boldsymbol{\eta}_2 = n^{-1/2}\Delta, \boldsymbol{\eta}_1 = \boldsymbol{\eta}_1$ *with* $\Delta = (\Delta_{\boldsymbol{\beta}}^t, \Delta_{\boldsymbol{\gamma}}^t)^t$ *any vector in* $\mathbb{R}^q$ *such that* $\left(\boldsymbol{\eta}_1^t, n^{-1/2}\Delta^t\right)^t$ *still belongs to* $\mathcal{O}$ *(see the conditions in the supplementary material for the definition of* $\mathcal{O}$*). Then, the statistic* Λ_{QM} *has asymptotic distribution*

$$\begin{aligned}&\sum_{i=1}^{q_1}\left(\sqrt{\lambda_i^{\boldsymbol{\beta}}} N_i + \left(P_{\boldsymbol{\beta}}^t \Delta_{\boldsymbol{\beta}}\right)_i\right)^2 + \sum_{j=1}^{q_2}\left(\sqrt{\lambda_j^{\boldsymbol{\gamma}}} N_j + \left(P_{\boldsymbol{\gamma}}^t \Delta_{\boldsymbol{\gamma}}\right)_j\right)^2 \\ =& \sum_{i=1}^{q_1} \lambda_i^{\boldsymbol{\beta}} \chi_1^2\left(\frac{\left(P_{\boldsymbol{\beta}}^t \Delta_{\boldsymbol{\beta}}\right)_i}{\sqrt{\lambda_i^{\boldsymbol{\beta}}}}\right) + \sum_{j=1}^{q_2} \lambda_j^{\boldsymbol{\gamma}} \chi_1^2\left(\frac{\left(P_{\boldsymbol{\gamma}}^t \Delta_{\boldsymbol{\gamma}}\right)_j}{\sqrt{\lambda_j^{\boldsymbol{\gamma}}}}\right),\end{aligned}$$

with $\chi_1^2(\cdot)$ a non-central χ^2-distribution. Here, $P_{\boldsymbol{\beta}}$ is a Choleski root of $M_{22.1}^{\boldsymbol{\beta}}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) = M_{(22)}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) - M_{(12)}^t\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) M_{(11)}^{-1}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) M_{(12)}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right)$ and $P_{\boldsymbol{\beta}}^t\left(M^{-1}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) Q\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right) M^{-1}\left(\Psi^{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}_0}\right)\right)_{(22)} P_{\boldsymbol{\beta}} = diag\left(\lambda_1^{\boldsymbol{\beta}}, \dots, \lambda_{q_1}^{\boldsymbol{\beta}}\right)$, and similarly for $P_{\boldsymbol{\gamma}}$.

Note that the eigenvalues $\lambda_1^{\boldsymbol{\beta}}, \dots, \lambda_{q_1}^{\boldsymbol{\beta}}$ and $\lambda_1^{\gamma}, \dots, \lambda_{q_2}^{\gamma}$ can be calculated by using the expressions for $M(\Psi, F)$ and $Q(\Psi, F)$ in the supplementary material. Part 1 of this proposition can then be used to obtain p-values for the test based on Λ_{QM}, see Davies (1980, 1990).

To investigate the robustness of the test, we study the influence of a small amount of contamination at a particular point on the asymptotic level of the test. We thus focus on the local stability of the test which is often the main concern at the inference stage. Following Heritier and Ronchetti (1994), define the sequence of ε-contaminations $F_{\varepsilon,n} = \left(1 - \frac{\varepsilon}{\sqrt{n}}\right) F_{\eta_0} + \frac{\varepsilon}{\sqrt{n}} G$, where G is an arbitrary distribution. The impact of such ε-contamination on likelihood ratio tests based on M-estimators was studied by Cantoni and Ronchetti (2001), generalizing the work of Heritier and Ronchetti (1994). They showed that a bounded influence function of the M-estimator of $\hat{\boldsymbol{\eta}}_{(2)}$ translates to a bound on the asymptotic level of the proposed test. Corollary 1, proven in the supplementary material, shows that this general result is also applicable to the RDE estimator. A similar result can be obtained for the power of the proposed likelihood ratio test using similar techniques.

Corollary 1. *Assume conditions (A1)-(A9) (see the supplementary material), then for any M-estimator $\hat{\boldsymbol{\eta}}_{(2)}$ with bounded influence function, the asymptotic level of the robust likelihood ratio test statistic Λ_{QM} in* (9) *under point mass contamination is given by*

$$\begin{aligned}\lim_{n\to\infty} \alpha(F_{\varepsilon,n}) = \alpha_0 &+ \varepsilon^2 \kappa_{\boldsymbol{\beta}}^t diag\left(P_{\boldsymbol{\beta}}\, IF\left(y; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}\right) IF\left(y; \hat{\boldsymbol{\beta}}_{(2)}, F_{\boldsymbol{\beta}_0}\right)^t P_{\boldsymbol{\beta}}^t\right)\\ &+ \varepsilon^2 \kappa_{\boldsymbol{\gamma}}^t diag\left(P_{\boldsymbol{\gamma}}\, IF\left(y; \hat{\boldsymbol{\gamma}}_{(2)}, F_{\boldsymbol{\gamma}_0}\right) IF\left(y; \hat{\boldsymbol{\gamma}}_{(2)}, F_{\boldsymbol{\gamma}_0}\right)^t P_{\boldsymbol{\gamma}}^t\right) + o(\varepsilon^2),\end{aligned} \tag{10}$$

where $P_{\boldsymbol{\beta}}$ is an orthogonal matrix such that $P_{\boldsymbol{\beta}}^t D_{\boldsymbol{\beta}} P_{\boldsymbol{\beta}} = \Omega_{22}^{\boldsymbol{\beta}} M_{22.1}^{\boldsymbol{\beta}}$, $\Omega^{\boldsymbol{\beta}}$ is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ and $D_{\boldsymbol{\beta}}$ is the diagonal matrix with elements $\lambda_1^{\boldsymbol{\beta}}, \dots, \lambda_{q_1}^{\boldsymbol{\beta}}$ and similarly for $P_{\boldsymbol{\gamma}}$.

Corollary 1 shows that when a bounded influence estimator $\hat{\boldsymbol{\eta}}_{(2)}$ is used, then also the effect of contamination on the asymptotic level (and power) of the robust generalized quasi-deviance test remains bounded. When model comparison or model selection is the main focus of an analysis, then the tuning constants in the RDE estimator can be chosen to control the maximal bias on the asymptotic level of the test in a neighborhood of the model according to (10), as explained in Ronchetti and Trojani (2001) and Cantoni and Ronchetti (2001).

4 Finite-sample Performance

Extensive simulation results showing the good estimation performance of the RDE estimator are provided in the supplementary material. Here, we investigate the performance of robust inference based on Λ_{QM}. To investigate the level of the test, we consider a model without dispersion. To this end we generate $N = 1000$ samples with responses $Y \mid (X, Z) \sim \text{DEP}(\mu, \theta)$ where $\mu = \exp(3 + 0X)$ and $\theta = \exp(0)$, where X is uniformly distributed on the interval $[-0.5, 0.5]$. We consider the hypothesis test $H_0 : (\beta_1, \gamma_0) = (0, 0)$ vs $H_1 : (\beta_1, \gamma_0) \neq (0, 0)$.

To investigate the influence of the sample size on the level of this test we compare the empirical rejection rate to the corresponding nominal level for samples of size $n \in \{50, 100, 250, 500\}$. From the results in the left part of Table 1 it can be seen that for small samples ($n = 50$) the empirical rejection rates are already close to their nominal values. When the sample size increases, the empirical rejection rates approximate the asymptotic level even better and their is little difference between tests based on RDE using a Huber (HRDE) or Tukey bisquare (TRDE) weight function tuned for 90% efficiency. More details about the estimators are given in the supplementary material.

