Efficient Robust Regression via Two-Stage Generalized Empirical Likelihood

Howard D. Bondell and Leonard A. Stefanski

Department of Statistics, North Carolina State University,
Box 8203, Raleigh, NC 27695, U.S.A.

Correspondence Author: Howard D. Bondell
E-mail: bondell@stat.ncsu.edu
Telephone: (919) 515-1914
Fax: (919) 515-1169

January 29, 2013

Abstract

Large- and finite-sample efficiency and resistance to outliers are the key goals of robust statistics. Although often not simultaneously attainable, we develop and study a linear regression estimator that comes close. Efficiency obtains from the estimator's close connection to generalized empirical likelihood, and its favorable robustness properties are obtained by constraining the associated sum of (weighted) squared residuals. We prove maximum attainable finite-sample replacement breakdown point, and full asymptotic efficiency for normal errors. Simulation evidence shows that compared to existing robust regression estimators, the new estimator has relatively high efficiency for small sample sizes, and comparable outlier resistance. The estimator is further illustrated and compared to existing methods via application to a real data set with purported outliers.

Key Words: Asymptotic efficiency; Breakdown point; Constrained optimization; Efficient estimation; Empirical likelihood; Exponential tilting; Least trimmed squares; Robust regression; Weighted least squares.

1 Introduction

Ordinary least squares (OLS) linear regression is a workhorse statistical method used in almost every scientific discipline. However, although OLS is fully efficient under the assumption of normally distributed errors, and best linear unbiased more generally, atypical or outlying observations can have a dramatic impact on estimated parameters. Consequently many robust alternatives to OLS have been developed with the twin goals of maintaining high efficiency when the errors are normally distributed, while also maintaining stability in the presence of outlying observations and heavy-tailed error distributions. However, robustness necessarily entails some loss of efficiency. Thus the Holy Grail of robust regression is efficiency in finite samples and asymptotically, and resistance to outliers in finite samples and heavy-tailed distributions asymptotically.

Asymptotic expansions for many robust estimators reveal that they are approximately weighted least squares estimators where the weights are implied via the form of the estimator, with small weights corresponding to outlying observations. More direct weighted least squares approaches for robust estimators are given by Ruppert and Carroll (1980), Rousseeuw and Leroy (1987), He and Portnoy (1992), Agostinelli and Markatou (1998), and Gervini and Yohai (2002). These approaches construct weights based on an initial measure of outlyingness. The method presented in this paper is also a weighted least squares approach; however, the weights are estimated directly within a generalized empirical likelihood framework.

We consider the usual linear regression model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, with independent observations for $i = 1, ..., n$, where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are both k dimensional vectors. The first component of each $\mathbf{x}$ is typically a 1 to represent an intercept, but need not be.

Early approaches to robustness based on M-estimation and generalized M-estimation (Huber, 1973; Mallows, 1975; Hampel, 1978; Krasker, 1980; Krasker and Welsch, 1982) were shown to have an asymptotic breakdown point of at most $1/(k+1)$ (Maronna,

Bustos, and Yohai, 1979; Donoho and Huber, 1983). The breakdown point is the largest fraction of data that can be contaminated while maintaining boundedness of the estimate. This notion is made precise later. The Least Median of Squares (Hampel, 1975; Rousseeuw, 1984) was the first equivariant estimator to achieve the maximum asymptotic breakdown point of 1/2. However, this estimator is only $n^{1/3}-$consistent, and hence has asymptotic relative efficiency of 0 with respect to OLS. S-estimators (Rousseeuw and Yohai, 1984) are high-breakdown, $\sqrt{n}-$consistent estimators. However, when tuned to achieve a breakdown point of 1/2, the resulting estimator is inefficient at the normal distribution. Regression estimators that can be tuned to simultaneously obtain breakdown point of 1/2 along with high efficiency were given by Yohai (1987) and Yohai and Zamar (1988), as MM-, and $\tau-$estimators, respectively. These estimators are typically tuned to obtain 95% efficiency at the normal distribution.

Ideally, an estimator would obtain full asymptotic efficiency while maintaining a breakdown point of 1/2. To this end, Agostinelli and Markatou (1998) and Gervini and Yohai (2002) each proposed a weighted least squares estimator in which weights are based on an adaptive measure of discrepancy between the empirical distribution of the errors from an initial robust estimator, and the assumed normal distribution. The discrepancies are based on either a smoothed density estimate of the residuals from the robust fit (Agostinelli and Markatou, 1998) or the empirical cumulative distribution function of these residuals (Gervini and Yohai, 2002). Both methods obtain full asymptotic efficiency while maintaining maximum breakdown. Although asymptotically efficient, the finite-sample efficiency of the weighted least squares constructed in this manner can be relatively low (see Gervini and Yohai, 2002, for example). He and Wang (1996) use a crosschecking approach to compromise between a high-breakdown estimator and a fully efficient estimator. This method is also asymptotically efficient, but its finite sample performance depends on the choice of tuning threshold.

Empirical likelihood and generalized empirical likelihood methods have been successfully used in various settings for estimation and inference (Owen, 1988, 2001; Kolaczyk, 1994; Qin and Lawless, 1994; Baggerly, 1998; Kitamura and Stutzer, 1997; Imbens, Spady, and Johnson, 1998; Lazar and Mykland, 1999; Newey and Smith, 2004; Schennach, 2007). Qin and Lawless (1994) discuss the use of an empirical likelihood approach to parameter estimation under moment restrictions. As a special case, if the moment specifications are correct and the likelihood score functions are included as a subset of the restrictions, the resulting estimator is asymptotically efficient.

Motivated by the efficiency of generalized empirical likelihood, we propose a class of empirical likelihood-type estimators that obtain robustness via moment restrictions. These estimators are shown to be $\sqrt{n}-$consistent for the true regression parameters under standard regularity conditions. In addition, the resulting estimators are fully efficient when the errors are normally distributed, while remaining highly robust under deviations from the central model. The estimators simultaneously obtain full efficiency under the normal distribution, while retaining the asymptotic breakdown point of 1/2. Furthermore, the proposed estimators are shown to obtain the maximum possible finite-sample breakdown point for any regression equivariant estimator. The proposed approach is also shown via simulation to remain highly efficient even in small samples, and it compares favorably to existing methods in both efficiency and robustness.

We define the estimator in Section 2. Computational considerations are discussed in Section 3. Finite-sample breakdown and asymptotic boundedness are studied in Section 4. Section 5 establishes consistency and asymptotic normality. Numerical results from simulation studies and applications to data are presented in Section 6, with concluding remarks given in Section 7.

2 Formulation of the Estimator

We assume the random sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ follows the linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, where the ε_i are i.i.d. random variables independent of the $\mathbf{x}_i$, with unknown distribution $F_0(\cdot/\sigma)$ for some $\sigma > 0$. Our asymptotic results assume that F_0 is unimodal and symmetric around zero, as is common in research on robust regression. However, unimodality and symmetry are not assumed for our finite-sample robustness results. To discuss robustness properties, we assume that the vector (X, Y) follows the central model if $(\mathrm{X}, Y) \sim G_0$, with $G_0(\mathbf{x}, y) = M_0(\mathbf{x})F_0\left\{\left(y - \mathbf{x}^T\boldsymbol{\beta}\right)/\sigma\right\}$, where F_0 determines the conditional distribution of Y given X, while M_0 is the marginal distribution of X. The general types of departures considered for robustness are ϵ-contamination models with $(\mathrm{X}, Y) \sim G_\epsilon$, where $G_\epsilon = (1-\epsilon)G_0 + \epsilon G$, where G is an arbitrary distribution on $\mathbb{R}^{k+1}$. Thus the distribution G_ϵ produces a fraction, ϵ, of outliers coming from G. The goal is to estimate $\boldsymbol{\beta}$, in the presence of the contamination.

Our weighted least squares estimator achieves high breakdown and high efficiency by simultaneously bounding the weighted sum of squared residuals and selecting the weights to minimize a Cressie-Read divergence (Cressie and Read, 1984). The basic idea is that weights are determined as close to equal weighting as possible (for efficiency purposes) yet still downweighting observations that do not fit the central model.

2.1 Generalized Empirical Likelihood Form

Specifically we choose $\boldsymbol{p} = (p_1, \ldots, p_n)$ and $\boldsymbol{\beta}$ to minimize

$$\sum_{i=1}^n H_\gamma(p_i) \quad \text{subject to:} \qquad \sum_{i=1}^n p_i = 1; \quad \sum_{i=1}^n p_i \boldsymbol{g}(\mathbf{x}_i, y_i, \boldsymbol{\beta}, \widetilde{\sigma}_{\mathrm{T}}^2) = \mathbf{0};$$

where:

$$H_\gamma(p) = \begin{cases} \left\{(np)^{\gamma+1} - 1\right\} / \left\{n\gamma(\gamma+1)\right\}, & \gamma \neq -1, 0 \\ -n^{-1}\log(np), & \gamma = -1 \\ p\log(np), & \gamma = 0; \end{cases}$$

$$\boldsymbol{g}(\mathbf{x}, y, \boldsymbol{\beta}, \sigma^2) = \left\{ \begin{array}{c} (y - \mathbf{x}^T\boldsymbol{\beta})\mathbf{x} \\ (y - \mathbf{x}^T\boldsymbol{\beta})^2 - \sigma^2 \end{array} \right\}; \tag{1}$$

and $\widetilde{\sigma}^2_{\mathrm{T}}$ is a *target* residual scale determined via an initial robust, *but not necessarily efficient*, fit. The resulting estimator of $\boldsymbol{\beta}$ is denoted $\widehat{\boldsymbol{\beta}}$.

First note that for any γ, if we were to remove the set of $k+1$ moment conditions, $\sum_{i=1}^n p_i \boldsymbol{g}(\mathbf{x}_i, y_i, \boldsymbol{\beta}, \widetilde{\sigma}^2_{\mathrm{T}}) = 0$, the function $\sum_{i=1}^n H_\gamma(p_i)$ takes its global minimum at $p_i = 1/n$ for all i. Now, for any vector of weights, $\boldsymbol{p}$, the first k moment conditions fully define the estimator as a weighted least squares. Furthermore, if the last moment condition were omitted, the set of weights $p_i = 1/n$ for all i coupled with the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ would then satisfy the constraints and be the solution. The addition of the final moment constraint involving the weighted residual sum of squares forces the weights to vary away from equal weighting in order to achieve the target residual scale, thus differing from ordinary (unweighted) least squares estimation.

Intuitively, outliers need to be downweighted in order to satisfy the target scale constraint. Although this is true for all choices of γ in (1), we focus exclusively on the choice of $\gamma = 0$ for reasons presented in the next subsection.

Although different from our approach, and geared towards efficiency rather than robustness, Wang and LeBlanc (2008) study an estimator defined by adding a second moment term to the loss function for regression. They show that the resulting estimator is more efficient than ordinary least squares (OLS) under asymmetric errors. However, their estimator is asymptotically equivalent to OLS under symmetric errors, and hence non-robust to heavy tailed distributions.

If the errors are independent and identically distributed $\mathrm{N}(0, \sigma_0^2)$ and $\widetilde{\sigma}_{\mathrm{T}}^2$ is replaced by σ_0^2, then $n^{-1} \sum \left\{ \left(y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right)^2 - \sigma^2 \right\} \to 0$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ satisfies the constraints asymptotically. It follows from the general theory of empirical likelihood-type estimators (Qin and Lawless, 1994; Newey and Smith, 2004) that if the errors are independent and identically distributed $\mathrm{N}(0, \sigma_0^2)$ and $\widetilde{\sigma}_{\mathrm{T}}^2$ is replaced by σ_0^2, then the resulting estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically equivalent to the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and thus fully efficient. This is a clue to the efficiency of our estimator. Of course, we do not know σ_0^2. However, we can obtain a highly-robust $\widetilde{\sigma}_{\mathrm{T}}^2$ that is $n^{1/2}$-consistent for σ_0^2 under normality, and it transpires that this is all that is needed to ensure full efficiency of $\widehat{\boldsymbol{\beta}}$ under normality, and high efficiency more generally.

