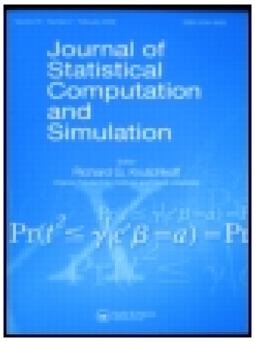

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Robust rank-based variable selection in double generalized linear models with diverging number of parameters under adaptive Lasso

Brice M. Nguelifack & Elizabeth Kemajou-Brown

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2019.1607346

Taylor & Francis
Taylor & Francis Group

Robust rank-based variable selection in double generalized linear models with diverging number of parameters under adaptive Lasso

Brice M. Nguelifack[a] and Elizabeth Kemajou-Brown[b]

[a]Mathematics, United States Naval Academy, Annapolis, MD, USA; [b]Mathematics, Morgan State University, Baltimore, MD, USA

ABSTRACT
We propose a robust rank-based estimation and variable selection in double generalized linear models when the number of parameters diverges with the sample size. The consistency of the variable selection procedure and asymptotic properties of the resulting estimators are established under appropriate selection of tuning parameters. Simulations are performed to assess the finite sample performance of the proposed estimation and variable selection procedure. In the presence of gross outliers, the proposed method is showing that the variable selection method works better. For practical application, a real data application is provided using nutritional epidemiology data, in which we explore the relationship between plasma beta-carotene levels and personal characteristics (e.g. age, gender, fat, etc.) as well as dietary factors (e.g. smoking status, intake of cholesterol, etc.).

ARTICLE HISTORY
Received 9 July 2018
Accepted 10 April 2019

KEYWORDS
Double generalized linear models; diverging number of parameters; high leverage point; rank-based; variable selection

JEL CLASSIFICATIONS
62G08; 62J12; 62K25

1. Introduction

In these days, many statistical problems are based on 'big data' which have raised the necessity to find a more robust way of determining which of the available factors are important to be considered in statistical modelling. For example, in generalized linear models (GLMs), determining important factors can be translated as identifying the covariates that are most needed in the prediction of the average outcome. A substantial amount of attention has been paid to GLMs which were first introduced by [1]. Recently, Miakonkana and Abebe [2] developed an iterative estimation approach using pseudo rank-based for GLM. This technique of modelling is now one of the most widely used techniques for statistical modelling [3]. It is also known as one of the most flexible generalization of linear regression. GLMs are mostly used in the situation where we allow the linear model to be related to the response variable via a link function. It also allows the magnitude of the variance of each measurement to be a function of the mean value. Models under GLMs usually allow for that magnitude of the variance also known as dispersion parameter to be constant. However,

CONTACT Brice M. Nguelifack ✉ nguelifa@usna.edu Mathematics, United States Naval Academy, 572-C Holloway Rd, Suite 300, Yard Mail Stop 9E, Annapolis, MD 21402, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2019.1607346

not all models satisfy the assumption under which the dispersion parameter is set to be constant. Thus the need to model the dispersion parameter as a function of some covariates. GLMs that also allow the dispersion parameter to be a function of some covariates via a link function are often called double generalized linear models (DGLMs).

Recent studies have been done in DGLMs with various applications such as insurance claims and losses. The maximum likelihood estimation for DGLMs was introduced by Smyth [4], where the authors considered several population distributions such as normal and inverse Gaussian. Wu and Li [5] looked at variable selection for joint mean and dispersion models of the inverse Gaussian distribution using the pseudo likelihood approach. Rigby and Stasinopoulos [6] considered smoothing spline modelling of both the mean and the dispersion and, [7,8] considered dispersion effects in designed experiments. Modelling using DGLMs while selecting important factors or covariates has been proven to be very useful in fitting more complex and potentially realistic models. As in [9], variable selection is fundamental to statistical modelling, especially for DGLMs with a diverging number of parameters. An effective variable selection can lead to better risk assessment and model interpretation. Concerning variable selection, the Lasso method introduced in [10] has generated significant attention in the past two decades. It is well known that the Lasso method is fundamentally very efficient, since it takes advantage of the singularity of the L_1 penalty to effectively select variables via a penalized procedure. Many other authors have worked on refining this selection procedure. For example, [11–14] had much of their work focusing in establishing the so-called 'oracle' property that consists of selection consistency and estimation efficiency.

Variable selection in DGLMs with a diverging number of parameters was recently studied by Xu et al. [9], in which they developed an efficient penalized pseudo-likelihood based method to select explanatory variables that make a significant contribution to DGLMs. They proposed a procedure that simultaneously selects significant variables in parametric components in mean model and parametric components in variance model. However, not too many results have addressed the lack of optimality of these variable selection procedures when the data contain gross outliers. Among many, the approach based on penalized Jaeckel-type rank regression was discussed in [15–19] in the low dimensional linear model case. The computation in this scenario is somehow complicated and these are only robust in the response space. Wang et al. [12] tried to discuss about a procedure that addresses the influence of high leverage points, but in the low dimensional linear model case. Wu et al. [20] proposed a robust variable selection to t-type joint GLMs via penalized t-type pseudo-likelihood that intent to model data containing extreme or outlying observations. But again their method does not include the case where we have diverging number of parameters. The approach we propose here which is based on minimization of a penalized weighted rank-based pseudo-norm is much simpler to compute and provides protection against outliers and high-leverage points in both the response and design space for DGLMs with a diverging number of parameters. Early results show that the penalized weighted rank-based pseudo-norm performs slightly better than the penalized weighted least absolute deviation norm and performs better than the penalized pseudo-likelihood in the presence of high leverage points.

The rest of the paper is organized as follows. In Section 2, we introduce some notations together with the robust rank-based estimation and variable selection techniques. Asymptotic results are presented in Section 3. The implementation of our techniques is outlined in Section 4. Simulation studies and a real data application using the plasma concentrations of Beta-carotene data to illustrate the procedure are given in Section 5. Section 6 concludes the article with a discussion. All proofs and technical details are provided in the Appendix.

2. Robust rank-based estimator and variable selection

2.1. Model and notation

We denote by $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^{\mathrm{T}}$ the vector of n independent responses, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ the vector of size n of predictors, and n the sample size. We recall that the superscript 'T' in the rest of this paper will denote the transpose of a vector (or matrix). We also let $\mu_i = E[\mathbf{y}_i|\mathbf{x}_i]$ the mean, $\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i) = \phi_i V(\mu_i)$ the variance for the ith observation, where ϕ_i is the dispersion and $V(\cdot)$ is a known variance function. We model the mean using a link-linear model,

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}, \tag{1}$$

where g is a monotonic differentiable link function, $\mathbf{x}_i$ is the ith component of a vector of covariates $\mathbf{x}$ relevant for predicting the mean and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_n})^{\mathrm{T}}$ is a vector of regression coefficients. In addition, we model the dispersion using a link-linear model,

$$h(\phi_i) = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\beta}, \tag{2}$$

where h is the another monotonic link function and $h^{-1}(\cdot) > 0$, $\mathbf{z}_i$ is the ith component of a vector of covariates $\mathbf{z}$ relevant for predicting the dispersion and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{q_n})^{\mathrm{T}}$ is the another vector of regression coefficients. It is important to notice that the subscripts p_n and q_n are used to show that both the covariates and the parameters are allowed to change with n.

2.2. Robust rank-based estimation

We first introduce the rank-based estimation technique and some notation that will be used throughout the rest of this paper. Let denote by Ω the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$. It is well known that the least squares (LS) estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ minimizes the euclidian distance between Y and Ω. We can replace the euclidian norm $\|\cdot\|_2$ with a rank pseudo-norm $\|\cdot\|_\varphi$ (see [21], Chapter 3) defined as

$$\|\mathbf{v}\|_\varphi = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{a}[\mathbf{R}(\mathbf{v_i})]\mathbf{v_i}, \tag{3}$$

where $R(\mathbf{v_i})$ is the rank of $\mathbf{v_i}$ among $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}$, and $\mathbf{a}(i) = \varphi(i/(n+1))$ for some nondecreasing score generating function φ standardized such that $\int_0^1 \varphi(u)\,\mathrm{d}u = 0$ and $\int_0^1 \varphi(u)^2\,\mathrm{d}u = 1$. Popular score generating functions are the Wilcoxon function $\varphi(u) = \sqrt{12}(u - 1/2)$, the sign function $\varphi(u) = \mathrm{sign}(u - 1/2)$ (yields the L_1 norm), the logistic function $\varphi(u) = 2u - 1$, the Bent score function $\varphi(u) = 4u - 1.5$, if $u \leq 0.5$, $\varphi(u) = 0.5$, if $u > 0.5$, and the normal function $\varphi(u) = \Phi^{-1}((u+1)/2)$ where Φ is the standard normal distribution function. Jaeckel [22] showed that the objective function $D_n(\cdot)$ induced by the above pseudo-norm is continuous, convex and almost everywhere differentiable function of $\boldsymbol{\beta}$. Thus a minimizer of the rank-based objective function exists and the resulting minimizer is the rank-based estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. More about the rank-based objective function can be found in [22].

Now, we define our robust rank-based objective function by including to $D_n(\boldsymbol{\beta})$ defined in [22], a positive function depending on the covariates called the weight function $w(\cdot)$. So choosing our Pearson residuals as in [2,9], we therefore have that

$$e_i = \frac{\mathbf{y}_i - \mu_i}{\sqrt{\phi_i V(\mu_i)}},$$

and given observations $\mathbf{y}_i$, $i = 1, \ldots, n$, we define the robust rank-based dispersion function using Wilcoxon scores as

$$Q_n(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i}^{n} w(\mathbf{x}_i) \left[\frac{R(e_i)}{n+1} - \frac{1}{2} \right] e_i, \tag{4}$$

where $R(e_i)$ is the rank of e_i, among $e_1, e_2, \ldots, e_n$, $w(\cdot)$ is a positive and symmetric weights used to down-weigh high leverage points.