To investigate the robustness of the level of the test, we fix the sample size at $n = 50$ and vary the contamination level. Vertical outliers are generated by multiplying the response with a factor 10 with the contamination fraction ε ranging from 0% to 25% in steps of 5%. From the results in the right part of Table 1 we can see that small to modest contamination levels ($\epsilon \leq 10\%$) have little impact on the level of the test based on the HRDE estimator, but larger fractions of contamination affect the level more heavily. On the other hand, all levels of contamination have little effect on the level of the test based on the TRDE estimator.

	sign.- level	n				ε					
		50	100	250	500	0%	5%	10%	15%	20%	25%
HRDE	10%	12.20	12.20	11.90	10.00	12.20	11.90	12.60	25.50	42.60	70.60
	5%	7.60	6.30	6.20	4.70	7.60	5.50	7.30	15.00	27.20	58.60
	2.5%	4.30	4.00	3.30	2.80	4.30	3.10	3.60	9.10	18.00	47.10
	1%	2.90	1.90	1.70	1.00	2.90	1.10	1.20	4.10	9.80	32.70
TRDE	10%	12.30	12.10	11.30	9.90	12.30	12.40	10.10	11.60	12.10	9.70
	5%	7.70	6.60	6.50	4.90	7.70	8.50	7.00	6.80	6.30	5.20
	2.5%	4.60	4.60	3.60	2.70	4.60	4.40	4.30	3.70	3.50	2.90
	1%	2.70	2.10	1.70	1.20	2.70	2.00	1.80	1.50	1.90	1.90

Table 1: Empirical rejection rates for different significance levels of the test (sign.-levels) for a Poisson model without dispersion. On the left, results for uncontaminated samples of different sizes. On the right, results for contaminated samples of size $n = 50$ for several contamination levels.

To investigate the power of the test, we now consider a model with constant dispersion. To this end we generate $N = 1000$ samples of size $n \in \{50, 100, 250, 500\}$ with responses $Y \mid (X, Z) \sim \text{DEP}(\mu, \theta)$ where $\mu = \exp(2 - X)$, where X is uniformly distributed on the interval $[-0.5, 0.5]$, and with constant dispersion θ which varies in the range [0.25, 3.5]. Hence, we consider both underdispersion and overdispersion. We test for presence of dispersion, i.e. $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$. For the setting without dispersion (i.e. $\theta = 1$) the results for the level of the test are similar as above and can be found in the supplementary material. The power curves in Figure 1 show that the power increases to 1 when the dispersion θ moves away from the null hypothesis. Clearly, the power increases faster when the sample size grows, as expected.

To investigate the robustness of the power, we again consider samples of size $n = 50$ with a varying fraction of vertical outliers, generated as before. The resulting power curves in Figure 2 clearly show that similarly as for the level, also the power of the test based on the HRDE estimator is affected more when the contamination level increases. On the other hand, the test based on the TRDE estimator again shows good behavior for all contamination levels. Overall we can conclude that robust inference based on the TRDE estimator yields reliable results in terms of both level and power for all contamination levels considered.

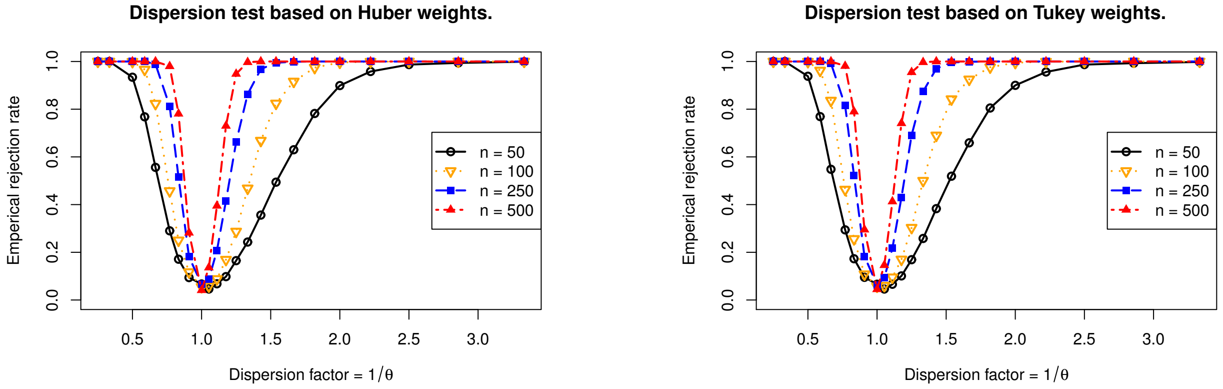

Figure 1: Power of the dispersion test for uncontaminated data with various sample sizes from a Poisson model with constant dispersion.

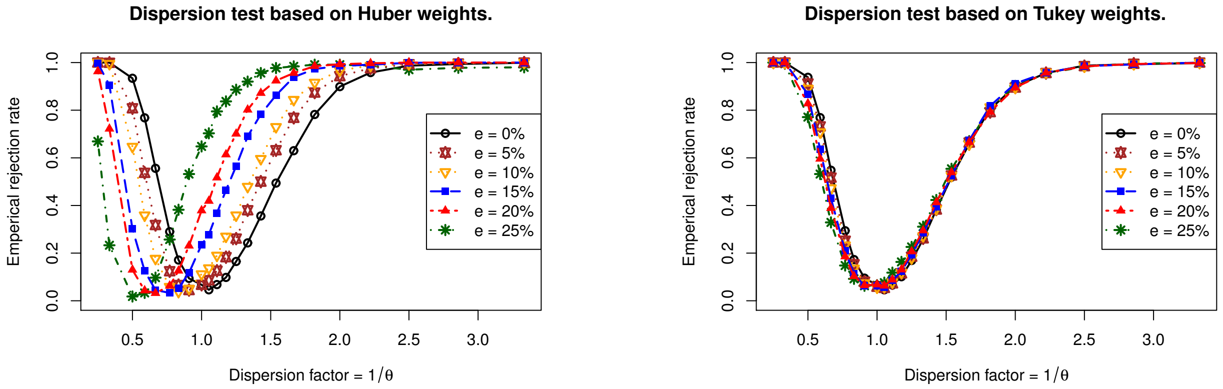

Figure 2: Power of the dispersion test for contaminated data from a Poisson model with constant dispersion.

5 Data Examples

In this section we illustrate our methodology on two real data examples. In the first example a double binomial model is used while a double Poisson model is used in the second example.

An additional example for the double binomial model can be found in the supplementary material. For these examples we apply the RDE estimators with 90% efficiency as before.

5.1 Double binomial model: UCB admissions data

We consider data of student admissions into UC Berkeley's graduate school of the year 1973 for the six largest departments (Bickel et al., 1975). These data, which are shown in Table 2, have been discussed by various authors as an illustration of Simpson's paradox. We consider a (double) binomial GLM with admissions rate as the response. Gender and department are used as covariates for the mean model and we consider a constant dispersion model. The model is thus given by

$$\operatorname{logit}(\mu) = \beta_0+\beta_1\text{Dept}_B+\beta_2\text{Dept}_C+\beta_3\text{Dept}_D+\beta_4\text{Dept}_E+\beta_5\text{Dept}_F+\beta_6\text{Female}, \quad \log(\theta) = \gamma_0$$

Table 2: UC Berkeley admissions proportions into graduate school for the year 1973.

Gender	Dept A	Dept B	Dept C	Dept D	Dept E	Dept F
Male	512/825	353/560	120/325	138/417	53/191	22/373
Female	89/108	17/25	202/593	131/375	94/393	24/341

The estimates and their corresponding p-values obtained by the CDE, CEQL, HRDE and TRDE estimator are shown in Table 3. From the results it can be seen that the gender effect is not significant. Moreover, the admission rates of department B do not differ significantly from those of department A (the baseline). We can also test the joint null hypothesis $H_0 : \beta_1 = \beta_6 = 0$. For the test based on the TRDE estimator, the resulting p-value is 0.51, which confirms that there is no evidence for these effects.