It is important to note that if the linear model holds but the error distribution is not normal, then, under typical regularity conditions, $\widetilde{\sigma}_{\mathrm{T}}^2$, converges to a scale functional of the residual distribution, σ_{T}^2, that is different from the variance. In this case, although the normal-equation moment conditions are correctly specified, the second moment condition is misspecified. We explain later that this possibility relates to our decision to consider the special case of exponential tilting ($\gamma = 0$) in order to maintain $n^{1/2}$-consistency of the resulting estimator under any symmetric unimodal error distribution.

2.2 Proposed Estimator: Exponential Tilting

A few special cases of the Cressie-Read divergence are worth mentioning. For $\gamma = -1$ the method is empirical likelihood (Qin and Lawless, 1994). For $\gamma = 1$ the objective function is a chi-square discrepancy and results in what is called the *continuously updated* estimator by Hanson, Heaton, and Yaron (1996), and the *Euclidean likelihood* estimator by Owen (2001). Whereas for $\gamma = 0$ the method is known as *exponential tilting* as defined in Kitamura and Stutzer (1997), Imbens, Spady, and Johnson, (1998), and Newey and Smith (2004). For a given γ, Choi, Hall, and Presnell (2000) consid-

ered contamination neighborhoods as discrete distributions within a fixed Cressie-Read divergence neighborhood of the empirical distribution. Optimization over this class of discrete distributions gives a robust estimator whose tradeoff between efficiency and robustness is governed by the size of the neighborhood.

For the remainder of the paper we focus exclusively on *exponential tilting*, the $\gamma = 0$ case, for the following reason. As noted in Baggerly (1998), Owen (2001), and Schennach (2007), the family of Cressie-Read divergences have a natural split with respect to the so-called *implied probabilities* at the minimizer, $p_i(\tilde{\boldsymbol{\beta}})$. Positive weights are obtained when $\gamma < 0$; while $\gamma = 0$ results in non-negative weights; and $\gamma > 0$ yields implied probabilities that can be negative (the likelihood of negative implied probabilities vanishes asymptotically under correct specification of the moments, but not necessarily if the moment conditions are misspecified).

For robustness considerations, having non-negative estimated weights is natural and desirable as they have an interpretation in terms of the outlyingness of the corresponding observation. As previously mentioned, when the error distribution is not normal, the limiting scale, σ_{T}^2, is different from the variance. In this case, although the normal-equation moment conditions are correctly specified, the residual variance moment condition is misspecified and negative weights cannot be ruled out even asymptotically if $\gamma > 0$. Thus $\gamma \leq 0$ ensures non-negative weights.

It is known that when the moment constraints are misspecified, the parameter estimates, in general, converge to a population pseudo-value, but with a non-zero limit for the Lagrange multiplier. However, Schennach (2007) has shown that the empirical likelihood estimator ($\gamma = -1$) may no longer be $\sqrt{n}-$consistent for the population version in the case that the functions defining the moment conditions are unbounded in some directions unless the misspecification aligns away from the unbounded directions. Schennach conjectured that this result also holds for the remainder of the Cressie-

Read family with $\gamma < 0$. In our situation, since the last component of the moment function is $\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 - \widetilde{\sigma}_{\mathrm{T}}^2$, the moment function is unbounded in every direction except the direction $(0, ..., 0, -1)$, since the last component always remains bounded from below, while all other components are unbounded in every direction. Although if the misspecification aligns with this direction, the sub-optimal convergence result of Schennach will not apply, we may avoid this issue by choosing $\gamma \geq 0$.

Consequently, for both robustness and estimation consistency, we focus exclusively on exponential titling ($\gamma = 0$), due to it being among the class of estimators remaining $\sqrt{n}$ consistent under non-normality, while also being the only member in the class admitting only non-negative weights.

3 Computation

3.1 Lagrangian Formulation

Here we specifically discuss the case of exponential tilting. We use Lagrange multipliers to solve the constrained optimization problem in (1). Define $\boldsymbol{\Lambda} = \left(\lambda_1, \boldsymbol{\lambda}_2^T, \lambda_3\right)^T$ and

$$\begin{aligned} L_0(\boldsymbol{p}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) &= \sum_{i=1}^n p_i \log(np_i) - \lambda_1 \left(\sum_{i=1}^n p_i - 1\right) - \boldsymbol{\lambda}_2^T \sum_{i=1}^n p_i \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)\mathbf{x}_i \\ &\quad -\lambda_3 \sum_{i=1}^n p_i \left\{\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 - \widetilde{\sigma}_{\mathrm{T}}^2\right\}. \end{aligned} \tag{2}$$

Setting $\partial L_0/\partial p_i = 0,\ i = 1, \ldots, n$ and $\partial L_0/\partial \lambda_1 = 0$ and solving for p_i in terms of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_2$, and λ_3 produces

$$p_i = \frac{p_i^*}{\sum_j p_j^*} \quad \text{where} \quad p_i^* = \exp\left[\boldsymbol{\lambda}_2^T \mathbf{x}_i r_i(\boldsymbol{\beta}) + \lambda_3 \left\{r_i^2(\boldsymbol{\beta}) - \widetilde{\sigma}_{\mathrm{T}}^2\right\}\right].$$

Setting $\partial L_0/\partial \boldsymbol{\lambda}_2 = 0$, $\partial L_0/\partial \lambda_3 = 0$, and $\partial L_0/\partial \boldsymbol{\beta} = 0$ results in the equations

$$\sum_i p_i^* r_i(\boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0},$$

$$\begin{aligned}\sum_i p_i^* \left\{ r_i^2(\boldsymbol{\beta}) - \widetilde{\sigma}_{\mathrm{T}}^2 \right\} &= 0, \\ \sum_i p_i^* \left\{ \mathbf{x}_i \mathbf{x}_i^T \boldsymbol{\lambda}_2 + 2 r_i(\boldsymbol{\beta}) \mathbf{x}_i \right\} &= \mathbf{0}.\end{aligned}$$

Note that the first and third sets of equations above imply that $\boldsymbol{\lambda}_2 = \mathbf{0}$, thus we are left with the equations defining $\widehat{\lambda}_3$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{p}_i^*$

$$\begin{aligned}\sum_i \widehat{p}_i^* \left\{ r_i^2(\widehat{\boldsymbol{\beta}}) - \widetilde{\sigma}_{\mathrm{T}}^2 \right\} &= 0, \\ \sum_i \widehat{p}_i^* r_i(\widehat{\boldsymbol{\beta}}) \mathbf{x}_i &= \mathbf{0}, \\ \widehat{p}_i^* - \exp\left[\widehat{\lambda}_3 \left\{ r_i^2(\widehat{\boldsymbol{\beta}}) - \widetilde{\sigma}_{\mathrm{T}}^2 \right\} \right] &= 0, \quad \text{for} \quad i = 1, \ldots, n, \end{aligned} \tag{3}$$

from which we get $\widehat{p}_i = \widehat{p}_i^* / \sum_j \widehat{p}_j^*$.

Equations (3) reveal how the weighting works. Suppose henceforth that $\widetilde{\sigma}_{\mathrm{T}}^2 \leq \widehat{\sigma}_{\mathrm{OLS}}^2$, which can always be satisfied by suitable choice of $\widetilde{\sigma}_{\mathrm{T}}^2$, and which is a natural condition when robustness of ordinary least squares is a concern. Then because $\sum_i \widehat{p}_i = 1$ it follows that

$$\begin{aligned}\sum_i \widehat{p}_i r_i^2 \left(\widehat{\boldsymbol{\beta}} \right) = \widetilde{\sigma}_{\mathrm{T}}^2 &\leq \widehat{\sigma}_{\mathrm{OLS}}^2 \\ &= (1/n) \sum_i r_i^2 \left(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} \right) \\ &\leq (1/n) \sum_i r_i^2 \left(\widehat{\boldsymbol{\beta}} \right),\end{aligned}$$

where the last inequality follows via the properties of least squares. Because $n^{-1} \sum_i \widehat{p}_i = 1/n$, it follows that $\sum_i (\widehat{p}_i - 1/n)\, r_i^2 \left(\widehat{\boldsymbol{\beta}} \right) \leq 0$, and thus the sample covariance between the estimated probabilities $\widehat{p}_i$ and the squared residuals $r_i^2 \left(\widehat{\boldsymbol{\beta}} \right)$ is never positive and is negative when $\widetilde{\sigma}_{\mathrm{T}}^2 < \widehat{\sigma}_{\mathrm{OLS}}^2$. The same is true for the sample covariance between the non-normalized estimated probabilities $\widehat{p}_i^*$ and the squared residuals $r_i^2 \left(\widehat{\boldsymbol{\beta}} \right)$ from which it is apparent that it must be that $\widehat{\lambda}_3 \leq 0$ and strictly negative when $\widetilde{\sigma}_{\mathrm{T}}^2 < \widehat{\sigma}_{\mathrm{OLS}}^2$. Thus the method tends to assign small weights to large squared residuals and vice versa.

3.2 Alternative Characterization

We now derive an alternative characterization of the estimator that allows straightforward analysis of its asymptotic distribution and finite-sample robustness, while also allowing for a more convenient computational algorithm. This alternative characterization is based on examining the first-order conditions of the proposed optimization problem and exhibiting a saddlepoint problem with the same stationary point.

Denote the sample data by $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and define

$$\begin{aligned} J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D}) &= \frac{1}{n} \sum_i \exp\left[-\tau \left\{ r_i^2(\boldsymbol{\beta}) - \widetilde{\sigma}_{\mathrm{T}}^2 \right\}\right] \\ &= \frac{e^{\tau \widetilde{\sigma}_{\mathrm{T}}^2}}{n} \sum_i \exp\left\{-\tau r_i^2(\boldsymbol{\beta})\right\}. \end{aligned} \tag{4}$$

Note that $J_n(0, \boldsymbol{\beta} \mid \mathcal{D}) = 1$ for all $\boldsymbol{\beta}$. Also define for any fixed $\boldsymbol{\beta}$,

$$\widehat{\sigma}^2(\boldsymbol{\beta}) = n^{-1} \sum_i r_i^2(\boldsymbol{\beta}). \tag{5}$$

Note that the assumption $\widetilde{\sigma}_{\mathrm{T}}^2 \leq \widehat{\sigma}_{\mathrm{OLS}}^2$ ensures that $\widetilde{\sigma}_{\mathrm{T}}^2 \leq \widehat{\sigma}^2(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$. Finally, define $J_{n,1}(\tau, \boldsymbol{\beta} \mid \mathcal{D}) = \partial J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})/\partial\tau$; $J_{n,11}(\tau, \boldsymbol{\beta} \mid \mathcal{D}) = \partial J_{n,1}(\tau, \boldsymbol{\beta} \mid \mathcal{D})/\partial\tau$; and $J_{n,2}(\tau, \boldsymbol{\beta} \mid \mathcal{D}) = \partial J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})/\partial\boldsymbol{\beta}$.

It is easy to see that $J_{n,11}(\tau, \boldsymbol{\beta} \mid \mathcal{D}) > 0$ and $J_{n,1}(0, \boldsymbol{\beta} \mid \mathcal{D}) \leq 0$. Thus for fixed $\boldsymbol{\beta}$, $J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})$ is convex in τ and decreasing in τ at $\tau = 0$. Furthermore, inspection reveals that if $\min_i r_i^2(\boldsymbol{\beta}) < \widetilde{\sigma}_{\mathrm{T}}^2$, then $\lim_{\tau \to \infty} J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D}) = \infty$. Thus the inequalities, $\min_i r_i^2(\boldsymbol{\beta}) < \widetilde{\sigma}_{\mathrm{T}}^2 \leq \widehat{\sigma}_{\mathrm{OLS}}^2$, ensure that as a function of τ, $J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})$ assumes its minimum at some $0 \leq \widehat{\tau}(\boldsymbol{\beta}) < \infty$.

