The Annals of Statistics
2005, Vol. 33, No. 3, 977–1035
DOI 10.1214/009053604000001138

DISCUSSION PAPER

BREAKDOWN AND GROUPS[1]

BY P. LAURIE DAVIES AND URSULA GATHER

University of Duisburg–Essen and Technical University Eindhoven, and University of Dortmund

The concept of breakdown point was introduced by Hampel [Ph.D. dissertation (1968), Univ. California, Berkeley; *Ann. Math. Statist.* **42** (1971) 1887–1896] and developed further by, among others, Huber [*Robust Statistics* (1981). Wiley, New York] and Donoho and Huber [In *A Festschrift for Erich L. Lehmann* (1983) 157–184. Wadsworth, Belmont, CA]. It has proved most successful in the context of location, scale and regression problems. Attempts to extend the concept to other situations have not met with general acceptance. In this paper we argue that this is connected to the fact that in the location, scale and regression problems the translation and affine groups give rise to a definition of equivariance for statistical functionals. Comparisons in terms of breakdown points seem only useful when restricted to equivariant functionals and even here the connection between breakdown and equivariance is a tenuous one.

1. Introduction.

1.1. *Contents.* In Section 1 we give a short overview of the concepts of breakdown and equivariance and a brief discussion of previous work. Section 2 contains notation and the standard definition of breakdown and in Section 3 we derive an upper bound for the breakdown points of equivariant statistical functionals. Section 4 contains some old and new examples in light of the results of Section 3. The attainability of the bound is discussed in Section 5 and finally in Section 6 we argue that the connection between breakdown and equivariance is fragile.

1.2. *Breakdown points and equivariance.* The notion of breakdown point was introduced by Hampel (1968, 1971). Huber (1981) took a functional analytical approach; a simplified version for finite samples was introduced by Donoho (1982) and Donoho and Huber (1983). To be of practical use a definition of breakdown should be simple, reflect behavior for finite samples and allow comparisons

Received November 2002; revised January 2004.
[1]Supported in part by Sonderforschungsbereich 475, University of Dortmund.
AMS 2000 subject classifications. Primary 62G07; secondary 65D10, 62G20.
Key words and phrases. Equivariance, breakdown point, robust statistics.

between relevant statistical functionals. With some proviso (see Section 6) these goals have been achieved for location, scale and regression problems in $\mathbb{R}^k$ [see, e.g., Hampel (1975), Rousseeuw (1984, 1985), Lopuhaä and Rousseeuw (1991), Davies (1993), Stahel (1981), Donoho (1982), Tyler (1994) and Gather and Hilker (1997)] and for related problems [see, e.g., Ellis and Morgenthaler (1992), Davies and Gather (1993), Becker and Gather (1999), Hubert (1997), Terbeck and Davies (1998), He and Fung (2000) and Müller and Uhlig (2001)]. This success has led many authors to develop definitions applicable in other situations. We mention nonlinear regression [Stromberg and Ruppert (1992)], time series [Martin and Jong (1977), Papantoni-Kazakos (1984), Tatum and Hurvich (1993), Lucas (1997), Mendes (2000), Ma and Genton (2000) and Genton (2003)], radial data [He and Simpson (1992)], the binomial distribution [Ruckstuhl and Welsh (2001)] and more general situations as in Sakata and White (1995), He and Simpson (1993) and Genton and Lucas (2003). An essential component of the theory of high breakdown location, scale and regression functionals is the idea of equivariance. With the exception of He and Simpson (1993), none of the above generalizations of breakdown point incorporates a concept of equivariance. It is as if the equivariance part has been relegated to the small print and then forgotten [see 't Hooft (1997) for the role of the small print in physics]. The main purpose of this paper is to emphasize the role of a group structure, to give some new examples and to point out the fragility of the connection.

2. A definition of breakdown point. We consider a measurable sample space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and the family $\mathcal{P}$ of all nondegenerate probability measures on this space. We assume that a pseudometric d is defined on $\mathcal{P}$ which satisfies

$$\sup_{P,Q\in\mathcal{P}} d(P, Q) = 1 \tag{2.1}$$

and for all P, Q_1, $Q_2 \in \mathcal{P}$ and α, $0 < \alpha < 1$,

$$d\big(\alpha P + (1-\alpha)Q_1, \alpha P + (1-\alpha)Q_2\big) \le 1 - \alpha. \tag{2.2}$$

We consider functionals T which map $\mathcal{P}$ into a parameter space Θ which is equipped with a pseudometric D on $\Theta \times \Theta$ satisfying

$$\sup_{\theta_1,\theta_2} D(\theta_1, \theta_2) = \infty. \tag{2.3}$$

The breakdown point $\varepsilon^*(T, P, d, D)$ of the functional T at the distribution P with respect to the pseudometrics d and D is defined by

$$\varepsilon^*(T, P, d, D) = \inf\left\{\varepsilon > 0 : \sup_{d(P,Q)<\varepsilon} D\big(T(P), T(Q)\big) = \infty\right\}. \tag{2.4}$$

The finite-sample replacement breakdown point of a functional T is defined as follows. If $\mathbf{x}_n = (x_1, \ldots, x_n)$ is a sample of size n, we denote its empirical

distribution by $P_n = \sum_{i=1}^n \delta_{x_i}/n$. Let $\mathbf{y}_{n,k}$ be a sample obtained from $\mathbf{x}_n$ by altering at most k of the x_i and denote the empirical distribution of $\mathbf{y}_{n,k}$ by $Q_{n,k}$. The finite-sample breakdown point (fsbp) of T at the sample $\mathbf{x}_n$ (or P_n) is then defined by [see Donoho and Huber (1983)]

$$\text{fsbp}(T, \mathbf{x}_n, D) = \frac{1}{n} \min\left\{k \in \{1, \ldots, n\} : \sup_{Q_{n,k}} D\big(T(P_n), T(Q_{n,k})\big) = \infty\right\}. \tag{2.5}$$