Remark 2.1: It is well known that when the weight function is identically equal to one, then $Q_n(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the usual Wilcoxon statistic. Now, since Wilcoxon estimates are only robust in regard to the response, they may not be appropriate in observational studies if the independent variables (i.e. $\mathbf{x}_i$) are contaminated. As an alternative, we consider an analysis based on weighted Wilcoxon (WW) estimates. It is shown in [23] that the objective function defined in (4) is equivalent to

$$\sum_{1 \le i \le j \le n} b_{ij} |e_j - e_i|, \tag{5}$$

where b_{ij} denotes a weight to be used in the (i, j)th components. Also note that the above representation (5) is essentially a weighted version of Gini's mean difference measure of scale.

Now we are ready to define our penalized weighted rank-based objective function. Thus suppose that $\{\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i\}$, $i = 1, \ldots, n$, constitute an independent and identically distributed sample. Then we propose the following penalized weighted objective function,

$$D_n(\boldsymbol{\theta}, \boldsymbol{\beta}) = Q_n(\boldsymbol{\theta}, \boldsymbol{\beta}) - n \sum_{j=1}^{p_n} p_{\lambda_j^{(1)}}(|\theta_j|) - n \sum_{k=1}^{q_n} p_{\lambda_k^{(2)}}(|\beta_k|), \tag{6}$$

where $p_{\lambda_j^{(l)}}(\cdot)(l = 1, 2)$ are given penalty functions, $\lambda_j^{(l)}(l = 1, 2)$ are regularization parameters. In this paper, the regularization parameters will be chosen via BIC-type tuning parameter selector [12]. Note that the penalty functions and tuning parameters are not necessarily the same for all j, since we are only considering the adaptive lasso case for our proposed method. Another example will be that, if we wish to keep some important variables in the final model, we therefore don't want to penalize their coefficients.

For the sake of simplicity, we rewrite (6) as follows:

$$D_n(\boldsymbol{\gamma}) = Q_n(\boldsymbol{\gamma}) - n\sum_{j=1}^{p_n} p_{\lambda_j^{(1)}}(|\theta_j|) - n\sum_{k=1}^{q_n} p_{\lambda_k^{(2)}}(|\beta_k|), \tag{7}$$

where $D_n(\boldsymbol{\gamma}) = D_n(\boldsymbol{\theta}, \boldsymbol{\beta})$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{s_n})^{\mathrm{T}} = (\theta_1, \ldots, \theta_{p_n}; \beta_1, \ldots, \beta_{q_n})^{\mathrm{T}}$ with $s_n = p_n + q_n$ and $p_{\lambda_l^{(1)}}(\cdot)$ is a given penalty function with the tuning parameter $\lambda^{(l)} (l = 1, 2)$.

We could consider few of the several penalty functions as suggested in [17] such as the Lasso penalty function $p_{\lambda_j}(|t|) = \lambda|t|$ (for all j). The smoothly clipped absolute deviation (SCAD) penalty presented as

$$p_\lambda(|t|) = \lambda|t| \left\{ I(|t| \leq \lambda) + \frac{(a\lambda - |t|/2\lambda)}{(a-1)\lambda} I(\lambda < |t| \leq a\lambda) + \frac{a^2\lambda}{(a-1)2|t|} I(|t| > a\lambda) \right\}$$

for some $a > 2$. But for this paper, we will only consider the case of Lasso in which we assume that all λ s are not the same. That is, we will only consider the adaptive lasso case.

Remark 2.2: Note that the above notation, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{s_n})^{\mathrm{T}} = (\theta_1, \ldots, \theta_{p_n}; \beta_1, \ldots, \beta_{q_n})^{\mathrm{T}}$ means that, if it is the first parameter say $\boldsymbol{\theta}$ the parameter of interest then only the first p_n parameters are considered the rest of the q_n parameters being replaced by the zeros. Alternatively, if it is the second parameter say $\boldsymbol{\beta}$ the parameter of interest then only the last q_n parameters are considered the rest of the first p_n parameters being replaced by the zeros.

Letting Θ representing the parameter space, the corresponding penalized rank-based estimator is therefore defined by

$$\widehat{\boldsymbol{\gamma}} = \operatorname*{Argmin}_{\Theta} D_n(\boldsymbol{\gamma}).$$

Remark 2.3: With appropriate penalty functions, minimizing $D_n(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ leads to certain parameter estimators vanishing from the initial models so that the corresponding explanatory variables are automatically removed. Hence, through minimizing $D_n(\boldsymbol{\gamma})$ we achieve the goal of selecting important variables and obtaining the parameter estimators simultaneously.

Before stating the main results of this paper, we first present below some technical details and later provide an algorithm for calculating the rank-based DGLM estimator $\widehat{\boldsymbol{\gamma}}$.

2.3. Some useful approximations and notations

The rank dispersion function $Q_n(\cdot)$ presented in Equation (4) was well studied by [22,23]. It was shown in [22] that, $Q_n(\cdot)$ is continuous, convex and almost everywhere differentiable function of $\boldsymbol{\gamma}_n$. Thus these guarantee that the first two derivatives of the rank dispersion function $Q_n(\boldsymbol{\gamma})$ are continuous. First we define the rank score function denoted by $\Psi_n(\boldsymbol{\gamma}_n)$

defined as

$$\Psi_n(\boldsymbol{\gamma}_n) = -\nabla Q_n(\boldsymbol{\gamma}) = \left(U_1^{\mathrm{T}}(\boldsymbol{\theta}_n), U_2^{\mathrm{T}}(\boldsymbol{\beta}_n)\right)^{\mathrm{T}},$$

where

$$U_1^{\mathrm{T}}(\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} Q_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \Lambda_i \left[-\frac{1}{\phi_i^{1/2} V^{1/2}} - \frac{1}{2}\frac{V'}{V} \times \frac{y_i - \mu_i}{\phi_i^{1/2} V^{1/2}} \right] g'(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta})\mathbf{x}_i,$$

where $\Lambda_i = w(\mathbf{x}_i)[(R(e_i)/n+1) - (1/2)]$ and $V' = (\mathrm{d}/\mathrm{d}\theta)V, g' = (\mathrm{d}/\mathrm{d}\theta)g$. We also have

$$U_1^{\mathrm{T}}(\boldsymbol{\beta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}} Q_n(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \Lambda_i \left[-\frac{1}{2}\frac{1}{\phi_i^{3/2} V^{1/2}} \right] g_d'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_i \,.$$

We also denote by $\dot{\Psi}_n = E[\nabla^2 Q_n(\boldsymbol{\gamma})]$ the expectation of the second-order derivative of the rank dispersion function also known as the Hessian matrix given by

$$\dot{\Psi}_n = E[\nabla^2 Q_n(\boldsymbol{\gamma})] = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix},$$

where $\{\mathcal{I}_{ij}\}$ are given by

$$\mathcal{I}_{11} = E\left(\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\theta}^{\mathrm{T}}} Q_n(\boldsymbol{\gamma})\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} \Lambda_i \left[\frac{V'}{2\phi_i^{1/2} V^{3/2}}\right] \rho_1^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right),$$

$$\mathcal{I}_{22} = 0,$$

$$\mathcal{I}_{12} = E\left(\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\beta}^{\mathrm{T}}} Q_n(\boldsymbol{\gamma})\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} \Lambda_i \left[\frac{V'}{2\phi_i^{1/2} V^{3/2}}\right] \rho_1\rho_2 \mathbf{x}_i \mathbf{z}_i^{\mathrm{T}}\right),$$

and

$$\mathcal{I}_{21} = E\left(\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\beta}^{\mathrm{T}}} Q_n(\boldsymbol{\gamma})\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} \Lambda_i \left[\frac{1}{2\phi_i^{3/2} V^{1/2}}\right] \rho_1\rho_2 \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}\right),$$

where $\rho_1 = (\mathrm{d}/\mathrm{d}t)g^{-1}(t)$ and $\rho_2 = (\mathrm{d}/\mathrm{d}t)h^{-1}(t)$.

Let $\eta_n = \sqrt{n(p_n + q_n)}$, the following is an extension of the quadratic approximation of $Q_n(\boldsymbol{\gamma}_n)$ in any η_n neighbourhood of $\boldsymbol{\gamma}_{n0}$ of [22] in the case of high dimension. Thus for any $C > 0$,

$$Q_n(\boldsymbol{\gamma}) = Q_n(\boldsymbol{\gamma}_{n0}) + \left[\nabla Q_n(\boldsymbol{\gamma}_{n0})\right]^{\mathrm{T}} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_{n0}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_{n0})^{\mathrm{T}} \left[\nabla^2 Q_n(\boldsymbol{\gamma}_{n0})\right] (\boldsymbol{\gamma} - \boldsymbol{\gamma}_{n0}),$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \gamma_{s_n})^{\mathrm{T}} = (\theta_1, \ldots, \theta_{p_n}; \beta_1, \ldots, \beta_{q_n})^{\mathrm{T}}$ and $\boldsymbol{\gamma}_{n0} = (\gamma_{n01}, \ldots, \gamma_{n0s_n})^{\mathrm{T}} = (\theta_{n01}, \ldots, \theta_{n0p_n}; \beta_{n01}, \ldots, \beta_{n0q_n})^{\mathrm{T}}$.