As noted previously we can always guarantee that $\widetilde{\sigma}_{\mathrm{T}}^2 \leq \widehat{\sigma}_{\mathrm{OLS}}^2$. Although $\min_i r_i^2(\boldsymbol{\beta}) < \widetilde{\sigma}_{\mathrm{T}}^2$ may not hold *for all* $\boldsymbol{\beta}$, inspection of the first equation in (3) reveals that, except for degenerate cases, the inequality holds in a neighborhood of the estimator $\widehat{\boldsymbol{\beta}}$ defined via the constrained optimization problem. Thus henceforth we use the fact $0 \leq \widehat{\tau}(\boldsymbol{\beta}) < \infty$ minimizes $J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})$ with respect to $\tau \geq 0$ for fixed $\boldsymbol{\beta}$.

Note that the $\boldsymbol{\beta}$-*profiled* function defined as

$$Q_n(\boldsymbol{\beta}) = J_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} \mid \mathcal{D}) \tag{6}$$

is a bounded continuous function of $\boldsymbol{\beta}$. Also as long as the design matrix is full rank it follows that $Q_n(\boldsymbol{\beta}) \to 0$ as $||\boldsymbol{\beta}|| \to \infty$. So $Q_n(\boldsymbol{\beta})$ attains its maximum for some finite $\boldsymbol{\beta}$ call it $\widehat{\boldsymbol{\beta}}_Q$. That is, $\widehat{\boldsymbol{\beta}}_Q$ maximizes $J_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} \mid \mathcal{D})$. Define $\widehat{\tau}_Q = \widehat{\tau}(\widehat{\boldsymbol{\beta}}_Q)$.

Except for the boundary case $\widehat{\tau}(\widehat{\boldsymbol{\beta}}) = 0$ (that we can always avoid by choosing $\widetilde{\sigma}^2_{\mathrm{T}}$ *strictly less than* $\widehat{\sigma}^2_{\mathrm{OLS}}$) note that:

Minimizing $J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})$ with respect to τ for fixed $\boldsymbol{\beta}$ results in $\widehat{\tau}(\boldsymbol{\beta})$ satisfying

$$J_{n,1}(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} \mid \mathcal{D}) \;\; = \;\; 0; \tag{7}$$

Maximizing $J_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta}|\mathcal{D})$ with respect to $\boldsymbol{\beta}$ results in $\widehat{\boldsymbol{\beta}}_Q$ satisfying

$$J_{n,1}(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_Q), \widehat{\boldsymbol{\beta}}_Q)\widehat{\tau}_2(\widehat{\boldsymbol{\beta}}_Q) + J_{n,2}(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_Q), \widehat{\boldsymbol{\beta}}_Q) \;\; = \;\; 0, \tag{8}$$

where $\widehat{\tau}_2(\boldsymbol{\beta}) = \partial\widehat{\tau}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. By definition $J_{n,1}(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_Q), \widehat{\boldsymbol{\beta}}_Q) = 0$ and thus (7) and (8) show that $\widehat{\tau}_Q$ and $\widehat{\boldsymbol{\beta}}_Q$ satisfy

$$\begin{aligned} J_{n,1}(\widehat{\tau}_Q, \widehat{\boldsymbol{\beta}}_Q|\mathcal{D}) \;\; &= \;\; 0, \\ J_{n,2}(\widehat{\tau}_Q, \widehat{\boldsymbol{\beta}}_Q|\mathcal{D}) \;\; &= \;\; \mathbf{0}. \end{aligned} \tag{9}$$

Expanding (9) and comparing to (3) reveals their equivalence, and that $\widehat{\tau}_Q = -\widehat{\lambda}_3$ and $\widehat{\boldsymbol{\beta}}_Q = \widehat{\boldsymbol{\beta}}$. Henceforth we drop the subscript 'Q' and use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\tau}$ exclusively.

Our analysis of the equivalence of (3) and (9) shows that any solution to the constrained optimization problem (1) when $\gamma = 0$ also solves the saddle-point equations (9). But it does not establish that (9) have a unique solution. However, simulation studies of estimators obtained by solving (1) and those obtained by solving (9) using alternating minimization-maximization revealed no difference in performance. The

fact that the alternating minimization-maximization algorithm is simpler and does not require constrained optimization routines adds to its appeal.

We implemented the estimator via the alternative saddlepoint characterization using R, iterating between the k-dimensional optimization of $\boldsymbol{\beta}$ and the univariate optimization of τ. Note that each optimization is unconstrained and either convex (for the minimization step) or concave (for the maximization step). The original optimization problem involves an n-dimensional nonlinear constrained optimization. Hence the alternative characterization yields a more convenient algorithm.

4 Breakdown and Boundedness

We now give conditions under which $\widehat{\boldsymbol{\beta}}$ attains the maximum achievable breakdown for a regression-equivariant estimator when $\widetilde{\sigma}^2_{\mathrm{T}}$ is suitably defined. Define $n_{\mathrm{G}} = \lfloor \frac{n+k+1}{2} \rfloor$ and assume that the data set is partitioned as

$$\mathcal{D} \;=\; \mathcal{G} \cup \mathcal{C}, \tag{10}$$

where $\mathcal{G}$ contains n_{G} "good" data points, and $\mathcal{C}$ contains $n_{\mathrm{C}} = n - n_{\mathrm{G}}$ possibly-contaminated data points. We prove that $\sup_{\mathcal{C}} ||\widehat{\boldsymbol{\beta}}|| < \infty$ where $\sup_{\mathcal{C}}$ denotes the supremum as all data points in $\mathcal{C}$ vary freely. Similarly, we write $\sum_{\mathcal{S}}$, to denote summation over the points in $\mathcal{S}$, where $\mathcal{S}$ can be $\mathcal{G}$, or $\mathcal{C}$, or either of the two sets defined in the next paragraph.

Our proof makes use of a second, related partition of the data defined in terms of squared *Least Trimmed Squares* residuals. Let $\widehat{\boldsymbol{\beta}}_{\mathrm{LTS}}$ denote Least Trimmed Squares estimator that minimizes $\sum_{i=1}^{n_{\mathrm{G}}} r^2_{(i)}(\boldsymbol{\beta})$ where $r^2_{(1)}(\boldsymbol{\beta}) \leq r^2_{(2)}(\boldsymbol{\beta}) \leq \cdots \leq r^2_{(n)}(\boldsymbol{\beta})$; see Rousseeuw (1984). The second partition we use in our proof is

$$\mathcal{D} \;=\; \mathcal{G}_{\mathrm{LTS}} \cup \mathcal{C}_{\mathrm{LTS}}, \tag{11}$$

where

$$\mathcal{G}_{\text{LTS}} = \left\{(\mathbf{x}_i, y_i) \ : \ r_i^2(\widehat{\boldsymbol{\beta}}_{\text{LTS}}) \leq r^2_{(n_{\text{G}})}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\right\} \quad \text{and} \quad \mathcal{C}_{\text{LTS}} = \mathcal{D} \setminus \mathcal{G}_{\text{LTS}}. \tag{12}$$

Our proof relies on the assumptions:

(FB1) The good data are in *general position*, thus implying that for any $k \times 1$ vector $\boldsymbol{v}$, the set $\left\{(\mathbf{x}_i, y_i) \ : \ (\mathbf{x}_i, y_i) \in \mathcal{G}, \ \text{ and } \ \mathbf{x}_i^T \boldsymbol{v} = 0\right\}$ contains at most $k-1$ points.

(FB2) $\widetilde{\sigma}^2_{\text{T}} \leq \widehat{\sigma}^2_{\text{OLS}}$.

(FB3) If $\widetilde{\sigma}^2_{\text{T}} < \widehat{\sigma}^2_{\text{OLS}}$, then $\widetilde{\sigma}^2_{\text{T}} \geq n_{\text{G}}^{-1} \sum_{\mathcal{G}_{\text{LTS}}} r_i^2(\widehat{\boldsymbol{\beta}}_{\text{LTS}})$.

(FB4) $\sup_{\mathcal{C}} \widetilde{\sigma}^2_{\text{T}} = \widetilde{\sigma}^2_{\text{MAX}} < \infty$.

Assumption (FB1) is standard in high-breakdown proofs. Assumptions (FB2) and (FB3) can always be satisfied. For example, define the Least Trimmed Squares variance estimator

$$\widehat{\sigma}^2_{\text{LTS}} = \frac{K_{\text{LTS}}}{n_{\text{G}}} \sum_{\mathcal{G}_{\text{LTS}}} r_i^2(\widehat{\boldsymbol{\beta}}_{\text{LTS}}), \tag{13}$$

where K_{LTS} is chosen to obtain consistency at the normal distribution and incorporates the finite-sample correction factor given in Pison et al. (2002), from which it follows that $K_{\text{LTS}} > 1$. Finally, define

$$\widetilde{\sigma}^2_{\text{T}} \ = \ \min\left(\widehat{\sigma}^2_{\text{LTS}}, \ \widehat{\sigma}^2_{\text{OLS}}\right). \tag{14}$$

Clearly (FB2) is satisfied, and (FB3) is satisfied by the fact that $K_{\text{LTS}} > 1$. Note that (FB3) also implies the existence of a $\boldsymbol{\beta}$ such that $r^2_{(1)}(\boldsymbol{\beta}) < \widetilde{\sigma}^2_{\text{T}}$, and thus the implicit function $\widehat{\tau}(\boldsymbol{\beta})$ defined in Section 3 is always finite. Finally, (FB4) simply states the fact that $\widetilde{\sigma}^2_{\text{T}}$ must be a high-breakdown estimator. A proof of the following theorem appears in the appendix.

THEOREM 1 (HIGH BREAKDOWN) *Under conditions given in (FB1)-(FB4),* $\sup_{\mathcal{C}} ||\widehat{\boldsymbol{\beta}}||$ *is finite. Hence the finite sample replacement breakdown point is* $n - \lfloor \frac{n+k+1}{2} \rfloor$*, which is the maximum attainable by an equivariant estimator.*

We end this section by noting that a proof similar to that of Theorem 1 shows that $\widehat{\boldsymbol{\beta}}$ is asymptotically bounded, which in turn allows for the proof of consistency in the next section. Conditions for asymptotic boundedness and its proof appear in the appendix.

5 Consistency and Asymptotic Normality

We now assume the data are distributed as $Y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + \epsilon_i$, $i = 1 \ldots, n$, where $(\mathbf{x}_i, \epsilon_i)$ are independent and identically distributed pairs with $\mathbf{x}_1$ independent of ϵ_1. We show that if the distribution of ϵ_1 is unimodal and symmetric around the origin then $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$.

5.1 Consistency

The asymptotic boundedness established in the appendix shows that $\widehat{\boldsymbol{\beta}}$ is confined to a compact set from which it follows that every sequence contains a convergent subsequence. Thus consistency follows once uniqueness of those limits is established. The latter is guaranteed provided that the finite-sample objective function $Q_n(\boldsymbol{\beta})$ in (6) converges pointwise to its population version $Q(\boldsymbol{\beta})$ having a unique maximum. In order to bypass some uninformative technical details that arise when $\widetilde{\sigma}_{\mathrm{T}}^2 / \widehat{\sigma}_{\mathrm{OLS}}^2 \xrightarrow{p} 1$, we work under the assumption that the target residual variance is defined as

$$\widetilde{\sigma}_{\mathrm{T}}^2 \;=\; \min\left(\widetilde{\sigma}_{\mathrm{LTS}}^2,\; C\widehat{\sigma}_{\mathrm{OLS}}^2\right), \tag{15}$$

where $C < 1$ is fixed positive constant. This mathematical 'trick' avoids dealing with special techniques for handling the case $\widehat{\tau} \xrightarrow{p} 0$ resulting when $\widetilde{\sigma}^2_{\mathrm{T}}/\widehat{\sigma}^2_{\mathrm{OLS}} \xrightarrow{p} 1$. This is an important case of course, because it arises when the estimator is asymptotically equivalent to ordinary least squares, as seen by examining the original optimization problem (1). However, we know the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, and hence can immediately deduce the distribution of the proposed estimator. Alternatively we can capture the results for the case that the estimator is asymptotically equivalent to least squares by letting $C \to 1$.

