Journal of the Korean Statistical Society (2007), 36: 2, pp 299–320

ROBUST REGRESSION ESTIMATION BASED ON DATA PARTITIONING

DONG-HEE LEE[1] AND YOUSUNG PARK[2]

ABSTRACT

We introduce a high breakdown point estimator referred to as data partitioning robust regression estimator (DPR). Since the DPR is obtained by partitioning observations into a finite number of subsets, it has no computational problem unlike the previous robust regression estimators. Empirical and extensive simulation studies show that the DPR is superior to the previous robust estimators. This is much so in large samples.

AMS 2000 subject classifications. Primary 62J05; Secondary 90C06.
Keywords. Computation problem, data partitioning, efficiency, high breakdown point, outlier detection, performance in large sample.

1. INTRODUCTION

Robust inference methods have long occupied a position of eminence among the statistical tools. By modest changes to likelihood based inference methods, robust inference procedures provide near optimality when compared to the original procedure while maintaining good performance over a wide range of departures from the central model. A robust procedure is expected to work well when there are outliers that are not easy to detect.

Robust regression methods have the desirable features to be efficient, to have a bounded influence function and to attain the highest possible breakdown point of 50%. The efficiency refers to how well a robust regression method performs on clean data, compared to least squares estimation. The influence function measures how much an estimator is influenced by a single outlying observation,

Received November 2006; accepted January 2007.
[1]BK21 Education and Research Center for Economics and Statistics, Korea University, 208 Seochang-Dong, Jochiwon-Eup, Chungnam 339-700, Korea (e-mail: ld0351@korea.ac.kr)
[2]Corresponding author. Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea (e-mail: yspark@korea.ac.kr)

while the breakdown point is the minimum fraction of outliers that renders an estimator useless.

The estimators known as high breakdown estimators (HBE), least median of squares (LMS) and least trimmed squares (LTS) by Rousseeuw (1984), and S-estimators by Rousseeuw and Yohai (1984), have the highest breakdown point of 50%. Other high breakdown regression estimators are τ-estimators (Yohai and Zamar, 1988), R-estimators (Hössjer, 1994), generalized S-estimators such as the least quartile difference (LQD) and the least trimmed difference regression estimator (LTD) (Croux *et al.*, 1994; Stromberg *et al.*, 2000). However, these HBEs are known to be inefficient or to have unbounded influence functions. To circumvent these deficiencies with keeping the highest breakdown point, several composite robust regression estimators have been developed by combining two of the above estimators. Typical examples are the MM-estimator (Yohai, 1987) for high breakdown point and high efficiency and the one-step GM-estimators (GM1) for high breakdown point and bounded influences against x-axis outliers (Simpson *et al.*, 1992; Coakley and Hettmansperger, 1993).

When we denote a regression estimator by $T(\mathbf{X}, \mathbf{y})$ where $\mathbf{X}$ is the $n \times p$ matrix for the independent variables and $\mathbf{y}$ is the $n \times 1$ vector for the dependent variable, $T(\mathbf{X}, \mathbf{y} + \mathbf{X}\mathbf{u}) = T(\mathbf{X}, \mathbf{y}) + \mathbf{u}$ by the regression equivariance for any $p \times 1$ vector $\mathbf{u}$ and $T(\mathbf{X}\mathbf{A}, \mathbf{y}) = \mathbf{A}^{-1}T(\mathbf{X}, \mathbf{y})$ by the affine equivariance for any $p \times p$ matrix $\mathbf{A}$. These two properties are useful to derive asymptotic notions of the influence function and the maxbias curve which are used to evaluate how robust an estimator is against outliers (see, Hampel (1974) for the influence function; Martin *et al.* (1989) for the maxbias curve). Most of the high breakdown estimators described above have the two equivariances. However, computing the high breakdown estimators is notoriously difficult for a large sample and even for a small sample for calculating the LQD and LTD (Stromberg *et al.*, 2000). To avoid this problem, high breakdown estimators typically use only a subset of the data which is randomly selected by a resampling technique. However, this resampling brings a lower convergence rate and a lower breakdown point than theoretically expected (Hawkins and Olive, 2002). Furthermore, resampling also causes the high breakdown estimators to lose the invariance under any permutation of observations which is another desired property for an estimator to possess.

Our data partitioning robust regression estimator (DPR) is also a high breakdown estimator but not affine and regression equivariant. Therefore, its influence and maxbias curves may be difficult to obtain in general situations. However, the DPR is computed from a finite number of partitions of the data with consequences

that it is permutation invariant and no computational problem arises.

To see how this contrast between the DPR and other high breakdown estimators affects on robustness in practice, empirical and simulation studies compare the DPR to eight representative high breakdown estimators in terms of efficiency, resistance from bad leverage points, capability of detecting outliers, influence of heavy tail and asymmetric error distributions and performance in large samples.

The remainder of this paper is divided into four sections. Section 2 describes the DPR. The DPR is compared to OLS and eight competitive robust estimators through four real data sets in Section 3 and through extensive simulation studies in Section 4. Section 5 provides concluding remarks.

2. DATA PARTITIONING ROBUST REGRESSION ESTIMATOR

We consider the linear regression model given by

$$y_j = \mathbf{x}_j^t \boldsymbol{\beta} + \epsilon_j, \quad 1 \le j \le n, \tag{2.1}$$

where $\boldsymbol{\beta}$ is the p-dimensional parameter including an intercept parameter. The random sample $\{\mathbf{x}_j, y_j\}$ are from a distribution F and the error ϵ has a distribution F_ϵ with mean 0 and a finite variance σ^2, which is independent with $\mathbf{x}_j$.

Define a statistical functional T whose domain is a class of probability distributions on $\mathbb{R}^{p+1}$ and whose range is a vector in $\mathbb{R}^p$ of regression coefficients. Under (2.1), the ordinary least squares (OLS) functional defined by $T_{LS}(F) = E_F^{-1}(\mathbf{x}\mathbf{x}^t)E_F(\mathbf{x}y)$ has the corresponding empirical version defined by $T_{LS}(F_n) = (\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^t)^{-1}\sum_{j=1}^n \mathbf{x}_j y_j$ where F_n is the empirical distribution with probability mass $1/n$ at $(\mathbf{x}_j, y_j)$ for $j = 1, \ldots, n$.

An M-estimator $T_\psi(F_n)$ is obtained from estimating equations simultaneously:

$$\begin{aligned} \sum_{j=1}^n \psi\left(\frac{r_j}{\hat{\sigma}}\right)\mathbf{x}_j &= \mathbf{0}, \\ \sum_{j=1}^n \chi\left(\frac{r_j}{\hat{\sigma}}\right) &= 0, \end{aligned} \tag{2.2}$$

where the residuals r_j equal $y_j - \mathbf{x}_j^t T_\psi(F_n)$, ψ is the derivative of ρ, which is a symmetric function with a unique minimum at zero, and χ is a symmetric function. However, such M-estimators have a zero breakdown point, as indicated by Rousseeuw (1984), because of the possibility of leverage points. The estimators

fail to achieve robustness. We now develop an initial estimator $T_0(F_n)$ and the DPR by combining $T_0(F)$ and the $T_\psi(F_n)$.