The main difference between the two robust estimators and the two non-robust estimators is obtained for the dispersion. Both the CDE and CEQL estimates indicate presence of overdispersion. On the other hand, both RDE estimates indicate underdispersion. For example, the TRDE estimate for θ is $\exp(1.29) = 3.63$ corresponding to underdispersion.

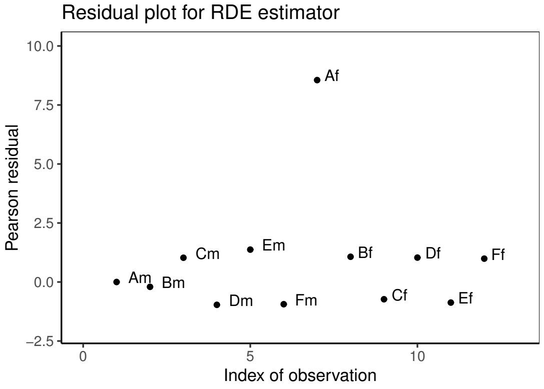

Figure 3: Residuals for the RDE fit on the UCB admissions data. The department has been indicated with a capital letter and the gender with a small letter.

The difference with the classical results can be explained by examining the residuals of the TRDE estimator in Figure 3. Observation 7, corresponding to the female admissions rate for department A, has a large positive residual. From the data in Table 2 it can indeed be seen that for department A the admissions percentage for female students is substantially higher than for male students. This deviating observation clearly influences the nonrobust estimates of dispersion while it has little effect on the RDE estimators. Indeed, when the data are refit without observation 7, also the nonrobust estimators indicate underdispersion which confirms the robustness of the results obtained by our methodology.

Estimator	β_0	β_1	β_2	β_3	β_4	β_5	β_6	γ_0
CDE	0.58 (0.00)	-0.04 (0.76)	-1.26 (0.00)	-1.29 (0.00)	-1.74 (0.00)	-3.31 (0.00)	0.10 (0.34)	-0.51 (0.20)
CEQL	0.58 (0.00)	-0.04 (0.86)	-1.26 (0.00)	-1.29 (0.00)	-1.74 (0.00)	-3.31 (0.00)	0.10 (0.58)	-1.59 (0.06)
HRDE	0.50 (0.00)	0.04 (0.56)	-1.10 (0.00)	-1.16 (0.00)	-1.58 (0.00)	-3.17 (0.00)	-0.02 (0.73)	0.86 (0.05)
TRDE	0.49 (0.00)	0.05 (0.39)	-1.09 (0.00)	-1.14 (0.00)	-1.57 (0.00)	-3.15 (0.00)	-0.03 (0.50)	1.29 (0.00)

Table 3: Parameter estimates for the UCB admissions data. P-values are shown between brackets.

Estimator	β_0	β_1	β_2	β_3	β_4	β_5	β_6	γ_0	γ_1
CDE	2.24 (0.00)	-1.18 (0.05)	-0.26 (0.60)	-0.24 (0.12)	-0.19 (0.24)	0.02 (0.00)	-0.41 (0.21)	-2.91 (0.00)	0.79 (0.17)
CEQL	2.31 (0.00)	-0.88 (0.02)	-0.32 (0.39)	-0.19 (0.07)	-0.14 (0.20)	0.02 (0.00)	-0.46 (0.03)	-2.39 (0.00)	0.38 (0.27)
HRDE	1.33 (0.00)	-0.29 (0.38)	-0.27 (0.53)	0.02 (0.85)	-0.17 (0.14)	0.05 (0.00)	-0.16 (0.44)	-1.36 (0.00)	0.51 (0.23)
TRDE	1.28 (0.00)	-0.30 (0.35)	-0.25 (0.55)	0.05 (0.65)	-0.15 (0.17)	0.05 (0.00)	-0.10 (0.62)	-1.21 (0.00)	0.36 (0.38)

Table 4: Parameter estimates for epilepsy data. P-values are shown between brackets.

5.2 Double Poisson model: epilepsy data

As an illustration for the Poisson model, we consider data from a double blind drug study comparing a new anti-epileptic drug called topiramate with a placebo (Faught et al., 1996). Patients suffering from epilepsy were randomized over the two groups. During the course of sixteen weeks, the number of seizures per week was recorded for each patient. We consider the total number of seizures during weeks nine through twelve. Patients dropping out of the study before this time were discarded. The resulting data set contains 40 patients that received a placebo and 39 patients that received the drug.

Next to the treatment (topiramate/placebo), 5 other predictor variables are available for each patient: sex, race, weight, height and its baseline seizure rate (base). This baseline consists of a 12-week period during which the number of seizures was measured before the start of treatment-placebo study. We have robustly standardized both weight and height. Since it is a priori unclear whether the drug could also have an impact on dispersion, we have included treatment in the dispersion model. This leads to the following models for the mean and dispersion:

$$\begin{aligned}\log(\mu_i) &= \beta_0 + \beta_1 \text{sex} + \beta_2 \text{race} + \beta_3 \text{height} + \beta_4 \text{weight} + \beta_5 \text{base} + \beta_6 \text{trt},\\ \log(\theta_i) &= \gamma_0 + \gamma_1 \text{trt}.\end{aligned}$$

Since there may be leverage points in the space of the three continuous predictors weight, height and base, we determine the weights $w_j(\boldsymbol{x}_i, \boldsymbol{z}_i)$ by applying the hard cutoff rule discussed in Section 2, where we use the MCD with tuning parameter α equal to 0.75 to obtain the robust location and scatter estimates.

The parameter estimates and corresponding p-values obtained by the CDE, CEQL and

RDE estimators are shown in Table 4. Note that the variable sex is found to be significant by the classical estimators, while it is not according to the robust estimators. We may use the robust inference to test the joint null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_6 = \gamma_1 = 0$. The test based on TRDE yields a p-value equal to 0.54, suggesting that these predictors may be excluded from the final model and thus there is no treatment effect, in particular. Note that according to the RDE estimate the constant dispersion parameter γ_0 is clearly significant.

The scaled Pearson residuals for the TRDE fit are shown in Figure 4. Clearly, there are two outliers with a large negative residual, which are the patients coded as 601731 and 601909. Given their huge deviation from the robust fit, it can be expected that the impact of these observations on the classical estimator is severe. These patients have an exceptionally high baseline seizure rate of respectively 198.3 and 117.0, whereas the remaining observations have a baseline seizure rate between 4 and 64. As both observations were recognized as a leverage point, they were down-weighted in the robust analyses. Furthermore, there are three observations in the control group with a moderately large positive residual. Since these residuals are positive, the observed number of seizures is larger than expected under the model. This illustrates that a robust analysis may provide useful extra information. For instance, based on this result it may be decided that an intervention is needed for these cases, e.g. the need of medication due to health concerns.

6 Penalized RDE Estimators

In this section, we consider two extensions of the RDE estimator. First, we consider high-dimensional regression where the number of predictors is large and may even exceed the sample size. To obtain stable estimates in this setting, we add a sparsity penalty to the RDE estimator. In the second extension we allow for more flexible models by replacing the GLM for the mean and/or dispersion in (1) by a generalized additive model (GAM).