We also assume there exists a positive constant σ^2_{T} such that $\{\widetilde{\sigma}^2_{\mathrm{T}} - \sigma^2_{\mathrm{T}}\} = o_p(1)$. These assumptions ensure pointwise convergence in probability of $J(\tau, \boldsymbol{\beta})$ and $Q_n(\boldsymbol{\beta})$.

The population versions of the key inequalities in (33) are

$$J(\tau, \boldsymbol{\beta}_*) \geq J(\tau_*, \boldsymbol{\beta}_*) \geq J(\tau(\boldsymbol{\beta}), \boldsymbol{\beta}), \tag{16}$$

where $\tau(\boldsymbol{\beta})$ is the population version of $\widehat{\tau}(\boldsymbol{\beta})$, $\boldsymbol{\beta}_*$ maximizes $Q(\boldsymbol{\beta}) = J(\tau(\boldsymbol{\beta}), \boldsymbol{\beta})$, and $\tau_* = \tau(\boldsymbol{\beta}_*)$.

Consider that

$$J(\tau, \boldsymbol{\beta}) = E\left\{E\left(\exp\left[-\tau\left\{r^2(\boldsymbol{\beta}) - \sigma^2_{\mathrm{T}}\right\}\right] \mid \mathbf{x}\right)\right\}$$

where

$$r^2(\boldsymbol{\beta}) = \left\{\epsilon_1 + \mathbf{x}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\right\}^2.$$

A symmetric unimodal error distribution is such that conditioned on $\mathbf{x}$, $r^2(\boldsymbol{\beta})$ is stochastically greater than $r^2(\boldsymbol{\beta}_0)$, thus the inner conditional expectation above is bounded above by

$$E\left(\exp\left[-\tau\left\{r^2(\boldsymbol{\beta}_0) - \sigma^2_{\mathrm{T}}\right\}\right] \mid \mathbf{x}\right).$$

It follows that for each fixed τ, $J(\tau, \boldsymbol{\beta}_*) < J(\tau, \boldsymbol{\beta}_0)$. Since this upper bound holds for all τ, it holds for $\tau_0 = \tau(\boldsymbol{\beta}_0)$ leading to the inequalities

$$J(\tau_0, \boldsymbol{\beta}_0) > J(\tau, \boldsymbol{\beta}_*) \geq J(\tau_*, \boldsymbol{\beta}_*) \geq J(\tau(\boldsymbol{\beta}), \boldsymbol{\beta}), \tag{17}$$

It follows that

$$Q(\boldsymbol{\beta}_0) = J(\tau_0, \boldsymbol{\beta}_0) > J(\tau_*, \boldsymbol{\beta}_*) = Q(\boldsymbol{\beta}_*).$$

In other words $Q(\boldsymbol{\beta})$ is maximized at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

5.2 Asymptotic Normality

We now sketch the proof of asymptotic normality under the additional assumption

(AN1) There exists a positive constant σ_{T}^2 such that $\sqrt{n}\,\{\widetilde{\sigma}_{\mathrm{T}}^2 - \sigma_{\mathrm{T}}^2\} = O_p(1)$.

Let $\lambda \in R^k$ be fixed but otherwise arbitrary. Then, since $\widetilde{\sigma}_{\mathrm{T}}^2 < \widehat{\sigma}_{\mathrm{OLS}}^2$, from (9) it follows that

$$0 = \exp\{-\widehat{\tau}\widetilde{\sigma}_{\mathrm{T}}^2\}\; J_{n,1}(\widehat{\tau}, \widehat{\boldsymbol{\beta}}) \stackrel{\mathrm{def}}{=} g(\widehat{\tau}, \widehat{\boldsymbol{\beta}}, \widetilde{\sigma}_{\mathrm{T}}^2), \tag{18}$$

$$0 = \widehat{\tau}^{-1} \exp\{-\widehat{\tau}\widetilde{\sigma}_{\mathrm{T}}^2\}\; \lambda^T J_{n,2}(\widehat{\tau}, \widehat{\boldsymbol{\beta}}) \stackrel{\mathrm{def}}{=} h(\widehat{\tau}, \widehat{\boldsymbol{\beta}}, \widetilde{\sigma}_{\mathrm{T}}^2). \tag{19}$$

Both g and h are scalar-valued and thus by the Mean Value Theorem

$$0 = g(\tau_0, \boldsymbol{\beta}_0, \widetilde{\sigma}_{\mathrm{T}}^2) + \widetilde{g}_\tau()(\widehat{\tau} - \tau_0) + \widetilde{g}_{\boldsymbol{\beta}}^T()(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \widetilde{g}_{\sigma_{\mathrm{T}}^2}()(\widetilde{\sigma}_{\mathrm{T}}^2 - \sigma_{\mathrm{T}}^2) \tag{20}$$

$$0 = h(\tau_0, \boldsymbol{\beta}_0, \widetilde{\sigma}_{\mathrm{T}}^2) + \widetilde{h}_\tau()(\widehat{\tau} - \tau_0) + \widetilde{h}_{\boldsymbol{\beta}}^T()(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \widetilde{h}_{\sigma_{\mathrm{T}}^2}()(\widetilde{\sigma}_{\mathrm{T}}^2 - \sigma_{\mathrm{T}}^2) \tag{21}$$

where $\widetilde{g}_\tau()$ denotes the partial derivative of g with respect to its first argument evaluated at a point, call it $(\widetilde{\tau}_g, \widetilde{\boldsymbol{\beta}}_g, \widetilde{\sigma}_{T,g}^2)$, lying on the line segment joining $(\widehat{\tau}, \widehat{\boldsymbol{\beta}}, \widetilde{\sigma}_{\mathrm{T}}^2)$ and $(\tau_0, \boldsymbol{\beta}_0, \sigma_{\mathrm{T}}^2)$. Similarly for the other partial derivatives of g and h. Although $(\widetilde{\tau}_g, \widetilde{\boldsymbol{\beta}}_g, \widetilde{\sigma}_{T,g}^2)$ and $(\widetilde{\tau}_h, \widetilde{\boldsymbol{\beta}}_h, \widetilde{\sigma}_{T,h}^2)$ differ, it matters only that both converge to the same limit $(\tau_0, \boldsymbol{\beta}_0, \sigma_T^2)$. Note that $\widetilde{h}_{\sigma_{\mathrm{T}}^2}() \equiv 0$.

Manipulating these equations reveals that

$$\begin{aligned}\left\{\widetilde{h}_\beta^T() - \frac{\widetilde{h}_\tau()}{\widetilde{g}_\tau()}\widetilde{g}_\beta^T()\right\}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) &= -h(\tau_0, \boldsymbol{\beta}_0, \widetilde{\sigma}_{\mathrm{T}}^2) \\ &\quad + \frac{\widetilde{h}_\tau()}{\widetilde{g}_\tau()}\left\{g(\tau_0, \boldsymbol{\beta}_0, \widetilde{\sigma}_{\mathrm{T}}^2) + \widetilde{g}_{\sigma_{\mathrm{T}}^2}()\left(\widetilde{\sigma}_{\mathrm{T}}^2 - \sigma_{\mathrm{T}}^2\right)\right\}. \end{aligned} \tag{22}$$

We note the following facts that are readily verified:

1.
$$\begin{aligned}\sqrt{n}\, h(\tau_0, \boldsymbol{\beta}_0, \sigma_{\mathrm{T}}^2) &= n^{-1/2} \sum_{i=1}^{n} \exp\left(-\tau_0 \epsilon_i^2\right) \epsilon_i \lambda^T \mathbf{x}_i \\ &\xrightarrow{d} \mathrm{N}\left(\mathbf{0},\ \lambda^T \boldsymbol{\Omega}_h \lambda\right) \end{aligned} \tag{23}$$

where $\boldsymbol{\Omega}_h = E\left\{\exp(-2\tau_0\epsilon^2)\epsilon^2\right\} E\left(\mathbf{x}\mathbf{x}^T\right)$;

2.
$$\begin{aligned}\widetilde{h}_\tau() &= n^{-1} \sum_{i=1}^{n} \exp\left(-\widetilde{\tau}_h \widetilde{r}_{h,i}^2\right) \widetilde{r}_{h,i}^3 \boldsymbol{\lambda}^T \mathbf{x}_i \\ &= o_p(1) \end{aligned} \tag{24}$$

where $\widetilde{r}_{h,i} = y_i - \mathbf{x}_i^T \widetilde{\boldsymbol{\beta}}_h$;

3.
$$\begin{aligned}\widetilde{g}_\tau() &= n^{-1} \sum_{i=1}^{n} \exp\left(-\widetilde{\tau}_g \widetilde{r}_{g,i}^{\,2}\right) \left(\widetilde{r}_{g,i}^{\,4} - r_{g,i}^{\,2} \widetilde{\sigma}_{\mathrm{Tg}}^2\right) \\ &\xrightarrow{p} E\left\{\exp(-\tau_0\epsilon^2)\left(\epsilon^4 - \epsilon^2 \sigma_{\mathrm{T}}^2\right)\right\} > 0, \end{aligned} \tag{25}$$

where $\widetilde{r}_{g,i} = y_i - \mathbf{x}_i^T \widetilde{\boldsymbol{\beta}}_g$;

4.
$$\sqrt{n}\, g(\tau_0, \boldsymbol{\beta}_0, \sigma_{\mathrm{T}}^2) = O_p(1); \tag{26}$$

5.
$$\widetilde{g}_\beta() = O(1); \tag{27}$$

6.
$$\begin{aligned}\widetilde{h}_\beta() &= -n^{-1} \sum_{i=1}^{n} \exp\left(-\widetilde{\tau}_h \widetilde{r}_{i,h}^2\right) \left\{1 - 2\widetilde{\tau}_h r_{i,h}^2\right\} \mathbf{x}_i \mathbf{x}_i^T \boldsymbol{\lambda} \\ &\xrightarrow{p} -E\left[\exp\left(-\tau_0\epsilon^2\right)\left\{1 - 2\tau_0\epsilon^2\right\}\right] E\left(\mathbf{x}\mathbf{x}^T\right) \boldsymbol{\lambda}. \end{aligned} \tag{28}$$

Since these results hold for any λ, it follows from Slutsky's Theorem and the Cramer-Wold device that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right),$$

where

$$\boldsymbol{V} = \frac{E\left\{\exp\left(-2\tau_0\epsilon^2\right)\epsilon^2\right\}}{\left[E\left\{\exp\left(-\tau_0\epsilon^2\right)\left(1 - 2\tau_0\epsilon^2\right)\right\}\right]^2} \left\{E\left(\mathbf{x}\mathbf{x}^T\right)\right\}^{-1}. \tag{29}$$

A natural estimator of $\boldsymbol{V}$ is $\widehat{\boldsymbol{V}} = \widehat{v}^2 \left(n^{-1} \sum_i^2 \mathbf{x}_i \mathbf{x}_i^T\right)^{-1}$ where

$$\widehat{v}^2 = \frac{n^{-1} \sum_1^n \exp(-2\widehat{\tau}\widehat{r}_i^2)\widehat{r}_i^2}{\left\{n^{-1} \sum_1^n \exp(-2\widehat{\tau}\widehat{r}_i^2)(1 - 2\widehat{\tau}\widehat{r}_i^2)\right\}^2}, \tag{30}$$

where $\widehat{r}_i = y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$.

5.2.1 Efficiency at the Normal-Error Model

The scalar multiplier of $E\left(\mathbf{x}\mathbf{x}^T\right)$ in (29) can be evaluated analytically for the normal-error model. Doing so is instructive for it reveals the manner in which the 'trick' constant C manifests itself.

For the normal-error Gauss-Markov model with error variance σ_0^2, both $\widetilde{\sigma}^2_{\text{LTS}}$ and $\widehat{\sigma}^2_{\text{OLS}}$ converge in probability to σ_0^2, and thus $\widetilde{\sigma}^2_{\text{T}}$ as defined in (15) converges to $C\sigma_0^2$. Tedious but straightforward calculation reveals that for normal errors

$$\tau_0 = \frac{1-C}{2C\sigma^2},$$

and that the efficiency of $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is

$$\text{eff}(C) = \left(2C - C^2\right)^{3/2}.$$

Note that taking $C = 1$ results in $\tau_0 = 0$ and full efficiency, $\text{eff}(1) = 1$.