Let $\tilde{\boldsymbol{\beta}}_n$ be a trial high breakdown estimator. If regressor coefficients $\tilde{\boldsymbol{\beta}}_n$ lead to a linear model fitting well with the data, residuals using $\tilde{\boldsymbol{\beta}}_n$ are small for the majority of the data while large for outliers. To find such $\tilde{\boldsymbol{\beta}}_n$, we first define a collection of subsets of n observations as follows. Based on response variable y, let $\tilde{y}_{q_1}$, $\tilde{y}_{q_2}$ and $\tilde{y}_{q_3}$ be the first quartile, the median and the third quartile of y, respectively. Then define

$$O_{01} = \{(\mathbf{x}_j, y_j) | \tilde{y}_{q_1} \le y_j < \tilde{y}_{q_2}\} \quad \text{and} \quad O_{02} = \{(\mathbf{x}_j, y_j) | \tilde{y}_{q_2} \le y_j \le \tilde{y}_{q_3}\}.$$

Next, based only on the i^{th} independent variable, partition n observations into four quadrants. To that end, set $\bar{y}_{u_i} = \left(\sum_{j=1}^n y_j I[x_{ij} \ge \bar{x}_i]\right) / \left(\sum_{j=1}^n I[x_{ij} \ge \bar{x}_i]\right)$ and $\bar{y}_{l_i} = \left(\sum_{j=1}^n y_j I[x_{ij} < \bar{x}_i]\right) / \left(\sum_{j=1}^n I[x_{ij} < \bar{x}_i]\right)$ where $\bar{x}_i = 1/n \sum_{j=1}^n x_{ij}$. Then, for $i = 1, \ldots, p$ where p is the number of independent variables, define

$$\begin{aligned}
O_{i1} &= \{(\mathbf{x}_j, y_j) \mid x_{ij} \ge \bar{x}_i \text{ and } y_j \ge \bar{y}_{u_i}\}, \\
O_{i2} &= \{(\mathbf{x}_j, y_j) \mid x_{ij} \ge \bar{x}_i \text{ and } y_j < \bar{y}_{u_i}\}, \\
O_{i3} &= \{(\mathbf{x}_j, y_j) \mid x_{ij} < \bar{x}_i \text{ and } y_j \ge \bar{y}_{l_i}\} \quad \text{and} \\
O_{i4} &= \{(\mathbf{x}_j, y_j) \mid x_{ij} < \bar{x}_i \text{ and } y_j < \bar{y}_{l_i}\}.
\end{aligned}$$

Let $\mathbf{O}_i$ be the closure of the class $\{O_{ij} : 1 \le j \le 4\}$ under the union operation,

$$\begin{aligned}
\mathbf{O}_i = \; & \{O_{i1}, \ldots, O_{i4}; O_{i1} \cup O_{i2}, \ldots, O_{i3} \cup O_{i4}; O_{i1} \cup O_{i2} \cup O_{i3}, \ldots, \\
& O_{i2} \cup O_{i3} \cup O_{i4}; \bigcup_{k=1}^{4} O_{ik}\}
\end{aligned}$$

and $\mathcal{C} = \bigcup_{i=0}^{p} \mathbf{O}_i$ where $\mathbf{O}_0 = \{O_{01}, O_{02}, O_{01} \cup O_{02}\}$. Note that there are at most $14p + 4$ non-empty different sets in $\mathcal{C}$. Let K denote the number of sets in $\mathcal{C}$ having cardinality at least $p + 2$ and further let $E_1, E_2, \ldots, E_K$, $K \le 14p + 4$ be an enumeration of these "thick" sets. We refer to an E_i as an elementary set for $i = 1, \ldots, K$.

EXAMPLE 2.1. Simpson and Montgomery (1998) used the satellite cost data to evaluate the robust estimators for bad leverage outlying observations. For a particular class of satellites, 19 observations were collected (see Table 2.1). Using these observations, we calculate $\tilde{y}_{q_1} = 1619$, $\tilde{y}_{q_2} = 2497$ and $\tilde{y}_{q_3} = 3989$ to obtain O_{0j}'s in Figure 2.1-(a) and $\bar{x}_1 = 43.42$, $\bar{y}_{l_1} = 3446.71$ and $\bar{y}_{u_1} = 2456.83$ to obtain O_{1j}'s in Figure 2.1-(b). The two figures also show that four valid observations (O_{12}) are not in-line with the other 15 observations.

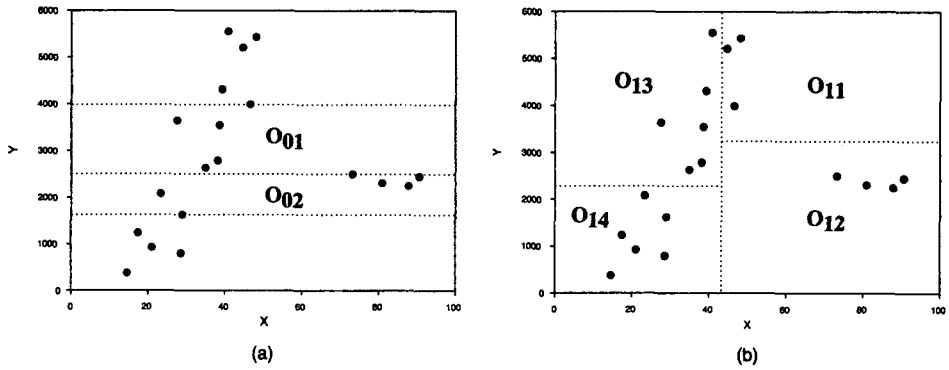

FIGURE 2.1 *Data partitioning: (a) $\tilde{y}_{q_1}$, $\tilde{y}_{q_2}$ and $\tilde{y}_{q_3}$ (b) $\bar{x}_1$, $\bar{y}_{u_1}$ and $\bar{y}_{l_1}$.*

TABLE 2.1 *The satellite cost data*

j	x_j	y_j	j	x_j	y_j	j	x_j	y_j	j	x_j	y_j
1	90.6	2449	6	23.3	2079	11	34.9	2628	16	38.1	2786
2	87.8	2248	7	21.1	918	12	46.6	3989	17	73.2	2497
3	38.6	3545	8	17.5	1231	13	80.9	2308	18	40.8	5551
4	28.6	794	9	27.6	3641	14	14.6	376	19	44.6	5208
5	28.9	1619	10	39.2	4314	15	48.1	5428			

Using the elementary sets, the following four steps precisely describe how to obtain an optimal $\tilde{\boldsymbol{\beta}}_n$ and the DPR.

Step 1. Obtain OLS estimate $\mathbf{b}_k$ from observations in E_k. With $\mathbf{b}_k$ as regressor coefficients calculate for all n data points standardized residuals $\tilde{r}_j(\mathbf{b}_k) = r_j(\mathbf{b}_k)/s(\mathbf{b}_k)$ where $r_j(\mathbf{b}_k) = y_j - \mathbf{x}_j^t\mathbf{b}_k$, $j = 1, 2, \ldots, n$ and $s(\mathbf{b}_k) = (0.6745)^{-1}\text{median}(|r_j(\mathbf{b}_k)|)$. Do this for each of the reduced samples $E_1, E_2, \ldots, E_K$ and then obtain, for $k = 1, 2, \ldots, K$,

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{b}_k} \sum_{|\tilde{r}_j(\mathbf{b}_k)|<c_1, |\tilde{r}_i(\mathbf{b}_k)|<c_1} \left[r_i^2(\mathbf{b}_k) - r_j^2(\mathbf{b}_k)\right]_+, \tag{2.3}$$

where $[x]_+ = \max(0, x)$ and c_1 is a cut-off value which is greater than 0.6745 to ensure median$\big(|r_j(\mathbf{b}_k)|\big)$ included in the above summation. Here

the summation is taken over all i, $j = 1, 2, \ldots, n$. Define $O(\tilde{\boldsymbol{\beta}}_n)$ to be the observations satisfying $|\tilde{r}_j(\tilde{\boldsymbol{\beta}}_n)| < c_1$ for $j = 1, 2, \ldots, n$.

Step 2. Only using observations in $O(\tilde{\boldsymbol{\beta}}_n)$, calculate $T_\psi(F_n)$ given in (2.2).

Step 3. Remove the observations with $|\tilde{r}_j(T_\psi(F_n))| > c_2$ for $j = 1, 2, \ldots, n$ where $c_2 \leq c_1$. We call the removed observations temporary outliers. When no temporary outlier is detected or the remaining number of observations is less than $p + 2$, the $T_\psi(F_n)$ in Step 2 is our DPR. Otherwise, denote the remaining observations by $\mathbb{S}_{n_1}$ where n_1 is the size of $\mathbb{S}_{n_1}$.

Step 4. Based only on $\mathbb{S}_{n_1}$, construct new E_k's and then repeat Step 1 and Step 2.

The estimate from Step 4 is the DPR denoted by $T_{DPR}(F_n)$. The cases with $|\tilde{r}_j(T_{DPR}(F_n))| > c_2$ for $j = 1, 2, \ldots, n$ are deemed to be outliers.