Both extensions rely on the weighted least squares representation of the RDE estimator in (3). Let X and Z denote the design matrices for the mean and dispersion model,

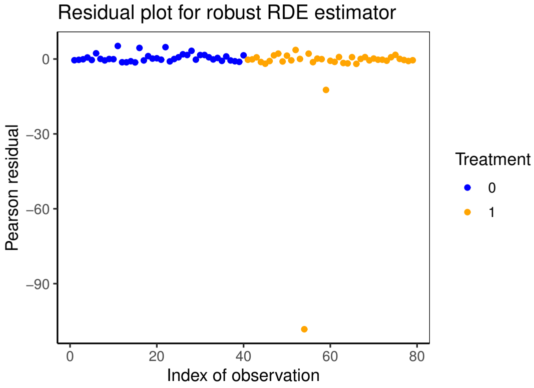

Figure 4: Scaled Pearson residuals for the RDE fit of the double Poisson model for the epilepsy data.

respectively. As shown in Appendix 8.2, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ jointly solve the weighted least squares problems

$$\min_{\boldsymbol{\beta}} (\tilde{y}_{\boldsymbol{\beta}} - X\boldsymbol{\beta})^t W_{\boldsymbol{\beta}} (\tilde{y}_{\boldsymbol{\beta}} - X\boldsymbol{\beta}) \tag{11}$$

$$\min_{\boldsymbol{\gamma}} (\tilde{y}_{\boldsymbol{\gamma}} - Z\boldsymbol{\gamma})^t W_{\boldsymbol{\gamma}} (\tilde{y}_{\boldsymbol{\gamma}} - Z\boldsymbol{\gamma}), \tag{12}$$

where the elements of the weight matrices $W_{\boldsymbol{\beta}} = \mathrm{diag}(w_{\boldsymbol{\beta},1}, \ldots, w_{\boldsymbol{\beta},n})$, $W_{\boldsymbol{\gamma}} = \mathrm{diag}(w_{\boldsymbol{\gamma},1}, \ldots, w_{\boldsymbol{\gamma},n})$ and pseudo response vectors $\tilde{y}_{\boldsymbol{\beta}} = (\tilde{y}_{\boldsymbol{\beta},1}, \ldots, \tilde{y}_{\boldsymbol{\beta},n})^t$, $\tilde{y}_{\boldsymbol{\gamma}} = (\tilde{y}_{\boldsymbol{\gamma},1}, \ldots, \tilde{y}_{\boldsymbol{\gamma},n})^t$ are given by

$$\begin{aligned} w_{\boldsymbol{\beta},i} &= \mathrm{E}[\nu_1(Y_i, \mu_i, \theta_i)\,\boldsymbol{U}_{\mu_i}] w_1(\boldsymbol{x}_i, \boldsymbol{z}_i) h'(\boldsymbol{x}_i^t\boldsymbol{\beta})^2, \\ \tilde{y}_{\boldsymbol{\beta},i} &= x_i^t\boldsymbol{\beta} + \frac{\nu_1(y_i, \mu_i, \theta_i) - E[\nu_1(y, \mu_i, \theta_i)]}{E[\nu_1(y, \mu_i, \theta_i) U_{\mu_i}] h'(\boldsymbol{x}_i^t\boldsymbol{\beta})}, \end{aligned} \tag{13}$$

$$\begin{aligned} w_{\boldsymbol{\gamma},i} &= \mathrm{E}[\nu_2(Y_i, \mu_i, \theta_i)\,\boldsymbol{U}_{\theta_i}] w_2(\boldsymbol{x}_i, \boldsymbol{z}_i) g'(\boldsymbol{z}_i^t\boldsymbol{\gamma})^2, \\ \tilde{y}_{\boldsymbol{\gamma},i} &= z_i^t\boldsymbol{\gamma} + \frac{\nu_2(y_i, \mu_i, \theta_i) - E[\nu_2(y, \mu_i, \theta_i)]}{E[\nu_2(y, \mu_i, \theta_i) U_{\theta_i}] h'(\boldsymbol{z}_i^t\boldsymbol{\gamma})}. \end{aligned} \tag{14}$$

This weighted least squares formulation of the RDE estimator allows to calculate the estimator via iteratively reweighted least squares, alternating between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. More importantly, this representation makes it possible to introduce penalized versions of the RDE estimator by iteratively calculating penalized least squares solutions as illustrated below.

6.1 Sparse RDE estimator

A popular way to select the most important predictors and estimate the parameters in high-dimensional regression problems is by adding a sparsity inducing penalty to the objective function of an estimator (Hastie et al., 2015). Avella-Medina and Ronchetti (2018) proposed a penalized version of the robust quasi-likelihood estimator of Cantoni and Ronchetti (2001). A similar approach can be used in the double exponential framework by adding a sparsity penalty in (11)-(12). For example, in case of a lasso penalty (Tibshirani, 1996), the sparse RDE estimator jointly solves

$$\min_{\boldsymbol{\beta}} \{(\tilde{y}_{\boldsymbol{\beta}} - X\boldsymbol{\beta})^t W_{\boldsymbol{\beta}}(\tilde{y}_{\boldsymbol{\beta}} - X\boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1\} \tag{15}$$

$$\min_{\boldsymbol{\gamma}} \{(\tilde{y}_{\boldsymbol{\gamma}} - Z\boldsymbol{\gamma})^t W_{\boldsymbol{\gamma}}(\tilde{y}_{\boldsymbol{\gamma}} - Z\boldsymbol{\gamma}) + \lambda_{\boldsymbol{\gamma}}\|\boldsymbol{\gamma}\|_1\}, \tag{16}$$

where $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$ are the usual sparsity tuning parameters. For given values of these tuning parameters, the sparse estimator can be calculated iteratively by alternately solving the penalized weighted least squares problems (15)-(16). These penalized weighted least squares problems can be solved efficiently via coordinate descent (Fu, 1998; Friedman et al., 2010), resulting in the lasso estimates.

Since lasso estimators tend to select more predictors than necessary (Meinshausen and Bühlmann, 2006), we also consider adaptive lasso estimators (Zou, 2006) based on the initial lasso estimates, as in Avella-Medina and Ronchetti (2018). Hence, in the second step the l_1 norm $\|\boldsymbol{\beta}\|_1$ in (15) is replaced by $\sum_{j=1}^{p_1} \tilde{w}(\tilde{\beta}_j)|\beta_j|$ with $\tilde{\boldsymbol{\beta}}$ the initial lasso estimate of $\boldsymbol{\beta}$ and where the weight function is given by

$$\tilde{w}(t) = \begin{cases} 1/|t|, & |t|> 0 \\ \infty, & |t|= 0 \end{cases} .$$

Analogously, $\|\boldsymbol{\gamma}\|_1$ in (16) is replaced by $\sum_{j=1}^{p_2} \tilde{w}(\tilde{\gamma}_j)|\gamma_j|$ with $\tilde{\boldsymbol{\gamma}}$ the initial lasso estimate of $\boldsymbol{\gamma}$. In the supplementary material, consistency of robust lasso RDE estimators and oracle properties of the corresponding adaptive lasso RDE estimators are studied by exploiting the results of Avella-Medina and Ronchetti (2018).

Selection of the tuning parameters $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$ is an important aspect of sparse estimation. A common approach is to select the optimal values from a grid according to a selection criterion. As selection criterion a computationally robust cross-validation (see Khan et al., 2010) or a robust extended Bayesian information criterion (EBIC) (see e.g. Avella-Medina and Ronchetti, 2018; Wang and Van Aelst, 2019) can be used. In particular, for any fixed value $\lambda_{\boldsymbol{\beta}}$ we determine the corresponding value of $\lambda_{\boldsymbol{\gamma}}$ that minimizes a robust extended BIC criterion of the form

$$\text{EBIC}_{\boldsymbol{\gamma}}(\lambda) = (\tilde{y}_{\boldsymbol{\gamma}_\lambda} - Z\boldsymbol{\gamma}_\lambda)^t W_{\boldsymbol{\gamma}_\lambda}(\tilde{y}_{\boldsymbol{\gamma}_\lambda} - Z\boldsymbol{\gamma}_\lambda) + (\log n + \tau \log p_2)\frac{|\boldsymbol{\gamma}_\lambda|}{n},$$

where $|\boldsymbol{\gamma}|$ is the number of nonzero coefficients in the vector $\boldsymbol{\gamma}$ and $0 \leq \tau \leq 1$ is a constant which we set equal to 0.5 by default. Then, we consider a set of $\lambda_{\boldsymbol{\beta}}$ values with corresponding optimal $\lambda_{\boldsymbol{\gamma}}$ values and select the solution that minimizes

$$\text{EBIC}_{\boldsymbol{\beta}}(\lambda) = (\tilde{y}_{\boldsymbol{\beta}_\lambda} - X\boldsymbol{\beta}_\lambda)^t W_{\boldsymbol{\beta}_\lambda}(\tilde{y}_{\boldsymbol{\beta}_\lambda} - X\boldsymbol{\beta}_\lambda) + (\log n + \tau \log p_1)\frac{|\boldsymbol{\beta}_\lambda|}{n}.$$

As usual, we consider a decreasing grid of $\lambda_{\boldsymbol{\beta}}$ values and use the solution corresponding to the larger $\lambda_{\boldsymbol{\beta}}$ value as initial values for the next $\lambda_{\boldsymbol{\beta}}$ value. Such warm starts speed up the computations considerably.