6 Numerical Results

6.1 Simulation Study

We present results from a simulation study designed to investigate the finite-sample efficiency and robustness of the new estimator. Comparisons are made with three well-known, asymptotic high-breakdown (= 1/2) regression estimators: S-estimator using Tukey's biweight function (Rousseeuw and Yohai, 1984); MM-estimator (Yohai, 1987) tuned to 95% asymptotic efficiency for normal errors; and the asymptotically efficient REweighted Least Squares estimator (REWLS) of Gervini and Yohai (2002). For the initial variance estimator required by our estimator we used $\widehat{\sigma}^2_{\text{LTS}}$. Although the focus was on the use of exponential tilting, we also included the use of empirical likelihood for comparison.

6.1.1 Efficiency

For assessing efficiency data were generated according to linear models with $k-1$ independent and identically distributed $N(0,1)$ predictors ($k = 2, 5, 24$) and sample sizes $n = 50, 100, 200, 500$. Because all of the estimators studied are regression equivariant, without loss of generality, all slopes and intercepts were set to 0. Error distributions studied included the normal and the t-distribution with three degrees of freedom (t_3). For each combination of k, n, and error distribution, 500 data sets were generated. Efficiency was assessed by the ratio of total mean squared errors (MSE)

$$\text{EFF} = \frac{\sum_{j=1}^{500} ||\widehat{\boldsymbol{\beta}}_j^{MLE}||^2}{\sum_{i=1}^{500} ||\widetilde{\boldsymbol{\beta}}_j||^2}, \tag{31}$$

where $\widehat{\boldsymbol{\beta}}_j^{MLE}$ and $\widetilde{\boldsymbol{\beta}}_j$ are the distribution-specific maximum likelihood estimator and the study estimator, respectively.

*** Table 1 goes here ***

*** Table 2 goes here ***

*** Table 3 goes here ***

Table 1 displays efficiency results for simple linear regression ($k = 2$), Table 2 for $k = 5$, and Table 3 for the high-dimensional case $k = 24$. A fair summary of the results is that the proposed exponential tilting (ET) estimator is highly efficient even with sample sizes as low as $n = 20$ or $n = 50$. Furthermore, it maintains the high efficiency even for the heavy-tailed t_3 error distribution. We note that in all cases the bias of each estimator is negligible relative to the variance, and hence the MSE almost exclusively captures the variance comparisons.

It is noteworthy that although both REWLS and the ET estimator are asymptotically efficient under the normal distribution, the results indicate that in finite samples

REWLS is farther from full efficiency than the ET estimator. The finite-sample deficiency of both ET and REWLS is more pronounced for larger k, but ET is much closer to fully efficient than REWLS for all k. Efficiencies for ET with $n = 50$ at $k = 2$, 5, and 24 are 99.2%, 96.5%, and 93.8%; whereas those for REWLS are 82.7%, 66.4%, and 36.9%. Comparable efficiencies for REWLS are reported in Gervini and Yohai (2002).

The approach based on using empirical likelihood is also shown for comparison, although the theory developed does not apply. While the efficiency results for the EL approach is slightly better than that for the exponential tilting, we shall see that under contamination, it is less robust.

Finally we note that the efficiencies in Table 3 show that the three benchmark estimators (S, REWLS, MM) are much more adversely impacted by the high dimension than the proposed ET estimator for both error distributions.

Although our theory on consistency and asymptotic normality of the proposed estimator assumes that the error distribution is symmetric and unimodal, we also compared efficiency for the exponential tilting proposal under asymmetric error distributions. Table 4 shows comparison of relative efficiency with respect to maximum likelihood under asymmetric Laplace distributions with varying degrees of skewness. Although the asymptotic theory does not apply in this case, the results for the proposed estimator shows much higher efficiency than OLS and is similar to the MM estimator.

*** Table 4 goes here ***

6.1.2 Robustness

We compared robustness of the estimators for the case of $n = 100$, $k = 5$, and normal errors using a contamination design similar to that in Gervini and Yohai (2002) wherein a percentage ($q = 5\%, 10\%, 20\%$) of randomly selected data points are replaced by outliers. The replacement outliers were of the form $(\mathbf{x}_*, y^*)$ for $\mathbf{x}_* = x_0\mathbf{1_{k-1}}/\sqrt{k-1}$,

with x_0 taking values of $1, 2, 3$, and 5 to examine a range of leverage points from low to high; and y^* varied on the grid $\{0.1jx_0 : j \text{ is a positive integer}\}$. For each estimator $\widetilde{\boldsymbol{\beta}}$ and each value of (q, x_0, y^*), the mean squared error, $MSE(\tilde{\beta}, q, x_0, y^*)$, was estimated based on 500 replications.

The same estimators as before are included in the robustness study. However, we omitted OLS as its contamination bias grows considerably larger than the rest.

*** Table 5 goes here ***

Table 5 reports $\max_{y^*} MSE(\tilde{\beta}, q, x_0, y^*)$, the worst performance of each estimator for a given contamination proportion and leverage value. In general the generalized empirical likelihood approach is competitive with the previously proposed estimators in terms of worst-case finite-sample robustness. The proposed estimators are generally more robust for low to moderate leverage, while typically worse for larger leverage. However, when also considering the high efficiency of the generalized empirical likelihood approach, the proposed estimators have an overall excellent performance and stability.

Exponential tilting was more robust in terms of contamination bias than was empirical likelihood, particularly under more extreme contamination. Comparisons between empirical likelihood and exponential tilting show that for 5% contamination, both maximum MSE values level off as the leverage point x_0 is increased. For exponential tilting, this value is approximately 15.25, whereas for empirical likelihood it is approximately 30.55.

The favorable robustness of exponential tilting over empirical likelihood can be anticipated from the form of the implied weight functions. For exponential tilting the weights decay exponentially in the squared residuals, whereas for empirical likelihood they decay linearly. The linear decay makes it difficult for empirical likelihood to sufficiently downweight extreme points without also downweighting observations in the

center of the distribution.

6.1.3 Variance Estimation and Confidence Intervals

We used data sets generated for the efficiency study to examine the finite-sample performance of the variance matrix estimator $\widehat{\boldsymbol{V}}$ in (30). Table 6 displays square-root trace ratios calculated as

$$\left\{ \frac{\text{Trace}(\overline{\widehat{\boldsymbol{V}}}_{\bullet})}{\text{Trace}(\widehat{\boldsymbol{V}}_{\text{MC}})} \right\}^{1/2} \tag{32}$$

where $\overline{\widehat{\boldsymbol{V}}}_{\bullet}$ is the average of 500 variance matrix estimates calculated via (30), and $\widehat{\boldsymbol{V}}_{\text{MC}}$ is the empirical (Monte Carlo) variance matrix calculated from the 500 replicate ET estimates. Except for cases of low sample size and high-dimension, the ratios are acceptably close to unity.

*** Table 6 goes here ***

In addition, we compared the coverage and width of 95% confidence intervals constructed using the asymptotic normality of our proposed estimator along with that of the MM-estimator, and the OLS estimator. Table 7 shows the actual coverage and the average width of the interval for a slope parameter for the multiple regression case with $k = 5$ under errors having normal distributions as well as t-distributions with 3 df.

*** Table 7 goes here ***

It is clear that the asymptotic distribution gives very close to nominal coverage even at $n = 50$ observations for both the proposed estimator and the MM-estimator. Note that the OLS intervals are exact intervals under normality, but asymptotic under the t-distribution. Furthermore, the width of the intervals for the proposed ET estimator are shorter than the MM-estimator in both distributional settings in almost all cases.

In addition, the intervals for the ET estimator are shorter than that of OLS under the t-distribution. Under normality, the OLS intervals are slightly wider, but that is likely due to the fact that the coverage is slightly higher.

6.1.4 Summary of Numerical Studies

The relatively uniformly high efficiency in finite samples of the exponential-tilting estimator is perhaps its most compelling feature. Furthermore, it achieves this while sacrificing little finite-sample resistance to outliers. Overall, the ET estimator's performance in terms of robustness, asymptotically and in finite samples, along with its asymptotic efficiency and high efficiency even in relatively small samples, shows that it is a valuable addition to the regression toolbox.

6.2 Real Data Analysis

We now examine the performance of the ET estimator on the aircraft data of Gray (1985). The data consist of five measured characteristics on each of 23 single-engine aircraft built in the years 1947-1979. The response is the Aircraft cost and the predictor variables are Aspect ratio, Lift-to-Drag ratio, Weight, and Thrust. Because these data are well documented as containing one extreme outlier, and other possibly moderate outliers and leverage points (Gray, 1985; Rousseeuw and Leroy, 1987; Pison et al., 2002), they provide a good basis for illustrating robust estimators in a real application. We calculated the same regression estimators used in the simulation study.

All of the robust regression estimates differ significantly from the normal-error MLE. The resulting estimates are in Table 8. It is clear that the proposed estimator is similar to the MM-estimator for the Aircraft data, although there are some differences, particularly in the statistical significance of the Thrust variable. Meanwhile, the S-estimation and REWLS result in estimates that differ markedly from the others. Further insight

into this result is accomplished by examining the resulting weights assigned to each of the 23 observations by the various robust estimators.

*** Table 8 goes here ***

Figure 1 plots the weights for each observation for the Exponential Tilting estimator, along with the MM-estimator (left panel). The right panel plots the weights given by the S-estimator and the REWLS estimator. Note that the particular choice of REWLS used is that given by Gervini and Yohai (2002) with 0/1 weights based on the initial S-estimator.

*** Figure 1 goes here ***

While all estimators heavily downweight observation #22, they differ in how the other observations are handled. The ET estimator and MM-estimator give relatively larger weight to all of the other observations. Meanwhile, the S-estimator significantly downweights numerous observations, while the REWLS also gives zero weight to observation #16. This additional downweighting by the S-estimator and REWLS is consistent with their lower efficiencies in the efficiency simulation study.

7 Conclusion

The impressive theoretical properties and finite-sample performance of the robust regression estimator introduced herein show the power of exponentially-tilted likelihood for tailoring estimators to achieve specific objectives. Our objectives were high breakdown and high efficiency, and we succeeded in developing an estimator that has finite-sample high breakdown, resistance to heavy-tailed error distributions asymptotically, and high efficiency both in finite samples and asymptotically. The proposed robust regression estimator exhibits highly competitive performance to existing robust methods,

and thus are a useful addition to the data analyst's tool box. The excellent performance of the estimator in finite samples may be expected due to the second-order behavior of exponentially-tilted likelihood and saddlepoint-based testing procedures (see, for example, Imbens et al., 1998; Ma and Ronchetti, 2011).

Although it appears that the use of empirical likelihood instead of exponential tilting may give better efficiency but less robustness, it remains an open question as to the conditions that will allow for $\sqrt{n}$ consistency for empirical likelihood due to the misspecification. Our simulation results show that the empirical likelihood suffers from worse bias under contamination, but whether or not the breakdown point is different remains an open question. There is a distinct difference in the forms of the weight functions, and thus it is unclear if the maximum breakdown property will hold for empirical likelihood.

To further analyze the robustness of the proposed estimator, it may be possible to derive its max bias functions along the lines of Berrendero and Zamar (2001) and Berrendero, Mendes, and Tyler (2007), but that remains an open question and a possible line of future research. The proposed estimator depends on an initial high breakdown estimate of scale. A way to tune this parameter without having to rely on this initial estimate would be desirable, however the choice of tuning method and properties of the resulting procedure are not straightforward.

The success of the exponentially-tilted, high-breakdown estimator in the homoscedastic linear regression model suggests future research adapting the approach to the heteroscedastic model via alteration of the moment constraints to incorporate either known heterogeneity, or heterogeneity estimated via variance function modeling. Another direction for future work is to enhance the estimator's robustness properties via the judicious choice of additional constraints bounding influence or leverage.

