Statistica Sinica **24** (2014), 653-673
doi:http://dx.doi.org/10.5705/ss.2012.022

ROBUST-BD ESTIMATION AND INFERENCE FOR VARYING-DIMENSIONAL GENERAL LINEAR MODELS

Chunming Zhang, Xiao Guo, Chen Cheng and Zhengjun Zhang

University of Wisconsin-Madison

Abstract: This paper investigates new aspects of robust inference for general linear models, calling for a broader array of error measures, beyond the conventional notion of quasi-likelihood, and allowing for a diverging number of parameters. We propose a class of robust error measures, called robust-BD, based on the notion of Bregman divergence (BD). That includes the (negative) quasi-likelihood and many other commonly used error measures as special cases, and we introduce the robust-BD estimators of parameters. We re-examine the classical likelihood ratio-type test statistic, constructed by replacing the negative log-likelihood with the robust-BD, and find that its asymptotic null distribution is a sum of weighted χ^2 with weights relying on unknown quantities, thus is not asymptotically distribution free. We propose a robust version of the Wald-type test statistic, based on the robust-BD estimator, and show that it is asymptotically χ^2 (central) under the null, thus distribution free, and χ^2 (noncentral) under the contiguous alternatives. Numerical examples are presented to illustrate the computational simplicity and effectiveness of the proposed estimator and test in the presence of outliers.

Key words and phrases: Generalized linear model, hypothesis test, influence function, quasi-likelihood, robustness.

1. Introduction

Robust inference for generalized linear models (GLM) plays an important role in statistical applications (McCullagh and Nelder (1989)); refer to Mebane and Sekhon (2004) for some interesting examples. The existing research on robust inference for GLM has some limitations. Robust inference is developed mainly for the logistic regression model, based on the deviance loss as the error measure. See Hauck and Donner (1977), Bianco and Yohai (1996), Croux and Haesbroeck (2003), Bianco and Martínez (2009), and references therein. The use of quasi-likelihood for GLM was studied in Cantoni and Ronchetti (2001). Adimari and Ventura (2001) devised a distribution free test statistic based on a suitable scale adjustment of the robust quasi-likelihood. However, the quasi-likelihood is not applicable to the exponential loss function (to be defined at the start of Section 3), commonly used in the machine learning and data mining literature. Further,

most of the research efforts focus on the finite-dimensional setting with the number p of parameters fixed or low. There is little work, either theoretically or empirically, when the dimension p can vary with the sample size n. Hence, existing results on robust inference are not directly applicable to situations calling for a broader array of error measures, beyond quasi-likelihood, in the presence of a diverging number of parameters.

As shown in Zhang, Jiang, and Shang (2009), the (negative) quasi-likelihood in regression, the deviance loss, the exponential loss in machine learning practice, and many other commonly used error measures belong to the class of Bregman divergence (BD). It is thus natural to develop the robust inference based on BD. We investigate new aspects of the robust inference for general linear models, described by (2.1)−(2.2), integrating

Case I : the error measure belongs to the class of BD,
Case II : the dimension p_n is relaxed to be either fixed or varying with n, where $p_n < n$.

By broadening the scope of robust estimation, the hope is to gain new insights into robust inference with applications to large-dimensional datasets.

Four issues are addressed for the varying-dimensional general linear model.

- We propose the robust version of BD, called robust-BD, and introduce a class of robust-BD estimators; see Section 3.
- We re-examine a classical likelihood ratio-type test statistic Λ_n, constructed by replacing the negative log-likelihood with the robust-BD. Theorem 3 finds the asymptotic null distribution of Λ_n is generally not χ^2, but a sum of weighted χ^2, with weights relying on unknown quantities, and holds under restrictive conditions. Even in the particular case of using classical (non-robust) BD, the limit distribution is not invariant with re-scaling of the generating function of the BD. Moreover, the limit null distribution of Λ_n (in either the non-robust or robust version) using the exponential loss, which does not belong to the (negative) quasi-likelihood but falls in BD, is always a weighted χ^2, thus limiting its use in applications. See Section 4.
- We propose a robust version of the Wald-type test statistic W_n, based on the robust-BD estimator, and its validity is justified in Theorems 4−6. It is asymptotically χ^2 (central) under the null, thus distribution free, and χ^2 (noncentral) under the contiguous alternatives. This result, when applied to the exponential loss as well as other loss functions in the wider class of BD, is practically feasible. See Section 4. Furthermore, it has computational advantages over Λ_n.

- We devise a robust-BD classifier based on the proposed robust-BD estimator and establish its classification consistency.

Simulation studies indicate that the proposed class of robust-BD estimators is either comparable or superior to the classical non-robust counterpart: the former is less sensitive to outliers than the latter, and they perform comparably in non-contaminated cases. The computational simplicity of the proposed estimator and detection effectiveness of the proposed test are illustrated through a dataset. The Appendix, for technical details, and Figures 7−10 in Section 6.2 are available on the online supplement.

2. Overview of Existing Methods

We start with a brief overview of robust inference for the general linear models. Let $\{(\boldsymbol{X}_{n1}, Y_1), \ldots, (\boldsymbol{X}_{nn}, Y_n)\}$ be independent observations from some underlying population, $(\boldsymbol{X}_n, Y)$, where $\boldsymbol{X}_n = (X_1, \ldots, X_{p_n})^T \in \mathbb{R}^{p_n}$ is the vector of explanatory variables and Y is the scalar response variable. We assume the general linear model

$$m(\boldsymbol{x}_n) = E(Y \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = F^{-1}(\beta_{0;0} + \boldsymbol{x}_n^T \boldsymbol{\beta}_0), \tag{2.1}$$

$$\operatorname{var}(Y \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = V(m(\boldsymbol{x}_n)), \tag{2.2}$$

where F is a known link function, $\beta_{0;0} \in \mathbb{R}^1$ and $\boldsymbol{\beta}_0 = (\beta_{1;0}, \ldots, \beta_{p_n;0})^T \in \mathbb{R}^{p_n}$ are the unknown true regression parameters, and the functional form of $V(\cdot)$ is known. It is worth noting that (2.1)–(2.2) include the GLM as a special case. Moreover, they allow the conditional distribution of $Y \mid \boldsymbol{X}_n$ to be incompletely (or partially) specified. For simplicity, we use $\widetilde{\boldsymbol{x}}_n = (1, x_1, \ldots, x_{p_n})^T$ and $\widetilde{\boldsymbol{\beta}} = (\beta_0, \beta_1, \ldots, \beta_{p_n})^T$.

2.1. Classical quasi-likelihood estimation

The classical quasi-likelihood estimator of the true parameter $\widetilde{\boldsymbol{\beta}}_0$ is

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{\text{QL}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \frac{1}{n} \sum_{i=1}^{n} \{-\mathtt{Q}_{\text{QL}}(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T \widetilde{\boldsymbol{\beta}}))\}, \tag{2.3}$$

where the classical quasi-likelihood function $\mathtt{Q}_{\text{QL}}(y, \mu)$ satisfies

$$\frac{\partial \mathtt{Q}_{\text{QL}}(y, \mu)}{\partial \mu} = \frac{y - \mu}{V(\mu)} = r(y, \mu) \times \frac{1}{\sqrt{V(\mu)}}, \tag{2.4}$$

with $r(y, \mu) = (y - \mu)/\sqrt{V(\mu)}$ denoting the Pearson residual. In general, $\mathtt{Q}_{\text{QL}}(y, \mu)$ can be recovered as $\mathtt{Q}_{\text{QL}}(y, \mu) = \int_y^\mu r(y, s)/\sqrt{V(s)} ds$.

Computationally, the quasi-likelihood estimator corresponding to (2.3) can be obtained by solving the estimating equation,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_{\mathrm{QL}}(\boldsymbol{X}_{ni}, Y_i; \widetilde{\boldsymbol{\beta}}) = \mathbf{0}, \tag{2.5}$$

where the score vector satisfies

$$\boldsymbol{\psi}_{\mathrm{QL}}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}) = \frac{\partial\{-\mathtt{Q}_{\mathrm{QL}}(y,\mu)\}}{\partial\widetilde{\boldsymbol{\beta}}} = r(y,\mu)\frac{-1}{\sqrt{V(\mu)}F'(\mu)}\widetilde{\boldsymbol{x}}, \tag{2.6}$$

with $\mu = F^{-1}(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}})$.

2.2. Robust quasi-likelihood estimation and inference

It is well-known that the classical maximum likelihood and quasi-likelihood estimators can be severely affected by outlying observations. Cantoni and Ronchetti (2001) (abbreviated as CR hereafter) formulated the robust quasi-likelihood estimator of $\widetilde{\boldsymbol{\beta}}_0$ as

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{\mathrm{RQL}} = \arg\min_{\widetilde{\boldsymbol{\beta}}}\frac{1}{n}\sum_{i=1}^{n}\{-\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{X}_{ni}, Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}))\}, \tag{2.7}$$

where the robust quasi-likelihood function $\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{x}, y, \mu)$ is

$$\begin{aligned}&\Big\{\int_{\mu_0}^{\mu}\psi(r(y,s))\frac{1}{\sqrt{V(s)}}ds\Big\}w(\boldsymbol{x})\\ &\quad-\frac{1}{n}\sum_{j=1}^{n}\int_{\mu_0}^{\mu_j}\Big[E\{\psi(r(Y_j,s))\mid\boldsymbol{X}_{nj}\}\frac{1}{\sqrt{V(s)}}ds\Big]w(\boldsymbol{X}_{nj}).\end{aligned} \tag{2.8}$$

Here $\mu_j = \mu_j(\widetilde{\boldsymbol{\beta}}) = F^{-1}(\widetilde{\boldsymbol{X}}_{nj}^T\widetilde{\boldsymbol{\beta}})$, $j = 1,\ldots,n$, where $\psi(r)$ is chosen to be a bounded, odd function, such as the Huber ψ-function (Huber (1964)), and $w(\cdot)\geq 0$ is a known bounded weight function that downweights high leverage points in the covariate space. It is easy to observe that

$$\frac{\partial\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{x}, y, \mu)}{\partial\mu} = \psi(r(y,\mu))\frac{1}{\sqrt{V(\mu)}}w(\boldsymbol{x}).$$

The estimating equation corresponding to (2.7) is given by

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_{\mathrm{RQL}}(\boldsymbol{X}_{ni}, Y_i; \widetilde{\boldsymbol{\beta}}) = \mathbf{0}, \tag{2.9}$$

with the score vector defined as

$$\boldsymbol{\psi}_{\mathrm{RQL}}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}) = \frac{\partial\{-\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{x}, y, \mu)\}}{\partial\widetilde{\boldsymbol{\beta}}} = \psi(r(y,\mu))\frac{-1}{\sqrt{V(\mu)}F'(\mu)}w(\boldsymbol{x})\widetilde{\boldsymbol{x}} - \alpha(\widetilde{\boldsymbol{\beta}}), \tag{2.10}$$

where

$$\alpha(\widetilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{j=1}^{n} E\{\psi(r(Y_j,\mu_j)) \mid \boldsymbol{X}_{nj}\}\frac{-1}{\sqrt{V(\mu_j)}F'(\mu_j)} w(\boldsymbol{X}_{nj})\widetilde{\boldsymbol{X}}_{nj}.$$

Thus $\boldsymbol{\psi}_{\mathrm{RQL}}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}})$ depends on the data observations $\{(\boldsymbol{X}_{ni}, Y_i)\}_{i=1}^{n}$ through $\alpha(\widetilde{\boldsymbol{\beta}})$.

Remark 1. The estimating equation (2.9) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}[\psi(r(Y_i,\mu_i)) - E\{\psi(r(Y_i,\mu_i)) \mid \boldsymbol{X}_{ni}\}]\frac{-1}{\sqrt{V(\mu_i)}F'(\mu_i)} w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni} = \mathbf{0}.$$

Remark 2. If $\psi(r) = r$ and $w(\boldsymbol{x}) \equiv 1$, then $\alpha(\widetilde{\boldsymbol{\beta}}) = \mathbf{0}$ and $\boldsymbol{\psi}_{\mathrm{RQL}}$ in (2.10) reduces to $\boldsymbol{\psi}_{\mathrm{QL}}$ in (2.6).

3. Robust Estimation Based on BD

We consider a class of error measures motivated by Bregman divergence. For a given concave q-function, Brègman (1967) introduced a device for constructing a bivariate function,

$$Q_q(\nu,\mu) = -q(\nu) + q(\mu) + (\nu-\mu)q'(\mu). \tag{3.1}$$

We call q the generating q-function of the BD. For example, $q(\mu) = a\mu - \mu^2$ for some constant a yields the quadratic loss $Q_q(Y,\mu) = (Y-\mu)^2$; for a binary response variable Y, $q(\mu) = \min\{\mu, (1-\mu)\}$ gives the misclassification loss $Q_q(Y,\mu) = \mathrm{I}\{Y \neq \mathrm{I}(\mu > 1/2)\}$; $q(\mu) = -2\{\mu\log(\mu) + (1-\mu)\log(1-\mu)\}$ gives the Bernoulli deviance loss $Q_q(Y,\mu) = -2\{Y\log(\mu) + (1-Y)\log(1-\mu)\}$; $q(\mu) = 2\min\{\mu, (1-\mu)\}$ results in the hinge loss $Q_q(Y,\mu) = \max\{1 - (2Y-1)\mathrm{sign}(\mu - 0.5), 0\}$ of the support vector machine; $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$ yields the exponential loss $Q_q(Y,\mu) = \exp[-(Y-0.5)\log\{\mu/(1-\mu)\}]$ used in AdaBoost (Hastie, Tibshirani, and Friedman (2001)). Moreover, Zhang, Jiang, and Shang (2009) showed that if

$$q(\mu) = \int_a^{\mu} \frac{s-\mu}{V(s)}ds, \tag{3.2}$$

where a is a finite constant such that the integral is well-defined, then $Q_q(y,\mu)$ gives the (negative) quasi-likelihood function $-\mathtt{Q}_{\mathrm{QL}}(y,\mu)$ in (2.4).

Now we can see clearly that the exponential loss $Q_q(y,\mu)$ yields $\frac{\partial\{-Q_q(y,\mu)\}}{\partial\mu}$ $= (y-\mu)/\{2\sqrt{V(\mu)}V(\mu)\}$, not in a form proportional to the right side of (2.4), thus is not a (negative) quasi-likelihood, but belongs to the class of BD.