Step 1 produces at most $14p + 4$ trial fits unlike other high breakdown regression estimators. This alleviates the computational burden arising from a large sample size n. When an E_k contains outliers, the OLS obtained from that E_k will generally be distorted resulting in large residuals as calculated in Step 1. Correspondingly, $s(\mathbf{b}_k) = (0.6745)^{-1}\text{median}\big(|r_j(\mathbf{b}_k)|\big)$ is large, and hence $\sum_{|\tilde{r}_j(\mathbf{b}_k)|<c_1, |\tilde{r}_i(\mathbf{b}_k)|<c_1} \big[r_i^2(\mathbf{b}_k) - r_j^2(\mathbf{b}_k)\big]_+$ for a given c_1 should be large. The upshot being that this $\mathbf{b}_k$ is unlikely to achieve the optimality criterion for our $\tilde{\boldsymbol{\beta}}_n$.

When outliers are roughly symmetrically located about the true regression plane, Step 2 is responsive to detecting outliers on that side of the regression plane positioned furthest out from the plane. Thus, an additional step to detect any outliers on the opposite side of the plane is required. Step 3 and Step 4 are used for this. Such an example is the Hawkins, Bradu and Kass data in Section 4.1 (*i.e.*, among 14 outliers, Step 2 detects only the first 10 outliers and then all the 14 outliers are detected by Step 3 and Step 4).

A high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary (Olive, 2003). Because the $T_\psi(F_n)$ in Step 2 and the DPR in the second round are based on $\tilde{\boldsymbol{\beta}}_n$ after removing all outliers, their median absolute residuals are always bounded whenever the median absolute residual obtained from $\tilde{\boldsymbol{\beta}}_n$ is bounded. Following Rousseeuw (1984), we shall say the observations are in general position when any p of them determines a unique $\boldsymbol{\beta}$. Under this condition, the following

shows that $\tilde{\beta}_n$ has a bounded median absolute residual for DPR to be a high breakdown estimator.

LEMMA 2.1. *If $n > 4p$ and the observations are in general position, $\tilde{\beta}_n$ has a bounded median absolute residual.*

PROOF. Since $d \leq [(n+1)/2] - 1$ where d is the number of contaminated observations, at least one elementary set in $\mathbf{O}_0$ consists of finite valued y_j's. The general position condition and $n > 4p$ ensure that there is a finite OLS estimator denoted by $\hat{\beta}_0$ in $\mathbf{O}_0$. The median of absolute residual from this $\hat{\beta}_0$ should be bounded because of $d \leq [(n+1)/2] - 1$ and thus

$$\sum_{|\tilde{r}_j(\hat{\beta}_0)|<c_1, |\tilde{r}_i(\hat{\beta}_0)|<c_1} \left[r_i^2(\hat{\beta}_0) - r_j^2(\hat{\beta}_0)\right]_+ < \infty.$$

When $\hat{\beta}_0 = \tilde{\beta}_n$, the claim follows.

When $\hat{\beta}_0 \neq \tilde{\beta}_n$, there should exist an OLS estimator $\hat{\beta}_1$ which is equal to $\tilde{\beta}_n$ with

$$\sum_{|\tilde{r}_j(\hat{\beta}_1)|<c_1, |\tilde{r}_i(\hat{\beta}_1)|<c_1} \left[r_i^2(\hat{\beta}_1) - r_j^2(\hat{\beta}_1)\right]_+ < \sum_{|\tilde{r}_j(\hat{\beta}_0)|<c_1, |\tilde{r}_i(\hat{\beta}_0)|<c_1} \left[r_i^2(\hat{\beta}_0) - r_j^2(\hat{\beta}_0)\right]_+.$$

This $\hat{\beta}_1$ has a finite median of absolute residuals. Otherwise,

$$\sum_{|\tilde{r}_j(\hat{\beta}_1)|<c_1, |\tilde{r}_i(\hat{\beta}_1)|<c_1} \left[r_i^2(\hat{\beta}_1) - r_j^2(\hat{\beta}_1)\right]_+ = \infty$$

since median of $|r_j(\hat{\beta}_1)|$ is always included in the summation. This completes the proof. □

EXAMPLE 2.2. We continue Example 2.1 to illustrate Step 1 through Step 4 with $c_1 = 4$ and $c_2 = 3$. These choices of c_1 and c_2 will be discussed in Section 3.

Step 1. We obtain $\tilde{\beta}_n = (-1738.60, 142.51)$ based on $O_{11} \cup O_{13} \cup O_{14}$ given in Figure 2.1-(b) by (2.3). Using $s(\tilde{\beta}_n) = 1128.14$, $O(\tilde{\beta}_n)$ contains except the four observations numbered by 1, 2, 13 and 17 (Figure 2.2-(a)).

Step 2. We obtain $T_\psi(F_n) = (-1701.76, 140.94)$ and $s(T_\psi(F_n)) = 873.34$ with $\delta = .001$, and the temporary outliers are observations 1, 2, 13 and 17 (observations outside of the two dashed lines in Figure 2.2-(b)).

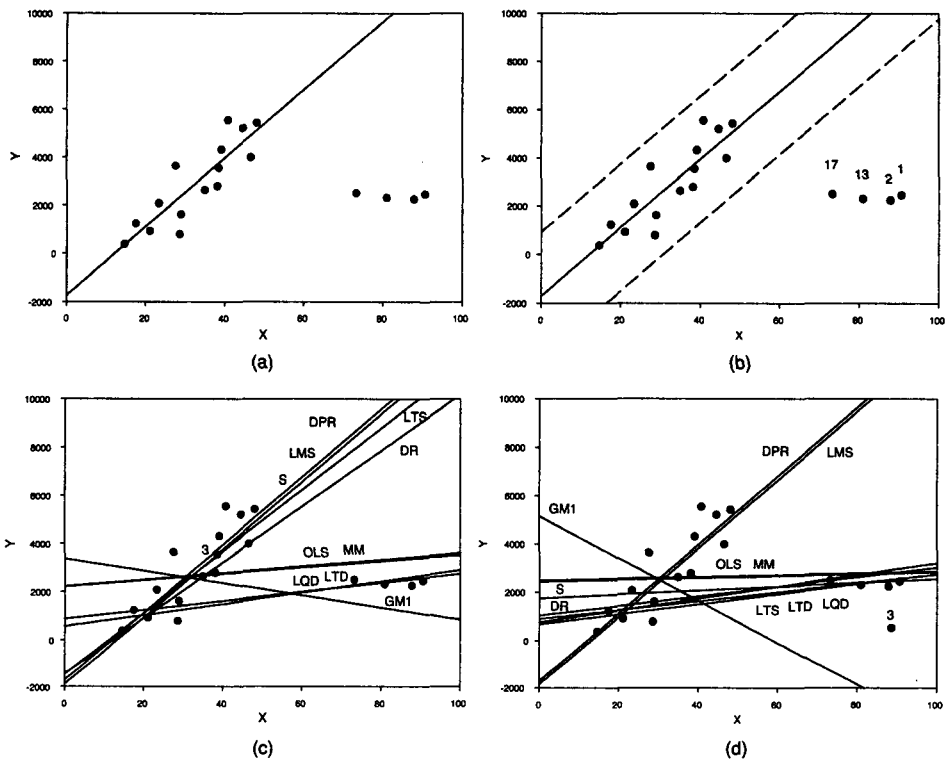

FIGURE 2.2 *Regression fits: (a) $\tilde{\beta}_n$, (b) $T_\psi(F_n)$ and temporary outliers at the first repetition, (c) DPR and competitive robust regression estimates, (d) DPR and competitive robust regression estimates after moving the case 3.*

Step 3 and Step 4. The DPR is the same as $T_\psi(F_n)$ and detects the same four outliers.

Figure 2.2-(c) compares the DPR to OLS, LTS, LMS, MM, GM1, LQD, LTD and DR in which the DR stands for the deepest regression estimator (van Aelst *et al.*, 2002). The DR is known to be useful for the regression model with skewness and heteroscedasticity although the DR is not a high breakdown estimator. The OLS, MM, GM1, LQD and LTD are seriously influenced by the four outliers while LTS and DR are slightly influenced. We intentionally move the observation numbered by 3 in Figure 2.2-(c) to a bad leverage position as shown in Figure 2.2-(d). This single movement increases the outlier percentage from 21% to 26% and

breaks down all estimators except the DPR and LMS as seen in Figure 2.2-(d).