The supplementary material contains the results of two small simulation studies which show the good performance of this sparse adaptive lasso RDE estimator in sparse settings. To illustrate the sparse RDE estimator we apply it on the epilepsy data using EBIC to select the final model. In Section 5.2 we applied a robust test to conclude that baseline seizure rate was the only significant predictor for the mean model while the dispersion is constant. Fitting the adaptive lasso HRDE and TRDE estimators to these data yields the coefficient estimates in Table 5. It can be seen that both estimators yield very similar results and are able to select the most important predictors. Moreover, the adaptive lasso RDE estimates in Table 5 resemble the corresponding RDE estimates in Table 4 very well.

Estimator	β_0	β_1	β_2	β_3	β_4	β_5	β_6	γ_0	γ_1
HRDE	1.18	0.00	0.00	0.00	0.00	0.05	0.00	-1.08	0.00
TRDE	1.16	0.00	0.00	0.00	0.00	0.05	0.00	-1.06	0.00

Table 5: The sparse RDE parameter estimates for the epilepsy data.

6.2 RDE estimator for generalized additive models

Sometimes the parametric assumption for the relation between the mean and/or dispersion and their predictor variables in the GLM setting given by (1) is too stringent. A popular way to make GLMs more flexible is by considering generalized additive models (GAMs) instead. Robust estimation methods for GAMs have been proposed by Alimadad and Salibian-Barrera (2011), Croux et al. (2012) and Wong et al. (2014). When both the mean and dispersion are modeled via GAMs, the parametric models in (1) are replaced by the more flexible relations

$$\mu_i = h\left(\sum_{j=1}^{p_1} f_{1j}(x_{ij})\right) \quad \text{and} \quad \theta_i = g\left(\sum_{j=1}^{p_2} f_{2j}(z_{ij})\right),$$

with $f_{11}, \ldots, f_{1p_1}$ and $f_{21}, \ldots, f_{2p_2}$ unknown smooth functions of the predictor variables $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, respectively. These smooth functions can be estimated via penalized basis expansion fitting. Similarly to Wong et al. (2014) we focus on the case $p_1 = p_2 = 1$ to ease notation and denote $f_{11} = f_1$, $f_{21} = f_2$, $x_{1i} = x_i$ and $z_{1i} = z_i$ for all i. The extension to $p_1 > 1$ and $p_2 > 1$ is straightforward. Consider two sets of prespecified basis functions $d_{11}(\cdot), \ldots, d_{1q_1}(\cdot)$ and $d_{21}(\cdot), \ldots, d_{2q_2}(\cdot)$, and assume that the smooth functions f_1 and f_2 can be represented as

$$f_1(x; \boldsymbol{\beta}) = \sum_{j=1}^{q_1} d_{1j}(x)\beta_j \quad \text{and} \quad f_2(z; \boldsymbol{\gamma}) = \sum_{j=1}^{q_2} d_{2j}(z)\gamma_j,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{q_1})^t$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q_2})^t$ are vectors of basis coefficients. To avoid overfitting the data, regularization is used to estimate the basis coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Let $S_{\boldsymbol{\beta}}$ and $S_{\boldsymbol{\gamma}}$ be prespecified penalty matrices and $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$ two strictly positive smoothing parameters, then $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be estimated by jointly solving

$$\min_{\boldsymbol{\beta}} \{(\tilde{y}_{\boldsymbol{\beta}} - D_1\boldsymbol{\beta})^t W_{\boldsymbol{\beta}}(\tilde{y}_{\boldsymbol{\beta}} - D_1\boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}}\boldsymbol{\beta}^T S_{\boldsymbol{\beta}}\boldsymbol{\beta}\} \tag{17}$$

$$\min_{\boldsymbol{\gamma}} \{(\tilde{y}_{\boldsymbol{\gamma}} - D_2\boldsymbol{\gamma})^t W_{\boldsymbol{\gamma}}(\tilde{y}_{\boldsymbol{\gamma}} - D_2\boldsymbol{\gamma}) + \lambda_{\boldsymbol{\gamma}}\boldsymbol{\gamma}^T S_{\boldsymbol{\gamma}}\boldsymbol{\gamma}\}, \tag{18}$$

where D_1 and D_2 are the design matrices corresponding to the two sets of basis functions. The GAM RDE estimates can again be calculated iteratively by alternately solving problems (17)-(18), which both are weighted additive model fits for which efficient algorithms are available. The tuning parameters $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$ can be selected via the same procedure as in the sparse setting based on either robust cross-validation or robust BIC (see e.g. Wong et al., 2014). In this setting, the robust BIC criteria to select $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\boldsymbol{\gamma}}$ are given by

$$\begin{aligned}
\mathrm{RBIC}_{\boldsymbol{\beta}}(\lambda) &= (\tilde{y}_{\boldsymbol{\beta}_\lambda} - D_1\boldsymbol{\beta}_\lambda)^t W_{\boldsymbol{\beta}_\lambda}(\tilde{y}_{\boldsymbol{\beta}_\lambda} - D_1\boldsymbol{\beta}_\lambda) + \log n \frac{\mathrm{tr}(P_{\boldsymbol{\beta}}^{-1}Q_{\boldsymbol{\beta}})}{n} \\
\mathrm{RBIC}_{\boldsymbol{\gamma}}(\lambda) &= (\tilde{y}_{\boldsymbol{\gamma}_\lambda} - D_2\boldsymbol{\gamma}_\lambda)^t W_{\boldsymbol{\gamma}_\lambda}(\tilde{y}_{\boldsymbol{\gamma}_\lambda} - D_2\boldsymbol{\gamma}_\lambda) + \log n \frac{\mathrm{tr}(P_{\boldsymbol{\gamma}}^{-1}Q_{\boldsymbol{\gamma}})}{n},
\end{aligned}$$

where for given λ we have that $P_{\boldsymbol{\beta}} = \frac{1}{n}(D_1^t W_{\boldsymbol{\beta}} D_1 + 2\lambda S_{\boldsymbol{\beta}})$ and $Q_{\boldsymbol{\beta}} = \frac{1}{n} D_1^t A_{\boldsymbol{\beta}} D_1$ with similar expressions for $P_{\boldsymbol{\gamma}}$ and $Q_{\boldsymbol{\gamma}}$. The matrix $A_{\boldsymbol{\beta}}$ is a diagonal matrix with elements

$$a_i = \mathrm{Var}[\nu_1(y_i, \mu_i, \theta_i)] w_1(\boldsymbol{x}_i, \boldsymbol{z}_i)^2 h'(\boldsymbol{x}_i^t\boldsymbol{\beta})^2,$$

on the diagonal.