The derivatives of the penalty functions play important roles in the operating characteristics of the penalized estimators. We use the dummy argument t for one of $\boldsymbol{\theta}_j$ or $\boldsymbol{\beta}_j$ for all j, and adopt the following notation:

$$\nabla p_{\lambda_j^{(l)}}(|t|) = q_{\lambda_j^{(l)}}(|t|)\mathrm{sgn}(t) \quad \text{and} \quad \nabla q_{\lambda_j^{(l)}}(|t|) = \dot{q}_{\lambda_j^{(l)}}(|t|)\mathrm{sgn}(t), \quad l = 1, 2.$$

3. Consistency and asymptotic distribution

This section studies the asymptotic properties of the proposed robust rank-based DGLM variable selection and estimation. We follow the same notation as in [9]. Let $\boldsymbol{\gamma}_{n0}$ denote the true values of $\boldsymbol{\gamma}_n$. Furthermore, let $\boldsymbol{\gamma}_{n0} = (\gamma_{n01}, \dots, \gamma_{n0s_n})^{\mathrm{T}} = (\boldsymbol{\gamma}_{n0}^{(1)^{\mathrm{T}}}, \boldsymbol{\gamma}_{n0}^{(2)^{\mathrm{T}}})^{\mathrm{T}}$. For ease of presentation and without loss of generality, it is assumed that $\boldsymbol{\gamma}_{n0}^{(1)}$ consists of all non-zero components of $\boldsymbol{\gamma}_{n0}$ and that $\boldsymbol{\gamma}_{n0}^{(2)} = 0$. Let

$$a_n = \max_{1 \le j \le s_n} \left\{ |q_{\lambda_j}(|\gamma_{n0j}|)|, \gamma_{n0j} \neq 0 \right\}$$

and

$$b_n = \max_{1 \le j \le s_n} \left\{ |\dot{q}_{\lambda_j}(|\gamma_{n0j}|)|, \gamma_{n0j} \neq 0 \right\},$$

where q_{λ_j} and $\dot{q}_{\lambda_j}$ represent the first and second derivatives of p_{λ_j} respectively.

The following assumptions will be made:

- (**A1:**) The parameter space Θ is compact and the true value $\boldsymbol{\gamma}_{n0}$ is in the interior of the parameter space.
- (**A2:**) There exists finite constants $n_0 > 0$ and $C > 0$ such that $(\delta_{\max}(\dot{\Psi}_n)/\delta_{\min}(\dot{\Psi}_n)) < C$ for all $n \ge n_0$, where $\delta_{\min}(\dot{\Psi}_n)$ and $\delta_{\max}(\dot{\Psi}_n)$ denote the minimum eigenvalue of matrix $\dot{\Psi}_n$ and the maximum eigenvalue of matrix $\dot{\Psi}_n$ respectively.
- (**A3:**) The non-zero components of true parameters, $\gamma_{n01}, \dots, \gamma_{n0s_{1n}}$ satisfy

$$\min_{1 \le j \le s_{1n}} \left\{ \frac{|\gamma_{n0j}|}{\lambda_n} \right\} \to \infty, \quad (n \to \infty).$$

Remark 3.1: Because $Q_n(\boldsymbol{\gamma})$ is continuous at $\boldsymbol{\gamma}$, using Lemma 2 of [22], assumption **A1** implies the existence of a minimizer of $Q_n(\boldsymbol{\gamma})$. Assumptions **A2** to **A3** follow the same ideas as in [9,12]. As noted in their paper, these assumptions guarantee the $\sqrt{n/s_n}$-consistency and the asymptotic normality of the unpenalized estimator, the asymptotic quadraticity of the unpenalized objective function and the asymptotic linearity of the corresponding score function.

The following theorem states the main theoretical result regarding the penalized weighted rank-based estimator, including the existence of an $\sqrt{n/s_n}$-consistent estimator, the sparsity of the estimator (that is shrinking some coefficient estimates exactly to zero) and the asymptotic normality of the estimator.

Theorem 3.1: *Assume $a_n = O_p(n^{-1/2})$, $b_n \to 0$, $\lambda_n \to 0$ and $s_n^4/n \to 0$ as $n \to \infty$. λ_n is equal to either $\lambda_n^{(1)}$ or $\lambda_n^{(2)}$ depending on whether γ_{n0j} is a component of $\boldsymbol{\theta}_{n0}$, $\boldsymbol{\beta}_{n0}$, $(1 \le j \le s_n)$. Under conditions* **A1–A3**, *with probability tending to 1 there must exist a local minimizer $\widehat{\boldsymbol{\gamma}}_n$ of the Wilcoxon-type penalized weighted rank-based dispersion function $D_n(\boldsymbol{\gamma}_n)$ in* (6) *such that $\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n0}\| = O_p(\sqrt{n/s_n})$.*

The following theorem gives the asymptotic normality property of $\widehat{\boldsymbol{\gamma}}_n$. Denote the number of non-zero components of $\boldsymbol{\gamma}_{n0}$ by $s_{1n}(< s_n)$. Let

$$\nabla \mathbf{p}_{\lambda_n}\left(\boldsymbol{\gamma}_{n0}^{(1)}\right) = \operatorname{diag}\left(q_{\lambda_n}\left(\boldsymbol{\gamma}_{n01}^{(1)}\right),\ldots,q_{\lambda_n}\left(\boldsymbol{\gamma}_{n0s_{1n}}^{(1)}\right)\right),$$

$$\nabla^2 \mathbf{p}_{\lambda_n}\left(\boldsymbol{\gamma}_{n0}^{(1)}\right) = \operatorname{diag}\left(\dot{q}_{\lambda_n}\left(\boldsymbol{\gamma}_{n01}^{(1)}\right),\ldots,\dot{q}_{\lambda_n}\left(\boldsymbol{\gamma}_{n0s_{1n}}^{(1)}\right)\right),$$

where λ_n has the same definition as that in Theorem 3.1 and $\boldsymbol{\gamma}_{n0j}^{(1)}$ is the jth component of $\boldsymbol{\gamma}_{n0}^{(1)}(1 \leq j \leq s_{1n})$ and $q_{\lambda_n}(t) = \mathrm{d}p_{\lambda_n}(t)/\mathrm{d}t$. Here $p_{\lambda_n}(t)$ is a component of $\mathbf{p}_{\lambda_n}(t)$. We also have

$$\Sigma_\lambda(\boldsymbol{\gamma}_n) = \operatorname{diag}\left\{\frac{q_{\lambda_1^{(1)}}(|\theta_{n1}|)}{|\epsilon+\theta_{n1}|},\ldots,\frac{q_{\lambda_{pn}^{(1)}}(|\theta_{np_n}|)}{|\epsilon+\theta_{np_n}|},\frac{q_{\lambda_1^{(2)}}(|\beta_{n1}|)}{|\epsilon+\beta_{n1}|},\ldots,\frac{q_{\lambda_{qn}^{(2)}}(|\beta_{nq_n}|)}{|\epsilon+\beta_{nq_n}|}\right\},$$

$$\mathbf{q}_\lambda(\boldsymbol{\gamma}_n) = \left\{q_{\lambda_1^{(1)}}(|\theta_{n1}|),\ldots,q_{\lambda_{pn}^{(1)}}(|\theta_{np_n}|),q_{\lambda_1^{(2)}}(|\beta_{n1}|),\ldots,q_{\lambda_{qn}^{(2)}}(|\beta_{nq_n}|)\right\}^{\mathrm{T}},$$

with $\boldsymbol{\gamma}_n = (\gamma_{n1},\ldots,\gamma_{ns_n}) = (\theta_{n1},\ldots,\theta_{np_n};\beta_{n1},\ldots,\beta_{nq_n})^{\mathrm{T}}$, $\boldsymbol{\gamma}_{n0} = (\gamma_{n01},\ldots,\gamma_{n0s_n}) = (\theta_{n01},\ldots,\theta_{n0p_n};\beta_{n01},\ldots,\beta_{n0q_n})^{\mathrm{T}}$ and with ϵ chosen to be a small number. The below theorem gives the oracle property of our variable selection procedure.

Theorem 3.2: *Assume that the penalty function $p_{\lambda_n(t)}$ satisfies*

$$\liminf_{n\to\infty}\ \liminf_{t\to 0^+}\frac{q_{\lambda_n(t)}}{\lambda_n} > 0,$$

and under the same assumptions as in Theorem 3.1, *if $\lambda_n \to 0$, $s_n^5/n \to 0$ and $\lambda_n\sqrt{n/s_n} \to \infty$, then the $\sqrt{n/s_n}$-consistent estimator $\widehat{\boldsymbol{\gamma}}_n = (\widehat{\boldsymbol{\gamma}}_n^{(1)^{\mathrm{T}}},\widehat{\boldsymbol{\gamma}}_n^{(2)^{\mathrm{T}}})^{\mathrm{T}}$ satisfies the following conclusion*:

(a) *If $\sqrt{n}a_n \xrightarrow{P} 0$, $\sqrt{n}b_n \xrightarrow{P} \infty$ and $\lambda_n\sqrt{n/s_n} \xrightarrow{P} \infty$, then $\widehat{\boldsymbol{\gamma}}_n^{(2)} \xrightarrow{P} \mathbf{0}$,*

(b) $\sqrt{n}(\dot{\Psi}_{n(1)} + \Sigma_\lambda^{11})\{(\widehat{\boldsymbol{\gamma}}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}) + (\dot{\Psi}_{n(1)} + \Sigma_\lambda^{11})^{-1}\mathbf{q}_{\lambda_n}(\boldsymbol{\gamma}_{n01}^{(1)})\} \xrightarrow{D} \mathcal{N}_k(0,\mathbf{G})$,

where $\dot{\Psi}_{n(1)}$, Σ_λ^{11} are the $s_{1n} \times s_{1n}$ submatrix in the upper left corner of $\dot{\Psi}_n$, $\Sigma_\lambda(\boldsymbol{\gamma}_n)$ respectively. $\mathbf{G} = \lim_n \tau^{-1}(\mathbf{V}_n^{\mathrm{T}}\mathbf{V}_n)/n$, $\mathbf{G}$ is a $k_0 \times k_0$ positive definite matrix, with $\mathbf{V}_n$ a $k_0 \times s_{1n}$ matrix, $\tau^2 = [\sqrt{12}\int f^2(u)\,\mathrm{d}u]^{-1}$, and $k_0(\leq s_{1n})$ is a constant.

The proofs for Theorems 3.1 and 3.2 are given in the Appendix. Finally, a consistent estimator for the asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\gamma}}_n^{(1)}$ is given by

$$n^{-1}\left(\widehat{\dot{\Psi}}_{n(1)} + n\widehat{\Sigma}_\lambda^{11}\right)^{-1}\widehat{\mathbf{G}}\left(\widehat{\dot{\Psi}}_{n(1)} + n\widehat{\Sigma}_\lambda^{11}\right)^{-1}.$$

Remark 3.2: In the above Theorem 3.2, we imposed assumptions on the penalty function and the regularization parameter in order to simultaneously achieve the $\sqrt{n/s_n}$-consistency of the regularized rank-based estimation and the consistency of the variable selection. It is crucial that the choice of our penality function satisfies those assumptions.