3. EMPIRICAL AND SIMULATION STUDIES

We consider four empirical data sets which are frequently used to evaluate the robustness of a regression estimator. They are the satellite cost data (Simpson and Montgomery, 1998) which was discussed in Section 2.2, the stack loss data (Rousseeuw and Leroy, 1987), the Hawkins, Bradu and Kass data (Hawkins *et al.*, 1984) and the Buxton data (Hawkins and Olive, 2002). We use $c_1 = 4$ and $c_2 = 3$ for our DPR. The choice of c_1 is a trade-off between efficiency and the degree of resistance from outliers. Although we do not present a simulation study for the choice of c_1, a general observation is that larger c_1 produces more efficiency but lower resistance against an outlier. The simulation study also suggests that an optimal choice of c_1 is between 3 and 5. The choice of $c_2 = 3$ to detect outliers simply follows the smooth rejection rule of Rousseeuw and Leroy (1987), which detects an outlier when its standardized residual is greater than 3. In addition, we use Huber's weight function as $T_\psi(F_n)$ in DPR,

$$\psi(u) = \min(k, \max(u, -k)), \tag{3.1}$$

where k is 1.5.

We compare the DPR to the nine estimators: the OLS to evaluate efficiency; the LMS in Rousseeuw and Leroy (1987); the LTS in Rousseeuw and Leroy (1987); S-estimator in Rousseeuw and Yohai (1984); the MM estimator which we use LTS as the initial estimate with $h = [n + p + 1]/2$ and Tukey's bisquare weight function with 95% efficiency to achieve high efficiency and high breakdown point (Yohai, 1987); the GM1, which uses LTS as the initial estimate and adjusts the estimates with empirically determined weights, proposed by Coakley and Hettmansperger (1993); the deepest regression estimator (DR) as one of the most recent robust regression estimation methods by Rousseeuw and Hubert (1999) and the estimates are obtained by the MEDSWEEP Fortran-coded program from Van Aelst *et al.* (2002); the LQD in Croux *et al.*, (1994); the LTD in Stromberg *et al.* (2000).

3.1. Empirical studies

The Buxton data were used by Hawkins and Olive (2002), which consists of 87 observations and 4 variables. Along with their work, we also predict stature using an intercept, head length, nasal height, bigonal breath and cephalic index. Five

TABLE 3.1 *Estimates for Buxton data*

Estimation	*OLS*	*OLS(clean)*	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
Intercept	74.7	147.6	152.6	183.9	149.2	147.8	152.1	155.6	152.7	165.0	147.8
$X1$	.006	.008	−.014	−.002	.007	.007	.013	.013	.006	.002	.007
$x2$	.006	−.499	−.285	−.600	−.502	−.498	−.590	−.536	−.506	−.473	−.498
$X3$	−.125	−.088	−.160	.021	−.030	−.079	−.021	−.234	−.046	−.012	−.074
$X4$	.047	.158	.150	.080	.135	.157	.144	.126	.127	.014	.156

TABLE 3.2 *Standardized residuals in stack loss data*

cases	*standardized residuals*										
	OLS	*OLS(clean)*	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
1	1.19	5.95	6.13	8.29	2.78	0.77	0.03	3.72	2.94	3.45	7.51
2	−0.72	1.08	2.17	3.42	0.44	−0.49	−2.54	0.69	0.04	0.77	2.80
3	1.55	5.65	5.93	7.77	2.87	1.08	0.90	3.76	3.15	3.49	7.20
4	1.89	7.04	6.91	8.81	3.79	1.48	3.14	4.92	4.66	4.51	8.25
21	−2.64	−8.16	−6.52	−8.18	−4.14	−1.95	−5.64	−5.58	−5.56	−4.71	−7.76

individuals, cases 62 to 66, appear to be clerical errors. Table 3.1 provides the ten estimates and the OLS obtained from 82 clean observations by removing the five cases with the clerical errors. We note that all ten robust estimates clearly identify the five outliers with big values of residuals. The coefficients of MM and DPR are surprisingly close to those of the OLS using only clean data (*i.e.*, OLS (clean) in the second column of Table 3.1).

The stack loss data have been intensively studied in the literature on robust estimation and applied linear regression models. It consists of 21 daily observations measured in a plant for the oxidation of ammonia to nitric acid. The response variable is ten times the percentage of ammonia lost. There are 3 independent variables: air flow, temperature of the cooling water and acid concentration. After a very careful analysis, researchers decided that cases 1, 3, 4 and 21 are outliers. Table 3.2 lists standardized residuals of two OLS and nine robust estimators for five cases 1, 2, 3, 4 and 21. All nine robust estimates correctly classify the other 16 cases not included in Table 3.2. By Rousseeuw's smooth rejection rule, LMS, LQD, DR and DPR completely detect the outliers without an erroneous detection.

The Hawkins, Bradu and Kass data was generated by Hawkins *et al.* (1984) for illustrating some of the merits of a robust technique. Such artificial data offer the advantage that at least the position of the bad points are known, cases

TABLE 3.3 *Standardized residuals in Hawkins-Bradu-Kass data*

cases	*standardized residuals*										
	OLS	*OLS(clean)*	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
1	1.55	21.7	24.4	23.0	0.15	10.1	13.5	0.63	12.9	7.23	14.6
2	1.83	22.6	25.7	24.3	0.39	10.6	14.0	0.96	13.5	7.64	15.2
3	1.40	23.6	26.1	24.7	0.12	10.8	14.4	0.50	13.8	7.72	15.7
4	1.19	22.2	24.4	23.3	−1.23	10.0	13.6	−0.75	12.8	6.96	14.9
5	1.41	22.9	25.4	24.1	−0.47	10.5	14.1	0.00	13.4	7.40	15.4
6	1.59	22.4	24.9	23.4	−0.03	10.4	13.9	0.45	13.3	7.31	14.9
7	2.08	23.8	27.0	25.4	0.88	11.2	14.8	1.45	14.3	8.11	15.9
8	1.76	23.0	26.1	24.7	0.86	10.8	14.2	1.33	13.8	7.85	15.4
9	1.26	22.3	24.7	23.5	−0.69	10.1	13.7	−0.23	13.0	7.16	15.0
10	1.41	22.8	25.7	24.4	0.20	10.5	14.0	0.63	13.4	7.61	15.3
11	−3.66	4.81	−0.53	0.38	−14.9	−0.06	2.71	−14.6	0.18	−3.10	3.96
12	−4.50	4.88	−1.03	−0.08	−15.4	−0.20	2.67	−15.3	0.00	−3.31	4.00
13	−2.88	6.13	0.89	1.42	−14.4	0.65	3.56	−13.9	1.04	−2.65	4.56
14	−2.56	6.98	−1.09	−0.29	−17.5	−0.22	2.79	−16.2	−0.11	−3.93	3.68

1 to 14. It is troublesome to detect outliers because they are located far away from the centroid of x-space, namely bad leverage points. We list standardized residuals of nine robust estimates in Table 3.3 for known outliers. According to Rousseeuw's smooth rejection rule, the DPR detects all 14 outliers and further kept correct classification for the remaining 61 cases while the DR fails to detect the 13^{th} case as an outlier. However, LMS, LTS, MM, GM1 and LTD detect only the first 10 outliers as outliers except GM1 which detect one more observation (*i.e.*, the 13^{th} case) as an outlier, whereas the S and LQD detect only the last 4 cases as outliers.