Acknowledgements

The authors are grateful to the editor, an associate editor, and three anonymous referees for their valuable comments. The authors' research was partially supported by: NSF grant DMS 1005612 and NIH grants P01-CA-142538, R01-MH-084022 (Bondell); and NSF grant DMS 0906421 and NIH grants R01-CA-085848, P01-CA-142538 (Stefanski).

8 Appendix

8.1 High Breakdown: Proof of Theorem 1

We prove high breakdown by assuming that $\sup_{\mathcal{C}} ||\widehat{\boldsymbol{\beta}}|| = \infty$ and arriving at a contradiction. By working with a subsequence (of $\widehat{\boldsymbol{\beta}}$ as points in $\mathcal{C}$ vary) if necessary, we assume without loss of generality that $\widehat{\boldsymbol{\beta}} = \boldsymbol{v}_0 + M\boldsymbol{v}_1$, with $||\boldsymbol{v}_0|| < \infty$, $||\boldsymbol{v}_1|| = 1$, and scalar M such that $\sup_{\mathcal{C}} M = \infty$. The general position assumption implies that $\mathbf{x}_i^T \boldsymbol{v}_1 = 0$ for at most $k-1$ points in $\mathcal{G}$, and hence it must be that $\sup_{\mathcal{C}} r_i^2(\widehat{\boldsymbol{\beta}}) = \infty$ for at least $n_{\mathcal{G}} - k + 1$ points in $\mathcal{G}$.

The facts that $\widehat{\boldsymbol{\beta}}$ maximizes the $\boldsymbol{\beta}$-*profiled* function $Q_n(\boldsymbol{\beta}) = J_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} \mid \mathcal{D})$, and for fixed $\boldsymbol{\beta}$, $\widehat{\tau}(\boldsymbol{\beta})$ minimizes $J_n(\tau, \boldsymbol{\beta} \mid \mathcal{D})$, imply the inequalities

$$J_n(\tau, \widehat{\boldsymbol{\beta}} \mid \mathcal{D}) \geq J_n(\widehat{\tau}, \widehat{\boldsymbol{\beta}} \mid \mathcal{D}) \geq J_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} \mid \mathcal{D}), \tag{33}$$

where $\widehat{\tau} = \widehat{\tau}(\widehat{\boldsymbol{\beta}})$. The endpoints of the key inequalities (33) show that for all $\tau > 0$ and any $\boldsymbol{\beta} \in R^k$,

$$nJ_n(\tau, \widehat{\boldsymbol{\beta}} \mid \mathcal{D}) \geq nJ_n(\widehat{\tau}(\boldsymbol{\beta}), \boldsymbol{\beta} | \mathcal{D}). \tag{34}$$

Using the fact that (34) holds for all $\boldsymbol{\beta} \in R^k$ we show, by judicious choices of $\boldsymbol{\beta}$ according to whether $\widehat{\sigma}^2_{\text{OLS}} \leq \widetilde{\sigma}^2_{\text{T}}$ or $\widehat{\sigma}^2_{\text{OLS}} > \widetilde{\sigma}^2_{\text{T}}$, that the right hand side of (34) is always bounded below by n_{G}.

For the case $\widehat{\sigma}^2_{\text{OLS}} \leq \widetilde{\sigma}^2_{\text{T}}$ replace $\boldsymbol{\beta}$ in the right hand side of (34) with $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ to get

$$\begin{aligned} nJ_n(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}), \widehat{\boldsymbol{\beta}}_{\text{OLS}} \mid \mathcal{D}) &= n\frac{1}{n}\sum_i \exp\left[\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\left\{\widetilde{\sigma}^2_{\text{T}} - r_i^2(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}\right] \\ &\geq n\exp\left[\frac{1}{n}\sum_i \widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\left\{\widetilde{\sigma}^2_{\text{T}} - r_i^2(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\}\right] \\ &= n\exp\left\{\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\left(\widetilde{\sigma}^2_{\text{T}} - \widehat{\sigma}^2_{\text{OLS}}\right)\right\} \\ &\geq n > n_{\text{G}}. \end{aligned} \tag{35}$$

The first inequality above follows from Jensen's inequality; the second from the fact that $\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\left(\widetilde{\sigma}^2_{\text{T}} - \widehat{\sigma}^2_{\text{OLS}}\right) \geq 0$; and the third is obvious.

Now consider the case $\widetilde{\sigma}^2_{\text{T}} < \widehat{\sigma}^2_{\text{OLS}}$, replacing $\boldsymbol{\beta}$ in the right hand side of (34) with $\widehat{\boldsymbol{\beta}}_{\text{LTS}}$. Then

$$\begin{aligned} nJ_n(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}}), \widehat{\boldsymbol{\beta}}_{\text{LTS}} \mid \mathcal{D}) &= \exp\left\{\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\widetilde{\sigma}^2_{\text{T}}\right\}\sum_{\mathcal{G}_{\text{LTS}}} \exp\left[-\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\text{LTS}}\right\}^2\right] \\ &\quad + \exp\left\{\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\widetilde{\sigma}^2_{\text{T}}\right\}\sum_{\mathcal{C}_{\text{LTS}}} \exp\left[-\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\text{LTS}}\right\}^2\right] \\ &> \exp\left\{\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\widetilde{\sigma}^2_{\text{T}}\right\}\sum_{\mathcal{G}_{\text{LTS}}} \exp\left[-\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\text{LTS}}\right\}^2\right] \\ &= n_{\text{G}}J_{n_{\text{G}}}\left(\widehat{\tau}(\widehat{\boldsymbol{\beta}}_{\text{LTS}}), \widehat{\boldsymbol{\beta}}_{\text{LTS}} \mid \mathcal{G}_{\text{LTS}}\right) \\ &> n_{\text{G}}J_{n_{\text{G}}}\left(\widehat{\tau}_{\text{LTS}}(\widehat{\boldsymbol{\beta}}_{\text{LTS}}), \widehat{\boldsymbol{\beta}}_{\text{LTS}} \mid \mathcal{G}_{\text{LTS}}\right) \\ &= n_{\text{G}}\frac{1}{n_{\text{G}}}\sum_{\mathcal{G}_{\text{LTS}}} \exp\left[\widehat{\tau}_{\text{LTS}}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{\widetilde{\sigma}^2_{\text{T}} - r_i^2(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\right\}\right] \\ &\geq n_{\text{G}}\exp\left[\frac{1}{n_{\text{G}}}\sum_{\mathcal{G}_{\text{LTS}}} \widehat{\tau}_{\text{LTS}}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{\widetilde{\sigma}^2_{\text{T}} - r_i^2(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\right\}\right] \\ &= n_{\text{G}}\exp\left[\widehat{\tau}_{\text{LTS}}(\widehat{\boldsymbol{\beta}}_{\text{LTS}})\left\{\widetilde{\sigma}^2_{\text{T}} - (\widehat{\sigma}^2_{\text{LTS}}/K_{\text{LTS}})\right\}\right] \\ &\geq n_{\text{G}}, \end{aligned} \tag{36}$$

where $\widehat{\tau}_{\text{LTS}}(\boldsymbol{\beta})$ minimizes $J_{n_{\text{G}}}(\tau, \boldsymbol{\beta} \mid \mathcal{G}_{\text{LTS}})$ with respect to τ for $\boldsymbol{\beta}$ fixed, and $\widehat{\sigma}^2_{\text{LTS}}$ is defined in (13). The first inequality above is straightforward; the second follows from the definition of $\widehat{\tau}_{\text{LTS}}(\boldsymbol{\beta})$; the third from Jensen's inequality; and the fourth via the facts

that $K_{\text{LTS}} > 1$ and (FB3) together ensure that $\widehat{\tau}_{\text{LTS}}(\widehat{\boldsymbol{\beta}}_{\text{LTS}}) \left\{ \widetilde{\sigma}_{\text{T}}^2 - (\widehat{\sigma}_{\text{LTS}}^2 / K_{\text{LTS}}) \right\} \geq 0$. This completes the proof that the right hand side of (34) is bounded below by n_{G}.

We now derive an upper bound for the left hand side of (34). For any $\tau > 0$,

$$\begin{aligned} nJ_n(\tau, \widehat{\boldsymbol{\beta}} \mid \mathcal{D}) &= e^{\tau \widetilde{\sigma}_{\text{T}}^2} \sum_{\mathcal{G}} \exp\left[-\tau \left\{ y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \right\}^2 \right] \\ &\quad + e^{\tau \widetilde{\sigma}_{\text{T}}^2} \sum_{\mathcal{C}} \exp\left[-\tau \left\{ y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \right\}^2 \right] \\ &< e^{\tau \widetilde{\sigma}_{\text{T}}^2} \sum_{\mathcal{G}_*} \exp\left[-\tau \left\{ y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \right\}^2 \right] + e^{\tau \widetilde{\sigma}_{\text{T}}^2} (k-1) \\ &\quad + e^{\tau \widetilde{\sigma}_{\text{T}}^2} (n - n_{\text{G}}), \end{aligned} \tag{37}$$

where

$$\mathcal{G}_* = \mathcal{G} \cap \left\{ (y_i, \mathbf{x}_i) \ : \ \mathbf{x}_i^T \boldsymbol{v}_1 \neq 0 \right\}.$$

The inequality in (37) uses the fact that $\exp\{-\tau r_i^2(\boldsymbol{\beta})\} \leq 1$, for all $\boldsymbol{\beta}$ and $\tau \geq 0$. Combining the upper and lower bounds shows that for any $\tau > 0$,

$$e^{\tau \widetilde{\sigma}_{\text{T}}^2} (n - n_{\text{G}} + k - 1) + e^{\tau \widetilde{\sigma}_{\text{T}}^2} \sum_{\mathcal{G}_*} \exp\left[-\tau \left\{ y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \right\}^2 \right] \ > \ n_{\text{G}}. \tag{38}$$

Upon taking $\sup_{\mathcal{C}}$ on the left hand side above, invoking (FB4), and noting that $\sup_{\mathcal{C}} \left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}} \right)^2 = \infty$ for all points in $\mathcal{G}_*$, shows that all of the exponential terms vanish, resulting in the inequality

$$\exp\left\{ \tau \widetilde{\sigma}_{\text{MAX}}^2 \right\} (n - n_{\text{G}} + k - 1) \geq n_{\text{G}}.$$

Because this holds for all $\tau > 0$ and because $\widetilde{\sigma}_{\text{MAX}}^2$ is finite, letting $\tau \to 0$ shows that $(n - n_{\text{G}} + k - 1) \geq n_{\text{G}}$ or equivalently that $n_{\text{G}} \leq (n + k - 1)/2$. However, $n_{\text{G}} = \lfloor \frac{n+k+1}{2} \rfloor > (n + k - 1)/2$, thus we have a contradiction. ●

8.2 Asymptotic Boundedness

The inequality (38) also allows us to prove that $\Pr\left(\limsup_n ||\widehat{\boldsymbol{\beta}}|| < \infty \right) = 1$. This fact enables the simple proof of consistency sketched in Section 5.1. We assume that as

$n \to \infty$ the data partition in (10) is such that n_{G} also diverges in a way to ensure that $n_{\text{G}} \geq \lfloor \frac{n+k+1}{2} \rfloor$ for all $n \geq k+1$. In addition we assume the following.

(AB1) The points $(\mathbf{x}_i, y_i)$ in $\mathcal{G}$ defined in (10) are independent and identically distributed from a distribution F satisfying $\Pr\left(\left|\mathbf{x}^T \boldsymbol{v}\right| = 0\right) = 0$, for all $k \times 1$ vectors $\boldsymbol{v} \neq \mathbf{0}$. We use $(\mathbf{x},\ y)$ to denote a generic random pair having distribution F.

(AB2) $\lim_{n\to\infty}(n_{\text{G}}/n) \to (1+\delta)/2$ for some $\delta > 0$.