In the supplementary material we present the results of a small simulation study that confirms the robustness of this GAM RDE estimator. To illustrate the GAM RDE estimator, we consider a data set on Influenza-Like Illness (ILI) Visits in the United States (see Alimadad and Salibian-Barrera, 2011). The response variable contains weekly counts of ILI visits in the United States while the predictor is the considered week in the influenza season. This season starts from week 40 and runs until the end of week 20 of the next year, so it lasts 33 weeks. Data are available for the influenza seasons of 2006/2007, 2007/2008 and 2008/2009 as shown in Figure 5. Note the 4 high counts at the end of season 2008-2009, which can be explained by the fact that the H1N1 flu started spreading. Moreover, also seasonal variation can be observed from Figure 5, so robust estimation of both mean and dispersion is advisable.

We compare the fits of four models. The first model assumes a GLM for both mean and dispersion. The second model assumes a GAM for the mean model and a GLM for the dispersion model. The third model reversely assumes a GLM for the mean model and a GAM for the dispersion model while the final model assumes a GAM for both mean and dispersion. All models are fitted via the penalized RDE estimator. For GAMs, we use cubic regression splines with 10 knots and the commonly used integrated square second derivative penalty. The resulting fits are shown in Figure 5. It is clear that a GLM for the mean is not flexible enough, leading to poor fits. On the other hand, much better fits are obtained when a GAM is used for the mean. There is little difference between the fit of model 2 which assumes a GLM for the dispersion and model 4 which assumes a GAM for the dispersion, so a GLM seems sufficiently flexible to model the dispersion in the data.

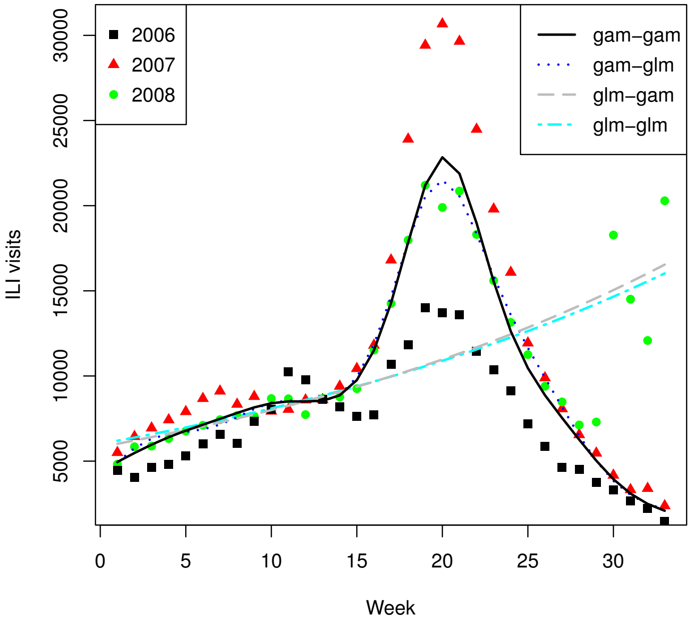

Figure 5: ILI visits data set with four fits obtained by different combinations of GLMs and GAMs estimated by (extended) RDE estimates.

7 Discussion

The double exponential family constitutes a powerful tool to model both mean and dispersion of the response in the context of generalized linear models. However, outliers in the data may heavily affect the classical estimates obtained by maximum likelihood. Therefore, we proposed the robust double exponential (RDE) estimator, which is less sensitive to outliers and allows to model simultaneously the mean and dispersion as a function of covariates. Moreover, we introduced a generalized quasi-deviance measure to develop robust inference which allows to test for the presence of dispersion among others.

While there is a natural choice for the link function in the mean model, it is well-known that there is no natural choice for the link function in the dispersion model. We have used the exponential function, which is a common choice. While reasonable choices for the link function lead to similar fits in our experience, a formal procedure to compare models based on different link functions would be desirable. This is a topic for further research.

The excellent performance of the RDE estimator and the corresponding robust tests was illustrated in an extensive simulation study. Especially the TRDE estimator based on the Tukey bisquare function combines a high level of robustness with a high accuracy of the inference. Real data applications illustrated that the proposed methodology can provide a better insight in the structure of the data and may lead to more reliable conclusions.

The weighted least squares representation of the RDE is exploited to develop penalized versions of the estimator. Sparse RDE estimators have been proposed to handle high-dimensional regression models while GAM RDE estimators have been introduced to allow for more flexible models. An implementation of RDE estimator together with its extensions will be made publicly available as an R package on CRAN.

References

Aeberhard, W. H., Cantoni, E., and Heritier, S. (2014). Robust inference in the negative binomial regression model with an application to falls data. Biometrics, 70(4):920–931.

(Cited on page 4.)

Aeberhard, W. H., Cantoni, E., and Heritier, S. (2017). Saddlepoint tests for accurate and robust inference on overdispersed count data. Computational Statistics & Data Analysis, 107:162 – 175. (Cited on page 4.)

Alimadad, A. and Salibian-Barrera, M. (2011). An outlier-robust fit for generalized additive models with applications to disease outbreak detection. Journal of the American Statistical Association, 106(494):719–731. (Cited on pages 23 and 24.)

Amiguet, M., Marazzi, A., Valdora, M., and Yohai, V. (2017). Robust estimators for generalized linear models with a dispersion parameter. ArXiv e-prints. (Cited on page 4.)

Antoniadis, A., Gijbels, I., Lambert-Lacroix, S., and Poggi, J. (2016). Joint estimation and variable selection for mean and dispersion in proper dispersion models. Electronic Journal of Statistics, 10(1):1630–1676. (Cited on page 2.)

Avella-Medina, M. and Ronchetti, E. (2018). Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika, 105(1):31–44. (Cited on pages 21 and 22.)

Bednarski, T. (1993). Fréchet differentiability of statistical functionals and implications to robust statistics. New Directions in Statistical Data Analysis and Robustness, pages 25–34. (Cited on page 8.)

Bergesio, A. and Yohai, V. J. (2011). Projection estimators for generalized linear models. Journal of the American Statistical Association, 106(494):661–671. (Cited on page 4.)

Bianco, A., Ben, M. G., and Yohai, V. J. (2005). Robust estimation for linear regression with asymmetric errors. Canadian Journal of Statistics, 33(4):511–528. (Cited on page 4.)

Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. In Rieder, H., editor, Robust Statistics, Data Analysis and Computer Intensive Methods, pages 17–34, New York. Lecture Notes in Statistics No. 109, Springer-Verlag. (Cited on page 4.)

Bickel, P. J., Hammel, E. A., O'Connell, J. W., et al. (1975). Sex bias in graduate admissions: Data from berkeley. Science, 187(4175):398–404. (Cited on page 16.)

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. Biometrika, 92(3):724–731. (Cited on page 4.)

Bondell, H. D. (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. (Cited on page 4.)

Cai, T. T., Wang, L., et al. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. The Annals of Statistics, 36(5):2025–2054. (Cited on page 2.)

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. Journal of the American Statistical Association, 96:1022–1030. (Cited on pages 4, 7, 8, 10, 12, 13, and 21.)

Cantoni, E. and Ronchetti, E. (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditure. Journal of Health Economics, 25:198–213. (Cited on pages 4 and 34.)

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. Journal of the Royal Statistical Society. Series B (Methodological), 55(3):693–706. (Cited on page 4.)

Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M functionals. Probability Theory and Related Fields, 73:197 – 209. (Cited on page 8.)

Croux, C., Gijbels, I., and Prosdocimi, I. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. Biometrics, 68(1):31–44. (Cited on pages 4 and 23.)

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. Computational Statistics and Data Analysis, 44:273–295. (Cited on page 4.)

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. Journal of the American Statistical Association, 82(400):1079–1091. (Cited on page 4.)

Davidian, M. and Carroll, R. J. (1988). A note on extended quasi-likelihood. Journal of the Royal Statistical Society. Series B (Methodological), 50(1):74–82. (Cited on page 4.)

Davies, R. B. (1980). Algorithm as 155: The distribution of a linear combination of χ^2 random variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(3):323–333. (Cited on page 12.)