3.1. Proposed robust-BD $\rho_q(y,\mu)$

In contrast to the BD, denoted by Q_q in (3.1), we propose the robust-BD

$$\rho_q(y,\mu) = \int_y^\mu \psi(r(y,s))\{q''(s)\sqrt{V(s)}\}ds - G(\mu), \tag{3.3}$$

where the bias-correction term, $G(\mu)$, entails the Fisher consistency of the parameter estimator (to be defined in (3.5)) and satisfies

$$G'(\mu) = G_1'(\mu)\{q''(\mu)\sqrt{V(\mu)}\},$$

with

$$G_1'(m(\boldsymbol{x})) = E\{\psi(r(Y, m(\boldsymbol{x}))) \mid \boldsymbol{X} = \boldsymbol{x}\}. \tag{3.4}$$

The following diagram illustrates the relation among the robust-BD, BD, and (negative) quasi-likelihood.

robust-BD		BD		(negative) quasi-likelihood
$\rho_q(y,\mu)$	$\xrightarrow{\psi(r)=r}$	$Q_q(y,\mu)$	$\xrightarrow{q \text{ in (3.2)}}$	$-\mathtt{Q}_{\mathrm{QL}}(y,\mu)$

3.2. Proposed robust-BD estimator

The robust-BD estimator $\widehat{\widetilde{\boldsymbol{\beta}}}$ of $\widetilde{\boldsymbol{\beta}}_0$ is defined as

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \Big\{ \frac{1}{n}\sum_{i=1}^n \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}))w(\boldsymbol{X}_{ni}) \Big\}. \tag{3.5}$$

It is easy to see that if $q(\mu) = \mu(1-\mu)$, $V(\mu) \equiv \sigma^2$, and $\{Y - m(\boldsymbol{X}_n)\} \mid \boldsymbol{X}_n$ is symmetrically distributed, then $\rho_q(y,\mu)$ reduces to the Huber loss function. As another example, for the q-function given in (3.2), $\rho_q(y,\mu)w(\boldsymbol{x})$ is equivalent to the robust quasi-likelihood in (2.8), up to an additive constant. Computationally, $\widehat{\widetilde{\boldsymbol{\beta}}}$ in (3.5) can be obtained by modifying the "coordinate-ascent updating algorithm" (see for e.g., Bickel and Doksum (2007)).

If the quantities

$$\mathrm{p}_j(y;\theta) = \frac{\partial^j}{\partial\theta^j}\rho_q(y, F^{-1}(\theta)), \qquad j = 0, 1, \ldots, \tag{3.6}$$

exist finitely up to any order required, then we have

$$\begin{aligned}
\mathrm{p}_1(y;\theta) &= \{\psi(r(y,\mu)) - G_1'(\mu)\}\frac{\{q''(\mu)\sqrt{V(\mu)}\}}{F'(\mu)},\\
\mathrm{p}_2(y;\theta) &= A_0(y,\mu) + \{\psi(r(y,\mu)) - G_1'(\mu)\}A_1(\mu),\\
\mathrm{p}_3(y;\theta) &= A_2(y,\mu) + \{\psi(r(y,\mu)) - G_1'(\mu)\}\frac{A_1'(\mu)}{F'(\mu)},
\end{aligned} \tag{3.7}$$

where $\mu = F^{-1}(\theta)$,

$$A_0(y,\mu) = -\Big[\psi'(r(y,\mu))\Big\{1 + \frac{y-\mu}{\sqrt{V(\mu)}} \times \frac{V'(\mu)}{2\sqrt{V(\mu)}}\Big\} + G_1''(\mu)\sqrt{V(\mu)}\Big]\frac{q''(\mu)}{\{F'(\mu)\}^2},$$

$$A_1(\mu) = \frac{\{q^{(3)}(\mu)\sqrt{V(\mu)} + 2^{-1}q''(\mu)V'(\mu)/\sqrt{V(\mu)}\}F'(\mu) - q''(\mu)\sqrt{V(\mu)}F''(\mu)}{\{F'(\mu)\}^3},$$

and

$$A_2(y,\mu) = \frac{\partial A_0(y,\mu)/\partial\mu + \partial\{\psi(r(y,\mu)) - G_1'(\mu)\}/\partial\mu\, A_1(\mu)}{F'(\mu)}.$$

The estimating equation corresponding to (3.5) is

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}_{\mathrm{RBD}}(\boldsymbol{X}_{ni}, Y_i; \widetilde{\boldsymbol{\beta}}) = \mathbf{0}, \tag{3.8}$$

and the score vector is

$$\boldsymbol{\psi}_{\mathrm{RBD}}(\boldsymbol{x}, y; \widetilde{\boldsymbol{\beta}}) = \frac{\partial\{\rho_{_q}(y,\mu)w(\boldsymbol{x})\}}{\partial\widetilde{\boldsymbol{\beta}}} = \mathrm{p}_1(y;\theta)w(\boldsymbol{x})\widetilde{\boldsymbol{x}}, \tag{3.9}$$

with $\theta = \widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}$.

The estimator $\widehat{\widetilde{\boldsymbol{\beta}}}$ is characterized by the score function and influence function,

$$\boldsymbol{\psi}_{\rho_{_q}}(Y, \boldsymbol{X}_n) = \mathrm{p}_1(Y; \widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)w(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n, \tag{3.10}$$

$$\mathrm{IF}(Y, \boldsymbol{X}_n; \boldsymbol{\psi}_{\rho_{_q}}) = \{M(\boldsymbol{\psi}_{\rho_{_q}})\}^{-1}\boldsymbol{\psi}_{\rho_{_q}}(Y, \boldsymbol{X}_n), \tag{3.11}$$

where $M(\boldsymbol{\psi}_{\rho_{_q}}) = -E[\partial\boldsymbol{\psi}_{\rho_{_q}}(Y, \boldsymbol{X}_n)/\partial\widetilde{\boldsymbol{\beta}}_0] = -E\{\mathrm{p}_2(Y; \widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)w(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T\}$; see Hampel (1974) and Hampel et al. (1986).

Remark 3. The format of the robust error measure, $\rho_{_q}(y,\mu)w(\boldsymbol{x})$, in (3.5), when applied to the generating q-function given in (3.2) for the quasi-likelihood, is

$$-\Big[\int_y^\mu \psi(r(y,s))\frac{1}{\sqrt{V(s)}}ds - \int_y^\mu E\{\psi(r(Y,s)) \mid \boldsymbol{X}\}\frac{1}{\sqrt{V(s)}}ds\Big]w(\boldsymbol{x}),$$

but is not identical to that of $-\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{x}, y, \mu)$, in (2.8); likewise, the score vectors (3.9) and (2.10) are different. Nonetheless, the estimating equations (3.8) and (2.9) coincide. This agreement can be verified from Remark 1, via a straightforward derivation.

The relationship among the criterion functions used in the three types of parameter estimation is summarized in Table 1.

Table 1. Relationship among the criterion functions associated with robust-BD, robust quasi-likelihood, and classical quasi-likelihood estimations.

criterion function	estimating equation	score vector
$\rho_q(y,\mu)w(\boldsymbol{x})$ in (3.5) $\Downarrow q$ in (3.2)	(3.8)	$\boldsymbol{\psi}_{\mathrm{RBD}}(\boldsymbol{x},y;\widetilde{\boldsymbol{\beta}})$ in (3.9)
$-\mathtt{Q}_{\mathrm{RQL}}(\boldsymbol{x},y,\mu)$ in (2.7) $\Downarrow \psi(r)=r,\ w(\boldsymbol{x})\equiv 1$	(2.9)	$\boldsymbol{\psi}_{\mathrm{RQL}}(\boldsymbol{x},y;\widetilde{\boldsymbol{\beta}})$ in (2.10)
$-\mathtt{Q}_{\mathrm{QL}}(y,\mu)$ in (2.3)	(2.5)	$\boldsymbol{\psi}_{\mathrm{QL}}(\boldsymbol{x},y;\widetilde{\boldsymbol{\beta}})$ in (2.6)

3.3. Asymptotic properties of the robust-BD estimator

The asymptotic distribution of $\widehat{\widetilde{\boldsymbol{\beta}}}$ involves two square matrices of size (p_n+1),

$$\Omega_n = E\{\mathrm{p}_1^2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)w^2(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T\},$$
$$\mathbf{H}_n = E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)w(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T\}.$$

Theorem 1 guarantees the existence of a $\sqrt{n/p_n}$-consistent minimizer of (3.5), with the dimension p_n allowed to be fixed or varying with n.

Theorem 1. *Assume* A0, A1, A2, A4, A5, A6, *and* A7 *in the Appendix. If* $p_n^4/n \to 0$ *as* $n\to\infty$, *then there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\ell_n(\widetilde{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^n \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}))w(\boldsymbol{X}_{ni})$ *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_0\| = O_P(\sqrt{p_n/n})$.

The proof of Theorem 1 relies on the positive definiteness of $\mathbf{H}_n = E[E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\}w(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T]$. We discuss conditions under which $E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\}\geq 0$ (and >0).

- The sign of $E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\}$ depends on the choice of BD only through $q''(\mu)$ (which is ≤ 0), thus is invariant with the choice of generating q-functions of BD.
- A sufficient condition for $E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\}\geq 0$ is that the conditional distribution of $Y\mid \boldsymbol{X}_n$ is symmetric about $m(\boldsymbol{X}_n)$.
- A sufficient condition for $E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\}\geq 0$ is that $E[\psi(r(Y,m(\boldsymbol{X}_n)))\frac{\partial}{\partial m(\boldsymbol{X}_n)}\log\{f(Y\mid \boldsymbol{X}_n, m(\boldsymbol{X}_n))\}\mid \boldsymbol{X}_n]\geq 0$, which holds when $\psi(r)r\geq 0$, and the conditional distribution of $Y\mid \boldsymbol{X}_n$ belongs to the exponential family, where f denotes the conditional density of $Y\mid \boldsymbol{X}_n$.
- If $\psi(r)=r$, a direct computation gives that $E\{\mathrm{p}_2(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_0)\mid \boldsymbol{X}_n\} = -q''(m(\boldsymbol{X}_n))/\{F'(m(\boldsymbol{X}_n))\}^2\geq 0$, for any conditional distribution of $Y\mid \boldsymbol{X}_n$.

Theorem 2. *Assume* A0, A1, A2, A4, A5, B5, A6, *and* A7 *in the Appendix. If* $p_n^5/n\to 0$ *as* $n\to\infty$, *then any* $\sqrt{n/p_n}$*-consistent minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *satisfies: for any*

fixed integer k *and any* $\mathsf{k} \times (p_n + 1)$ *matrix* A_n *such that* $A_n A_n^T \to \mathbb{G}$ *with* $\mathbb{G}$ *a* $\mathsf{k} \times \mathsf{k}$ *nonnegative-definite matrix,* $\sqrt{n} A_n \Omega_n^{-1/2} \mathbf{H}_n (\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbb{G})$.

It is then clear that for the robust-BD estimators, the choice of a bounded score function ensures robustness by putting a bound on the influence function. As observed from (3.10)−(3.11), a bounded function $\mathrm{p}_1(y;\theta)$ is introduced from a bounded function $\psi(r)$ to control large deviations in the Y-space, and high leverage points are down-weighted by the weight function $w(\boldsymbol{X}_n)$.

4. Robust Inference Based on BD

In many applications, we wish to test whether a subset of explanatory variables used is statistically significant. Specific examples include

$$H_0 : \beta_{j;0} = 0, \text{ for } j = j_0, \tag{4.1}$$

$$H_0 : \beta_{j;0} = 0, \text{ for } j = j_1, \ldots, j_2. \tag{4.2}$$

These hypotheses can be more generally formulated as

$$H_0 : A_n \widetilde{\boldsymbol{\beta}}_0 = \boldsymbol{g}_0 \longleftrightarrow H_1 : A_n \widetilde{\boldsymbol{\beta}}_0 \neq \boldsymbol{g}_0, \tag{4.3}$$

where A_n is a given $\mathsf{k} \times (p_n + 1)$ matrix such that $A_n A_n^T = \mathbb{G}$ with $\mathbb{G}$ a $\mathsf{k} \times \mathsf{k}$ positive-definite matrix, and $\boldsymbol{g}_0$ a known $\mathsf{k} \times 1$ vector. The hypothesis studied in the CR paper corresponds to the choice $A_n = [\mathbf{0}_{\mathsf{k}, p_n + 1 - \mathsf{k}}, \mathbf{I}_{\mathsf{k}}]$ with $A_n A_n^T = \mathbf{I}_{\mathsf{k}}$, where $p_n = p$, and $\boldsymbol{g}_0 = \mathbf{0}$.

4.1. A re-examination of the likelihood ratio-type test

It is well-known that the likelihood ratio-type test statistic, based on the maximum likelihood estimation, is asymptotically χ^2 under the null. When p_n is fixed, Heritier and Ronchetti (1994) developed robust versions of the likelihood ratio, Wald, and score tests, and their asymptotic distributions under the null and the alternative hypotheses.

This section explores the extent to which the likelihood ratio-type test can feasibly be extended to the robust-BD in the presence of a diverging number p_n of parameters. In such setting, the robust-BD test statistic takes the form

$$\Lambda_n = 2n \Big\{ \min_{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p_n+1} : A_n \widetilde{\boldsymbol{\beta}} = \boldsymbol{g}_0} \ell_n(\widetilde{\boldsymbol{\beta}}) - \min_{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p_n+1}} \ell_n(\widetilde{\boldsymbol{\beta}}) \Big\},$$

where $\ell_n(\widetilde{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^n \rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T \widetilde{\boldsymbol{\beta}})) w(\boldsymbol{X}_{ni})$ is the criterion function in (3.5) to be minimized. Clearly, when ℓ_n is replaced by the negative log-likelihood, Λ_n is the classical likelihood ratio statistic. Likewise, when ℓ_n is replaced by the

negative robust quasi-likelihood, Λ_n reduces to a test statistic which, while not identical, is asymptotically equivalent to that of the CR paper.