4. MONTE CARLO SIMULATION STUDIES

We compare DPR to OLS, LMS, LTS, S, MM, GM1, LTD, LQD and DR estimators by extensive Monte Carlo simulations. A goal of the study is to obtain information concerning efficiency, high breakdown point and bounded influence of estimators using 48 simulation scenarios. Efficiency of an estimator is assessed by a comparison of it to least squares under normal errors with no outliers. High breakdown point and bounded influence of an estimator are evaluated by its performance against highly contaminated data and bad leverage outlying observations. Our simulation studies basically follow the simulation designs by Wisnowski *et al.* (2001). In Section 4.1, we examine the ten estimators in terms

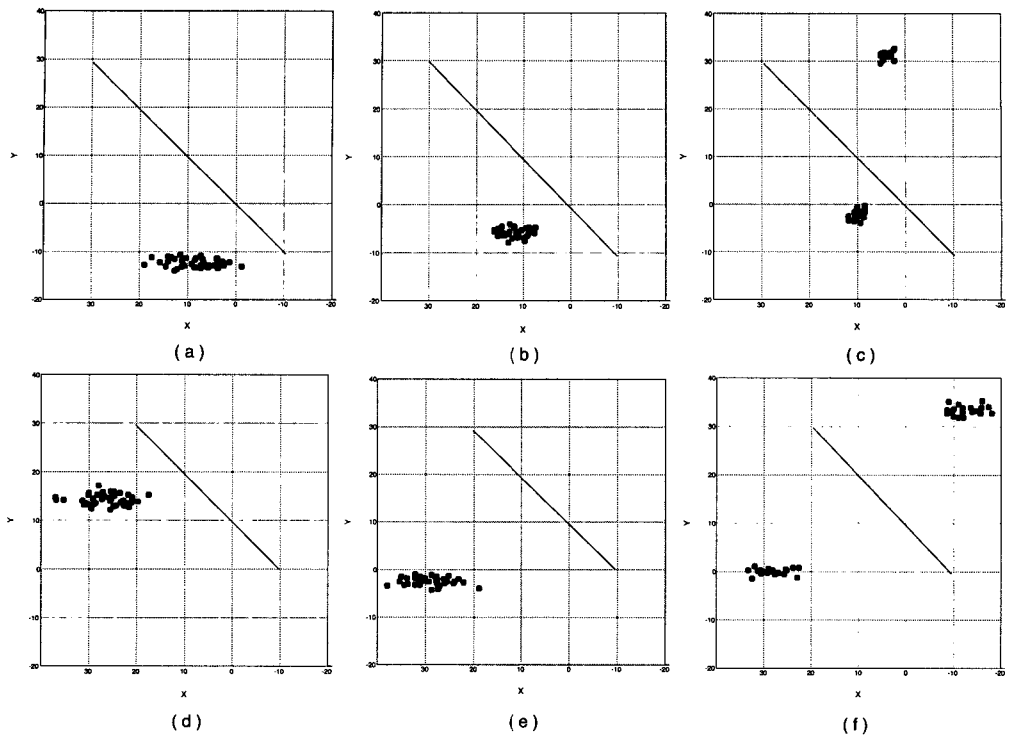

FIGURE 4.1 *Outlier patterns in Monte Carlo study.*

of MSE and bias. We apply robust estimators to Cauchy and extreme value distributed random samples to see how the estimators behave under heavy tail and asymmetric error distributions in Section 4.2. We also examine the influence of large sample size on these robust estimators.

4.1. Simulation scenarios

Consider the regression model with p independent variables.

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \epsilon_j, \quad j = 1, \ldots, n. \tag{4.1}$$

All simulations generate a fixed percentage of clean observations and plant outliers at locations specified by factor settings. As mentioned earlier, the DPR does not possess the regression and affine equivariances unlike all other competitive estimators discussed in this paper. Thus, for fair comparison, we generate clean observations $\{x_{ij}, y_j\}$ under the regression model (4.1) with $\beta_0 = 0$ and other β's equal to 1 and $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})$ from $N(7.51_p, \Omega_x)$ where Ω_x has the following

four types:

(i) $\Omega_x = 16\mathbf{I}$,

(ii) $(\Omega_x)_{ii}$ is independently from Uniform$(8, 24)$ and corr(x_{ij}, x_{kj}) is also independently from Uniform$(0.1, 0.2)$,

(iii) $(\Omega_x)_{ii}$ is the same as (ii) but the correlations are chosen from Uniform$(0.4, 0.5)$ and

(iv) 20 cases of Ω_x's in which each case has independent Ω_x with the same $(\Omega_x)_{ii}$ as (ii) but the correlations are from Uniform$(-0.75, 0.75)$.

The three types (i), (ii) and (iii) are designed to examine robustness for almost equi-correlated x-variables (*i.e.*, (i) for uncorrelated x, (ii) for low correlated x and (iii) for moderately correlated x-variables), while (iv) is to evaluate the effect of the correlation structure randomly ranged from $-.75$ to $.75$. Finally, ϵ_j is generated from *i.i.d.* $N(0, 1)$. Let $\bar{x}_i$ and $\bar{y}$ be the sample means of x_{ij} and y_j for $i = 1, \ldots, p$ to define outliers as below. Notify that simulation studies for other specification of $\boldsymbol{\beta}$'s also showed a similar fashion as the results given below. Thus we do not present the simulation results other than the specification of regression parameters; $\beta_0 = 0$ and $\beta_j = 1$, for $j = 1, \ldots, p$.

We consider two sample sizes $n = 60$ and 100 and outlier percentages from 0% to 45%. After the clean observations are generated, the remainder of the whole sample is filled with outliers. We are concerned with outliers likely to be encountered in practice: the three types of outliers in the y-direction are:

(a) $y_j = N(-6, 1)$ over all range of x_{ij} (Figure 4.1-(a)),

(b) $x_{ij} \in [\bar{x}_i, \bar{x}_i + 2.5\sigma_{x_i}]$ where σ_{x_i} is the standard deviation of the i^{th} independent variable (Figure 4.1-(b)) and

(c) $x_{ij} = \bar{x}_i \pm$ Uniform$(1, 5)$ and $y_j = \min\{y_1, \ldots, y_n\} - 3\sigma_\epsilon$ for $-(1, 5)$ and $y_j = \max\{y_1, \ldots, y_n\}$ for $(1, 5)$ (Figure 4.1-(c)).

Similarly, the three types of outliers in the x-direction are:

(d) $x_{ij} = \bar{x}_i + 5\sigma_{x_i} + N(0, \sigma_{x_i}^2)$ and $y_j = \bar{y} + N(0, 1)$ (Figure 4.1-(d)),

(e) $x_{ij} = \bar{x}_i + 5\sigma_{x_i} + N(0, \sigma_{x_i}^2)$ and $y_j = \min\{y_1, \ldots, y_n\} - 3\sigma_\epsilon + N(0, 1)$ (Figure 4.1-(e)) and

TABLE 4.1 *Maximum MSE when no outlier is present under normally distributed error*

n	p	*OLS*	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
60	5	.22	1.29	1.46	.48	.23	.54	.57	.51	3.60	.24
	10	.50	2.76	2.24	1.18	.51	1.34	1.48	1.49	8.95	.60
100	5	.12	.82	.85	.26	.13	.17	.33	.23	1.89	.17
	10	.27	1.84	1.38	.61	.28	.55	1.09	.71	4.67	.29

(f) $x_{ij} = \bar{x}_i + 5\sigma_x + N(0, \sigma_x^2)$ and $y_j = \min\{y_1, \ldots, y_n\} - 3\sigma_\epsilon + N(0,1)$ or $y_j = \max\{y_1, \ldots, y_n\} + 3\sigma_\epsilon + N(0,1)$ (Figure 4.1-(f)).

We also performed the same six simulation studies as above when outliers are in a subset of p independent variables to see how estimators are affected in such cases. However, we do not attach the result since there is no distinction between simulations.

Therefore, we have 12 simulation designs in which each design has 4 types of variances of x-variables, Ω_x. However, the results for outliers in a subset of independent variables were the same patterns as the corresponding six simulation designs from (a) to (f). Moreover, the six simulation designs also showed the similar results in relative efficiency to the DPR. For economy of presentation, we only present (c) which provides a representative study. A complete simulation study will be provided upon request to the first author.

We use the average of mean squared error (MSE) and absolute values of biases

$$\text{MSE} = \frac{1}{m(p+1)} \sum_{i=0}^{p} \sum_{j=1}^{m} (T_{ij} - \beta_i)^2 \quad \text{and} \quad \text{BIAS} = \frac{1}{p+1} \sum_{i=0}^{p} \left| \frac{1}{m} \sum_{j=1}^{m} T_{ij} - \beta_i \right|, \tag{4.2}$$

where m is the replication number and T_{ij} is the i^{th} coefficient estimated in the j^{th} replication with T_{0j} being the estimated intercept. We also compute outlier detection percentages (DP) which give the fraction of outliers that are classified as such and wrong detection percentages (WP) which give the fraction of clean observations that were misclassified as an outlier. Therefore, a good robust estimate has a high outlier detection percentage near 100% and a low wrong detection percentage near 0%. The 1000 replicates of each treatment combination were run to obtain MSEs and outlier detection percentages for three types of Ω_x's (i), (ii) and (iii), while 500 replicates for each of 20 cases for the type (iv).