(AB3) The target variance estimator $\widetilde{\sigma}^2_{\text{T}}$ is such that under (AB1) and (AB2), there exists a nonrandom constant $0 < \widetilde{\sigma}^2_{\infty} < \infty$, such that $\Pr\left(\sup_n \widetilde{\sigma}^2_{\text{T}} < \widetilde{\sigma}^2_{\infty}\right) = 1$.

Assumptions (AB1) and (AB2) are the population analogs of the general-position assumption (FB1) and the requirement that n_{G} strictly exceed 1/2. Assumption (AB3) requires that target variance $\sup_n \widetilde{\sigma}^2_{\text{T}}$ remain bounded asymptotically.

THEOREM 2 *Under (AB1)-(AB3) it follows that* $Pr\left(\limsup_n ||\widehat{\boldsymbol{\beta}}|| < \infty\right) = 1$.

Proof. Because our argument uses subsequences we add the subscript n to $\widehat{\boldsymbol{\beta}}_n$ to enhance clarity. We prove asymptotic boundedness by assuming that for every $B > 0$,

$$\Pr\left(||\widehat{\boldsymbol{\beta}}_n|| > B \text{ infinitely often}\right) = 1$$

and arriving at a contradiction. Note that if for every $B > 0$, $\Pr\left(||\widehat{\boldsymbol{\beta}}_n|| > B \text{ i.o.}\right) = 1$, then there exists a random subsequence $n^{(1)} < n^{(2)} < ...$ such that $n^{(j)} \to \infty$ almost surely, and $||\widehat{\boldsymbol{\beta}}_{n^{(j)}}|| \to \infty$ almost surely.

Starting with (38), replacing $\widetilde{\sigma}^2_{\text{T}}$ by $\widetilde{\sigma}^2_{\infty}$ and $\mathcal{G}_*$ by $\mathcal{G}$, and dividing both sides by n_{G} results in

$$e^{\tau\widetilde{\sigma}^2_{\infty}}\left(\frac{n - n_{\text{G}} + k - 1}{n_{\text{G}}}\right) + e^{\tau\widetilde{\sigma}^2_{\infty}}\frac{1}{n_{\text{G}}}\sum_{\mathcal{G}}\exp\left[-\tau\left\{y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\right\}^2\right] \ > \ 1. \qquad (39)$$

Let $\epsilon > 0$ and $\eta > 0$ be given, and let $B > 0$ be such that $2/B^{1/2} < \epsilon$. Define $\widehat{\boldsymbol{V}}_n = \widehat{\boldsymbol{\beta}}_n/||\widehat{\boldsymbol{\beta}}_n||$, and the indicator functions

$$I_{i,n} = I\left(|\mathbf{x}_i^T\widehat{\boldsymbol{V}}| > \epsilon,\ |y_i| \leq B^{1/2}\right) \quad \text{and} \quad I_{n,B} = I\left(||\widehat{\boldsymbol{\beta}}_n|| > B\right).$$

Note that $\widehat{\boldsymbol{V}}_n$ is trapped in a compact set and thus we can assume without loss of generality that there exists a $\boldsymbol{V}$ with $||\boldsymbol{V}|| = 1$ such $\widehat{\boldsymbol{V}}_{n^{(j)}} \to \boldsymbol{V}$ almost surely.

Consider the inequality

$$\begin{aligned} I_{i,n}I_{n,B}\left(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\right)^2 &= I_{i,n}I_{n,B}\left\{y_i^2 + \left(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\right)^2 - 2y_i\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\right\} \\ &\geq I_{i,n}I_{n,B}\left\{||\widehat{\boldsymbol{\beta}}_n||^2\left(\mathbf{x}_i^T\widehat{\boldsymbol{V}}\right)^2 - 2|y_i|\left|\mathbf{x}_i^T\widehat{\boldsymbol{V}}\right|||\widehat{\boldsymbol{\beta}}_n||\right\} \\ &\geq I_{i,n}I_{n,B}\inf_{z>\epsilon}\left\{||\widehat{\boldsymbol{\beta}}_n||^2z^2 - 2B^{1/2}||\widehat{\boldsymbol{\beta}}_n||z\right\}. \end{aligned} \tag{40}$$

The function $g(z) = ||\widehat{\boldsymbol{\beta}}_n||^2z^2 - 2B^{1/2}||\widehat{\boldsymbol{\beta}}_n||z$ is such that $g(0) = 0$, $g\left(2B^{1/2}/||\widehat{\boldsymbol{\beta}}_n||\right) = 0$, and $g'(0) = -2B^{1/2}||\widehat{\boldsymbol{\beta}}_n|| < 0$. When $I_{n,B} = 1$, $2B^{1/2}/||\widehat{\boldsymbol{\beta}}_n|| \leq 2B^{1/2}/B = 2/B^{1/2} < \epsilon$, and it follows that $\inf_{z>\epsilon} g(z) = g(\epsilon) = ||\widehat{\boldsymbol{\beta}}_n||^2\epsilon^2 - 2B^{1/2}||\widehat{\boldsymbol{\beta}}_n||\epsilon = M_n$. Note that M_n diverges as $||\widehat{\boldsymbol{\beta}}_n||$ does.

These inequalities imply that the exponential terms in (39) are bounded as

$$\exp\left\{-\tau r_i^2\left(\widehat{\boldsymbol{\beta}}_n\right)\right\} \leq (1 - I_{i,n}I_{n,B}) + I_{i,n}I_{n,B}\exp\left\{-\tau M_n\right\},$$

leading to the inequality

$$e^{\tau\widetilde{\sigma}^2_\infty}\left\{\left(\frac{n - n_{\mathrm{G}} + k - 1}{n_{\mathrm{G}}}\right) + \frac{1}{n_{\mathrm{G}}}\sum_{\mathcal{G}}(1 - I_{i,n}I_{n,B}) + \exp\left\{-\tau M_n\right\}\right\} > 1. \tag{41}$$

We now analyze the left-hand-side of (41) as $n \to \infty$ through the subsequence $n^{(j)}$. Because $M_{n^{(j)}} \to \infty$ a.s., the exponential term vanishes a.s. Assumption (AB2) ensures that the ratio $(n - n_{\mathrm{G}} + k - 1)/n_{\mathrm{G}} \to (1 - \delta)/(1 + \delta)$ along *all* subsequences. The nonnegative term

$$T_n^{**} = \frac{1}{n_{\mathrm{G}}}\sum_{\mathcal{G}}(1 - I_{i,n}I_{n,B}),$$

requires some finesse. Using the identity $1 - I_{i,n}I_{n,B} = 1 - I_{i,n} + I_{i,n}\left(1 - I_{n,B}\right)$ shows that $T_n^{**} = T_n^* + R_n^*$ where

$$T_n^* = \frac{1}{n_{\mathrm{G}}} \sum_{\mathcal{G}} \left(1 - I_{i,n}\right),$$

and $|R_n^*| \leq 1 - I_{n,B}$ which converges to zero a.s. along the subsequence $n^{(j)}$. Thus we now work with T_n^*.

Define the indicator $I_{n,\boldsymbol{V}} = I\left(||\widehat{\boldsymbol{V}} - \boldsymbol{V}|| \leq \eta\right)$. When $||\widehat{\boldsymbol{V}} - \boldsymbol{V}|| \leq \eta$ it follows that

$$\begin{aligned} |\mathbf{x}_i^T\widehat{\boldsymbol{V}}| = |\mathbf{x}_i^T\boldsymbol{V} + \mathbf{x}_i^T(\widehat{\boldsymbol{V}} - \boldsymbol{V})| &\geq |\mathbf{x}_i^T\boldsymbol{V}| - (||\mathbf{x}_i||)\left(||\widehat{\boldsymbol{V}} - \boldsymbol{V}||\right) \\ &\geq |\mathbf{x}_i^T\boldsymbol{V}| - ||\mathbf{x}_i||\eta. \end{aligned} \tag{42}$$

Writing $T_n^* = T_n^* I_{n,\boldsymbol{V}} + T_n^*\left(1 - I_{n,\boldsymbol{V}}\right)$ it is apparent that $0 \leq T_n^*\left(1 - I_{n,\boldsymbol{V}}\right) \leq \left(1 - I_{n,\boldsymbol{V}}\right) \to 0$ a.s. along the subsequence $n^{(j)}$. Furthermore, using the inequality (42) shows that $0 \leq T_n^* I_{n,\boldsymbol{V}} \leq T_n$ where

$$T_n = \frac{1}{n_{\mathrm{G}}} \sum_{\mathcal{G}} \left\{1 - I\left(|\mathbf{x}_i^T\boldsymbol{V}| > \epsilon + \eta||\mathbf{x}_i||,\ |y_i| \leq B^{1/2}\right)\right\}.$$

Note that T_n is an average of independent and identically distributed Bernoulli random variables and thus converges almost surely as $n \to \infty$ (and thus also for all subsequences) to $p(\boldsymbol{V}, B, \eta, \epsilon)$ where

$$p(\mathrm{w}, B, \eta, \epsilon) = 1 - \Pr\left(|\mathbf{x}^T\mathrm{w}| > \epsilon + \eta||\mathbf{x}||,\ |y| \leq B^{1/2}\right).$$

Under Assumption (AB1), $p(\boldsymbol{V}, B, \eta, \epsilon)$ can be made arbitrarily small choosing B large enough and η and ϵ small enough. Summarizing thus far, we have shown that consideration of (41) along the subsequence $n^{(j)}$ leads to the inequality

$$e^{\tau\widetilde{\sigma}_\infty^2}\left\{\frac{1-\delta}{1+\delta}\right\} > 1.$$

Letting τ converge to zero, we arrive at the contradiction $(1-\delta)/(1+\delta) > 1$. ●

References

Agostinelli, C. and Markatou, M. (1998). A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. *Statistics and Probability Letters* **37**, 341-350.

Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85**, 535-547.

Berrendero, J. R. and Zamar, R. H. (2001). Maximum bias curves for robust regression with non-elliptical regressors. *Annals of Statistics* **29**, 224-251.

Berrendero, J. R., Mendes, B. V. M., and Tyler, D. E. (2007). On the maximum bias functions of MM-estimates and constrained M-estimates of regression. *Annals of Statistics* **35**, 13-40.

Choi, E., Hall, P., and Presnell, B. (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika* **87**, 453-465.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B* **46**, 440-464.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr., eds.), 157-184. Wadsworth, Belmont, CA.

Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *Annals of Statistics* **30**, 583-616.

Gray, J.B. (1985). Graphics for Regression Diagnostics. In: *American Statistical Association Proceedings of the Statistical Computing Section*, 102-107. American Statistical Association, Washington, DC.

Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bulletin of the International Statistics Institute* **46**, 375-391.

Hampel, F. R. (1978). Optimally bounding the gross-error sensitivity and the influence of position in factor space. *1978 Proceedings of the ASA Statistical Computing Section*, 59-64.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* **14**, 262-280.

He, X. and Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *Annals of Statistics* **20**, 2161-2167.

He, X. and Wang, G. (1996). Cross-checking using the minimum volume ellipsoid estimator. *Statistica Sinica* **6**, 367-374.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221-233.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics* **1**, 799-821.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* **66**, 333-357.

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**, 861-874.

Kolaczyk, E. D. (1994). Empirical likelihood and generalized linear models. *Statistica Sinica* **4**, 199-218.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. *Econometrica* **48**, 1333-1346.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association* **77**, 595-604.

Lazar, N. A. and Mykland, P. (1999). Empirical likelihood in the presence of nuisance parameters. *Biometrika* **86**, 203-211.

Ma, Y. and Ronchetti, E. (2011). Saddlepoint Test in Measurement Error Models. *Journal of the American Statistical Association* **106**, 147-156.

Mallows, C. L. (1975). On some topics in robustness. Bell Laboratories Technical Report, Murray Hill, New Jersey.

Maronna, R. A., Bustos, O., and Yohai, V. J. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.) 91-116. Springer Verlag, New York.

Newey, W. and Smith, R. J. (2004). Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219-255.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall / CRC, New York.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika* **55**, 111-123.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300-325.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871-880.

Rousseeuw, P. J. and Leroy (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics* **26**, 256-272. Springer, New York.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association* **75**, 828-838.