4. Implementation

4.1. Algorithm

The following algorithm summarizes the computation of the penalized weighted Wilcoxon-type rank-based estimators of the parameters in DGLMs.

Step 1 Take the weighted Wilcoxon-type rank-based estimator without penalty say $\boldsymbol{\theta}_0$, $\boldsymbol{\beta}_0$ of $\boldsymbol{\theta}_n$ and $\boldsymbol{\beta}_n$ as their initial values.

Step 2 Given current values $\{\boldsymbol{\theta}_n^{[k]}, \boldsymbol{\beta}_n^{[k]}\}$ or $\boldsymbol{\gamma}_n^{[k]} = \{\boldsymbol{\theta}_n^{[k]^{\mathrm{T}}}, \boldsymbol{\beta}_n^{[k]^{\mathrm{T}}}\}^{\mathrm{T}}$, update as in [17] using the iterative majorize–minorize (MM) algorithm given by

$$\boldsymbol{\gamma}_n^{[k+1]} = \boldsymbol{\gamma}_n^{[k]} - \rho_k \left[\dot{\Psi}_n^{[k]}\left(\boldsymbol{\gamma}_n^{[k]}\right) - n\Sigma_\lambda(\boldsymbol{\gamma}_n^{[k]})\right]^{-1} \times \left[\Psi_n^{[k]}\left(\boldsymbol{\gamma}_n^{[k]}\right) - n\mathbf{q}_\lambda(\boldsymbol{\gamma}_n^{[k]})\right], \quad k > 0, \tag{8}$$

where $\Psi_n(\boldsymbol{\gamma}_n)$ and $\dot{\Psi}_n(\boldsymbol{\gamma}_n)$ are also known as the gradient vector and Hessian matrix, respectively, ρ_k is some positive scalar.

Step 3 Repeat step 2 above until certain convergence criteria are satisfied. In our case, we continue until $\max_{1\le j\le s_n} |\boldsymbol{\gamma}_{nj}^{[k+1]} - \boldsymbol{\gamma}_{nj}^{[k]}| < 10^{-8}$.

4.2. Choosing the tuning parameter

The implementation of our method involves estimating the tuning parameters $\lambda^{(l)}(l = 1, 2)$, since the penalty function $p_{\lambda^{(l)}}(\cdot)$ involves the tuning parameter $\lambda^{(l)}(l = 1, 2)$ that controls the amount of penalty. One way to tune our estimators could be by minimizing a generalized cross-validation (GCV) statistic to choose the optimal $\{\lambda_i,\ i = 1, \dots, s_n\}$ which is equal to either $\{\lambda_j^{(1)},\ j = 1, \dots, p_n\}$ or $\{\lambda_k^{(2)},\ k = 1, \dots, q_n\}$. However, in real application, how to simultaneously select a total of s_n shrinkage parameters $\{\lambda_i,\ i = 1, \dots, s_n\}$ could be a bit challenging. So, to avoid the hardship of the many tuning parameters, we follow the idea of [11–13] and simplify the tuning parameters as

(i) $\lambda_{1j} = (\lambda_1/|\tilde{\boldsymbol{\theta}}_j^{(0)}|)$, $j = 1, \dots, p_n$;

(ii) $\lambda_{2k} = (\lambda_2/|\tilde{\boldsymbol{\beta}}_k^{(0)}|)$, $k = 1, \dots, q_n$;

where $\tilde{\boldsymbol{\theta}}_j^{(0)}$ and $\tilde{\boldsymbol{\beta}}_k^{(0)}$ are respectively the jth element and kth element of the unpenalized estimates $\tilde{\boldsymbol{\theta}}_j^{(0)}$ and $\tilde{\boldsymbol{\beta}}_k^{(0)}$. Consequently, the original s_n dimensional problem about λ_i becomes a binary problem about $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. For completeness, we now describe the details of the generalized cross-validation procedures. In the penalized least squares setup, the GCV statistic is given by

$$\mathrm{GCV}_{LS}(\boldsymbol{\lambda}) = \frac{\mathrm{RSS}(\boldsymbol{\lambda})/n}{\{1 - e(\boldsymbol{\lambda})/n\}^2}, \tag{9}$$

where RSS is the residual sum of squares and $e(\boldsymbol{\lambda})$ is the effective number of parameters given by $e(\boldsymbol{\lambda}) = tr[\mathbf{X}\{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \Sigma_\lambda(\widehat{\boldsymbol{\gamma}}(\boldsymbol{\lambda}))\}^{-1}\mathbf{X}^{\mathrm{T}}]$. Johnson and Peng [17] proposed a different cross-validation statistic using rank-based convex dispersion function. In this later scenario, the final regularization parameter is chosen such that

$$\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{Argmin}} \operatorname{GCV}(\boldsymbol{\lambda}).$$

But, since we are restricting ourselves to penalized weighted rank-based with adaptive lasso penalty, another alternative to estimating $\boldsymbol{\lambda}$ is to consider the AIC and BIC approaches discussed in [12] based on the considered objective function. We will consider only the BIC in this paper. That is, obtain $\widehat{\boldsymbol{\lambda}}$ as

$$\widehat{\boldsymbol{\lambda}} = \underset{\lambda}{\operatorname{Argmin}} \left\{ D(\tilde{\boldsymbol{\gamma}}) + df_{\boldsymbol{\lambda}} \times \frac{\log(n)}{n} \times C_n \right\},$$

where $0 \le df_{\boldsymbol{\lambda}} \le s_n$ is simply the number of non-zero coefficients of $\widehat{\boldsymbol{\gamma}}_n$ and $C_n \to \infty$ as $n \to \infty$. For example, C_n could be chosen as $\log\log s_n$.

4.3. Choice of weights

In our analysis, we choose the weight function $w(\mathbf{x})$ to be

$$w(\mathbf{x}) = \min\left[1, \frac{b}{(\mathbf{x} - \widehat{\mu})^{\mathrm{T}} S^{-1} (\mathbf{x} - \widehat{\mu})}\right],$$

with $(\widehat{\mu}, S)$ being the robust minimum volume ellipsoid estimators of the location and scatter, and b the 95th percentile of $\chi^2(p)$. Under this choice, it is shown in [24] that the resulting estimator has a bounded influence function.

5. Applications

In this section, we demonstrate the performance of our proposed method. To this end, several simulation scenarios could be considered. However, we will be investigating one of the scenario borrowed from [9]. A real data application using nutritional data is also investigated.

5.1. Simulation example

So we consider the Extra-Poisson model. That is conditionally on m_i, the response $y_i|m_i$ is Poisson with parameter m_i and m_i itself is $Gamma(\nu_i, \alpha)$, so that

$$E(y_i) = \mu_i = \nu_i, \quad \operatorname{Var}(y_i) = \nu_i\alpha_i + \nu_i\alpha_i^2 = \mu_i(1 + \alpha_i)$$

and the dispersion parameter is $\phi_i = 1 + \alpha_i$. We also decided to choose the structure of the mean model as $E(y_i) = \mu_i = \nu_i = \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}_{n0})$, where $\mathbf{x}_i$ is a $p_n \times 1$ vector with elements independently generated in two different scenarios:

Scenario 1: as $\mathbf{x} \sim (1 - \epsilon)\mathcal{N}(\mathbf{0}, V) + \epsilon\mathcal{N}(\mathbf{1}\mu, V)$ for several levels of contamination ϵ, where $V = (v_{ij})$ and $v_{ij} = 0.5^{|i-j|}$ and

Scenario 2: as $\mathbf{x} \sim (1-\epsilon)\mathcal{N}(\mathbf{0}, V) + \epsilon\mathcal{T}(\mathrm{d}f)$ for several levels of contamination ϵ, where $V = (v_{ij})$ and $v_{ij} = 0.5^{|i-j|}$ and $\mathcal{T}(\mathrm{d}f)$ represent the t-distribution with various degrees of freedom df. We will only investigate the case where d$f = 3$ and d$f = 5$. These two scenarios will certainly enable us to study the effect of contamination (such as gross outliers and leverage points) in the design space.

The regression coefficient vector is set at $\boldsymbol{\theta} = (2, 2, 2, , 0, 0, 0, \ldots)$, in which the first three components are non-zero. The structure of the dispersion parameter is $\phi_i = 1 + \alpha_i = \exp(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_{n0})$ with $\boldsymbol{\beta} = (1, 1, 0, 0, 0, \ldots)$, where the first two components are non-zero and $\mathbf{z}_i$ is a $q_n \times 1$ vector with elements generated in the same fashion as the $\mathbf{x}$'s. The sample sizes are chosen to be $n =$ 100, 200, 400, 600, with $p_n = [4n^{1/5}]$ and $q_n = [4n^{1/5}] - 5$ (Here $[\cdot]$ represent the ceiling function). It is worthwhile mentioning that the choice of p_n and q_n here is solely based on computation expenses. One could definitely consider the case where $p_n, q_n > n$. However, this choice still gives us the great fact that the size of the parameter growth with the sample size n. As in [9], we take $C_n = 1$ in our simulation setting. In all cases, we considered the robust penalized rank-based under the adaptive Lasso penalty