TABLE 4.2 *Maximum mean squared error and bias over three types of Ω_x (i), (ii) and (iii) (maximum biases in the parentheses)*

n	p	e	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
60	5	.05	.35(.01)	.37(.01)	.13(.00)	.08(.00)	.28(.17)	.16(.01)	.15(.01)	.60(.24)	.08(.00)
		.15	.34(.01)	.36(.01)	.11(.00)	.08(.00)	2.0(.60)	.21(.02)	.13(.01)	225(5.3)	.09(.00)
		.25	.38(.01)	.34(.01)	.11(.00)	436(9.6)	8.4(1.3)	.32(.02)	.13(.01)	1082(15)	.12(.01)
		.35	.37(.01)	.24(.01)	1191(16)	629(12)	26(2.2)	1005(14)	765(11)	1397(17)	.15(.00)
		.45	5.3(.06)	.17(.01)	1961(20)	770(13)	134(4.8)	1481(16)	1632(18)	1825(20)	.22(.01)
	10	.05	.38(.01)	.34(.01)	.16(.01)	.08(.00)	.56(.19)	.28(.01)	.20(.01)	3.1(.37)	.10(.01)
		.15	.40(.01)	.32(.00)	.12(.00)	.09(.00)	5.4(.73)	.45(.02)	.18(.01)	1289(14)	.12(.00)
		.25	.56(.01)	.26(.01)	.11(.00)	1160(12)	30(1.7)	5.4(.05)	15(.12)	2090(16)	.20(.01)
		.35	905(7.0)	2249(16)	2804(17)	1486(13)	1042(10)	2140(16)	2240(16)	3011(18)	.34(.02)
		.45	4481(23)	4460(23)	3987(21)	1769(14)	2588(17)	3400(19)	3955(21)	3669(20)	1.1(.02)
100	5	.05	.24(.01)	.25(.01)	.08(.00)	.04(.00)	.19(.17)	.11(.01)	.08(.01)	.32(.20)	.04(.00)
		.15	.25(.01)	.25(.01)	.07(.00)	.05(.00)	1.8(.59)	.18(.01)	.08(.01)	181(4.7)	.05(.00)
		.25	.25(.01)	.21(.01)	.06(.01)	460(9.9)	6.5(1.1)	.29(.02)	.07(.00)	1239(16)	.06(.00)
		.35	.28(.01)	.15(.00)	23(.32)	667(12)	20(2.0)	1064(15)	314(5.1)	1580(18)	.06(.00)
		.45	.40(.01)	.10(.01)	2232(22)	829(13)	83(4.0)	1511(17)	1751(18)	2074(21)	.09(.01)
	10	.05	.24(.01)	.21(.00)	.07(.00)	.04(.00)	.37(.18)	.21(.01)	.08(.01)	.75(.25)	.04(.00)
		.15	.28(.01)	.20(.00)	.06(.00)	.05(.00)	4.1(.66)	.43(.01)	.08(.01)	1415(15)	.05(.00)
		.25	.42(.01)	.15(.01)	.05(.00)	1195(12)	20(1.5)	1.8(.01)	85(.70)	2303(16)	.06(.00)
		.35	7.1(.03)	151(1.1)	2391(17)	1557(13)	202(4.3)	2248(16)	2409(17)	3360(19)	.07(.01)
		.45	5168(24)	5181(24)	4588(23)	1875(15)	2843(18)	3175(17)	4254(22)	4230(22)	.09(.00)

Table 4.1 shows maximum MSEs of ten estimators over four types of Ω_x to evaluate the efficiency over the OLS when no outlier is present. The MM and DPR have efficiency comparable to OLS regardless of sample size n and the number of independent variables p. The GM1, S, LQD and LTD are less efficient than MM and DPR, whereas the LMS, LTS and DR are least efficient. This matches the results reported in Hinloopen and Wagenvoort (1997) and Rousseeuw (1984).

Table 4.2 shows maximum mean squared error and bias for simulation design (c) where the maximum for each estimator is taken over the three types of Ω_x (i), (ii) and (iii). Except DPR, all robust estimators are demolished when contamination percentages are 35% and more for $p = 10$. The DR is worst and the high breakdown property of the GM1 diminishes faster than its initial estimator LTS, which was pointed out by Simpson and Yohai (1998). The S and MM heavily depend on outlier percentage although the S and MM are good up to 25% and contamination percentages. Their MSEs and biases suddenly increase at the 35% and 25% outlier percentages and increase thereafter. This effect is more pronounced as p increases. Recently, Salibian-Barrera and Zamar (2004) studied

TABLE 4.3 *Maximum mean squared error and bias over 20 cases of the variances of x-variables in (iv) (maximum biases in the parentheses)*

n	p	e	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
60	5	.05	5.9(.17)	5.7(.18)	2.3 (.07)	1.2(.06)	3.9(.63)	3.2(.19)	2.8(.13)	92(3.9)	1.2(.06)
		.15	4.4(.11)	4.3(.10)	1.8 (.04)	1.4(.03)	88(4.1)	3.3(.15)	2.1(.10)	941(14)	1.5(.04)
		.25	5.7(.12)	4.0(.10)	1.8 (.06)	1082(16)	1903(19)	5.0(.11)	1.9(.07)	1386(18)	1.8(.03)
		.35	5.7(.23)	3.9(.22)	1282 (17)	1141(16)	7732(40)	1130(16)	1099(16)	1561(18)	2.8(.14)
		.45	323(5.4)	112(1.9)	1650 (19)	1167(17)	20931(63)	1504(18)	1626(18)	1671(19)	3.2(.06)
	10	.05	8.4(.08)	8.3(.15)	3.3(.10)	1.9(.06)	22(1.0)	7.8(.23)	4.9(.14)	244(4.4)	2.4(.07)
		.15	9.8(.21)	8.5(.15)	2.7(.08)	2.2(.06)	2289(16)	9.3(.23)	4.3(.20)	1813(16)	2.4(.08)
		.25	15(.14)	6.9(.11)	2.5(.05)	2040(17)	19339(48)	138(1.4)	63(.56)	2272(17)	4.5(.09)
		.35	1054(10)	2187(16)	2213(16)	2124(17)	42867(52)	2247(17)	2248(16)	2340(17)	6.7(.07)
		.45	2328(16)	2359(17)	2247(17)	2088(17)	2118(17)	2201(17)	2212(16)	2339(16)	20(.17)
100	5	.05	3.7(.05)	3.2(.14)	1.0(.04)	.58(.02)	2.6(.66)	2.0(.09)	.90(.05)	54(3.1)	.63(.02)
		.15	3.2(.09)	3.5(.05)	1.0(.04)	.64(.04)	69(3.7)	2.6(.17)	1.1(.14)	1027(15)	.69(.03)
		.25	3.6(.07)	4.0(.06)	1.2(.03)	1179(16)	1482(17)	4.7(.25)	1.4(.13)	1443(19)	1.1(.04)
		.35	4.3(.12)	3.0(.09)	1071(15)	1210(17)	5227(34)	1220(17)	1181(17)	1794(20)	1.1(.03)
		.45	73(1.2)	1.9(.03)	1907(20)	1244(17)	21276(67)	1663(18)	1750(19)	1937(20)	1.7(.02)
	10	.05	5.2(.04)	4.5(.13)	1.8(.09)	1.0(.04)	7.5(.83)	5.2(.07)	2.1(.14)	192(4.0)	1.1(.04)
		.15	7.7(.09)	5.1(.20)	1.6(.12)	1.3(.06)	1995(15)	10(.21)	2.2(.15)	2087(17)	1.4(.07)
		.25	12(.12)	3.9(.06)	1.6(.05)	2201(17)	11195(40)	179(1.3)	504(4.8)	2357(18)	1.6(.06)
		.35	167(1.1)	249(1.7)	2437(17)	2260(17)	85363(109)	2434(17)	2473(17)	2558(17)	1.7(.06)
		.45	2701(18)	2704(18)	2610(17)	2282(17)	2347(17)	2514(17)	2603(17)	2666(18)	2.3(.11)

a uniform asymptotic property of MM-estimator and reported the situation such as this result.