We require a convexity condition on the robust-BD:

$$\mathrm{p}_2(y;\theta) > 0 \text{ for all } \theta \in \mathbb{R} \text{ and all } y \text{ in the range of } Y. \tag{4.4}$$

Under this assumption, $\ell_n(\widetilde{\boldsymbol{\beta}})$ is strictly convex in $\widetilde{\boldsymbol{\beta}}$, and thus the minimizer of (3.5) is globally unique.

Theorem 3. *Assume* (4.4) *and* A0, A1, A2, C4, A5, B5, A6, A7, *and* D5 *in the Appendix.*

(i) *If* $p_n^5/n \to 0$ *as* $n \to \infty$, *then under* H_0 *in* (4.3),

$$\Lambda_n = nL_n^T \mathbf{H}_n^{-1/2} P_{\mathbf{H}_n^{-1/2} A_n^T} \mathbf{H}_n^{-1/2} L_n + o_P(1),$$

where $\sqrt{n}L_n \sim N(\mathbf{0}, \Omega_n)$, *and* $P_X = X(X^TX)^{-1}X^T$ *is defined for a matrix* X *such that* $(X^TX)^{-1}$ *exists.*

(ii) *If* $\psi(r) = r$ *and the generating* q*-function of* BD *satisfies*

$$q''(m(\boldsymbol{x}))w(\boldsymbol{x}) = -\frac{C}{V(m(\boldsymbol{x}))}, \quad \text{for a constant } C > 0, \tag{4.5}$$

then under H_0 *in* (4.3), *we have* $\Lambda_n/C \xrightarrow{\mathcal{L}} \chi^2_k$ *for any* $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\widetilde{\boldsymbol{\beta}}_0$.

Here, from part (i), Λ_n is not asymptotically distribution free. From part (ii), the restriction (4.5) on the q-function limits the application domain of Λ_n. For the quasi-likelihood function associated with the q-function in (3.2), (4.5) holds with $C = 1$ and $\Lambda_n \xrightarrow{\mathcal{L}} \chi^2_k$; in other cases, C is operationally a nuisance parameter. Further, the asymptotic distribution of Λ_n is not invariant with re-scaling the q-function. In the particular case of binary responses, the Bernoulli deviance loss satisfies (4.5), but the quadratic and exponential losses violate (4.5). These limitations reflect that, under the general framework of BD, the likelihood ratio-type test statistic Λ_n may not be valid.

4.2. Proposed Wald-type test based on the robust-BD

We propose a robust version of the Wald-type test statistic based on the robust-BD estimator as,

$$W_n = n(A_n\widehat{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{g}_0)^T (A_n \widehat{\mathbf{H}}_n^{-1} \widehat{\Omega}_n \widehat{\mathbf{H}}_n^{-1} A_n^T)^{-1} (A_n \widehat{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{g}_0),$$

where $\widehat{\widetilde{\boldsymbol{\beta}}}$ is the proposed robust-BD estimator of $\widetilde{\boldsymbol{\beta}}_0$, and

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1^2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}})w^2(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T,$$
$$\widehat{\mathbf{H}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T.$$

Other types of estimates $\widehat{\Omega}_n$ and $\widehat{\mathbf{H}}_n$ can also work provided that $A_n(\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1} - \mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1})A_n^T \xrightarrow{P} \mathbf{0}$. We find under the null, that W_n is asymptotically distribution-free.

Theorem 4. *Assume* A0, A1, A2, C4, A5, B5, A6, *and* A7 *in the Appendix. If* $p_n^5/n \to 0$ *as* $n \to \infty$, *then under* H_0 *in* (4.3), $W_n \xrightarrow{\mathcal{L}} \chi_k^2$ *for any* $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\widetilde{\boldsymbol{\beta}}_0$.

We assess the discriminating power of the test based on W_n.

Theorem 5. *Assume* A0, A1, A2, A5, B5, A6, *and* A7 *in the Appendix, and* $A_n\mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}A_n^T \xrightarrow{P} \mathbf{M}$ *where* $\mathbf{M}$ *is a* $k \times k$ *positive definite matrix. If* $p_n^5/n \to 0$ *as* $n \to \infty$, *then under the fixed alternative* H_1 *in* (4.3) *with* $\|A_n\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{g}_0\|$ *independent of* n,

$$n^{-1}W_n \geq \lambda_{\max}^{-1}(\mathbf{M})\|A_n\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{g}_0\|^2 + o_P(1)$$

for any $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\widetilde{\boldsymbol{\beta}}_0$.

Consider a sequence of contiguous alternatives of the form,

$$H_{1n} : A_n\widetilde{\boldsymbol{\beta}}_0 - \boldsymbol{g}_0 = \delta_n\boldsymbol{c}\{1 + o(1)\}, \tag{4.6}$$

where $\delta_n = n^{-1/2}$ and $\boldsymbol{c} = (c_1, \ldots, c_k)^T \neq \mathbf{0}$ is fixed. We explore the local power of W_n for detecting contiguous alternatives.

Theorem 6. *Assume* A0, A1, A2, A5, B5, A6, *and* A7 *in the Appendix, and* $A_n\mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}A_n^T \xrightarrow{P} \mathbf{M}$ *where* $\mathbf{M}$ *is a* $k \times k$ *positive definite matrix. If* $p_n^5/n \to 0$ *as* $n \to \infty$, *then under* H_{1n} *in* (4.6), $W_n \xrightarrow{\mathcal{L}} \chi_k^2(\tau^2)$ *for any* $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\widetilde{\boldsymbol{\beta}}_0$, *with the noncentrality parameter* $\tau^2 = \boldsymbol{c}^T\mathbf{M}^{-1}\boldsymbol{c}$.

4.3. Advantages of W_n over Λ_n

The test based on W_n offers some obvious advantages over the test based on Λ_n: W_n is asymptotically distribution-free, while Λ_n is not; W_n removes the

convexity condition (4.4) required for Λ_n, and W_n is invariant under re-scaling of the generating q-function of the BD. The computational cost of W_n is much reduced from that of Λ_n: integration operations are involved in Λ_n but not in W_n; Λ_n requires both unrestricted and restricted parameter estimates, while W_n is useful when restricted parameter estimates are difficult to compute. Numerical studies in Section 6 will focus on W_n.

5. Classification Consistency

For a binary response variable Y, the mean regression function $m(\boldsymbol{x}_n)$ in (2.1) is the class probability $P(Y = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}_n)$. From the robust-BD estimator $(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}^T)^T$ of Section 3.2, we construct the robust-BD classifier,

$$\widehat{\phi}_n(\boldsymbol{x}_n) = \mathrm{I}\{\widehat{m}(\boldsymbol{x}_n) > 1/2\},$$

for a future input $\boldsymbol{x}_n$, where $\mathrm{I}(\cdot)$ is an indicator function and $\widehat{m}(\boldsymbol{x}_n) = F^{-1}(\widehat{\beta}_0 + \boldsymbol{x}_n^T\widehat{\boldsymbol{\beta}})$. Details on binary classification can be found in Devroye, Györfi, and Lugosi (1996).

To emphasize the dependence of the dimension p_n on n in our current setting, the optimal Bayes rule is written as $\phi_{n,\mathrm{B}}(\boldsymbol{x}_n) = \mathrm{I}\{m(\boldsymbol{x}_n) > 1/2\}$. For a test sample $(\boldsymbol{X}_n^o, Y^o)$, which is an i.i.d. copy of samples in the training set $\mathcal{T}_n = \{(\boldsymbol{X}_{ni}, Y_{ni}), i = 1, \ldots, n\}$, the optimal Bayes risk is $R(\phi_{n,\mathrm{B}}) = P\{\phi_{n,\mathrm{B}}(\boldsymbol{X}_n^o) \neq Y^o\}$ and the conditional risk of the robust-BD classification rule $\widehat{\phi}_n$ is $R(\widehat{\phi}_n) = P\{\widehat{\phi}_n(\boldsymbol{X}_n^o) \neq Y^o \mid \mathcal{T}_n\}$. For $\widehat{\phi}_n$ induced by the robust-BD regression estimation using a range of loss functions, we have classification consistency preserved by $\widehat{\phi}_n$.

Theorem 7. *Assume* A1 *and* A4 *in the Appendix. If* $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\| = O_P(r_n)$ *and* $r_n\sqrt{p_n} = o(1)$*, then the classification rule* $\widehat{\phi}_n$ *constructed from* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *satisfies* $E\{R(\widehat{\phi}_n)\} - R(\phi_{n,\mathrm{B}}) \to 0$ *as* $n \to \infty$.

6. Simulation Study

We conducted simulation studies to evaluate the performance of the robust-BD estimator and the robust Wald-type test statistic W_n in the absence and presence of outliers. The robust-BD estimation utilized the Huber ψ-function $\psi(\cdot)$ with $c = 1.345$. For count responses, the weight function was of the form $w(\boldsymbol{x}) = 1/\{1 + \sum_{j=1}^{p_n}(\frac{x_j - m_j}{s_j})^2\}^{1/2}$, where $\boldsymbol{x} = (x_1, \ldots, x_{p_n})^T$, m_j denotes the median of $\{X_{i,j} : i = 1, \ldots, n\}$ and s_j denotes the median absolute deviation from the median respectively, $j = 1, \ldots, p_n$. See Maronna, Martin, and Yohai (2006) for robust estimates of location and scale. This form of $w(\boldsymbol{x})$ is a generalization of a weight function used in Boente, He, and Zhou (2006) for a one-dimensional

covariate. For binary responses, the weight function was $w(\boldsymbol{x}) = 1/\{1 + (\boldsymbol{x} - \widehat{\boldsymbol{m}})^T\widehat{\Sigma}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{m}})\}^{1/2}$, where $\widehat{\boldsymbol{m}}$ and $\widehat{\Sigma}$ denote the robust estimates of the location vector and scatter matrix of $\boldsymbol{X}_n$, implemented using the fast S-estimator with default set-ups from the function "`CovSest.R`" in the R package "`rrcov`". Other options can be found in Heritier et al. (2009). Comparisons are made with the classical non-robust counterparts with $\psi(r) = r$ and $w(\boldsymbol{x}) \equiv 1$.

6.1. Overdispersed Poisson responses

We generated overdispersed Poisson counts Y_i with $\text{var}(Y_i \mid \boldsymbol{X}_{ni} = \boldsymbol{x}_i) = 2\,m(\boldsymbol{x}_i)$, via a negative Binomial$(m(\boldsymbol{X}_{ni}), 1/2)$ distribution. In the predictor $\boldsymbol{X}_{ni} = (X_{i,1}, X_{i,2}, \ldots, X_{i,p_n})^T$, $X_{i,1} \sim$ Uniform$(-0.5, 0.5)$; for $j = 2, \ldots, p_n$, $X_{i,j} = \Phi(Z_{i,j}) - 0.5$, where Φ is the standard normal distribution function, and $(Z_{i,2}, \ldots, Z_{i,p_n})^T \sim N(\boldsymbol{0}, \Sigma_{p_n-1})$, with $\Sigma_{p_n-1}(j,k) = 0.2^{|j-k|}$, for $j, k = 1, \ldots, p_n - 1$ and $p_n = [6.5(n^{1/5.5} - 1)]$, where $[x]$ is the largest integer that is less than or equal to x. The link function was $\log\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T\boldsymbol{\beta}_0$ with $\beta_{0;0} = 2.5$ and $\boldsymbol{\beta}_0 = (2, 2, 0, \ldots, 0)^T$. The (negative) quasi-likelihood was utilized as the BD, generated by the q-function in (3.2) with $V(\mu) = \mu$, and (3.4) for the robust estimator was calculated assuming the Poisson$(m(\boldsymbol{X}))$ distribution. The sample size n was 400, and 500 replications were conducted.

For each data set generated from the model, we created a contaminated data set, where 20 data points $(X_{i,1}, \ldots, X_{i,p_n}, Y_i)$ were subject to contamination. (i) In the first 10 points, Y_i was replaced by $Y_i^* = Y_i\,\text{I}(Y_i > 100) + 100\,\text{I}(Y_i \le 100)$, and $X_{1,1}$, $X_{2,2}$, $X_{3,3}$, $X_{4,5}$, $X_{5,7}$, $X_{6,8}$, $X_{7,9}$ by

$$X_{1,1}^* = .5\,\text{sign}(U_1 - 0.5),\ X_{2,2}^* = 3\,\text{sign}(U_2 - 0.5),\ X_{3,3}^* = 3\,\text{sign}(U_3 - 0.5),$$
$$X_{4,5}^* = 3\,\text{sign}(U_4 - 0.5),\ X_{5,7}^* = 3\,\text{sign}(U_5 - 0.5),\ X_{6,8}^* = 3\,\text{sign}(U_6 - 0.5),$$
$$X_{7,9}^* = 3\,\text{sign}(U_7 - 0.5),$$

respectively, with $\{U_i\} \overset{\text{i.i.d.}}{\sim}$ Uniform$(0, 1)$; (ii) $X_{i,5}$ was replaced in the next 10 points by $X_{i,5}^* = X_{i,5} + e_i$, where $\{e_i\}$ were independent $N(0, 0.02^2)$ variables.

Part 1: Parameter estimation. Figure 1 compares the boxplots of $\widehat{\beta}_j - \beta_{j;0}$, $j = 0, 1, \ldots, p_n$, using the non-robust and robust quasi-likelihood estimates. The non-robust estimates are more sensitive to outliers than the robust counterparts. Simulation results (omitted for lack of space) also indicate that the robust estimator performs as well as the non-robust estimator for non-contaminated cases.

Part 2: Null distribution of the robust test statistic W_n**.** Figure 2 displays the QQ plots of the (1st to 99th) percentiles of the robust test statistic W_n versus those of the χ^2_{k} distribution for the null hypothesis (4.1) with $j_0 = 6$, and the null hypothesis (4.2) with $j_1 = 9$ and $j_2 = 12$. It is clear that the asymptotic χ^2

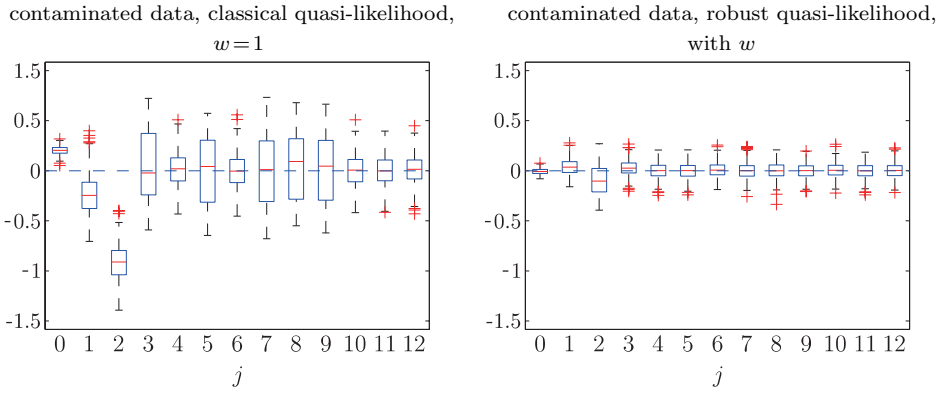

Figure 1. (Simulated overdispersed Poisson response data with contamination) Boxplots of $\widehat{\beta}_j - \beta_{j;0}$, $j = 0, 1, \ldots, p_n$ (from left to right). Left panel: the non-robust estimates; right panel: the robust estimates.