Davies, R. B. (1990). Algorithm as 256: The distribution of a quadratic form in normal variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 39:294–309. (Cited on page 12.)

Efron, B. (1986). Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association, 81(395):709–721. (Cited on pages 2 and 3.)

Faught, E., Wilder, B., Ramsay, R., Reife, R., Kramer, L., Pledger, G., and Karim, R. (1996). Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. Neurology, 46(6):1684–1690. (Cited on page 18.)

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22. (Cited on page 21.)

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3):397–416. (Cited on page 21.)

Ghosh, A. and Basu, A. (2016). Robust estimation in generalized linear models: the density power divergence approach. TEST, 25(2):269–290. (Cited on page 4.)

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley, New York. (Cited on page 6.)

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC. (Cited on page 21.)

Heritier, S. (1993). Contributions to Robustness in Nonlinear Models : Application to Economic Data. PhD thesis, Université de Genève. (Cited on page 34.)

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). Robust methods in Biostatistics. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester. (Cited on page 8.)

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. Journal of the American Statistical Association, 89(427):897–904. (Cited on page 12.)

Hosseinian, S. and Morgenthaler, S. (2011). Robust binary regression. Journal of Statistical Planning and Inference, 141(4):1497 – 1509. (Cited on page 4.)

Huber, P. (1981). Robust Statistics. Wiley, New York. (Cited on pages 6 and 8.)

Jørgensen, B. (1987). Exponential dispersion models. Journal of the Royal Statistical Society. Series B (Methodological), 49(2):127–162. (Cited on page 4.)

Jørgensen, B. (1997). The Theory of Dispersion Models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. (Cited on page 4.)

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2010). Fast robust estimation of prediction error based on resampling. Computational Statistics & Data Analysis, 54(12):3121–3130. (Cited on page 22.)

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association, 84(406):460–466. (Cited on page 4.)

Lee, Y. and Nelder, J. A. (1998). Generalized linear models for the analysis of quality-improvement experiments. Canadian Journal of Statistics, 26(1):95–105. (Cited on page 3.)

Lee, Y. and Nelder, J. A. (2000). The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling geissler's human sex ratio data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 49(3):413–419. (Cited on pages 3 and 4.)

Lian, H., Liang, H., and Carroll, R. J. (2015). Variance function partially linear single-index models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(1):171–194. (Cited on page 2.)

Lopuhaä, H. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. The Annals of Statistics, 17(4):1662–1683. (Cited on page 7.)

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman and Hall, London. (Cited on page 2.)

Meinshausen, N. and Bühlmann, P. (2006). Variable selection and high-dimensional graphs with the lasso. Ann Stat, 34:1436–1462. (Cited on page 21.)

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

Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function. Biometrika, 74(2):221–232. (Cited on page 3.)

Nelder, J. A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. Journal of the Royal Statistical Society. Series B (Methodological), 54(1):273–284. (Cited on page 4.)

Neykov, N., Filzmoser, P., and Neytchev, P. (2012). Robust joint modeling of mean and dispersion through trimming. Computational Statistics & Data Analysis, 56(1):34 – 48. (Cited on page 4.)

Ronchetti, E. and Trojani, F. (2001). Robust inference with gmm estimators. Journal of econometrics, 101(1):37–69. (Cited on page 13.)

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79:871–880. (Cited on page 7.)

Rousseeuw, P. J. and Leroy, A. M. (2005). Robust regression and outlier detection, volume 589. John wiley & sons. (Cited on page 7.)

Small, C. G., Christopher, G., Wang, J., et al. (2003). Numerical methods for nonlinear estimating equations, volume 29. Oxford University Press on Demand. (Cited on page 34.)

Smyth, G. K. (1989). Generalized linear models with varying dispersion. Journal of the Royal Statistical Society. Series B (Methodological), 51(1):47–60. (Cited on pages 2 and 4.)

Smyth, G. K. and Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics, 10(6):695–709. (Cited on page 2.)

Tatsuoka, K. and Tyler, D. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. The Annals of Statistics, 28:1219–1243. (Cited on page 7.)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288. (Cited on page 21.)

Valdora, M. and Yohai, V. J. (2014). Robust estimators for generalized linear models. Journal of Statistical Planning and Inference, 146:31 – 48. (Cited on page 4.)

Wang, Y. and Van Aelst, S. (2019). Robust variable screening for regression using factor profiling. Statistical Analysis and Data Mining: The ASA Data Science Journal, 12(2):70–87. (Cited on page 22.)

Wong, R. K., Yao, F., and Lee, T. C. (2014). Robust estimation for generalized additive models. Journal of Computational and Graphical Statistics, 23(1):270–289. (Cited on pages 23 and 24.)

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476):1418–1429. (Cited on page 21.)

8 Appendix

8.1 Proof of Theorem 1

As before, let $\hat{\boldsymbol{\eta}}$ be the RDE estimator of $\boldsymbol{\eta}$ in the full model and $\hat{\boldsymbol{\eta}}_1^{(0)}$ the RDE estimator of $\boldsymbol{\eta}_1$ in the reduced model under the null hypothesis. Denote by $\hat{\boldsymbol{\eta}}^{(0)} = ((\hat{\boldsymbol{\eta}}_1^{(0)})^t, 0^t)^t$ the corresponding estimate of $\boldsymbol{\eta}$ under H_0. Let us define

$$\begin{aligned}\Lambda_{QM}^{\boldsymbol{\beta}} &= 2\sum_{i=1}^n \Big[Q_M^{\boldsymbol{\beta}}(y_i,\hat{\mu}_i,\hat{\theta}_i,\hat{\mu}_i,\hat{\theta}_i) - Q_M^{\boldsymbol{\beta}}(y_i,\hat{\mu}_i^{(0)},\hat{\theta}_i^{(0)},\hat{\mu}_i,\hat{\theta}_i)\Big],\\ \Lambda_{QM}^{\gamma} &= 2\sum_{i=1}^n \Big[Q_M^{\gamma}(y_i,\hat{\mu}_i,\hat{\theta}_i,\hat{\mu}_i,\hat{\theta}_i) - Q_M^{\gamma}(y_i,\hat{\mu}_i^{(0)},\hat{\theta}_i^{(0)},\hat{\mu}_i,\hat{\theta}_i)\Big],\end{aligned}$$

with

$$\begin{aligned}Q_M^{\boldsymbol{\beta}}(y_i,\tilde{\mu}_i,\tilde{\theta}_i,\mu_i,\theta_i) &= \int_{\tilde{s}_{1i}}^{\tilde{\mu}_i} \nu_1(y_i,s,\theta_i)w_1(\boldsymbol{x}_i,\boldsymbol{z}_i)\,ds - \frac{1}{n}\sum_{j=1}^n \int_{\tilde{s}_{2j}}^{\tilde{\mu}_j} \mathrm{E}\left[\nu_1(y,s,\theta_j)\right] w_1(\boldsymbol{x}_j,\boldsymbol{z}_j)\,ds,\\ Q_M^{\gamma}(y_i,\tilde{\mu}_i,\tilde{\theta}_i,\mu_i,\theta_i) &= \int_{\tilde{t}_{1i}}^{\tilde{\theta}_i} \nu_2(y_i,\mu_i,t)w_2(\boldsymbol{x}_i,\boldsymbol{z}_i)\,dt - \frac{1}{n}\sum_{j=1}^n \int_{\tilde{t}_{2j}}^{\tilde{\theta}_j} \mathrm{E}\left[\nu_2(y,\mu_j,t)\right] w_2(\boldsymbol{x}_j,\boldsymbol{z}_j)\,dt.\end{aligned}$$