Table 1. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	7.8 (98.75%)	0.0	0.0250
			200	12	9.0 (100%)	0.0	0.0083
0%	R-AL		400	14	11.0 (100%)	0.0	0.0070
			600	15	12.0(100%)	0.0	0.0029
		$\boldsymbol{\gamma}_n$	100	6	3.8(95.00%)	0.0	0.0086
			200	7	5.0(100%)	0.0	0.0066
0%	R-AL		400	9	7.0(100%)	0.0	0.0040
			600	10	8.0(100%)	0.0	0.0029
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0262
			200	12	9.0(100%)	0.0	0.0085
0%	WR-AL		400	14	11.0(100%)	0.0	0.0071
			600	15	12.0(100%)	0.0	0.0031
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0080
			200	7	5.0(100%)	0.0	0.0064
0%	WR-AL		400	9	7.0(100%)	0.0	0.0040
			600	10	8.0(100%)	0.0	0.0025
		$\boldsymbol{\theta}_n$	100	11	7.7(96.25%)	0.0	2.1564
			200	12	8.9(98.89%)	0.0	0.7285
0%	MLE-AL		400	14	11.0(100%)	0.0	0.3251
			600	15	12.0(100%)	0.0	0.3612
		$\boldsymbol{\gamma}_n$	100	6	3.8(95.00%)	0.0	0.1616
			200	7	4.6(92.00%)	0.0	0.2944
0%	MLE-AL		400	9	7.0(100%)	0.0	0.2760
			600	10	8.0(100%)	0.0	0.2284
		$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0450
			200	12	9.0(100%)	0.0	0.0083
0%	WLAD-AL		400	14	11.0(100%)	0.0	0.0305
			600	15	12.0(100%)	0.0	0.0209
		$\boldsymbol{\gamma}_n$	100	6	3.8(95.00%)	0.0	0.016
			200	7	5.0(100%)	0.0	0.0126
0%	WLAD-AL		400	9	7.0(100%)	0.0	0.0031
			600	10	8.0(100%)	0.0	0.0022

Different sample sizes $n = 100, 200, 400$ and 500 are considered. The size of the parameters are p_n and q_n for the mean and the variance respectively. The design space was not contaminated. That is the design space was generated using scenario 1 with a 0% contamination.

where the unknown tuning parameters $\lambda^{(l)} (l = 1, 2)$ are computed using the BIC criterion. The estimators studied were maximum likelihood (MLE-AL), rank-based (R-AL) and weighted rank-based (WR-AL). The weights were computed as discussed in Section 4.3 using minimum covariance determinant (MCD).

We performed 1000 replications and calculated the average number of correct zeros which mean here the average number of zero regression coefficients that are correctly estimated as zeroes. The average number of incorrect zeros, which are the number of non-zero regression coefficients that are erroneously set to zero, and the percentage of correct models identified which represents here the proportion in which the method correctly identifies the correct zeroes. The average number of correct zeroes will be denoted as 'Correct' and the average of incorrect zeroes will be denoted as 'Incorrect'. The simulation results of Scenario 1 are given in Tables 2 and 3 while the results of Scenarios 2 are given in Tables 4–7, respectively. Contamination from 0% to 10% was considered. But for simplicity of our presentation, we only reported on the given tables, results when the design space has 0%, 5% and 10% contamination respectively. The proportion of contamination is represented on

Table 2. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	R-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0016
			200	7	5.0(100%)	0.0	0.0006
5%	R-AL		400	9	7.0(100%)	0.0	0.0030
			600	10	8.0(100%)	0.0	0.0001
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	WR-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0070
			200	7	5.0(100%)	0.0	0.0004
5%	WR-AL		400	9	7.0(100%)	0.0	0.0002
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	7.3(91.25%)	0.1	0.0720
			200	12	8.1(90.00%)	0.0	0.1406
5%	MLE-AL		400	14	10.2(92.27%)	0.0	0.0771
			600	15	11.1(92.50%)	0.0	0.0397
		$\boldsymbol{\gamma}_n$	100	6	3.2(80.00%)	0.0	0.4616
			200	7	4.5(90.00%)	0.1	0.3844
5%	MLE-AL		400	9	5.6(80.00%)	0.0	0.2660
			600	10	6.7(83.75%)	0.0	0.2184
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0020
			200	12	9.0(100%)	0.0	0.0014
5%	WLAD-AL		400	14	11.0(100%)	0.0	0.0012
			600	15	12.0(100%)	0.0	0.0010
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0086
			200	7	5.0(100%)	0.0	0.0071
5%	WLAD-AL		400	9	7.0(100%)	0.0	0.0034
			600	10	8.0(100%)	0.0	0.0002

Different sample sizes $n = 100$, 200, 400 and 500 are considered. The size of the parameters is p_n and q_n for the mean and the variance respectively. The design space was contaminated using 5% contamination from a normal distribution under scenario 1.

all tables by ‘prop x cont’. A summary of all other percentage of contamination (including the one presented on the tables) was plotted and presented as a supplementary material. Interested readers are encouraged to look at the supplemental material for more about our simulation results.

Table 1 shows that in the case of non-contamination in the design space (which correspond to the 0% contamination case) and when the sample size is set to 100, WR-AL performed better in detecting the true zeroes compared to the MLE-AL and R-AL in both mean and variance cases. But as the sample size increases, all methods perform fairly the same in detecting the true zeroes. A quick look on the results from Tables 2 through 7 suggests that the MLE-AL gets over-penalized as the proportion of high leverage points increases. While R-AL and WR-AL provide superior performance in high leverage situations. The percentage of correctly estimated models deteriorates with increasing design vector contamination for the MLE method while the rank-based and weighted rank-based stay constant and correctly estimate the model. Taken together, these indicate that the MLE-AL are not very good at identifying the true zeroes as the proportion of

Table 3. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
10%	R-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0000
			200	7	5.0(100%)	0.0	0.0000
10%	R-AL		400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
10%	WR-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0000
			200	7	5.0(100%)	0.0	0.0000
10%	WR-AL		400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	7.0(87.50%)	0.4	0.7257
			200	12	7.3(81.11%)	0.2	0.2094
10%	MLE-AL		400	14	9.6(87.27%)	0.3	0.1974
			600	15	10.1(84.17%)	0.1	0.3186
		$\boldsymbol{\gamma}_n$	100	6	3.2(80.00%)	0.2	0.2970
			200	7	3.7(74.00%)	0.2	0.1772
10%	MLE-AL		400	9	5.4(77.14%)	0.1	0.0525
			600	10	5.8(72.50%)	0.0	0.0596
		$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
10%	WLAD-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0003
			200	7	5.0(100%)	0.0	0.0001
10%	WLAD-AL		400	9	7.0(100%)	0.0	0.0001
			600	10	8.0(100%)	0.0	0.0001

Different sample sizes $n = 100, 200, 400$ and 500 are considered. The size of the parameters is p_n and q_n for the mean and the variance respectively. The design space was contaminated using 10% contamination from a normal distribution under scenario 1.

contamination in the design space increases. Moreover, when the contamination is from heavy-tail distribution such as t-distribution, even when the model is correctly specified, high leverage points have a detrimental effect on model selection. While the number of correct zeroes decreases for the MLE-AL, the weighted rank-based cases appear to provide good resistance for both low percentage (5% contamination) and high percentage (10% contamination) of high-leverage points. These results can be seen from Tables 4 to 7. Overall, Tables 4–7 suggest that as the proportion of **x** contamination increases, the average number of correct zeroes decreases in the case of MLE-AL, while the average number of incorrect zeroes increases. This shows that the presence of gross outliers and/or high leverage point has a damage effect on the performance of MLE-AL, while the R-AL and WR-AL are not influenced by those high leverage point, due to their constant behaviour in all cases. Moreover, the WR-AL has a significantly smaller mean squares error compared to the WLAD-AL specially when estimating the dispersion parameters. The results obtained throughout different scenarios in our simulation studies clearly suggest that both the R-AL and WR-AL outperform the MLE-AL when there is a presence of gross outliers and

Table 4. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	R-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0395
			200	7	5.0(100%)	0.0	0.0111
5%	R-AL		400	9	7.0(100%)	0.0	0.0086
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	WR-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0282
			200	7	5.0(100%)	0.0	0.0094
5%	WR-AL		400	9	7.0(100%)	0.0	0.0048
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	7.0(87.50%)	0.1	0.1953
			200	12	8.1(90.00%)	0.1	0.1910
5%	MLE-AL		400	14	10.2(92.72%)	0.0	0.0000
			600	15	10.6(85.00%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.4(85.00%)	0.3	2.5040
			200	7	4.0(80.00%)	0.0	0.8808
5%	MLE-AL		400	9	5.4(77.14%)	0.1	0.6061
			600	10	5.7(71.25%)	0.2	0.3222
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	WLAD-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	4.0(100%)	0.0	0.0795
			200	7	5.0(100%)	0.0	0.0551
5%	WLAD-AL		400	9	7.0(100%)	0.0	0.0098
			600	10	8.0(100%)	0.0	0.0001

Different sample sizes $n = 100$, 200, 400 and 500 are considered. The size of the parameters is p_n and q_n for the mean and the variance respectively. The design space was contaminated using 5% contamination from a t-distribution with df = 3 under scenario 2.

Table 5. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
10%	R-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0000
			200	7	5.0(100%)	0.0	0.0000
10%	R-AL		400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
10%	WR-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0000
			200	7	5.0(100%)	0.0	0.0000
10%	WR-AL		400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	7.0(87.50%)	0.4	0.3230
			200	12	7.5(83.33%)	0.1	0.3311
10%	MLE-AL		400	14	9.3(84.54%)	0.0	0.2216
			600	15	9.5(79.16%)	0.0	0.0176
		$\boldsymbol{\gamma}_n$	100	6	3.0(75.00%)	0.3	0.9886
			200	7	4.0(80.00%)	0.2	0.9129
10%	MLE-AL		400	9	4.5(64.28%)	0.5	0.4454
			600	10	5.5(68.75%)	0.2	0.2034
		$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0000
			200	12	9.0(100%)	0.1	0.0000
10%	WLAD-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.0(75.00%)	0.3	0.0006
			200	7	4.0(80.00%)	0.2	0.0002
10%	WLAD-AL		400	9	4.5(64.28%)	0.5	0.0004
			600	10	5.5(68.75%)	0.2	0.0001

Different sample sizes $n = 100$, 200, 400 and 500 are considered. The size of the parameters are p_n and q_n for the mean and the variance respectively. The design space was contaminated using 10% contamination from a t-distribution with df = 3 under scenario 2.

perform better than the WLAD-AL with the presence of gross outliers. Thus the robustness of our proposed method is compared to the other method such as MLE.