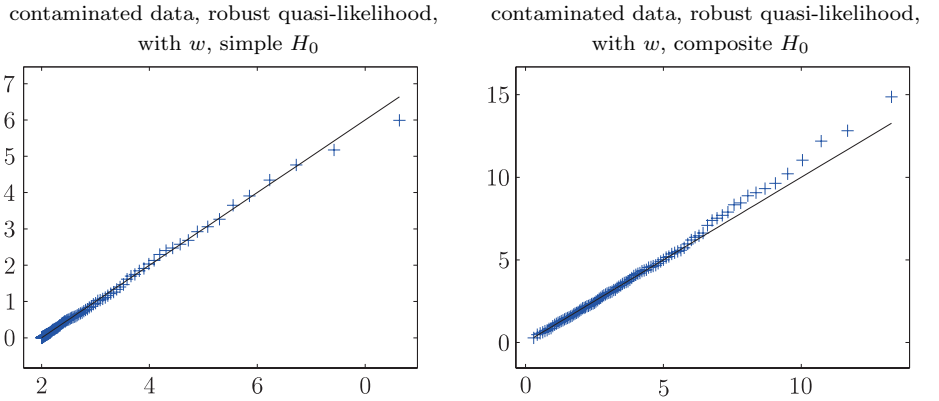

Figure 2. (Simulated overdispersed Poisson response data with contamination) Empirical quantiles (on the y-axis) of the robust test statistics W_n versus quantiles (on the x-axis) of the χ^2 distribution. Solid line: the 45 degree reference line. Left panel: for testing (4.1); right panel: for testing (4.2).

distribution well-approximates the finite sampling null distribution of W_n, and that the test for the composite null models can be made as precise as the test for the simple null.

Part 3: Level of tests. To investigate the stability of the level of the robust test, we replaced the previous 20 data points by $(X_{i,j}^{**}, Y_i^{**})$, with $X_{i,j}^{**} = (1-\theta)X_{i,j} + \theta X_{i,j}^{*}$ and $Y_i^{**} = [(1-\theta)Y_i + \theta Y_i^{*}]$. As θ varies from 0 to 1, each sample ranges from the non-contaminated case to the contaminated case. Figure 3 shows the observed rates of rejecting H_0. We observe that the asymptotic nominal level 0.05 is approximately retained by the robust Wald-type test. On the other hand,

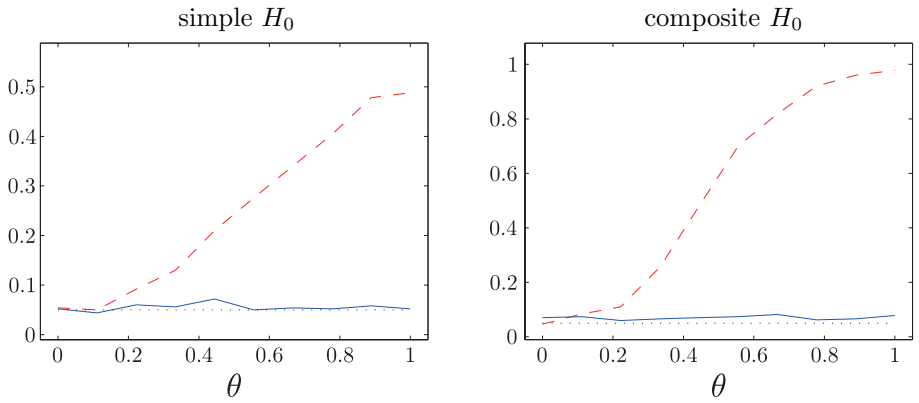

Figure 3. Level of tests for the overdispersed Poisson response data. The dashed line corresponds to the non-robust Wald-type test; the solid line corresponds to the robust Wald-type test; the dotted line indicates the 5% nominal level. Left panel: for testing (4.1); right panel: for testing (4.2).

under contamination, the non-robust Wald-type test breaks in level, showing high sensitivity to the presence of outliers.

Part 4: Power of tests. To assess the stability of the power of the tests, we generated the original data from the true model, but with the true parameter $\widetilde{\boldsymbol{\beta}}_0$ replaced by $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_0 + \Delta \boldsymbol{c}$, with $\boldsymbol{c} = (1, \ldots, 1)^T$ a vector of ones. Figure 4 plots the empirical rejection rates of the null model in the non-contaminated and contaminated cases. The price to pay for the robust Wald-type tests is a little loss of power in the non-contaminated cases. Under contamination, the observed power function of the robust test is close to that achieved in the non-contaminated case, while the non-robust test is less informative, since the power function is not higher than that of the robust test under the alternative hypotheses with $\Delta \neq 0$, but higher than the nominal level under the null hypotheses with $\Delta = 0$.

6.2. Bernoulli responses

We generated data with two classes from the model, $Y \mid \boldsymbol{X}_n = \boldsymbol{x} \sim$ Bernoulli $\{m(\boldsymbol{x})\}$, where $\text{logit}\{m(\boldsymbol{x})\} = \beta_{0;0} + \boldsymbol{x}^T \boldsymbol{\beta}_0$ with $\beta_{0;0} = 0$ and $\boldsymbol{\beta}_0 = (2, 2, 0, \ldots, 0)^T$. The predictor $\boldsymbol{X}_n$ was $N(\boldsymbol{0}, \Sigma_{p_n})$. For illustrative purposes, we set $\Sigma_{p_n} = \mathbf{I}_{p_n}$, and $p_n = 2$. For each data set generated from the model, we created a contaminated data set with 45 contaminated points. (i) For the original first 5 data points $(X_{i,1}, X_{i,2}, Y_i)$, we replaced $X_{i,1}$ by $X^*_{i,1} = 2 + i/2$, and Y_i by $Y^*_i = 0$, misclassified observations on a hyperplane parallel to the true discriminating hyperplane, the orthogonal distance between the two hyperplanes being $d = \sqrt{2}$; (ii) we replaced $X_{i,1}$ in the next 19 points with $X^*_{i,1} = X_{i,1} + e_i$, where $\{e_i\}$ were independent $N(0, 0.02^2)$ variables; (iii) we replaced $X_{i,2}$ in the next 19 points

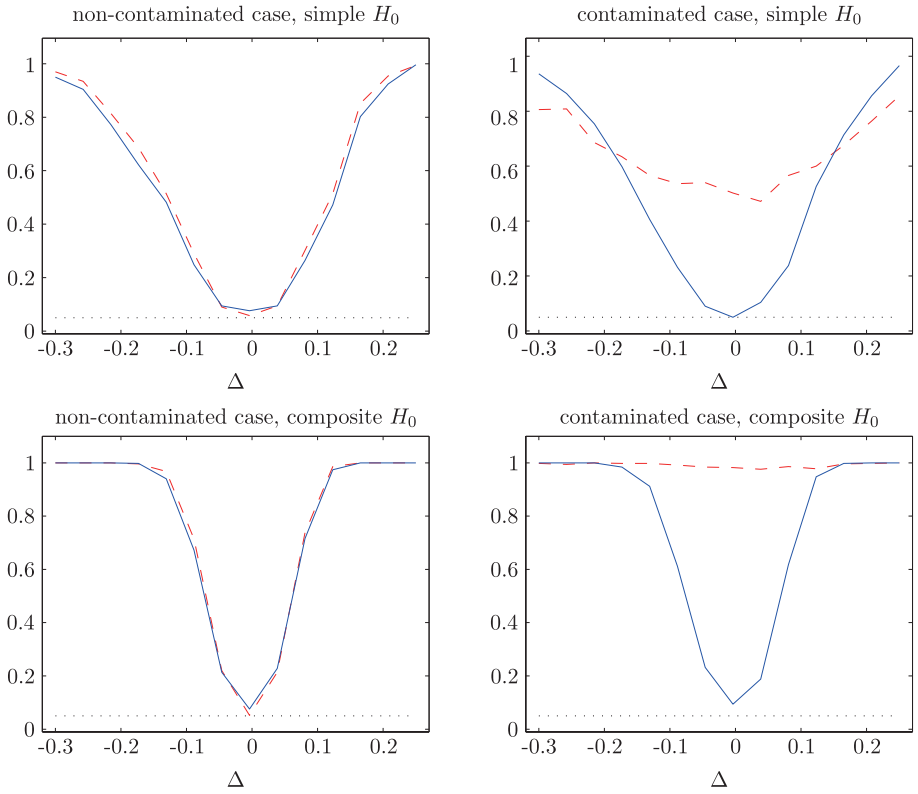

Figure 4. Observed power functions of tests for the overdispersed Poisson response data. The dashed line corresponds to the non-robust Wald-type test; the solid line corresponds to the robust Wald-type test; the dotted line indicates the 5% nominal level. Top panels: for testing (4.1); bottom panels: for testing (4.2). Left panels: without contamination; right panels: with contamination.

with $X^*_{i,2} = X_{i,2} + e_i$, where $\{e_i\}$ were independent $N(0, 0.02^2)$ variables; (iv) we replaced Y_i in the next 2 points with $Y_i^* = 1 - Y_i$. Figure 5 shows the two hyperplanes and locations $(X^*_{i,1}, X^*_{i,2})$ of contaminated points. Both the deviance loss and the exponential loss were employed as the BD. The sample size n was 800, and 500 replications were conducted.

Part 1: Parameter estimation. Figure 7 compares the boxplots of $\widehat{\beta}_j - \beta_{j;0}$, $j = 0, 1, \ldots, p_n$, using the non-robust and robust BD estimates, where the deviance loss and the exponential loss are used as the BD in the top and bottom panels respectively. In the presence of contamination, robust procedures are effective in reducing the estimation bias without excessively inflating the variance.

Part 2: Null distribution of the robust test statistic W_n**.** Two types of

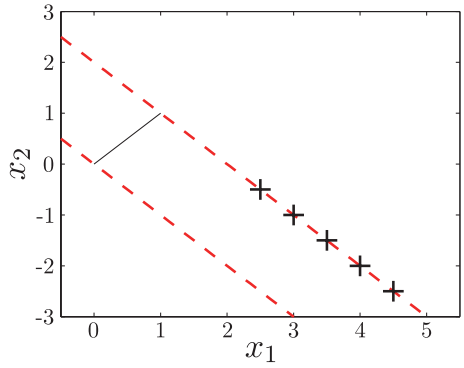

Figure 5. Lower dashed line: true discriminating hyperplane; upper dashed line: the hyperplane for the contaminated points marked by the plus "+" signs; solid line with length d: orthogonal to the parallel dashed lines.

null hypotheses were considered:

$$H_0^{(\mathrm{I})} : \beta_{0;0} = 0,\ \beta_{1;0} = 2,\ \beta_{2;0} = 2, \qquad \text{and} \quad H_0^{(\mathrm{II})} : \beta_{1;0} = \beta_{2;0}.$$

We checked the agreement between the asymptotic χ^2_{k} distribution and the finite sampling distribution of the robust test statistic W_n under the null hypotheses. The QQ plots of the (1st to 99th) percentiles of W_n against those of the χ^2_{k} distribution are displayed in Figure 8. The left panels test for $H_0^{(\mathrm{I})}$, and the right panels test for $H_0^{(\mathrm{II})}$. We observe that the finite sampling null distribution of W_n, using the deviance loss as the BD, agrees reasonably well with the χ^2 distribution. As a comparison, results using the exponential loss as the BD can be improved with a reduced number (for example 5) of contamination points or more refined estimation of the covariance matrices.

Part 3: Level of tests. To investigate the stability of the level of the robust test, we followed the previous contamination scheme, but with d varying from 2 to 10. The larger the d, the more severe the contamination. Figure 9 shows the empirical rejection rates of H_0. We observe that, under contamination, the level of the robust Wald-type test is more stable. In contrast, the non-robust Wald-type test shows high sensitivity to the presence of outliers.

Part 4: Power of tests. To assess the stability of the power of the tests, we generated the original data from the true model, but with the true parameter $\widetilde{\boldsymbol{\beta}}_0$ replaced by $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_0 + \Delta \boldsymbol{c}$, where $\boldsymbol{c} = (1, 3, 4)^T$. Figure 10 plots the empirical rejection rates of the null hypothesis $H_0^{(\mathrm{I})}$ in the non-contaminated and contaminated cases. The non-robust and robust Wald-type tests perform comparably in the non-contaminated cases. Under contamination, the observed power function of the robust test is close to that achieved in the non-contaminated case, while

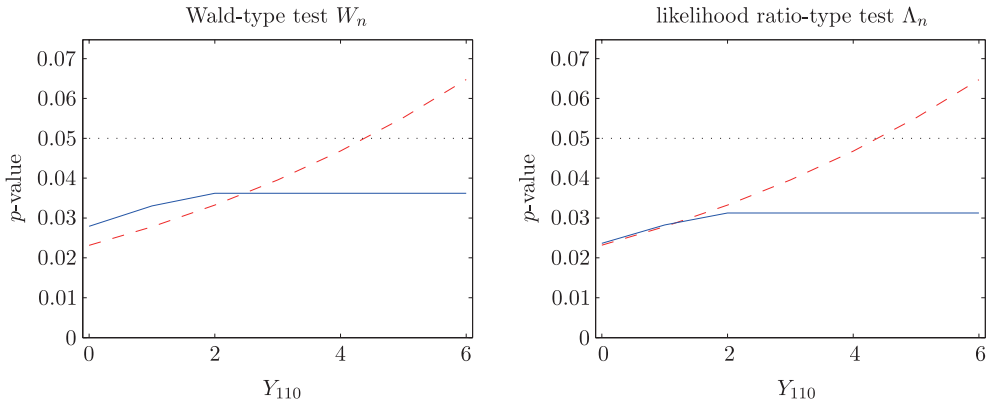

Figure 6. Sensitivity curves of the p-values for the Wald-type test W_n (left panel) and likelihood ratio-type test Λ_n (right panel). The dashed line corresponds to the non-robust test; the solid line corresponds to the robust test; the dotted line indicates the 5% significance level.

the non-robust test is less informative, since the power function is higher than the nominal level under the null hypotheses with $\Delta = 0$. Results for the null hypothesis $H_0^{(\mathrm{II})}$ were similar and are omitted.