Hence, $\Lambda_{QM} = \Lambda^{\boldsymbol{\beta}}_{QM} + \Lambda^{\boldsymbol{\gamma}}_{QM}$. Next, consider $\mathcal{D}_n$ as defined in condition B1. A second order Taylor expansion at any point of $\mathcal{D}_n$ leads to

$$\begin{aligned} Q_M(y_i, \hat{\mu}_i^{(0)}, \hat{\theta}_i^{(0)}, \hat{\mu}_i, \hat{\theta}_i) =& \, Q_M(y_i, \hat{\mu}_i, \hat{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) + (\hat{\boldsymbol{\eta}}^{(0)} - \hat{\boldsymbol{\eta}})^t \frac{\partial}{\partial \tilde{\boldsymbol{\eta}}} \left[Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) \right]_{\hat{\boldsymbol{\eta}}} \\ & + \frac{1}{2} (\hat{\boldsymbol{\eta}}^{(0)} - \hat{\boldsymbol{\eta}})^t \frac{\partial^2}{\partial \tilde{\eta}_j \partial \tilde{\eta}_k} \left[Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) \right]_{\boldsymbol{\eta}^*} (\hat{\boldsymbol{\eta}}^{(0)} - \hat{\boldsymbol{\eta}}), \end{aligned}$$

where $\boldsymbol{\eta}^*$ is on the line joining $\hat{\boldsymbol{\eta}}^{(0)}$ and $\hat{\boldsymbol{\eta}}$. Using the Leibniz integral rule, one can show that:

$$\frac{\partial}{\partial \tilde{\boldsymbol{\eta}}} Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) = \begin{pmatrix} \Psi_{\boldsymbol{\beta}}(y_i, \tilde{\mu}_i, \hat{\theta}_i) \\ \Psi_{\boldsymbol{\gamma}}(y_i, \hat{\mu}_i, \tilde{\theta}_i) \end{pmatrix},$$

which is zero when evaluated in $\hat{\boldsymbol{\eta}}$. By differentiating this expression once more, we can see that $\frac{\partial^2}{\partial \tilde{\eta}_j \partial \tilde{\eta}_k} Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \hat{\mu}_i, \hat{\theta}_i)$ is a block diagonal matrix:

$$\frac{\partial^2}{\partial \tilde{\eta}_j \partial \tilde{\eta}_k} Q_M(y_i, \tilde{\mu}_i, \tilde{\theta}_i, \hat{\mu}_i, \hat{\theta}_i) = \begin{pmatrix} \frac{\partial}{\partial \tilde{\boldsymbol{\beta}}} \Psi_{\boldsymbol{\beta}}(y_i, \tilde{\mu}_i, \hat{\theta}_i) & \mathbf{0} \\ \mathbf{0} & \frac{\partial}{\partial \tilde{\boldsymbol{\gamma}}} \Psi_{\boldsymbol{\gamma}}(y_i, \hat{\mu}_i, \tilde{\theta}_i) \end{pmatrix}.$$

Combining the latter results, we obtain:

$$\begin{aligned} \Lambda^{\boldsymbol{\beta}}_{QM} + \Lambda^{\boldsymbol{\gamma}}_{QM} =& - \sum_{i=1}^n (\hat{\boldsymbol{\beta}}^{(0)} - \hat{\boldsymbol{\beta}})^t \left[\frac{\partial}{\partial \tilde{\boldsymbol{\beta}}} \Psi_{\boldsymbol{\beta}}(y_i, \tilde{\mu}_i, \hat{\theta}_i) \right]_{\boldsymbol{\beta}^*} (\hat{\boldsymbol{\beta}}^{(0)} - \hat{\boldsymbol{\beta}}) \\ & - \sum_{i=1}^n (\hat{\boldsymbol{\gamma}}^{(0)} - \hat{\boldsymbol{\gamma}})^t \left[\frac{\partial}{\partial \tilde{\boldsymbol{\gamma}}} \Psi_{\boldsymbol{\gamma}}(y_i, \hat{\mu}_i, \tilde{\theta}_i) \right]_{\boldsymbol{\gamma}^*} (\hat{\boldsymbol{\gamma}}^{(0)} - \hat{\boldsymbol{\gamma}}). \end{aligned}$$

Using this expression, we can employ a similar reasoning as in Cantoni and Ronchetti (2006) and Proposition 4.1 in Heritier (1993) for $\Lambda^{\boldsymbol{\beta}}_{QM}$ and $\Lambda^{\boldsymbol{\gamma}}_{QM}$ separately to obtain the desired result.

8.2 Fisher scoring and weighted least squares representation

The two estimating equations (3) for the RDE estimator can easily be solved alternately via Fisher scoring (Small et al., 2003, p. 50-52). For $\boldsymbol{\beta}$ we obtain that

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\beta} + (nM_{\boldsymbol{\beta}})^{-1} \sum_{i=1}^n \Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= \boldsymbol{\beta} + (X^t B_{11} X)^{-1} \sum_{i=1}^n \Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{\beta}, \boldsymbol{\gamma}), \qquad (19) \end{aligned}$$

with B_{11} defined in the supplementary material. Similarly, $\boldsymbol{\gamma}$ satisfies

$$\boldsymbol{\gamma} \;=\; \boldsymbol{\gamma} + (Z^t B_{22} Z)^{-1} \sum_{i=1}^{n} \Psi_{\boldsymbol{\gamma}}(y_i, \boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{20}$$

with B_{22} defined in the supplementary material. These two equations can be solved alternately until convergence.

The two equations (19)-(20) can be rewritten in weighted least squares form. If we multiply both sides of (19) by $X^t B_{11} X$, then its j-th component becomes

$$\begin{aligned}
\big[X^t B_{11} X \boldsymbol{\beta}\big]_j &= \Big[X^t B_{11} X \boldsymbol{\beta} + \sum_{i=1}^{n} \Psi_{\boldsymbol{\beta}}(y_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\Big]_j \\
&= \sum_{l=1}^{p} \sum_{i=1}^{n} x_{ij} b_{11,i} x_{il} \beta_l + \sum_{i=1}^{n} \left(\nu_1(y_i, \mu_i, \theta_i) w_1(\boldsymbol{x}_i, \boldsymbol{z}_i) \mu'_{i,j} - a_{1,j}\right) \\
&= \sum_{l=1}^{p} \sum_{i=1}^{n} x_{ij} b_{11,i} x_{il} \beta_l + \sum_{i=1}^{n} (\nu_1(y_i, \mu_i, \theta_i) - E[\nu_1(y, \mu_i, \theta_i)]) w_1(\boldsymbol{x}_i, \boldsymbol{z}_i) \mu'_{i,j} \\
&= \sum_{i=1}^{n} \Big(\boldsymbol{x}_i^t \boldsymbol{\beta} + \frac{1}{b_{11,i}} (\nu_1(y_i, \mu_i, \theta_i) - E[\nu_1(y, \mu_i, \theta_i)]) w_1(\boldsymbol{x}_i, \boldsymbol{z}_i) h'(\boldsymbol{x}_i^t \boldsymbol{\beta})\Big) b_{11,i} x_{ij} \\
&= \sum_{i=1}^{n} \left(\boldsymbol{x}_i^t \boldsymbol{\beta} + \frac{\nu_1(y_i, \mu_i, \theta_i) - E[\nu_1(y, \mu_i, \theta_i)]}{E[\nu_1(y, \mu_i, \theta_i) U_{\mu_i}] h'(\boldsymbol{x}_i^t \boldsymbol{\beta})}\right) b_{11,i} x_{ij} \\
&= [X^t B_{11} \tilde{y}_{\boldsymbol{\beta}}]_j,
\end{aligned}$$

with $\tilde{y}_{\boldsymbol{\beta}}$ defined in (13). Hence, $\boldsymbol{\beta}$ is a solution of weighted least squares problem (11).

Similarly, the fixed point solution for $\boldsymbol{\gamma}$ in (20) can be rewritten as

$$Z^t B_{22} Z \boldsymbol{\gamma} = Z^t B_{22} \tilde{y}_{\boldsymbol{\gamma}},$$

which implies that $\boldsymbol{\gamma}$ solves the weighted least squares problem (12).