To assess the performance of estimators $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\beta}}_n$, we compute the generalized mean square error (GMSE), defined as

$$\mathrm{GMSE}(\widehat{\boldsymbol{\theta}}) = E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} E(\mathbf{X}\mathbf{X}^{\mathrm{T}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right]$$

and

$$\mathrm{GMSE}(\widehat{\boldsymbol{\beta}}) = E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}} E(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right],$$

using the MLE-AL, R-AL, and WR-AL estimators for both the mean and variance. Tables 1 through 7 contain a report of the results obtained from our simulations. It shows that the R-AL and the WR-AL have a relatively smaller mean squares error compared to the MLE-AL. We can therefore conclude that the R-AL and WR-AL method correctly identify relevant variables, remove irrelevant ones and estimate efficiently the corresponding coefficients as if the true model was known in advance.

Table 6. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	R-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0158
			200	7	4.9(98.75%)	0.0	0.0053
5%	R-AL		400	9	7.0(100%)	0.0	0.0022
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	WR-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0147
			200	7	5.0(100%)	0.0	0.0048
5%	WR-AL		400	9	7.0(100%)	0.0	0.0020
			600	10	8.0(100%)	0.0	0.0000
		$\boldsymbol{\theta}_n$	100	11	6.8(96.25%)	0.0	0.1953
			200	12	9.0(100%)	0.0	0.1910
5%	MLE-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.4(85.00%)	0.4	0.7112
			200	7	4.2(84.75%)	0.4	0.2992
5%	MLE-AL		400	9	5.0(71.75%)	0.1	0.8189
			600	10	6.2(77.50%)	0.1	0.0878
		$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
5%	WLAD-AL		400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
		$\boldsymbol{\gamma}_n$	100	6	3.4(85.00%)	0.4	0.2221
			200	7	4.2(84.75%)	0.4	0.1245
5%	WLAD-AL		400	9	5.0(71.75%)	0.1	0.0130
			600	10	6.2(77.50%)	0.1	0.0022

Different sample sizes $n = 100$, 200, 400 and 500 are considered. The size of the parameters is p_n and q_n for the mean and the variance respectively. The design space was contaminated using 5% contamination from a t-distribution with df = 5 under scenario 2.

5.2. Real data: plasma concentrations of beta-carotene

A dataset from the nutritional epidemiological study is considered in which there were 315 observations and 14 variables. Patients were the subjects of studies and they had an elective surgical procedure during a 3-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus. The related reference to this data set is [25]. The interest here is the relationship between plasma beta-carotene level and some other characteristics such as x_1 = calories (number of calories consumed per day), x_2 = quetelet, x_3 = fibre (grams of fibre consumed per day), x_4 = gender, x_5 = smokesta2 (smoking status [1 = formersmoker, 0 = neversmoked], x_6 = smokestat3 (smoking status [1 = currentsmoker, 0 = neversmoked]), x_7 = vituse (Vitamin2 = Yes, not often (82 (26.0%))) and x_8 = vituse (Vitamin Use (1 = Yes, fairly often (122 (38.7%))), x_9 = age, and x_{10} = cholesterol (cholesterol consumed mg per day). Contrary to [25], none of the high leverage point was removed from the data before analysing them.

Table 7. Average number of correct and incorrect zeroes, GMSE for both the mean $\boldsymbol{\theta}_n$ and variance $\boldsymbol{\gamma}_n$.

Prop. of x cont.	Method	Parameter	n	$p_n(q_n)$	Correct (%)	Incorrect	GMSE
10%	R-AL	$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
			400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
10%	R-AL	$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0008
			200	7	5.0(100%)	0.0	0.0000
			400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
10%	WR-AL	$\boldsymbol{\theta}_n$	100	11	8.0(100%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
			400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
10%	WR-AL	$\boldsymbol{\gamma}_n$	100	6	3.9(97.50%)	0.0	0.0005
			200	7	5.0(100%)	0.0	0.0000
			400	9	7.0(100%)	0.0	0.0000
			600	10	8.0(100%)	0.0	0.0000
10%	MLE-AL	$\boldsymbol{\theta}_n$	100	11	6.6(82.50%)	0.6	0.3230
			200	12	7.7(85.56%)	0.1	0.3311
			400	14	9.6(87.27%)	0.0	0.2216
			600	15	10.2(85.00%)	0.2	0.0176
10%	MLE-AL	$\boldsymbol{\gamma}_n$	100	6	2.8(70.00%)	0.7	0.4109
			200	7	3.3(66.75%)	0.4	0.4515
			400	9	4.8(68.57%)	0.2	0.3656
			600	10	5.4(72.50%)	0.1	0.2240
10%	WLAD-AL	$\boldsymbol{\theta}_n$	100	11	7.9(98.75%)	0.0	0.0000
			200	12	9.0(100%)	0.0	0.0000
			400	14	11.0(100%)	0.0	0.0000
			600	15	12.0(100%)	0.0	0.0000
10%	WLAD-AL	$\boldsymbol{\gamma}_n$	100	6	2.8(70.00%)	0.7	0.0073
			200	7	3.3(66.75%)	0.4	0.0060
			400	9	4.8(68.57%)	0.2	0.0002
			600	10	5.4(72.50%)	0.1	0.0001

Different sample sizes $n = 100, 200, 400$ and 500 are considered. The size of the parameters is p_n and q_n for the mean and the variance respectively. The design space was contaminated using 10% contamination from a t-distribution with df = 5 under scenario 2.

We first fit unpenalized regression models using the MLE and R procedures. Furthermore, to estimate the dispersion parameters using only the information provided by the covariates, we assume the variance function $V(\cdot) \equiv 1$ and apply the proposed variable selection procedure to the following model as proposed in [9]

$$\begin{cases} E(y_i) = \mu_i, \mathrm{Var}(y_i) = \phi_i V(\mu_i), \\ \log(\mu_i) = \sum_{j=1}^{10} x_{ij}\boldsymbol{\theta}_j, \\ \log(\phi_i) = \boldsymbol{\beta}_0 + \sum_{j=1}^{10} x_{ij}\boldsymbol{\beta}_j, \\ i = 1, 2, \ldots, 315. \end{cases}$$

We can see from Table 8 that both the R and MLE indicate that the variables x_1, x_9 and x_{10} are not important in predicting the mean plasma beta-carotene. On the other hand,

Table 8. MLE, R MLE-AL, R-AL and WR-AL for both the mean and the variance are computed using the nutritional epidemiological data. Here $\boldsymbol{\theta}$ represents the mean, $\boldsymbol{\beta}$ represents the variance and (mse) are the corresponding mean squares errors.

Coefficient	MLE (mse)	R (mse)	MLE-AL (mse)	R-AL (mse)	WR-AL (mse)
$\boldsymbol{\theta}_1$	0.000 (0.000)	0.0000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_2$	−0.032 (0.009)	−0.004 (0.000)	−0.030 (0.007)	−0.092 (0.003)	−0.128 (0.002)
$\boldsymbol{\theta}_3$	0.028 (0.012)	0.004 (0.008)	0.032 (0.002)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_4$	0.044 (0.186)	0.043 (0.093)	0.000 (0.000)	0.000(0.000)	0.000 (0.000)
$\boldsymbol{\theta}_5$	0.181 (0.184)	0.027 (0.001)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_6$	0.171 (0.183)	0.034 (0.015)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_7$	−0.105 (0.145)	−0.006 (0.074)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_8$	−0.297 (0.149)	−0.035 (0.002)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\theta}_9$	0.006 (0.004)	0.001 (0.001)	0.004 (0.022)	0.032 (0.001)	0.017 (0.001)
$\boldsymbol{\theta}_{10}$	−0.002 (0.000)	0.000 (0.000)	0.001(0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_1$	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_2$	0.026 (0.013)	0.093 (0.003)	0.000 (0.000)	0.000 (0.009)	0.000 (0.003)
$\boldsymbol{\beta}_3$	−0.043 (0.018)	0.032 (0.002)	0.000 (0.000)	0.001 (0.002)	0.000 (0.000)
$\boldsymbol{\beta}_4$	−0.185 (0.262)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_5$	−0.067 (0.260)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_6$	−0.247 (0.258)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_7$	0.449 (0.205)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_8$	0.028 (0.210)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\boldsymbol{\beta}_9$	0.005 (0.006)	0.032 (0.001)	0.000 (0.000)	0.002 (0.002)	0.000 (0.001)
$\boldsymbol{\beta}_{10}$	0.001 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

variables like x_2, x_3 and x_7 are mildly important in predicting the mean under the rank-based method. For the variance case, both methods have x_1 and x_{10} as non-important factors in predicting the variance. The rank-based shows that they are only three factors that are important in predicting the variance which are x_2, x_3 and x_9. However, the MLE-AL shows that x_2, x_3, x_9 and x_{10} (in which x_9 and x_{10} are mildly important) are important in the model, while R-AL and WR-AL show that only x_2 and x_9 are important to the model. This is in accordance with the many results using the nutritional epidemiology data. MLE-AL and WR-AL show that none of the factor are important in predicting the variance. However, R-AL shows x_2 and x_9 are mildly important in predicting the variance. It is important to notice that Table 8 shows that the mean squares errors (mse) of the penalized and unpenalized rank-based were consistently smaller compared to MLEs mean squares error in both the penalized and unpenalized cases.

6. Discussion

This paper considered variable selection for DGLMs in high dimension using penalized weighted rank-based technique. We only studied the case of adaptive Lasso penalty and the robustness in the design space. We demonstrated that the method still provides consistency and asymptotic normality even when we have the presence of gross outliers and/or high leverage points. The results obtained under simulations show clearly that the proposed method performs better compared to the MLE in the presence of high leverage points. That is, our proposed method is robust in the design space in estimation and selection, for both the mean and dispersion parameters. The results obtained certainly look promising. However, there are various extension of this result that could be considered. One is to consider the case where data are ultra-high dimensional; that is, the dimension of the predictor also goes to infinity. We anticipated that this extension does not look straightforward and require further efforts. The main challenging thing here will be to establish asymptotic properties of the methods and to give theoretical justifications. This is currently under investigation by the authors.