7. Data Analysis

For the sake of comparison with the published results on robust methods in the CR paper, we analyze the data from a study of the diversity of possum (arboreal marsupials) in the Montane ash forest (Australia), described in Lindenmayer et al. (1990, 1991). Refer to the CR paper for a detailed description of the dataset and analysis using robust methods.

A Poisson generalized linear regression model with log-link was fitted to the response variable `diversity` with explanatory variables (`Shrubs`, `Stumps`, `Stags`, `Bark`, `Habitat`, `BAcacia`, `E.regnans`, `E.delegatensis`, `E.nitens`, `NW-NE`, `NW-SE`, `SE-SW`, and `SW-NW`), involving the estimation of 12 parameters (including one for the intercept term). The (negative) quasi-likelihood was utilized as the BD, generated by the q-function in (3.2) with $V(\mu) = \mu$. For the sake of comparison, we took the tuning constant $c = 1.6$ and weights $w(\boldsymbol{x}_i) = \sqrt{1 - h_i}$, with h_i as the ith diagonal element of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{X}$ the design matrix with the ith row $\widetilde{\boldsymbol{X}}_{ni}^T$. Table 2 tabulates the robust-BD estimates of parameters in (3.5) and their standard errors (indicated in parentheses), together with the p-values calculated from the proposed Wald-type tests. It is seen that the robust-BD estimates are nearly identical to the robust quasi-likelihood estimates given in Table 3 of the CR paper, but have slightly smaller standard errors.

Table 3 presents parameter estimation and p-values calculated from W_n for the final reduced model, where variables retained were taken from the CR paper.

Table 2. Coefficient estimation and p-values for Poisson model with log-link of the Possum dataset.

	Non-robust BD estimation			Robust-BD estimation		
Variable	estimate	(s.e.)	p-value of W_n	estimate	(s.e.)	p-value of W_n
Intercept	-0.9469	(0.2655)	0.0004	-0.8974	(0.2680)	0.0008
`Shrubs`	0.0119	(0.0219)	0.5867	0.0099	(0.0222)	0.6542
`Stumps`	-0.2724	(0.2859)	0.3408	-0.2515	(0.2872)	0.3811
`Stags`	0.0402	(0.0112)	0.0003	0.0401	(0.0113)	0.0004
`Bark`	0.0399	(0.0144)	0.0056	0.0400	(0.0145)	0.0058
`Habitat`	0.0717	(0.0381)	0.0600	0.0714	(0.0385)	0.0633
`BAcacia`	0.0176	(0.0106)	0.0961	0.0178	(0.0107)	0.0964
`E.regnans`	0	(—)	—	0	(—)	—
`E.delegatensis`	-0.0154	(0.1916)	0.9361	-0.0203	(0.1935)	0.9164
`E.nitens`	0.1150	(0.2724)	0.6730	0.1268	(0.2734)	0.6429
`NW-NE`	0	(—)	—	0	(—)	—
`NW-SE`	0.0668	(0.1902)	0.7254	0.0601	(0.1910)	0.7529
`SE-SW`	0.1170	(0.1903)	0.5388	0.0950	(0.1918)	0.6202
`SW-NW`	-0.4889	(0.2475)	0.0482	-0.5077	(0.2502)	0.0424

Again, the standard errors of the robust-BD estimates are slightly smaller than those of the robust quasi-likelihood estimates given in Table 5 of the CR paper. To compare the sensitivity of the Wald-type test W_n and the likelihood ratio-type test Λ_n, we let Y_{110} span the range of integer values from 0 to 6. In each situation, the null hypothesis for the insignificance of the variable `Habitat` was tested. Figure 6 plots the p-values of both tests. The p-value of the robust Wald-type test is very stable, but the p-value of its non-robust version varies much more, yielding a different model choice if the level is set at 5%. The p-value of the robust Wald-type test is bounded, whereas the p-value of the non-robust Wald-type test increases as the value of Y_{110} grows. The proposed robust Wald-type test is as stable to perturbation as the robust likelihood ratio-type test, but the gain in computational simplicity is substantial, especially when the number of simultaneous hypotheses increases. The sensitivity curve of the robust likelihood ratio-type test Λ_n appears to be different from that of the robust quasi-likelihood test displayed in Figure 1 of the CR paper. This is due to the fact that Λ_n and the test in the CR paper are asymptotically equivalent, but not identical; moreover, numerical procedures involved in estimating covariance matrices as well as numerical integrations performed in computing $\rho_q(\cdot, \cdot)$ differ.

Acknowledgements

The authors thank the Co-Editor, an associate editor and two referees for insightful comments and suggestions. The research is supported by the NSF grant DMS-1106586 and DMS-1308872, and Wisconsin Alumni Research Foundation.

Table 3. Coefficient estimation and p-values for Poisson model with log-link of the Possum dataset.

	Non-robust BD estimation			Robust-BD estimation		
Variable	estimate	(s.e.)	p-value of W_n	estimate	(s.e.)	p-value of W_n
Intercept	-0.8212	(0.2001)	0.0000	-0.7976	(0.2028)	0.0001
`Stags`	0.0410	(0.0103)	0.0001	0.0406	(0.0104)	0.0001
`Bark`	0.0406	(0.0125)	0.0011	0.0410	(0.0126)	0.0011
`Habitat`	0.0782	(0.0367)	0.0332	0.0776	(0.0370)	0.0361
`BAcacia`	0.0136	(0.0097)	0.1609	0.0143	(0.0098)	0.1449
`SW-NW`	-0.5967	(0.2086)	0.0042	-0.6043	(0.2118)	0.0043

References

Adimari, G. and Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statist. Probab. Lett.* **55**, 413-419.

Bianco, A. M. and Martínez, E. (2009). Robust testing in the logistic regression model. *Comput. Statist. Data Anal.* **53**, 4095-4105.

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods* (Edit by Rieder), 17-34. Springer, New York.

Bickel, P. and Doksum, K. (2007). *Mathematical Statistics, Basic Ideas and Selected Topics*. Second edition. Pearson Prentice Hall.

Boente, G., He, X., and Zhou, J. (2006). Robust estimates in generalized partially linear models. *Ann. Statist.* **34**, 2856-2878.

Brègman, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.* **7**, 620-631.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96**, 1022-1030.

Chow, Y. S. and Teicher, H. (1989). *Probability Theory*. 2nd edition. Springer-Verlag.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Statist. Data Anal.* **44**, 273-295.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.* **72**, 851-853.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89**, 897-904.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. Wiley-Interscience.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Smith, A. P., and Nix, H. A. (1990). The conservation of arboreal marsupials in the montane ash forests of the victoria, southeast australia, I. Factors influencing the occupancy of trees with hollows, *Biological Conservation* **54**, 111-131.

Lindenmayer, D. B., Cunningham, R. B., Tanton, M. T., Nix, H. A. and Smith, A. P. (1991). The conservation of arboreal marsupials in the montane ash forests of the central highlands of victoria, south-east australia: III. The habitat requirements of leadbeater's possum gymnobelideus leadbeateri and models of the diversity and abundance of arboreal marsupials. *Biological Conservation* **56**, 295-315.

Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, London.

Mebane, W. R. and Sekhon, J. S. (2004). Robust estimation and outlier detection for overdispersed multinomial models of count data. *Amer. J. Political Sci.* **48**, 392-411.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Zhang, C. M., Jiang, Y. and Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canad. J. Statist.* **37**, 119-139.

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: cmzhang@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: xguo@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: ccheng@stat.wisc.edu

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA.

E-mail: zjz@stat.wisc.edu

(Received January 2012; accepted March 2013)

Robust-BD Estimation and Inference for Varying-Dimensional General Linear Models

Chunming Zhang Xiao Guo Chen Cheng Zhengjun Zhang

University of Wisconsin-Madison

Supplementary Material

S1 Notation and Assumptions

For a matrix M, its eigenvalues, minimum eigenvalue, maximum eigenvalue and trace are labeled by $\lambda_j(M)$, $\lambda_{\min}(M)$, $\lambda_{\max}(M)$ and $\mathrm{tr}(M)$ respectively. Let $\|M\| = \sup_{\|\boldsymbol{x}_n\|=1} \|M\boldsymbol{x}_n\| = \{\lambda_{\max}(M^TM)\}^{1/2}$ be the matrix L_2 norm; let $\|M\|_F = \{\mathrm{tr}(M^TM)\}^{1/2}$ be the Frobenius norm. See Golub and Van Loan (1996) for details. Throughout the proof, C is used as a generic finite constant.

We first impose some regularity conditions, which are not the weakest possible but facilitate the technical derivations.

Condition A:

A0. $\sup_{n\geq 1} \|\widetilde{\boldsymbol{\beta}}_0\|_1 < \infty$.

A1. $\|\boldsymbol{X}_n\|_\infty = \max_{1\leq j\leq p_n} |X_j|$ is bounded almost surely.

A2. $E(\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T)$ exists and is nonsingular.

A4. There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_0$, such that $F^{-1}(\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}})$ is bounded almost surely for all $\widetilde{\boldsymbol{\beta}}$ in the subset.

A5. $w(\cdot) \geq 0$ is a bounded function. Assume that $\psi(r)$ is a bounded, odd function, and twice differentiable, such that $\psi'(r)$, $\psi'(r)r$, $\psi''(r)$, $\psi''(r)r$ and $\psi''(r)r^2$ are bounded; $V(\cdot) > 0$, $V^{(2)}(\cdot)$ is continuous. The matrix $\mathbf{H}_n$ is positive definite, with eigenvalues uniformly bounded away from 0.

A6. $q^{(4)}(\cdot)$ is continuous, and $q^{(2)}(\cdot) < 0$. $G_1^{(3)}(\cdot)$ is continuous.

A7. $F(\cdot)$ is monotone and a bijection, $F^{(3)}(\cdot)$ is continuous, and $F^{(1)}(\cdot) \neq 0$.

Condition B:

B5. The matrices Ω_n and $\mathbf{H}_n$ are positive definite, with eigenvalues uniformly bounded away from 0. Also, $\|\mathbf{H}_n^{-1}\Omega_n\|$ is bounded away from ∞.

Condition C:

C4. There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_0$, such that $A_n\widetilde{\boldsymbol{\beta}}_0 = \boldsymbol{g}_0$, and $F^{-1}(\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}})$ is bounded almost surely for all $\widetilde{\boldsymbol{\beta}}$ in the subset.

Condition D:

D5. The eigenvalues of $\mathbf{H}_n$ are uniformly bounded away from 0. Also, $\|\mathbf{H}_n^{-1/2}\Omega_n^{1/2}\|$ is bounded away from ∞.

S2 Proofs of Main Results

Proof of Theorem 1

We follow the idea of the proof in Fan and Peng (2004). Let $r_n = \sqrt{p_n/n}$ and $\widetilde{\mathbf{u}}_n = (u_0, u_1, \ldots, u_{p_n})^T \in \mathbb{R}^{p_n+1}$. It suffices to show that for any given $\epsilon > 0$, there exists a sufficiently large constant C_ϵ such that, for large n we have

$$P\Big\{ \inf_{\|\widetilde{\mathbf{u}}_n\|=C_\epsilon} \ell_n(\widetilde{\boldsymbol{\beta}}_0 + r_n\widetilde{\mathbf{u}}_n) > \ell_n(\widetilde{\boldsymbol{\beta}}_0)\Big\} \geq 1-\epsilon. \tag{S2.1}$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}$ of $\ell_n(\widetilde{\boldsymbol{\beta}})$ in the ball $\{\widetilde{\boldsymbol{\beta}}_0 + r_n\widetilde{\mathbf{u}}_n : \|\widetilde{\mathbf{u}}_n\| \leq C_\epsilon\}$ such that $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_0\| = O_P(r_n)$. To show (S2.1), consider

$$\begin{aligned} \ell_n(\widetilde{\boldsymbol{\beta}}_0 + r_n\widetilde{\mathbf{u}}_n) - \ell_n(\widetilde{\boldsymbol{\beta}}_0) &= \frac{1}{n}\sum_{i=1}^n \{\rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T(\widetilde{\boldsymbol{\beta}}_0 + r_n\widetilde{\mathbf{u}}_n)))w(\boldsymbol{X}_{ni}) \\ &\quad -\rho_q(Y_i, F^{-1}(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_0))w(\boldsymbol{X}_{ni})\} \\ &\equiv I_1, \end{aligned} \tag{S2.2}$$

where $\|\widetilde{\mathbf{u}}_n\| = C_\epsilon$.