On the other hand, the DPR does not manifest any of the drawbacks which were noted in other estimators. Its MSE and bias are stable for p, comparing favorably to the MM for 5% and 15% contamination percentages and the S estimator for 25% contamination percentage, and achieve minimum values among all estimators for 25% $\sim$ 45% contamination percentages.

Table 4.3 shows maximum MSE and bias for the variance (iv) where the maximum is taken among 20 cases of the variances Ω_x. The MSEs and biases are more than 10 times compared to those of Table 4.2, keeping very similar pattern as that of Table 4.2. A general observation from the 20 cases is that the more different in absolute value correlations for different pairs of x-variables, the larger MSEs and biases.

The observations made in Table 4.2 and Table 4.3 are also true for simulation designs (a) through (f) whose results are not presented. Hence, we see that the DPR has provided a significantly superior performance in this comparison of

TABLE 4.4 *Average of outlier detection percentages over the four types of variances among independent variables*

n	p	e	*LMS*	*LTS*	*S*	*GM1*	*MM*	*LQD*	*LTD*	*DR*	*DPR*
			DP(WP)	*DP(WP)*	*DP(WP)*	*DP(WP)*	*DP(WP)*	*DP(WP)*	*DP(WP)*	*DP(WP)*	*DP(WP)*
60	5	.05	100(7.2)	100(11)	100(.37)	100(.53)	100(.02)	100(8.1)	100(13)	100(9.5)	100(.66)
		.15	100(4.0)	100(6.5)	100(.03)	100(.19)	100(.00)	100(2.2)	100(4.8)	82(31)	100(.37)
		.25	100(2.0)	100(3.2)	100(.00)	96(.05)	.00(.00)	100(.20)	100(.67)	48(53)	100(.13)
		.35	100(1.3)	100(1.8)	.26(5.0)	75(.08)	.00(.00)	4.0(.18)	26(.72)	32(66)	100(.02)
		.45	75(13)	80(13)	.12(13)	36(7.3)	.00(.73)	.96(2.1)	.02(5.8)	27(71)	100(.00)
	10	.05	100(8.5)	100(23)	100(.72)	100(.63)	100(.01)	100(8.2)	100(26)	100(16)	100(1.6)
		.15	100(4.1)	100(14)	100(.04)	98(.27)	100(.00)	100(1.6)	100(12)	70(47)	100(1.4)
		.25	100(1.5)	100(6.1)	100(.00)	84(.12)	.00(.00)	97(.09)	100(1.8)	52(60)	100(1.0)
		.35	61(11)	24(37)	1.2(7.7)	36(2.5)	.00(.22)	4.8(.00)	.69(.76)	40(67)	100(.36)
		.45	1.5(40)	6.2(55)	0.3(15)	.10(12)	.01(.75)	1.1(.18)	.02(6.2)	34(72)	100(.01)
100	5	.05	100(4.2)	100(5.8)	100(.24)	100(.33)	100(.02)	100(5.4)	100(8.3)	100(8.8)	100(.37)
		.15	100(2.4)	100(3.5)	100(.01)	100(.10)	100(.00)	100(1.2)	100(2.4)	83(31)	100(.13)
		.25	100(1.6)	100(2.0)	100(.00)	97(.01)	.01(.00)	100(.05)	100(.23)	48(53)	100(.03)
		.35	100(1.3)	100(1.5)	63(2.0)	75(.07)	.00(.00)	6.5(.07)	51(.37)	33(69)	100(.00)
		.45	78(13)	84(12)	.08(13)	43(7.1)	.00(.77)	1.5(1.7)	.00(4.8)	27(74)	100(.00)
	10	.05	100(4.6)	100(12)	100(.39)	100(.35)	100(.01)	100(4.4)	100(16)	100(12)	100(.52)
		.15	100(2.5)	100(7.0)	100(.02)	98(.12)	100(.00)	100(.51)	100(5.6)	71(47)	100(.29)
		.25	100(1.4)	100(3.2)	100(.00)	82(.08)	.00(.00)	98(.01)	95(.55)	52(61)	100(.08)
		.35	98(1.4)	97(3.2)	1.5(5.9)	78(.22)	.00(.37)	7.8(.00)	1.5(.28)	39(72)	100(.02)
		.45	.86(39)	2.7(57)	0.4(15)	.05(12)	.01(1.1)	2.7(.04)	.01(4.4)	32(77)	100(.00)

estimators.

Table 4.4 shows average outlier detection percentages over the four types of the variances of x-variables. The percentage that classifies an outlier as an outlier (*i.e.*, DP) of the DPR is 100% for all cases while the percentage that misclassifies a clean observation as an outlier (*i.e.*, WP) of the DPR ranges from 0% to 1.6%. This means that the DPR almost perfectly classifies the observations for all the four types of Ω_x. However, this is not true for other robust estimates. The MM does not detect any outlier when the contamination percentage is larger than 15% (*i.e.*, $e > .15$). The DR is not reliable at all when $e > .15$ because of its lower DP and higher WP, whereas S, LQD and LTD are not reliable for $e \geq .35$ because of their low DP's. The LMS, LTS and GM1 are not reliable for their outlier detection ability for 35% or 45% contamination percentages. Similar statements can be addressed for other simulations designs (a) through (f).

TABLE 4.5 *Maximum mean squared errors over four types of x-variables under Cauchy distributed errors*

n	p	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
	2	.985	.805	.405	.637	.625	.854	.478	.965	.498
60	5	7.17	8.11	3.96	6.69	6.64	7.43	2.95	48.4	4.85
	10	15.9	10.5	6.08	9.13	14.7	29.9	8.33	64.2	7.20
	2	.467	.456	.226	.404	.347	.330	.225	.388	.278
100	5	4.74	4.02	1.41	2.51	2.54	4.69	1.44	22.8	2.27
	10	11.1	7.30	3.32	5.23	6.14	20.6	3.39	36.9	3.82

4.2. Robustness for heavy tail and asymmetric distributions and for large sample with high dimension

We take errors, ϵ_i from Cauchy(0) distribution in the regression model $y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \epsilon_j$ where $i = 1, \ldots, n$ and x_{ij} are the same as before (*i.e.*, the four types of variances Ω_x). This process yields extremely large y's observed over the full range of the x-space. Although these large y's are not outliers under the Cauchy distribution, a good robust estimator should recognize such large y's as outliers to resist errors observed by a heavy tail distribution. Table 4.5 shows maximum MSEs of the nine robust estimators where the maximum for each estimator is taken over the four types of variances of x-variables for $n = 60, 100$ and $p = 2, 5$ and 10. The S is the best and LTD and DPR are the next showing that the DPR worked satisfactorily for a heavy tail error distribution as well.

We also investigate how the nine estimators are affected by asymmetric errors, using errors from an extreme value distribution with mean 0 and variance 1 with other conditions to be the same as in the heavy tail distribution. Table 4.6 presents maximum MSEs of the nine estimators where the maximum for each estimator is taken over the four types of variances of x-variables for $n = 60, 100$ and $p = 2, 5$ and 10. The table shows that MM and DPR are the best in this extreme value distribution. We also note from Table 4.6 that DR not as good as van Aelst *et al.* (2002) indicated.