Acknowledgments

The authors are grateful to the associate editor and two anonymous referees for their constructive comments that substantially improved this paper. The first author gratefully acknowledges support from ONR (Office of Naval Research).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A General. 1972;135:370–384.

[2] Miakonkana GM, Abebe A. Iterative rank estimation for generalized linear models. J Stat Plann Inference. 2014;151/152:60–72.

[3] McCullagh P, Nelder JA. Generalized linear models, 2nd ed [of MR0727836]. London: Chapman & Hall; 1989. (Monographs on statistics and applied probability)

[4] Smyth GK. Generalized linear models with varying dispersion. J R Stat Soc Ser B. 1989;51(1):47–60.

[5] Wu L, Li H. Variable selection for joint mean and dispersion models of the inverse Gaussian distribution. Metrika. 2012;75(6):795–808.

[6] Rigby RA, Stasinopoulos MD. Mean and dispersion additive models. In: Statistical theory and computational aspects of smoothing. Springer: 1996. p. 215–230.

[7] Bergman B, Hynén A. Dispersion effects from unreplicated designs in the 2^{k-p} series. Technometrics. 1997;39(2):191–198.

[8] Nair VN, Pregibon D. Analyzing dispersion effects from replicated factorial experiments. Technometrics. 1988;30(3):247–257.

[9] Xu D, Zhang Z, Wu L. Variable selection in high-dimensional double generalized linear models. Stat Pap. 2014;55(2):327–347.

[10] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267–288.

[11] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.

[12] Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B Stat Methodol. 2009;71(3):671–683.

[13] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429.

[14] Zou H, Li R. Rejoinder: 'One-step sparse estimates in nonconcave penalized likelihood models'. Ann Stat. 2008;36(4):1561–1566.

[15] Chung M, Long Q, Johnson BA. A tutorial on rank-based coefficient estimation for censored data in small- and large-scale problems. Stat Comput. 2013;23(5):601–614.

[16] Johnson BA, Lin DY, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. J Am Stat Assoc. 2008;103(482):672–680.

[17] Johnson BA, Peng L. Rank-based variable selection. J Nonparametr Stat. 2008;20(3):241–252.

[18] Leng C. Variable selection and coefficient estimation via regularized rank regression. Stat Sin. 2010;20(1):167–181.

[19] Xu J, Leng C, Ying Z. Rank-based variable selection with censored data. Stat Comput. 2010;20(2):165–176.

[20] Wu L-C, Zhang Z-Z, Tian G-L, Xu D-K. A robust variable selection to t-type joint generalized linear models via penalized t-type pseudo-likelihood. Commun Stat Simul Comput. 2016;45(7):2320–2337.

[21] Hettmansperger TP, McKean JW. Robust nonparametric statistical methods, 2nd ed. Boca Raton, FL: CRC Press; 2011. (Monographs on statistics and applied probability; 119).

[22] Jaeckel LA. Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Stat. 1972;43(5):1449–1458.

[23] Terpstra JT, McKean JW, Naranjo JD. Weighted Wilcoxon estimates for autoregression. Aust N Z J Stat. 2001;43(4):399–419.

[24] Bindele HF, Abebe A. Bounded influence nonlinear signed-rank regression. Canad J Statist. 2012;40(1):172–189.

[25] Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. Am J Epidemiol. 1989;130(3):511.

Appendix. Preliminaries

It is well known from [22] that the rank dispersion function has a quadratic approximation. We explore an extension of that approximation in the context of this paper. For any $\sqrt{n/s_n}$ neighbourhood of $\boldsymbol{\gamma}_{n0}$, the rank dispersion function has a quadratic approximation given by, for any constant $M > 0$,

$$\sup_{\|\boldsymbol{\gamma}_{n0}-\boldsymbol{\gamma}_n\|\leq M\cdot(n/s_n)^{-1/2}} \Big| Q_n(\boldsymbol{\gamma}_n) - Q_n(\boldsymbol{\gamma}_{n0}) + (n/s_n)^{-1/2}\Psi_n(\boldsymbol{\gamma}_{n0})^{\mathrm{T}}\left\{(n/s_n)^{-1/2}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n0})\right\}$$
$$- \frac{1}{2}\left\{(n/s_n)^{-1/2}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n0})\right\}^{\mathrm{T}} \mathbf{G}\left\{(n/s_n)^{-1/2}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n0})\right\}\Big| \rightarrow_p 0,$$

Where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \gamma_{s_n})^{\mathrm{T}} = (\theta_1, \dots, \theta_{p_n}; \beta_1, \dots, \beta_{q_n})^{\mathrm{T}}$ and $\boldsymbol{\gamma}_{n0} = (\gamma_{n01}, \dots, \gamma_{n0s_n})^{\mathrm{T}} = (\theta_{n01}, \dots, \theta_{n0p_n}; \beta_{n01}, \dots, \beta_{n0q_n})^{\mathrm{T}}$.

We also have the usual asymptotic linearity for rank statistics

$$\sup_{\|\boldsymbol{\gamma}_{n0}-\boldsymbol{\gamma}_n\|\leq M\cdot(n/s_n)^{-1/2}} \Big\| (n/s_n)^{-1/2}\Psi_n(\boldsymbol{\gamma}_n) - (n/s_n)^{-1/2}\Psi_n(\boldsymbol{\gamma}_{n0})$$
$$+ \mathbf{G}\left\{(n/s_n)^{-1/2}(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_{n0})\right\}\Big\| \rightarrow_p 0.$$

In addition

$$(n/s_n)^{-1/2}\Psi_n(\boldsymbol{\gamma}_{n0}) \rightarrow_d \mathcal{N}\left(0, \kappa\mathbf{G}^{-1}\right),$$

where $\kappa = [\int_0^1 \{\phi(u) - \int_0^1 \phi(v)\}^2 \,\mathrm{d}v]^{1/2}$. In the special case of the Wilcoxon score (i.e. $\phi(u) = (u - 1/2)$), it is easy to check that $\kappa = 12^{-1/2}$, $\tau = [\int_{-\infty}^{\infty} \{f(u)\}^2 \,\mathrm{d}u]^{-1}$.

Proof of Theorem 3.1.: Let $\alpha_n = \sqrt{p_n + q_n}(n^{-1/2} + a_n)$. Using the continuity of $Q_n(\boldsymbol{\gamma}_n)$, it suffices to show that, for a given $\epsilon > 0$, there exist a large enough constant C such that, for a large n, we have

$$Pr\left\{\inf_{\|\mathbf{u}\|=\mathbf{C}} \left(D_n(\boldsymbol{\gamma}_{n0} + \alpha_n\mathbf{u}) - D_n(\boldsymbol{\gamma}_{n0})\right) > 0\right\} \geq 1 - \epsilon.$$

This implies the existence of a local minimizer in the C-ball around $\boldsymbol{\gamma}_{n0}$ and thus completes the proof of Theorem 3.1. We note that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) > 0$. Moreover, if we denote $\Gamma_n(\mathbf{u})$ by the quantity, $\Gamma_n(\mathbf{u}) = D_n(\boldsymbol{\gamma}_{n0} + \alpha_n\mathbf{u}) - D_n(\boldsymbol{\gamma}_{n0})$, we have that

$$\Gamma_n(\mathbf{u}) = \left[Q_n(\boldsymbol{\gamma}_{n0} + \alpha_n\mathbf{u}) - n\sum_{j=1}^{s_n} p_{\lambda_j}(|\gamma_{n0j} + \alpha_n u_j|)\right] - \left[Q_n(\boldsymbol{\gamma}_{n0}) - n\sum_{j=1}^{s_n} p_{\lambda_j}(|\gamma_{n0j}|)\right]$$

$$= \left[Q_n(\boldsymbol{\gamma}_{n0} + \alpha_n\mathbf{u}) - Q_n(\boldsymbol{\gamma}_{n0})\right] - n\sum_{j=1}^{s_n}\left[p_{\lambda_j}(|\gamma_{n0j} + \alpha_n u_j|) - p_{\lambda_j}(|\gamma_{n0j}|)\right].$$

Using the above approximation, we have that around a neighbourhood of $\boldsymbol{\gamma}_{n0}$,

$$\Gamma_n(\mathbf{u}) \geq -\alpha_n\Psi_n(\boldsymbol{\gamma}_{n0})^{\mathrm{T}}\mathbf{u} + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\dot{\Psi}_n(\boldsymbol{\gamma}_{n0})\mathbf{u}\alpha_n^2 + n\sum_{j=1}^{s_{1n}}\left[p_{\lambda_j}(|\gamma_{n0j} + \alpha_n u_j|) - p_{\lambda_j}(|\gamma_{n0j}|)\right]$$

$$= \Gamma_1 + \Gamma_2 + \Gamma_3.$$

And by using Taylor expansion, on the penalty terms we have that around a neighbourhood of $\boldsymbol{\gamma}_{n0}$,

$$\Gamma_3 = -\sum_{j=1}^{s_{1n}}\left[n\alpha_n q_{\lambda_j}(|\gamma_{n0j}|)\mathrm{sgn}(\gamma_{n0j})u_j + \frac{n\alpha_n^2}{2}\dot{q}_{\lambda_j}(|\gamma_{n0j}|)u_j^2\{1 + o(1)\}\right] = I_1 + I_2.$$

We note from the above approximation that $\|\Psi_n(\boldsymbol{\gamma}_{n0})\| = O_p(\sqrt{n(p_n + q_n)})$. So applying the Cauchy Schwarz inequality, on $\Gamma_1 = -\alpha_n\Psi_n(\boldsymbol{\gamma}_{n0})^{\mathrm{T}}\mathbf{u}$, we obtain

$$|\Gamma_1| = |\alpha_n\Psi_n(\boldsymbol{\gamma}_{n0})^{\mathrm{T}}\mathbf{u}| \leq \alpha_n\|\Psi_n(\boldsymbol{\gamma}_{n0})\|\|\mathbf{u}\|^2 = O_p(\alpha_n\sqrt{n(p_n + q_n)})\|\mathbf{u}\| = O_p(\alpha_n^2 n)\|\mathbf{u}\|^2.$$