By Taylor's expansion,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \tag{S2.3}$$

where

$$\begin{aligned} I_{1,1} &= r_n/n\sum_{i=1}^n \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_0)w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n, \\ I_{1,2} &= r_n^2/(2n)\sum_{i=1}^n \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2, \end{aligned}$$

$$I_{1,3} \quad = \quad r_n^3/(6n)\sum_{i=1}^{n} \mathrm{p}_3(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)w(\boldsymbol{X}_{ni})(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^3$$

for $\widetilde{\boldsymbol{\beta}}_n^*$ located between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widetilde{\boldsymbol{\beta}}_{n;0} + r_n\widetilde{\mathbf{u}}_n$. Hence

$$|I_{1,1}| \le r_n\left\|\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\|\|\widetilde{\mathbf{u}}_n\| = O_P(r_n\sqrt{p_n/n})\|\widetilde{\mathbf{u}}_n\|. \qquad \text{(S2.4)}$$

For $I_{1,2}$ in (S2.3),

$$\begin{aligned} I_{1,2} \quad &= \quad \frac{r_n^2}{2n}\sum_{i=1}^{n} E\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2\} \\ &\quad + \frac{r_n^2}{2n}\sum_{i=1}^{n}[\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2 - E\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2\}] \\ &\equiv \quad I_{1,2,1} + I_{1,2,2}, \end{aligned}$$

where $I_{1,2,1} = 2^{-1}r_n^2\widetilde{\mathbf{u}}_n^T\mathbf{H}_n\widetilde{\mathbf{u}}_n$. Meanwhile, we have

$$\begin{aligned} |I_{1,2,2}| \quad &\le \quad r_n^2\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T - E\{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T\}\right]\right\|_F\|\widetilde{\mathbf{u}}_n\|^2 \\ &= \quad r_n^2 O_P(p_n/\sqrt{n})\|\widetilde{\mathbf{u}}_n\|^2. \end{aligned}$$

Thus,

$$I_{1,2} = \frac{r_n^2}{2}\widetilde{\mathbf{u}}_n^T\mathbf{H}_n\widetilde{\mathbf{u}}_n + O_P(r_n^2 p_n/\sqrt{n})\|\widetilde{\mathbf{u}}_n\|^2. \qquad \text{(S2.5)}$$

For $I_{1,3}$ in (S2.3), we observe that

$$|I_{1,3}| \le r_n^3\frac{1}{n}\sum_{i=1}^{n}|\mathrm{p}_3(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)|w(\boldsymbol{X}_{ni})|\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\mathbf{u}}_n|^3 = O_P(r_n^3 p_n^{3/2})\|\widetilde{\mathbf{u}}_n\|^3,$$

which follows from Conditions A0, A1, A4 and A5.

By (S2.4) and $p_n^4/n \to 0$, we can choose some large C_ϵ such that $I_{1,1}$ and $I_{1,3}$ are all dominated by the first term of $I_{1,2}$ in (S2.5), which is positive by the eigenvalue assumption on $\mathbf{H}_n$. This implies (S2.1). ■

Proof of Theorem 2

Notice the estimating equations $\frac{\partial \ell_n(\widetilde{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}}}|_{\widetilde{\boldsymbol{\beta}}=\widehat{\widetilde{\boldsymbol{\beta}}}} = \mathbf{0}$, since $\widehat{\widetilde{\boldsymbol{\beta}}}$ is a local minimizer of $\ell_n(\widetilde{\boldsymbol{\beta}})$. Taylor's expansion applied to the left side of the estimation equations yields

$$\mathbf{0} = \left\{\frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\}$$

$$
\begin{aligned}
&+\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T\right\}(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})\\
&+\frac{1}{2n}\sum_{i=1}^{n}\mathrm{p}_3(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)w(\boldsymbol{X}_{ni})\{\widetilde{\boldsymbol{X}}_{ni}^T(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})\}^2\widetilde{\boldsymbol{X}}_{ni}\\
&\equiv\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\}+K_2(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})+K_3, \qquad \text{(S2.6)}
\end{aligned}
$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ lies between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}$. Below, we will show

$$
\begin{aligned}
\|K_2-\mathbf{H}_n\| &= O_P(p_n/\sqrt{n}), &\text{(S2.7)}\\
\|K_3\| &= O_P(p_n^{5/2}/n). &\text{(S2.8)}
\end{aligned}
$$

First, to show (S2.7), note that $K_2-\mathbf{H}_n=K_2-E(K_2)\equiv L_1$. Similar arguments for the proof of $I_{1,2,2}$ in Theorem 1 give $\|L_1\|=O_P(p_n/\sqrt{n})$.

Second, a similar proof used for $I_{1,3}$ in (S2.3) completes (S2.8).

Third, by (S2.6)–(S2.8) and $\|\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0}\|=O_P(\sqrt{p_n/n})$, we see that

$$
\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})=-\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}+\mathbf{u}_n, \qquad \text{(S2.9)}
$$

where $\|\mathbf{u}_n\|=O_P(p_n^{5/2}/n)$. Note that by Condition B5,

$$
\begin{aligned}
\|\sqrt{n}A_n\Omega_n^{-1/2}\mathbf{u}_n\| &\le \sqrt{n}\|A_n\|_F\lambda_{\max}(\Omega_n^{-1/2})\|\mathbf{u}_n\|\\
&=\sqrt{n}\{\mathrm{tr}(A_nA_n^T)\}^{1/2}/\lambda_{\min}^{1/2}(\Omega_n)\|\mathbf{u}_n\|=O_P(p_n^{5/2}/\sqrt{n})=o_P(1).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\sqrt{n}A_n\Omega_n^{-1/2}\{\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})\}\\
&\quad=-\frac{1}{\sqrt{n}}A_n\Omega_n^{-1/2}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}+o_P(1).
\end{aligned}
$$

To complete proving Theorem 2, we apply the Lindeberg-Feller central limit theorem (van der Vaart, 1998) to $\sum_{i=1}^n\boldsymbol{Z}_{ni}$, where $\boldsymbol{Z}_{ni}=-n^{-1/2}A_n\Omega_n^{-1/2}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}$. It suffices to check (Ⅰ) $\sum_{i=1}^n\mathrm{cov}(\boldsymbol{Z}_{ni})\to\mathbb{G}$; (Ⅱ) $\sum_{i=1}^nE(\|\boldsymbol{Z}_{ni}\|^{2+\delta})=o(1)$ for some $\delta>0$. Condition (Ⅰ) follows from the fact that $\mathrm{var}\{\mathrm{p}_1(Y;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_n)\widetilde{\boldsymbol{X}}_n\}=\Omega_n$. To verify condition (Ⅱ), notice that using Conditions B5 and A5,

$$
\begin{aligned}
E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) &\le n^{-(2+\delta)/2}E\Big\{\|A_n\|_F^{2+\delta}\Big[\|\Omega_n^{-1/2}\widetilde{\boldsymbol{X}}_n\|\\
&\quad\Big|\{\psi(r(Y,m(\boldsymbol{X}_n)))-G_1'(m(\boldsymbol{X}_n))\}\frac{\{q''(m(\boldsymbol{X}_n))\sqrt{V(m(\boldsymbol{X}_n))}\}}{F'(m(\boldsymbol{X}_n))}w(\boldsymbol{X}_n)\Big|\Big]^{2+\delta}\Big\}\\
&\le Cn^{-(2+\delta)/2}E[\{\lambda_{\min}^{-1/2}(\Omega_n)\|\widetilde{\boldsymbol{X}}_n\|\}^{2+\delta}|\{\psi(r(Y,m(\boldsymbol{X}_n)))-G_1'(m(\boldsymbol{X}_n))\}\times
\end{aligned}
$$

$$
\begin{aligned}
& \{q''(m(\boldsymbol{X}_n))\sqrt{V(m(\boldsymbol{X}_n))}\}/F'(m(\boldsymbol{X}_n))|^{2+\delta}] \\
\leq\ & Cp_n^{(2+\delta)/2}n^{-(2+\delta)/2}E[|\{\psi(r(Y,m(\boldsymbol{X}_n))) - G_1'(m(\boldsymbol{X}_n))\}\times \\
& \{q''(m(\boldsymbol{X}_n))\sqrt{V(m(\boldsymbol{X}_n))}\}/F'(m(\boldsymbol{X}_n))|^{2+\delta}] \\
\leq\ & O((p_n/n)^{(2+\delta)/2}).
\end{aligned}
$$

Thus, we get $\sum_{i=1}^n E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) \leq O(n(p_n/n)^{(2+\delta)/2}) = O(p_n^{(2+\delta)/2}/n^{\delta/2})$, which is $o(1)$. This verifies Condition (II). ■

Proposition 1 (covariance matrix estimation) *Assume* A0, A1, A2, A4, A5, B5, A6, *and* A7 *in the Appendix. Let* $V_n = \mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}$ *and* $\widehat{V}_n = \widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}$. *If* $p_n^4/n \to 0$ *as* $n \to \infty$, *then for any* $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}$ *of* $\widetilde{\boldsymbol{\beta}}_{n;0}$, *we have* $A_n(\widehat{V}_n - V_n)A_n^T \xrightarrow{P} \mathbf{0}$ *for any* $k \times (p_n+1)$ *matrix* A_n *satisfying* $A_nA_n^T \to \mathbb{G}$, *where* $\mathbb{G}$ *is a* $k \times k$ *matrix and* k *is any fixed integer.*

Proof: Note $\|A_n(\widehat{V}_n - V_n)A_n^T\| \leq \|\widehat{V}_n - V_n\|\|A_n\|_F^2$. Since $\|A_n\|_F^2 \to \mathrm{tr}(\mathbb{G})$, it suffices to prove that $\|\widehat{V}_n - V_n\| = o_P(1)$.

First, we prove $\|\widehat{\mathbf{H}}_n - \mathbf{H}_n\| = o_P(1)$. Note that

$$
\begin{aligned}
\widehat{\mathbf{H}}_n - \mathbf{H}_n &= \frac{1}{n}\sum_{i=1}^n \{\mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}) - \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T \\
&\quad + \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{p}_2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T - \mathbf{H}_n\right\} \\
&\equiv I_1 + I_2.
\end{aligned}
$$

From the proof of (S2.7) in Theorem 2, we know that $\|I_2\| = O_P(p_n/\sqrt{n}) = o_P(1)$. We only need to consider the term I_1. Let $\widehat{m}_i = \widehat{m}(\boldsymbol{X}_{ni})$, $m_i = m(\boldsymbol{X}_{ni})$, $\widehat{r}_i = r(Y_i, \widehat{m}_i)$ and $r_i = r(Y_i, m_i)$. Then

$$
\begin{aligned}
I_1 &= \frac{1}{n}\sum_{i=1}^n [A_0(Y_i,\widehat{m}_i) + \{\psi(\widehat{r}_i) - G_1'(\widehat{m}_i)\}A_1(\widehat{m}_i) \\
&\quad - A_0(Y_i,m_i) - \{\psi(r_i) - G_1'(m_i)\}A_1(m_i)]w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T \\
&= -\frac{1}{n}\sum_{i=1}^n \{G_1'(\widehat{m}_i)A_1(\widehat{m}_i) - G_1'(m_i)A_1(m_i)\}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{A_0(Y_i,\widehat{m}_i) + \psi(\widehat{r}_i)A_1(\widehat{m}_i) - A_0(Y_i,m_i) - \psi(r_i)A_1(m_i)\}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T \\
&\equiv I_{1,1} + I_{1,2}.
\end{aligned}
$$

Let $g(\cdot) = G_1'(\cdot)A_1(\cdot)$. By the assumptions, $g(\cdot)$ is differentiable. Thus

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n |g(\widehat{m}_i) - g(m_i)| &= \frac{1}{n}\sum_{i=1}^n |(g\circ F^{-1})'(\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)\boldsymbol{X}_{ni}^T(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0})| \\
&= O_P(1)O_P(\sqrt{p_n})O_P(\sqrt{p_n/n}) = O_P(p_n/\sqrt{n}),
\end{aligned}
$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ is between $\widehat{\widetilde{\boldsymbol{\beta}}}$ and $\widetilde{\boldsymbol{\beta}}_{n;0}$. Thus

$$\left\| \frac{1}{n}\sum_{i=1}^{n} |g(\widehat{m}(\boldsymbol{X}_{ni}))-g(m(\boldsymbol{X}_{ni}))| w(\boldsymbol{X}_{ni}) \widetilde{\boldsymbol{X}}_{ni} \widetilde{\boldsymbol{X}}_{ni}^T \right\|_F = O_P(p_n/\sqrt{n}) O_P(p_n) = O_P(p_n^2/\sqrt{n}).$$

Similar arguments give $\|I_{1,1}\| = O_P(p_n^2/\sqrt{n})$ and $\|I_{1,2}\| = O_P(p_n^2/\sqrt{n})$. Thus $\|I_1\| = O_P(p_n^2/\sqrt{n}) = o_P(1)$.

Second, we show $\|\widehat{\Omega}_n - \Omega_n\| = o_P(1)$. It is easy to see that

$$\begin{aligned} \widehat{\Omega}_n - \Omega_n &= \frac{1}{n}\sum_{i=1}^{n} \{\mathrm{p}_1^2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T \widehat{\widetilde{\boldsymbol{\beta}}}) - \mathrm{p}_1^2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T \widetilde{\boldsymbol{\beta}}_{n;0})\} w^2(\boldsymbol{X}_{ni}) \widetilde{\boldsymbol{X}}_{ni} \widetilde{\boldsymbol{X}}_{ni}^T \\ &\quad + \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}_1^2(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T \widetilde{\boldsymbol{\beta}}_{n;0}) w^2(\boldsymbol{X}_{ni}) \widetilde{\boldsymbol{X}}_{ni} \widetilde{\boldsymbol{X}}_{ni}^T - \Omega_n \right\} \\ &= \Delta_{1,1} + \Delta_{1,2}, \end{aligned}$$

where $\|\Delta_{1,1}\| = O_P(p_n^2/\sqrt{n})$ and $\|\Delta_{1,2}\| = O_P(p_n/\sqrt{n})$. We observe that $\|\widehat{\Omega}_n - \Omega_n\| = O_P(p_n^2/\sqrt{n}) = o_P(1)$.