Schennach, S. M. (2007). Point estimation with exponentially tilted likelihood. *Annals of Statistics* **35**, 634-672.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Econometric Journal* **107**, 503-510.

Wang, L. and Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics* **60**, 883-900.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics* **15**, 642-656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83**, 406-413.

Table 1: Efficiency with respect to maximum likelihood for $k = 2$ for both normal and t_3 errors for sample sizes $n = 20, 50, 100, 200, 500$. Efficiencies calculated as in (31). Exponential Tilting (ET) and Empirical Likelihood (EL) are compared with S-estimation (S; Rousseeuw and Yohai, 1984), Re-weighted Least Squares (REWLS; Gervini and Yohai, 2002), MM-estimator (MM; Yohai, 1987), and Ordinary Least Squares (OLS).

		ET	EL	S	REWLS	MM	OLS
Normal	$n = 20$	91.8	93.7	33.5	69.4	86.7	100.0
	$n = 50$	99.2	99.3	34.1	82.7	94.8	100.0
	$n = 100$	99.3	99.3	26.5	83.9	94.7	100.0
	$n = 200$	99.7	99.8	28.6	91.7	95.5	100.0
	$n = 500$	100.0	100.0	28.2	94.6	94.2	100.0
t_3	$n = 20$	89.5	89.5	56.0	85.2	94.5	50.7
	$n = 50$	92.4	93.4	48.5	88.6	96.5	56.2
	$n = 100$	91.1	92.3	51.4	92.8	97.7	52.3
	$n = 200$	93.7	94.2	54.5	87.5	97.2	52.0
	$n = 500$	92.9	92.6	54.4	90.7	96.1	53.2

Table 2: Efficiency with respect to maximum likelihood for $k = 5$ for both normal and t_3 errors for sample sizes $n = 20, 50, 100, 200, 500$. Efficiencies calculated as in (31). Exponential Tilting (ET) and Empirical Likelihood (EL) are compared with S-estimation (S; Rousseeuw and Yohai, 1984), Re-weighted Least Squares (REWLS; Gervini and Yohai, 2002), MM-estimator (MM; Yohai, 1987), and Ordinary Least Squares (OLS).

		ET	EL	S	REWLS	MM	OLS
Normal	$n = 20$	90.9	91.1	28.0	44.1	81.8	100.0
	$n = 50$	96.5	96.6	28.1	66.4	91.8	100.0
	$n = 100$	98.4	98.4	25.7	79.1	94.6	100.0
	$n = 200$	98.9	99.0	29.2	89.9	94.5	100.0
	$n = 500$	99.8	99.8	27.1	94.0	94.3	100.0
t_3	$n = 20$	86.8	87.2	43.1	62.4	89.2	66.3
	$n = 50$	93.8	94.8	46.5	81.5	95.4	59.4
	$n = 100$	95.9	96.3	49.0	86.9	96.0	53.2
	$n = 200$	95.3	95.3	52.2	88.9	95.4	50.6
	$n = 500$	96.8	97.0	54.7	88.4	95.2	53.9

Table 3: Efficiency with respect to maximum likelihood for $k = 24$ for both normal and t_3 errors for sample sizes $n = 50, 100, 200, 500$. Efficiencies calculated as in (31). Exponential Tilting (ET) and Empirical Likelihood (EL) are compared with S-estimation (S; Rousseeuw and Yohai, 1984), Re-weighted Least Squares (REWLS; Gervini and Yohai, 2002), MM-estimator (MM; Yohai, 1987), and Ordinary Least Squares (OLS).

		ET	EL	S	REWLS	MM	OLS
Normal	$n = 50$	93.8	93.9	32.6	36.9	56.3	100.0
	$n = 100$	95.1	95.0	28.0	47.4	82.9	100.0
	$n = 200$	97.4	97.5	25.3	65.9	92.9	100.0
	$n = 500$	99.1	99.1	24.1	84.8	94.6	100.0
t_3	$n = 50$	88.8	89.7	47.9	54.7	76.7	68.4
	$n = 100$	91.9	92.3	38.7	63.7	90.4	62.3
	$n = 200$	94.1	94.2	40.6	81.0	95.9	54.3
	$n = 500$	95.6	96.0	44.0	88.2	96.3	54.0

Table 4: Efficiency of slope parameter estimates with respect to maximum likelihood for $k = 2$ for Asymmetric Laplace errors with $n = 200$ for various choices of skewness. Efficiencies calculated as in (31). Exponential Tilting (ET) is compared with S-estimation (S; Rousseeuw and Yohai, 1984), Re-weighted Least Squares (REWLS; Gervini and Yohai, 2002), MM-estimator (MM; Yohai, 1987), and Ordinary Least Squares (OLS).

	ET	S	REWLS	MM	OLS
$\tau = 0.1$	24.5	28.3	28.2	22.2	12.5
$\tau = 0.2$	46.0	44.7	48.3	43.0	25.8
$\tau = 0.3$	69.3	61.1	68.1	69.1	45.0
$\tau = 0.4$	81.9	62.1	82.6	82.5	58.5
$\tau = 0.5$	91.3	66.5	86.3	92.1	69.5

Table 5: Maximum mean squared error ($100 \times$ MSE) for the slope parameters over contaminating locations y for various choices of $\mathbf{x} = x_0\mathbf{1_{k-1}}/\sqrt{k-1}$ and contamination proportions (q), with $n = 100$, $k = 5$ and normally distributed errors. The MSE was computed on a grid of y values with $B = 500$ simulated data sets for each grid point. The predictors for the uncontaminated data are generated from $N(\mathbf{0}, I)$. The proposed Exponential Tilting procedure (ET) and Empirical Likelihood (EL) are compared with the S-estimator (S; Rousseeuw and Yohai, 1984), Re-weighted Least Squares (REWLS; Gervini and Yohai, 2002), and the MM-estimator (MM; Yohai, 1987).

		ET	EL	S	REWLS	MM
$x_0 = 1$	$q = .05$	0.82	0.84	2.58	0.86	0.71
	$q = .10$	1.61	1.65	5.54	2.05	1.50
	$q = .20$	6.85	6.95	23.30	7.53	9.59
$x_0 = 2$	$q = .05$	1.53	1.59	4.39	1.25	1.13
	$q = .10$	5.15	5.30	13.53	5.53	4.97
	$q = .20$	51.56	53.64	51.89	28.29	41.20
$x_0 = 3$	$q = .05$	2.96	3.12	6.53	2.94	2.24
	$q = .10$	10.88	11.92	20.05	11.61	10.74
	$q = .20$	61.12	72.25	60.82	41.21	43.76
$x_0 = 5$	$q = .05$	5.66	6.07	6.04	3.64	4.53
	$q = .10$	18.19	21.76	15.27	9.92	10.96
	$q = .20$	63.70	73.48	60.30	43.53	41.64

Table 6: Assessment of variance estimation for the ET estimator for normal and t_3 errors and various sample sizes and dimensions. Table entries calculated as in (32) indicate underestimation (< 1) or overestimation (> 1).

		$k = 2$	$k = 5$	$k = 24$
Normal	$n = 20$	0.899	0.782	—
	$n = 50$	0.978	0.893	0.642
	$n = 100$	0.977	0.957	0.858
	$n = 200$	1.006	0.969	0.881
	$n = 500$	1.005	1.009	0.962
t_3	$n = 20$	0.935	0.772	—
	$n = 50$	0.968	0.941	0.678
	$n = 100$	0.979	0.940	0.802
	$n = 200$	0.978	0.975	0.921
	$n = 500$	0.981	0.982	0.973

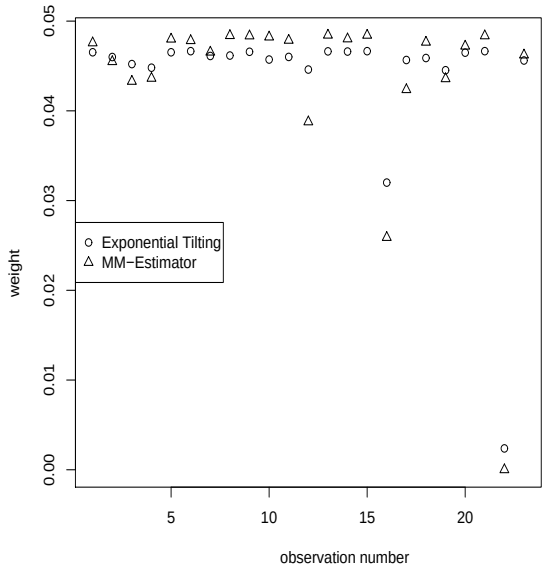
Table 7: Comparison of mean width and actual coverage (in parentheses) of 95% confidence intervals based on asymptotic distributions with $k = 5$.

		ET	MM	OLS
Normal	$n = 20$	0.841(84.2%)	0.990(85.4%)	1.070(95.4%)
	$n = 50$	0.546(92.3%)	0.596(92.4%)	0.593(95.4%)
	$n = 100$	0.390(94.8%)	0.410(93.9%)	0.406(95.9%)
	$n = 200$	0.277(94.6%)	0.288(94.3%)	0.282(95.1%)
	$n = 500$	0.175(94.5%)	0.181(94.6%)	0.176(94.6%)
t_3	$n = 20$	1.152(82.5%)	1.336(84.3%)	1.614(92.0%)
	$n = 50$	0.735(92.3%)	0.776(92.7%)	0.960(95.1%)
	$n = 100$	0.515(93.9%)	0.519(94.4%)	0.669(94.3%)
	$n = 200$	0.364(94.7%)	0.358(94.6%)	0.469(95.5%)
	$n = 500$	0.229(94.2%)	0.225(94.2%)	0.300(94.9%)

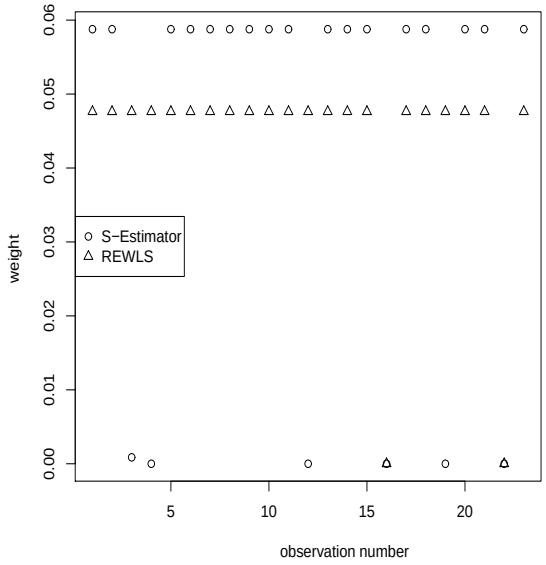
Table 8: Coefficient estimates for the Aircraft data with standard errors in parentheses. Significance (two-sided) at level 0.05 (*) or level 0.01 (**).

	Intercept	Aspect Ratio	Lift-to-Drag	Weight (in 1000s)	Thrust (in 1000s)
ET	3.20 (6.76)	-3.39 (1.18) **	1.94 (0.79) *	2.37 (0.32) **	-1.22 (0.33) **
S	13.37 (4.47) **	-4.02 (1.16) **	1.54 (0.44) **	1.70 (0.34) **	-0.98 (0.29) **
REWLS	9.50 (5.58)	-3.05 (0.92) **	1.21 (0.65)	1.38 (0.39) **	-0.55 (0.33)
MM	6.14 (8.31)	-3.23 (0.86) **	1.67 (0.70) *	1.92 (0.79) *	-0.93 (0.51)
OLS	-3.79 (10.12)	-3.85 (1.76) *	2.49 (1.19)	3.50 (0.48) **	-1.95 (0.50) **

Figure 1: Resulting weights for each of the 23 observations in the Aircraft data. The proposed exponential tilting estimator is shown along with the MM-estimator (left panel). The right panel shows the REWLS estimator along with the S-estimator.



(a) Weights resulting from Exponential Tilting and the MM-estimator.



(b) Weights resulting from REWLS and the S-estimator.