For term I_1 and I_2 it holds that

$$|I_1| \leq \sum_{j=1}^{s_{1n}}\left|n\alpha_n q_{\lambda_j}(|\gamma_{n0j}|)\mathrm{sgn}(\gamma_{n0j})u_j\right| \leq \sqrt{s_{1n}}\cdot n\alpha_n a_n\|\mathbf{u}\|^2 \leq n\alpha_n^2\|\mathbf{u}\|^2$$

and

$$I_2 = \sum_{j=1}^{s_{1n}}\frac{n\alpha_n^2}{2}\dot{q}_{\lambda_j}(|\gamma_{n0j}|)u_j^2\{1 + o(1)\} \leq \max_{1\leq j\leq s_{1n}}\dot{q}_{\lambda_j}(|\gamma_{n0j}|)\cdot n\alpha_n^2\|\mathbf{u}\|^2.$$

Therefore, we have that

$$\Gamma(\mathbf{u}) \geq -O_p(n\alpha_n^2)\|\mathbf{u}\|^2 + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\dot{\Psi}_n(\boldsymbol{\gamma}_{n0})\mathbf{u}\alpha_n^2 \geq -O_p(n\alpha_n^2)\|\mathbf{u}\|^2 + \frac{1}{2}n\alpha_n^2\delta_0\|\mathbf{u}\|^2,$$

where δ_0 is the smallest eigenvalue of $\dot{\Psi}_n(\boldsymbol{\gamma}_{n0})$. Since $\dot{\Psi}_n(\boldsymbol{\gamma}_{n0})$ is positive definite, $\delta_0 > 0$. Therefore $\Pr\{\inf_{\|\mathbf{u}\|=C}(D_n(\boldsymbol{\gamma}_{n0} + \alpha_n\mathbf{u}) - D_n(\boldsymbol{\gamma}_{n0})) > 0\} \geq 1 - \epsilon$ holds by choosing a sufficiently large C. The proof of Theorem 3.1 is completed. ■

Proof of Theorem 3.2.: First, we prove that under the conditions of Theorem 3.2, it is sufficient to show that with probability tending to 1, for any $C > 0$, for any given $\boldsymbol{\gamma}_n^{(1)}$ satisfying $\|\boldsymbol{\gamma}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}\| = O_p((n/s_n)^{-(1/2)})$ and $\gamma_{nj}^{(1)} \in (-C(n/s_n)^{-(1/2)}, C(n/s_n)^{-(1/2)})$ with $j = s_{1n}, \ldots, s_n$,

$$\left.\frac{\partial D_n(\boldsymbol{\gamma}_n)}{\partial\gamma_{nj}}\right|_{\boldsymbol{\gamma}_n=\boldsymbol{\gamma}_n^{(1)}} \quad \text{and} \quad \boldsymbol{\gamma}_n^{(1)} \quad \text{have the same sign.}$$

In fact we have that

$$\frac{\partial D_n(\boldsymbol{\gamma}_n)}{\partial\gamma_{nj}} = \frac{\partial Q_n(\boldsymbol{\gamma}_n)}{\partial\gamma_{nj}} - nq_{\lambda_j}(|\gamma_{nj}|)\mathrm{sgn}(\gamma_{nj}) = \Psi_{nj}(\boldsymbol{\gamma}_n) - nq_{\lambda_j}(|\gamma_{nj}|)\mathrm{sgn}(\gamma_{nj}),$$

where $(\partial Q_n(\boldsymbol{\gamma}_n)/\partial \gamma_{nj}) = \Psi_{nj}(\boldsymbol{\gamma}_n)$. By the asymptotic linearity of $\Psi_n(\boldsymbol{\gamma}_n)$, we have that

$$\frac{\partial D_n(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} = \Psi_{nj}(\boldsymbol{\gamma}_{n0}) + \mathbf{G}_{(j)}\{(\boldsymbol{\gamma}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)})\} - nq_{\lambda_j}(|\gamma_{nj}|)\text{sgn}(\gamma_{nj}),$$

where $\mathbf{G}_{(j)}$ denotes the jth row of $\mathbf{G}$. But we have that $\Psi_{nj}(\boldsymbol{\gamma}_{n0}) = O_p((n/s_n)^{-(1/2)})$ and $\|\boldsymbol{\gamma}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}\| = O_p((n/s_n)^{-(1/2)})$, thus

$$\frac{\partial D_n(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} = n\lambda_n \left\{ -\lambda_n^{-1} q_{\lambda_j}(|\gamma_{nj}|)\text{sgn}(\gamma_{nj}) + O_p\left(\sqrt{\frac{p_n + q_n}{n}} / \lambda_n \right) \right\}$$

and

$$\liminf_{n\to\infty} \liminf_{\boldsymbol{\gamma}_n \to 0^+} \frac{q_{\lambda_n}(\boldsymbol{\gamma}_n)}{\lambda_n} > 0, \quad \sqrt{\frac{p_n + q_n}{n}} / \lambda_n \to 0,$$

it is easy to see that as n is large enough the sign of $\boldsymbol{\gamma}_n^{(1)}$ completely determines the sign of $(\partial D_n(\boldsymbol{\gamma}_n)/\partial \gamma_{nj})|_{\boldsymbol{\gamma}_n = \boldsymbol{\gamma}_n^{(1)}}$. That is

$$\left. \frac{\partial D_n(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} \right|_{\boldsymbol{\gamma}_n = \boldsymbol{\gamma}_n^{(1)}} = \begin{cases} < 0, & \text{for } 0 < \boldsymbol{\gamma}_{nj}^{(1)} < C(n/s_n)^{-\frac{1}{2}}, \\ > 0, & \text{for } -C(n/s_n)^{-\frac{1}{2}} < \boldsymbol{\gamma}_{nj}^{(1)} < 0. \end{cases}$$

By the definition of $\widehat{\boldsymbol{\gamma}}_n$, we see from previous arguments that $D_n(\boldsymbol{\gamma}_n)$ achieves its minimum at $\boldsymbol{\gamma}_n = ((\boldsymbol{\gamma}_n^{(1)})^{\mathrm{T}}, 0^{\mathrm{T}})$ and the first part of Theorem 3.2 has been proved.

The second part of Theorem 3.2 discusses the asymptotic normality of $\widehat{\boldsymbol{\gamma}}_n^{(1)}$. From Theorem 3.1 and the first part of Theorem 3.2, there exist a rank-based penalized estimator $\widehat{\boldsymbol{\gamma}}_n^{(1)}$ that is the $\sqrt{n/s_n}$-consistent local minimizer of the penalized rank-based dispersion function $D_n(\boldsymbol{\gamma}_n)$.

If we can show that

$$\left(\dot{\Psi}_{n(1)} + \nabla \mathbf{q}_{\lambda_n}\left(\boldsymbol{\gamma}_{n01}^{(1)}\right)\right)\left(\widehat{\boldsymbol{\gamma}}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}\right) + \mathbf{q}_{\lambda_n}\left(\boldsymbol{\gamma}_{n01}\right) = \frac{1}{n}\Psi_n\left(\boldsymbol{\gamma}_{n01}^{(1)}\right) + o_p\left(\frac{1}{\sqrt{n/s_n}}\right).$$

In fact by the definition of $\widehat{\boldsymbol{\gamma}}_n$, we see from previous arguments that $(\partial D_n(\boldsymbol{\gamma}_n)/\partial \gamma_{nj})|_{\boldsymbol{\gamma}_n = (\boldsymbol{\gamma}_n^{(1)\mathrm{T}}, 0^T)^T} = o_p((1/\sqrt{n}))$. It follows from the asymptotic normality of $\Psi_n(\boldsymbol{\gamma}_n)$ that

$$o_p\left(\frac{1}{\sqrt{n}}\right) = \Psi_n\left(\boldsymbol{\gamma}_{n01}^{(1)}\right) - \dot{\Psi}_{n(1)}\left(\widehat{\boldsymbol{\gamma}}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}\right) - q_{\lambda_n}(|\widehat{\boldsymbol{\gamma}}_n^{(1)}|)\text{sgn}(\widehat{\boldsymbol{\gamma}}_n^{(1)}).$$

After the Taylor series expansion of the last term on the right-hand side of the above equality, we conclude that

$$(\dot{\Psi}_{n(1)} + \Sigma_\lambda^{11})\left\{\left(\widehat{\boldsymbol{\gamma}}_n^{(1)} - \boldsymbol{\gamma}_{n0}^{(1)}\right) + (\dot{\Psi}_{n(1)} + \Sigma_\lambda^{11})^{-1}\mathbf{q}_{\lambda_n}\left(\boldsymbol{\gamma}_{n01}\right)\right\} = \begin{pmatrix} \Psi_{n1} \\ \Psi_{n2} \\ \vdots \\ \Psi_{ns_{1n}} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n/s_n}}\right),$$

where $\mathbf{q}_{\lambda_n}(\boldsymbol{\gamma}_{n01}) = (q_{\lambda_n}(|\gamma_{n01}|)\text{sgn}(\gamma_{n01}), \ldots, q_{\lambda_n}(|\gamma_{n0s_{1n}}|)\text{sgn}(\gamma_{n0s_{1n}}))^{\mathrm{T}}$ and Σ_λ^{11} is the first $s_{1n} \times s_{1n}$ submatrices of $\Sigma_\lambda(\boldsymbol{\gamma}_n)$. Clearly we can have that

$$\begin{pmatrix} \Psi_{n1} \\ \Psi_{n2} \\ \vdots \\ \Psi_{ns_{1n}} \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n/s_n}}\right) \xrightarrow{D} \mathcal{N}(0, \mathbf{G}),$$

where $\mathbf{G}$ is the first $s_{1n} \times s_{1n}$ submatrices of $\dot{\Psi}_n$. Hence, this completes the proof of Theorem 3.2. ■