Third, we show $\|\widehat{V}_n - V_n\| = o_P(1)$. Note $\widehat{V}_n - V_n = L_1 + L_2 + L_3$, where $L_1 = \widehat{\mathbf{H}}_n^{-1}(\widehat{\Omega}_n - \Omega_n)\widehat{\mathbf{H}}_n^{-1}$, $L_2 = \widehat{\mathbf{H}}_n^{-1}(\mathbf{H}_n - \widehat{\mathbf{H}}_n)\mathbf{H}_n^{-1}\Omega_n\widehat{\mathbf{H}}_n^{-1}$ and $L_3 = \mathbf{H}_n^{-1}\Omega_n\widehat{\mathbf{H}}_n^{-1}(\mathbf{H}_n - \widehat{\mathbf{H}}_n)\mathbf{H}_n^{-1}$. By Assumption B5, it is straightforward to verify that $\|\mathbf{H}_n^{-1}\| \le O(1)$, $\|\widehat{\mathbf{H}}_n^{-1}\| \le O_P(1)$ and $\|\mathbf{H}_n^{-1}\Omega_n\| \le O(1)$. Since $\|L_1\| \le \|\widehat{\mathbf{H}}_n^{-1}\| \|\widehat{\Omega}_n - \Omega_n\| \|\widehat{\mathbf{H}}_n^{-1}\|$, we conclude $\|L_1\| = o_P(1)$, and similarly $\|L_2\| = o_P(1)$ and $\|L_3\| = o_P(1)$. Hence $\widehat{V}_n - V_n = o_P(1)$. ■

Proof of Theorem 3

For the matrix A_n in (4.3), there exists a $(p_n + 1 - \mathsf{k}) \times (p_n + 1)$ matrix B_n satisfying $B_n B_n^T = \mathbf{I}_{p_n+1-\mathsf{k}}$ and $A_n B_n^T = \mathbf{0}$. Therefore, $A_n\widetilde{\boldsymbol{\beta}}_n = \boldsymbol{g}_0$ is equivalent to $\widetilde{\boldsymbol{\beta}}_n = B_n^T\boldsymbol{\gamma}_n + \boldsymbol{c}_0$, where $\boldsymbol{\gamma}_n$ is a $(p_n + 1 - \mathsf{k}) \times 1$ vector and $\boldsymbol{c}_0 = A_n^T\mathbb{G}^{-1}\boldsymbol{g}_0$. Thus under H_0 in (4.3), we have $\widetilde{\boldsymbol{\beta}}_{n;0} = B_n^T\boldsymbol{\gamma}_{n;0} + \boldsymbol{c}_0$. Then minimizing $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$ subject to $A_n\widetilde{\boldsymbol{\beta}}_n = \boldsymbol{g}_0$ is equivalent to minimizing $\ell_n(B_n^T\boldsymbol{\gamma}_n + \boldsymbol{c}_0)$ with respect to $\boldsymbol{\gamma}_n$, and we denote by $\widehat{\boldsymbol{\gamma}}_n$ the minimizer. Note that under (4.4), $\widehat{\widetilde{\boldsymbol{\beta}}}$ is the unique minimizer of $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$. Hence $\Lambda_n = 2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})\}$. Before showing Theorem 3, we need Lemma 1.

Lemma 1 *Assume conditions of Theorem* 3. *Then under H_0 in (4.3), we have that* $B_n^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n;0}) = -n^{-1}B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n\sum_{i=1}^{n}\mathrm{p}_1(Y_i; \widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni} + o_P(n^{-1/2})$, *and* $2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})\} = n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}})^T\mathbf{H}_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}}) + o_P(1)$.

Proof: To obtain the first part, following the proof of (S2.9) in Theorem 2, we have a similar expression for $\widehat{\boldsymbol{\gamma}}_n$,

$$B_n\mathbf{H}_nB_n^T(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_{n;0})=-\frac{1}{n}B_n\sum_{i=1}^n \mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}+\mathbf{w}_n,$$

with $\|\mathbf{w}_n\|=o_P(n^{-1/2})$. As a result,

$$B_n^T(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_{n;0})=-\frac{1}{n}B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n\sum_{i=1}^n \mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}+B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}\mathbf{w}_n.$$

We notice that

$$\|B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}\mathbf{w}_n\|\le\|(B_n\mathbf{H}_nB_n^T)^{-1}\|\|\mathbf{w}_n\|\le\|\mathbf{w}_n\|/\lambda_{\min}(\mathbf{H}_n)=o_P(n^{-1/2}),$$

in which the fact $\lambda_{\min}(B_n\mathbf{H}_nB_n^T)\ge\lambda_{\min}(\mathbf{H}_n)$ is used.

The proof of the second part proceeds in three steps. In Step 1, we use the following Taylor expansion for $\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0)-\ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})$,

$$\begin{aligned}\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0)-\ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}) &= \frac{1}{2n}\sum_{i=1}^n \mathrm{p}_2(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}})w(\boldsymbol{X}_{ni})\{\widetilde{\boldsymbol{X}}_{ni}^T(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}})\}^2\\ &\quad+\frac{1}{6n}\sum_{i=1}^n \mathrm{p}_3(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)w(\boldsymbol{X}_{ni})\{\widetilde{\boldsymbol{X}}_{ni}^T(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}})\}^3\\ &\equiv I_1+I_2,\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ lies between $\widehat{\widetilde{\boldsymbol{\beta}}}$ and $B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0$.

In Step 2, we analyze the stochastic order of $B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}}$. For a matrix X whose column vectors are linearly independent, set $P_X=X(X^TX)^{-1}X^T$. Define $H_n=\mathbf{I}_{p_n+1}-P_{\mathbf{H}_n^{1/2}B_n^T}=P_{\mathbf{H}_n^{-1/2}A_n^T}$. Then $\mathbf{H}_n^{-1}-B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n=\mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}$. By (S2.9) and the first part of Lemma 1, we see immediately that

$$\begin{aligned}B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}} &= B_n^T(\widehat{\boldsymbol{\gamma}}_n-\boldsymbol{\gamma}_{n;0})-(\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0})\\ &= \mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^n \mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\}+o_P(n^{-1/2}),\end{aligned}\tag{S2.10}$$

where $\mathrm{p}_{1,i}=\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})$. Note that $\|\mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}\{n^{-1}\sum_{i=1}^n \mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\}\|=O_P(1/\sqrt{n})$. This gives

$$\|B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}}\|=O_P(1/\sqrt{n}).\tag{S2.11}$$

In Step 3, we conclude from (S2.11) that $I_2=O_P\{(p_n/n)^{3/2}\}=o_P(1/n)$. Then $2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0)-\ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})\}=2nI_1+o_P(1)$. Similar to the proof of Proposition 1, it is straightforward to see that

$$2nI_1 = n(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}})^T\left\{\frac{1}{n}\sum_{i=1}^n \mathrm{p}_2(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T\right\}(B_n^T\widehat{\boldsymbol{\gamma}}_n+\boldsymbol{c}_0-\widehat{\widetilde{\boldsymbol{\beta}}})$$

$$
\begin{aligned}
&= n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}})^T\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{p}_2(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\widetilde{\boldsymbol{X}}_{ni}^T\right\}(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}}) + o_P(1)\\
&= n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}})^T E\{\mathrm{p}_2(Y_n;\widetilde{\boldsymbol{X}}_n^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_n\widetilde{\boldsymbol{X}}_n^T\}(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}}) + o_P(1)\\
&= n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}})^T\mathbf{H}_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0 - \widehat{\widetilde{\boldsymbol{\beta}}}) + o_P(1).
\end{aligned}
$$

Then the second part of Lemma 1 is proved. ■

We now show Theorem 3. For part (i), a direct use of Lemma 1 and (S2.10) leads to

$$
\begin{aligned}
&2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})\}\\
&\quad= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\}^T\mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\} + o_P(1).
\end{aligned}
$$

Since H_n is idempotent of rank k, it can be written as $H_n = C_n^TC_n$, where C_n is a $\mathsf{k}\times(p_n+1)$ matrix satisfying $C_nC_n^T = \mathbf{I}_\mathsf{k}$. Then

$$
\begin{aligned}
&2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n + \boldsymbol{c}_0) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}})\}\\
&= \left\{\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^{n}\mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\}^T\left\{\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^{n}\mathrm{p}_{1,i}w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}\right\} + o_P(1).
\end{aligned}
$$

Now consider part (ii). If $\psi(r) = r$ and the q-function satisfies (4.5), then $\mathrm{p}_1(y;\theta) = \mathrm{q}_1(y;\theta)$, $\mathrm{p}_2(y;\theta) = \mathrm{q}_2(y;\theta)$ and $\mathbf{H}_n = \Omega_n/C$, where $\mathrm{q}_j(y;\theta) = \frac{\partial^j}{\partial\theta^j}Q_q(y, F^{-1}(\theta))$. In this case, similar arguments for Theorem 2 yield

$$
\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^{n}\mathrm{q}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni} \xrightarrow{\mathcal{L}} N(\mathbf{0}, C\mathbf{I}_\mathsf{k}),
$$

which completes the proof. ■

Proof of Theorem 4

Before showing Theorem 4, Lemma 2 is needed.

Lemma 2 *Assume conditions of Theorem* 4. *Then*

$$
\begin{aligned}
\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0} &= -\frac{1}{n}\mathbf{H}_n^{-1}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni} + o_{_P}(n^{-1/2}),\\
\sqrt{n}(A_n\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}A_n^T)^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) &\xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_k).
\end{aligned}
$$

Proof: Following (S2.9) in the proof of Theorem 2, we observe that $\|\mathbf{u}_n\| = O_P(p_n^{5/2}/n) = o_{_P}(n^{-1/2})$. Condition B5 completes the proof for the first part.

To show the second part, denote $U_n = A_n\mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}A_n^T$ and $\widehat{U}_n = A_n\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}A_n^T$. Notice that the eigenvalues of $\mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}$ are uniformly bounded away from 0. So are the eigenvalues of U_n. From the first part, we see that

$$A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) = -\frac{1}{n}A_n\mathbf{H}_n^{-1}\sum_{i=1}^{n}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni} + o_{P}(n^{-1/2}).$$

It follows that

$$\sqrt{n}U_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) = \sum_{i=1}^{n}\boldsymbol{Z}_{ni} + o_{P}(1),$$

where $\boldsymbol{Z}_{ni} = -n^{-1/2}U_n^{-1/2}A_n\mathbf{H}_n^{-1}\mathrm{p}_1(Y_i;\widetilde{\boldsymbol{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})w(\boldsymbol{X}_{ni})\widetilde{\boldsymbol{X}}_{ni}$. To show $\sum_{i=1}^{n}\boldsymbol{Z}_{ni} \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I}_{\mathsf{k}})$, similar to the proof for Theorem 2, we check (III) $\sum_{i=1}^{n}\mathrm{cov}(\boldsymbol{Z}_{ni}) \to \mathbf{I}_{\mathsf{k}}$; (IV) $\sum_{i=1}^{n}E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = o(1)$ for some $\delta > 0$. Condition (III) is straightforward since $\sum_{i=1}^{n}\mathrm{cov}(\boldsymbol{Z}_{ni}) = U_n^{-1/2}U_nU_n^{-1/2} = \mathbf{I}_{\mathsf{k}}$. To check condition (IV), similar arguments used in the proof of Theorem 2 give that $E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = O((p_n/n)^{(2+\delta)/2})$. This and the boundedness of ψ yield $\sum_{i=1}^{n}E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) \le O(p_n^{(2+\delta)/2}/n^{\delta/2}) = o(1)$. Hence

$$\sqrt{n}U_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I}_{\mathsf{k}}). \tag{S2.12}$$

From the proof of Proposition 1, it can be concluded that $\|\widehat{U}_n - U_n\| = o_P(1)$ and that the eigenvalues of $\widehat{U}_n$ are uniformly bounded away from 0 and ∞ with probability tending to one. Consequently,

$$\|\widehat{U}_n^{-1/2}U_n^{1/2} - \mathbf{I}_{\mathsf{k}}\| = o_{P}(1). \tag{S2.13}$$

Combining (S2.12), (S2.13) and Slutsky's theorem completes the proof that $\sqrt{n}\widehat{U}_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0},\mathbf{I}_{\mathsf{k}})$. ■

We now show Theorem 4, which follows directly from H_0 in (4.3) and the second part of Lemma 2. This completes the proof. ■

Proof of Theorem 5

Note that W_n can be decomposed into three additive terms,

$$\begin{aligned}
I_1 &= n\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0})\}^T(A_n\widehat{V}_nA_n^T)^{-1}\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0})\},\\
I_2 &= 2n(A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0)^T(A_n\widehat{V}_nA_n^T)^{-1}\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0})\},\\
I_3 &= n(A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0)^T(A_n\widehat{V}_nA_n^T)^{-1}(A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0),
\end{aligned}$$

where $\widehat{V}_n = \widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}$. We observe that $I_1 \xrightarrow{\mathcal{L}} \chi^2_{\mathsf{k}}$ following the second part of Lemma 2; $I_3 = n(A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0)^T\mathbf{M}^{-1}(A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0)\{1 + o_P(1)\}$ by Proposition 1; $I_2 = O_P(\sqrt{n})$ by Cauchy-Schwartz inequality. Thus

$$n^{-1}I_3 \ge \lambda_{\min}(\mathbf{M}^{-1})\|A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0\|^2\{1 + o_P(1)\} = \lambda_{\max}^{-1}(\mathbf{M})\|A_n\widetilde{\boldsymbol{\beta}}_{n;0} - \boldsymbol{g}_0\|^2 + o_P(1).$$

These complete the proof for W_n. ■

Proof of Theorem 6

Following the second part of Lemma 2, we observe that $\sqrt{n}(A_n\widehat{V}_nA_n^T)^{-1/2}(A_n\widehat{\widetilde{\boldsymbol{\beta}}}-\boldsymbol{g}_0)\xrightarrow{\mathcal{L}} N(\mathbf{M}^{-1/2}\boldsymbol{c},\mathbf{I}_{\mathtt{k}})$, which completes the proof. ■

Proof of Theorem 7

We first need to show Lemma 3.