Most robust estimators use resampling methods to avoid impractical computations in large samples. Thus, they are obtained using only a small portion of the sample and produce lower consistency rates and breakdown values than theoretically expected as the sample size grows (Hawkins and Olive, 2002). We exclude the LQD and LTD in this simulation study because calculating them takes too long. However, since the LQD and LTD have similar performance as

TABLE 4.6 *Maximum mean squared errors over four types of x-variables under extreme distributed errors*

n	p	*LMS*	*LTS*	*S*	*MM*	*GM1*	*LQD*	*LTD*	*DR*	*DPR*
	2	.738	.889	.318	.230	.260	.469	.407	.481	.232
60	5	6.63	6.94	2.87	1.99	2.64	2.80	2.67	20.4	2.21
	10	12.1	11.4	4.13	2.52	6.86	9.37	5.36	27.1	2.47
	2	.618	.721	.217	.129	.155	.313	.308	.276	.129
100	5	4.48	3.81	1.48	.997	1.25	2.17	1.36	17.3	1.17
	10	5.71	5.88	1.68	1.25	2.42	6.03	2.52	22.2	1.26

TABLE 4.7 *MSE and outlier detection percentage in large data set for* $n = 1,000$ *and* $p = 50$

e	*LMS*	*LTS*	*S*	*GM1*	*MM*	*DR*	*DPR*
	MSE (DP)	*MSE (DP)*	*MSE (DP)*	*MSE (DP)*	*MSE (DP)*	*MSE (DP)*	*MSE (DP)*
.05	.45(100)	.02(100)	.01(100)	6.27(100)	.01(100)	606(100)	.00(100)
.10	1.14(100)	.03(100)	.01(100)	62.7(100)	.00(100)	6862(31)	.00(100)
.15	761(89)	79(99)	670(90)	4702(64)	.00(100)	7041(22)	.00(100)
.20	953(87)	7139(.00)	6869(.00)	6221(41)	5871(.00)	7067(18)	.00(100)
.25	1176(84)	7089(.00)	6923(.00)	7883(32)	6154(.00)	7041(19)	.01(100)
.30	1324(82)	7200(.00)	7090(.00)	13191(12)	6482(.00)	7150(19)	.01(100)
.35	1419(80)	7083(.00)	7014(.00)	23397(3.0)	6550(.00)	7037(21)	.01(100)
.40	1413(48)	2738(.00)	7120(.00)	62003(.00)	6730(.00)	7122(24)	.01(100)
.45	3620(48)	6966(.05)	6952(.00)	70606(.00)	6679(.00)	6955(25)	.01(100)

the S estimator as shown in the four tables from Table 4.1 through Table 4.4, we may infer the behavior of the LQD and LTD by the S estimator.

To examine the effect of large samples on the seven robust estimators, we again consider the simulation design (c) in Section 4.1 with $n = 1,000$ and $p = 50$, *i.e.*, values which are 10 times those of Section 4.1 but keeping the same ratio of $n/p = 20$. We only consider $x_{ij} \sim$ *i.i.d.* $N(7.5, 4^2)$ (*i.e.*, the Ω_x in (i)) since the remaining three types of Ω_x give the same pattern as shown in Table 4.7. We replicate 500 times for each case.

Table 4.7 shows MSEs and outlier detection percentages in parentheses. The GM1 and DR are almost useless at the 5% or higher contamination percentages while the LMS, LTS and S are useless at a 15% or higher contamination percentage. It is interesting to compare these results with those observed from Table 4.2 and Table 4.3 where GM1 and DR were acceptable till 15%, LMS and LTS were not bad except at 35% or more contaminations with $n = 100$, and S was good up to 25%. This implies that LMS, LTS, S, GM1 and DR are inefficient for large

samples. Similar conclusions can be made for outlier detection percentages (*i.e.*, DP). Note that LMS has relatively high outlier detection percentages although its MSEs are very large. This means that a robust estimator with high outlier detection percentage generally does not imply its reliability.

However, as we expected, the DPR becomes better as the sample size increases and perfectly detects outliers. The MSEs remain almost the same for all contamination percentages. Furthermore, the DPR has the efficiency comparable to the OLS whose MSE is .0039 under the 0% contamination.

5. CONCLUSION REMARKS

We introduced a high breakdown point estimator called the data partitioning robust estimator (DPR). No computational problem arises in calculating the DPR. This is achieved by partitioning the observations into a finite number of subsets based on the means of independent and dependent variables. However, this partitioning makes the DPR not to possess the regression equivariance and the affine equivariance although the DPR still retain the permutation equivariance.

On the other hand, the previous high breakdown point estimators suffer from the computational problem although they are regression and affine equivariant. To overcome the computational problem, a resampling technique is adapted. Because of resampling, they are calculated using only partial observations and lose the permutation equivariance. This may damage the robustness theoretically expected.

Through empirical and simulation study, we showed how this contrast between the DPR and the previous robust estimators acts on their theoretical robustness in practice. The empirical study using the four real data sets showed that the DPR perfectly detects known outliers while the eight competitive robust estimators fail to detect outliers in at least one of the data sets. Extensive simulation studies showed that the DPR is almost unbiased, as efficient as OLS, has the smallest MSE in most cases, the most accurate outlier detection ability for all 48 simulation scenarios. In particular, the DPR is not only much better than other robust estimators in large sample but also works well for a heavy tail distribution and an asymmetric tail distribution.

Mainly because of lack of regression and affine equivariances, the asymptotic property, the influence function, and the maxbias curve of the DPR are not theoretically resolved. It remains as open problems.

ACKNOWLEDGEMENTS

We are grateful to the editor and anonymous referees for useful comments.

REFERENCES

VAN AELST, S., ROUSSEEUW, P. J., HUBERT, M. AND STRUYF, A. (2002). "The deepest regression method", *Journal of Multivariate Analysis*, **81**, 138–166.

COAKLEY, C. W. AND HETTMANSPERGER, T. P. (1993). "A bounded influence, high breakdown, efficient regression estimator", *Journal of the American Statistical Association*, **88**, 872–880.

CROUX, C., ROUSSEEUW, P. J. AND HÖSSJER, O. (1994). "Generalized S-estimators", *Journal of the American Statistical Association*, **89**, 1271–1281.

HAMPEL, F. R. (1974). "The influence curve and its role in robust estimation", *Journal of the American Statistical Association*, **69**, 383–393.

HAWKINS, D. M., BRADU, D. AND KASS, G. V. (1984). "Location of several outliers in multiple-regression data using elemental sets", *Technometrics*, **26**, 197–208.

HAWKINS, D. M. AND OLIVE, D. J. (2002). "Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm", *Journal of the American Statistical Association*, **97**, 136–159.

HINLOOPEN, J. AND WAGENVOORT, R. (1997). "On the computation and efficiency of a HBP-GM estimator some simulation results", *Computational Statistics & Data Analysis*, **25**, 1–15.

HÖSSJER, O. (1994). "Rank-based estimates in the linear model with high breakdown point", *Journal of the American Statistical Association*, **89**, 149–158.

MARTIN, R. D., YOHAI, V. J. AND ZAMAR, R. H. (1989). "Min-max bias robust regression", *The Annals of Statistics*, **17**, 1608–1630.

OLIVE, D. J. (2003). *"Applied Robust Statistics"*, Southern Illinois University.

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

ROUSSEEUW, P. J. AND HUBERT, M. (1999). "Regression depth", *Journal of the American Statistical Association*, **94**, 388–402.

ROUSSEEUW, P. J. AND LEROY, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

ROUSSEEUW, P. J. AND YOHAI, V. J. (1984). "Robust regression by means of S-estimators", *In Robust and Nonlinear Time Series Analysis* (Franke, J., Hardle, W. and Martin, R. D., eds), 256–272, Springer-Verlag, New York.

SALIBIAN-BARRERA, M. AND ZAMAR, R. H. (2004). "Uniform asymptotics for robust location estimates when the scale is unknown", *The Annals of Statistics*, **32**, 1434–1447.

SIMPSON, J. R. AND MONTGOMERY, D. C. (1998). "A robust regression technique using compound estimation", *Naval Research Logistics*, **45**, 125–139.

SIMPSON, D. G., RUPPERT, D. AND CARROLL, R. J. (1992). "On one-step GM estimates and stability of inferences in linear regression", *Journal of the American Statistical Association*, **87**, 439–450.

SIMPSON, D. G. AND YOHAI, V. J. (1998). "Functional stability of one-step GM-estimators in approximately linear regression", *The Annals of Statistics*, **26**, 1147–1169.

STROMBERG, A. J., HÖSSJER, O. AND HAWKINS, D. M. (2000). "The least trimmed differences regression estimator and alternatives", *Journal of the American Statistical Association*, **95**, 853–864.

WISNOWSKI, J. W., MONTGOMERY, D. C. AND SIMPSON, J. R. (2001). "A comparative analysis of multiple outlier detection procedures in the linear regression model", *Computational Statistics & Data Analysis*, **36**, 351–382.

YOHAI, V. J. (1987). "High breakdown-point and high efficiency robust estimates for regression", *The 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.