Lemma 3 *Suppose that* $(\boldsymbol{X}_n^o,Y^o)$ *follows the distribution of* $(\boldsymbol{X}_n,Y)$ *and is independent of the training set* $\mathcal{T}_n$*. If* $\mathtt{Q}$ *is a* BD*, then*

$$E\{\mathtt{Q}(Y^o,\widehat{m}(\boldsymbol{X}_n^o))\}=E\{\mathtt{Q}(Y^o,m(\boldsymbol{X}_n^o))\}+E\{\mathtt{Q}(m(\boldsymbol{X}_n^o),\widehat{m}(\boldsymbol{X}_n^o))\}.$$

Proof: Let $\mathtt{q}$ be the generating function of $\mathtt{Q}$. Then

$$\begin{aligned}\mathtt{Q}(Y^o,\widehat{m}(\boldsymbol{X}_n^o))=&[\mathtt{q}(m(\boldsymbol{X}_n^o))-E\{\mathtt{q}(Y^o)\mid\mathcal{T}_n,\boldsymbol{X}_n^o\}]+[E\{\mathtt{q}(Y^o)\mid\mathcal{T}_n,\boldsymbol{X}_n^o\}\\&-\mathtt{q}(Y^o)]-\mathtt{q}(m(\boldsymbol{X}_n^o))+\mathtt{q}(\widehat{m}(\boldsymbol{X}_n^o))+\{Y^o-\widehat{m}(\boldsymbol{X}_n^o)\}\mathtt{q}'(\widehat{m}(\boldsymbol{X}_n^o)).\end{aligned}\quad\text{(S2.14)}$$

Since $(\boldsymbol{X}_n^o,Y^o)$ is independent of $\mathcal{T}_n$, we deduce from Chow and Teicher (1989, Corollary 3, p. 223) that

$$E\{\mathtt{q}(Y^o)\mid\mathcal{T}_n,\boldsymbol{X}_n^o\}=E\{\mathtt{q}(Y^o)\mid\boldsymbol{X}_n^o\}.\quad\text{(S2.15)}$$

Similarly,

$$E\{Y^o\mathtt{q}'(\widehat{m}(\boldsymbol{X}_n^o))\mid\mathcal{T}_n,\boldsymbol{X}_n^o\}=E(Y^o\mid\boldsymbol{X}_n^o)\mathtt{q}'(\widehat{m}(\boldsymbol{X}_n^o))=m(\boldsymbol{X}_n^o)\mathtt{q}'(\widehat{m}(\boldsymbol{X}_n^o)).\quad\text{(S2.16)}$$

Applying (S2.15) and (S2.16) to (S2.14) results in

$$E\{\mathtt{Q}(Y^o,\widehat{m}(\boldsymbol{X}_n^o))\mid\mathcal{T}_n,\boldsymbol{X}_n^o\}=E\{\mathtt{Q}(Y^o,m(\boldsymbol{X}_n^o))\mid\boldsymbol{X}_n^o\}+\mathtt{Q}(m(\boldsymbol{X}_n^o),\widehat{m}(\boldsymbol{X}_n^o))$$

and thus the conclusion. ■

Now show Theorem 7. Setting $\mathtt{Q}$ in Lemma 3 to be the misclassification loss gives

$$\begin{aligned}1/2[E\{R(\widehat{\phi}_n)\}-R(\phi_{n,\mathrm{B}})]&\le E[|m(\boldsymbol{X}_n^o)-.5|\mathrm{I}\{m(\boldsymbol{X}_n^o)\le.5,\ \widehat{m}(\boldsymbol{X}_n^o)>.5\}]\\&\quad+E[|m(\boldsymbol{X}_n^o)-.5|\mathrm{I}\{m(\boldsymbol{X}_n^o)>.5,\ \widehat{m}(\boldsymbol{X}_n^o)\le.5\}]\\&=I_1+I_2.\end{aligned}$$

For any $\epsilon>0$, it follows that

$$\begin{aligned}I_1&=E[|m(\boldsymbol{X}_n^o)-.5|\mathrm{I}\{m(\boldsymbol{X}_n^o)<.5-\epsilon,\ \widehat{m}(\boldsymbol{X}_n^o)>.5\}]\\&\quad+E[|m(\boldsymbol{X}_n^o)-.5|\mathrm{I}\{.5-\epsilon\le m(\boldsymbol{X}_n^o)\le.5,\ \widehat{m}(\boldsymbol{X}_n^o)>.5\}]\\&\le P\{|\widehat{m}(\boldsymbol{X}_n^o)-m(\boldsymbol{X}_n^o)|>\epsilon\}+\epsilon\end{aligned}$$

and similarly, $I_2\le\epsilon+P\{|\widehat{m}(\boldsymbol{X}_n^o)-m(\boldsymbol{X}_n^o)|\ge\epsilon\}$. Recall that

$$|\widehat{m}(\boldsymbol{X}_n^o)-m(\boldsymbol{X}_n^o)|=|F^{-1}(\widetilde{\boldsymbol{X}_n^o}^T\widehat{\widetilde{\boldsymbol{\beta}}})-F^{-1}(\widetilde{\boldsymbol{X}_n^o}^T\widetilde{\boldsymbol{\beta}}_{n;0})|\le|(F^{-1})'(\widetilde{\boldsymbol{X}_n^o}^T\widetilde{\boldsymbol{\beta}}_n^*)|\|\boldsymbol{X}_n^o\|\|\widehat{\widetilde{\boldsymbol{\beta}}}-\widetilde{\boldsymbol{\beta}}_{n;0}\|,$$

for some $\widetilde{\boldsymbol{\beta}}_n^*$ between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}$, where $\widetilde{\boldsymbol{X}_n^o} = (1, \boldsymbol{X}_n^{o\,T})^T$. By Condition A4, we conclude that $(F^{-1})'(\widetilde{\boldsymbol{X}_n^o}^T \widetilde{\boldsymbol{\beta}}_n^*) = O_P(1)$. This along with $\|\widehat{\widetilde{\boldsymbol{\beta}}} - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(1)$ and $\|\widetilde{\boldsymbol{X}_n^o}\| = O_P(\sqrt{p_n})$ implies that $|\widehat{m}(\boldsymbol{X}_n^o) - m(\boldsymbol{X}_n^o)| = O_P(r_n\sqrt{p_n}) = o_P(1)$. Therefore $I_1 \to 0$ and $I_2 \to 0$, which completes the proof. ■

S3 Figures 7–10 in Section 6.2

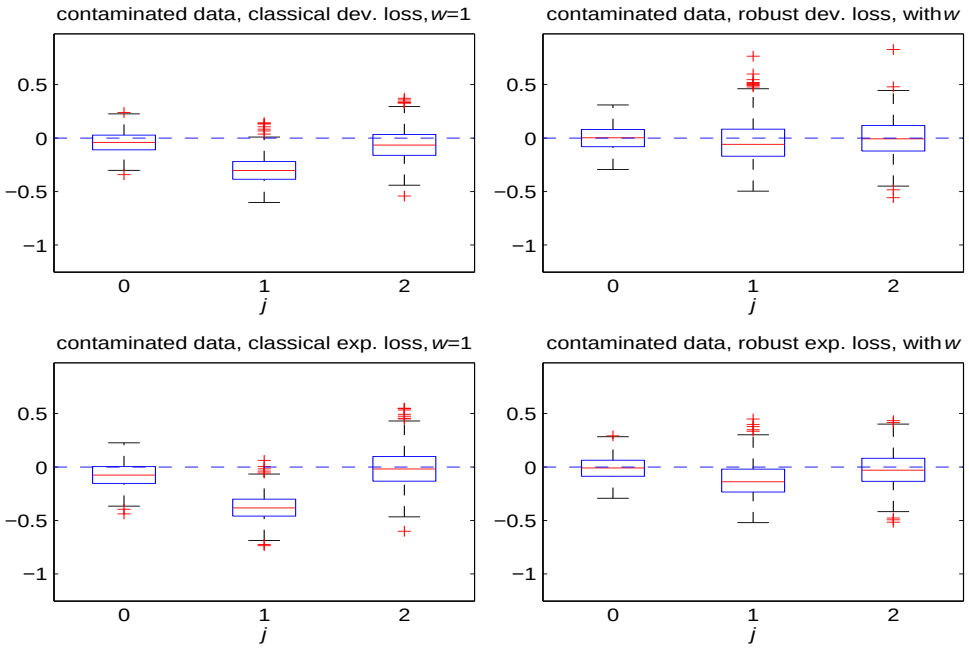

Figure 7: (Simulated Bernoulli response data with contamination) Boxplots of $\widehat{\beta}_j - \beta_{j;0}$, $j = 0, 1, \ldots, p_n$ (from left to right in each panel). Left panels: the non-robust estimates; right panels: the robust estimates.

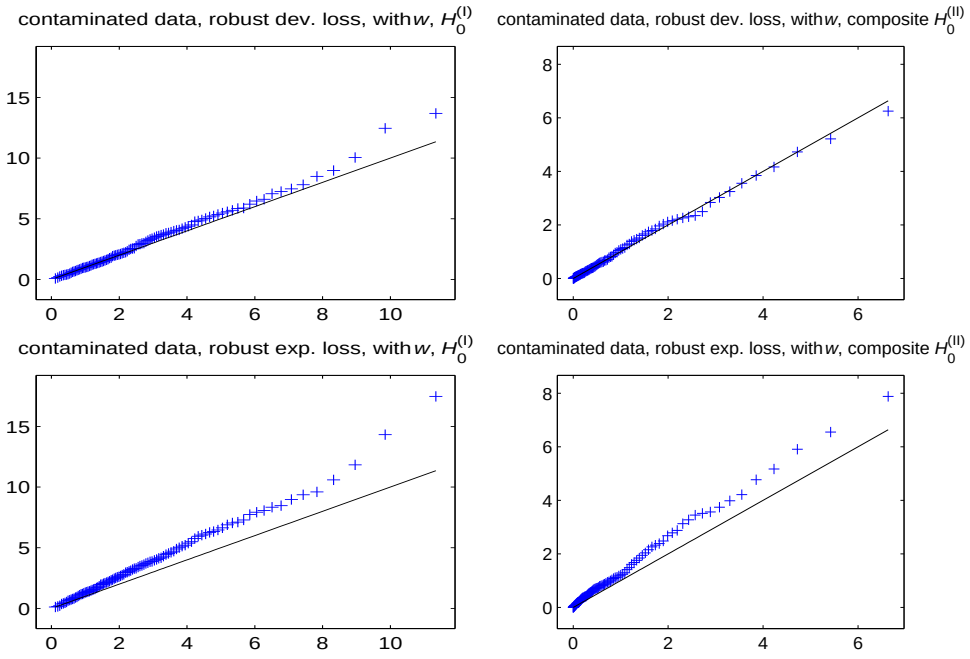

Figure 8: (Simulated Bernoulli response data with contamination) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2_{k} distribution. Solid line: the 45 degree reference line. Left panels: for testing $H_0^{(\mathrm{I})}$; right panels: for testing $H_0^{(\mathrm{II})}$.

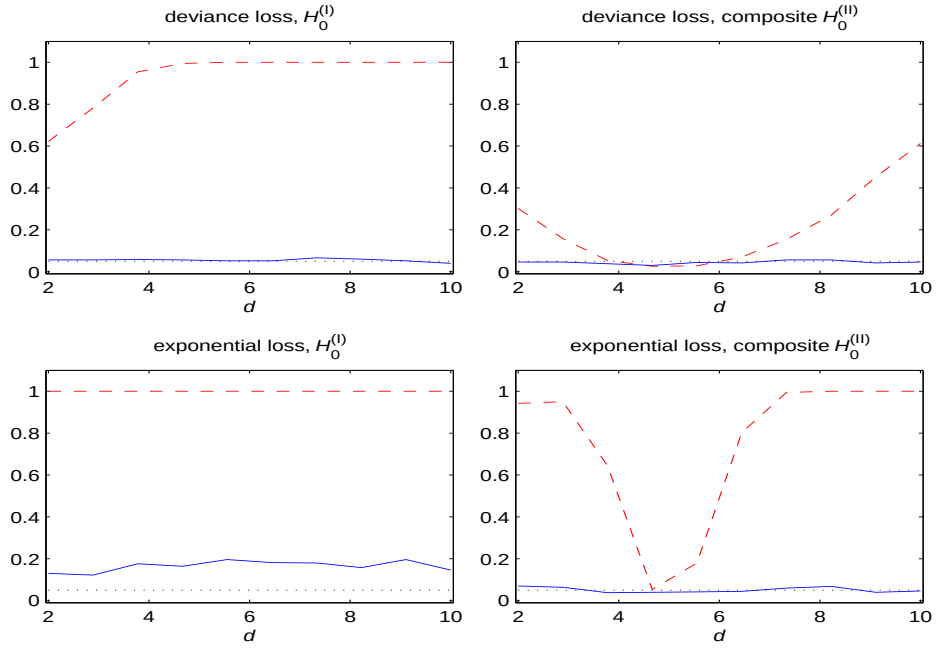

Figure 9: Level of tests for the Bernoulli response data. The dashed line corresponds to the non-robust Wald-type test; the solid line corresponds to the robust Wald-type test; the dotted line indicates the 5% nominal level. Left panels: for testing $H_0^{(\mathrm{I})}$; right panels: for testing $H_0^{(\mathrm{II})}$.

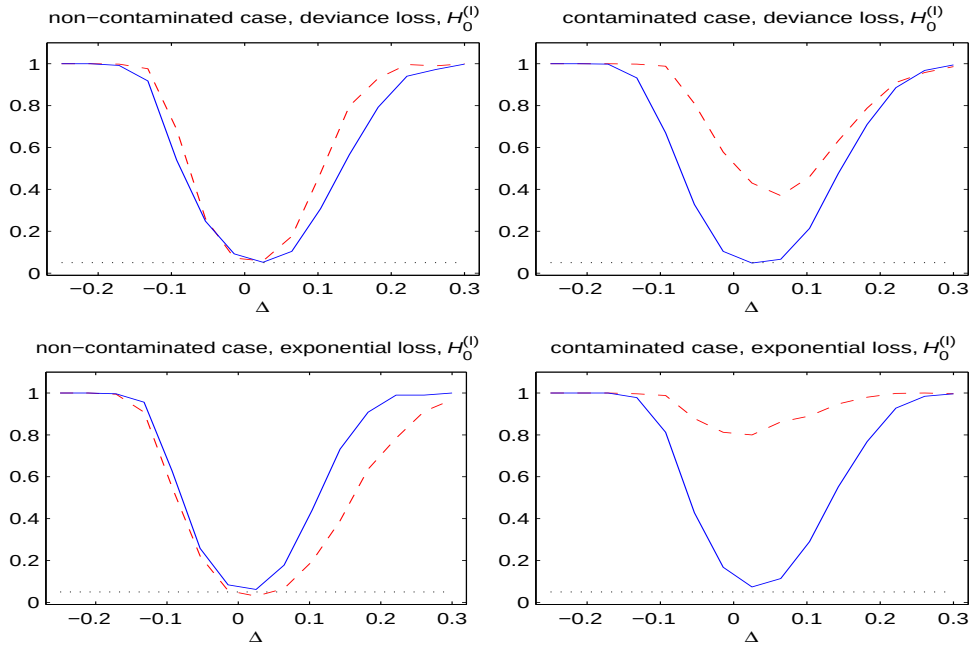

Figure 10: Observed power functions of tests for the Bernoulli response data. The dashed line corresponds to the non-robust Wald-type test; the solid line corresponds to the robust Wald-type test; the dotted line indicates the 5% nominal level. Left panels: non-contaminated case; right panels: contaminated case.