3. Groups and equivariance.

3.1. *An upper bound for the breakdown point.* Let G be a group of measurable transformations g of $\mathcal{X}$ onto itself with unit element ι. For any $P \in \mathcal{P}$ and any $g \in G$ we define P^g by $P^g(B) = P(g^{-1}(B))$. The group G induces a group $H_G = \{h_g : g \in G\}$ of transformations $h_g : \Theta \to \Theta$ and a functional $T : \mathcal{P} \to \Theta$ is called equivariant with respect to G if

$$T(P^g) = h_g(T(P)) \qquad \text{for all } g \in G,\ P \in \mathcal{P}. \tag{3.1}$$

We set

$$G_1 = \left\{g \in G : \lim_{n\to\infty} \inf_\theta D\big(\theta, h_{g^n}(\theta)\big) = \infty\right\}. \tag{3.2}$$

The restriction of $g \in G$ to a set $B \in \mathcal{B}$ will be denoted by $g_{|B}$. Given this we define

$$\Delta(P) = \sup\{P(B) : B \in \mathcal{B},\ g_{|B} = \iota_{|B} \text{ for some } g \in G_1\}. \tag{3.3}$$

The functional $\Delta(P)$ appears explicitly in the expression for the highest possible breakdown point. We give two examples. If G is the translation group on $\mathbb{R}^k$, then the defining set in (3.3) is empty so that $\Delta(P) = 0$. For affine transformations $Ax + b = x$ for $x \in B$ and consequently $\Delta(P)$ is the greatest measure of a lower-dimensional hyperplane.

THEOREM 3.1. *With the above notation and under the assumption that* $G_1 \neq \varnothing$ *we have*

$$\varepsilon^*(T, P, d, D) \leq \big(1 - \Delta(P)\big)/2 \tag{3.4}$$

for all G*-equivariant functionals* T*, for all* $P \in \mathcal{P}$*, for all pseudometrics* d *and* D *satisfying* (2.1)–(2.3).

PROOF. Let B_0 and $g \in G_1$ be such that $g_{|B_0} = \iota_{|B_0}$. Consider the measures defined by $Q_1(B) = P(B \cap B_0)$, $Q_2(B) = P(B) - Q_1(B)$ and $Q_n(B) = (Q_2(B) + Q_2^{g^n}(B))/2 + Q_1(B)$ for $B \in \mathcal{B}$. As $Q_1^g = Q_1^{g^{-1}} = Q_1$ we have

$Q_n^{g^{-n}} = (Q_2^{g^{-n}} + Q_2)/2 + Q_1$ and on using (2.2) it follows that $d(Q_n^{g^{-n}}, P) \leq (1 - P(B_0))/2$ and $d(Q_n, P) \leq (1 - P(B_0))/2$. Clearly

$$D\big(T(Q_n^{g^{-n}}), T(Q_n)\big) \leq D\big(T(P), T(Q_n^{g^{-n}})\big) + D\big(T(P), T(Q_n)\big).$$

The definition of G_1 implies

$$\lim_{n\to\infty} \big(D\big(T(P), T(Q_n^{g^{-n}})\big) + D\big(T(P), T(Q_n)\big)\big) = \infty$$

and we deduce that for any $\varepsilon > (1 - P(B_0))/2$

$$\sup_{d(P,Q)<\varepsilon} D\big(T(P), T(Q)\big) = \infty.$$

The claim of the theorem follows. □

THEOREM 3.2. *With the above notation and under the assumption $G_1 \neq \varnothing$ we have*

$$\text{fsbp}(T, \mathbf{x}_n, D) \leq \left\lfloor \frac{n - n\Delta(P_n) + 1}{2} \right\rfloor \Big/ n. \tag{3.5}$$

PROOF. The proof follows the lines of the proof of Theorem 3.1. For the details we refer to Davies and Gather (2002). □

4. Examples.

4.1. *Location functionals and the translation group.* We take $\mathcal{X}$ to be k-dimensional Euclidean space $\mathbb{R}^k$ and G the translation group. The parameter space Θ is $\mathbb{R}^k$ and the group H_G is again the translation group. The pseudometric D on Θ is the Euclidean metric. Any pseudometric d which satisfies (2.1) and (2.2) will suffice. This applies for all other examples so we no longer specify d. As mentioned just after (3.3), we have $\Delta(P) = 0$ for all P and Theorem 3.1 now states that $\varepsilon^*(T, P, d, D) \leq 1/2$ for any translation equivariant functional.

4.2. *Scatter functionals and the affine group.* $\mathcal{X}$ is k-dimensional Euclidean space $\mathbb{R}^k$ and G is the affine group, the parameter space Θ is the space Σ_k of nonsingular symmetric $(k \times k)$-matrices and the elements h_g of H_G are defined by

$$h_g(\sigma) = A\sigma A^t, \qquad \sigma \in \Sigma_k, \tag{4.1}$$

where $g(x) = Ax + b$. The pseudometric on Σ_k is given by

$$D(\sigma_1, \sigma_2) = \big|\log\big(\det(\sigma_1\sigma_2^{-1})\big)\big|, \qquad \sigma_1, \sigma_2 \in \Sigma_k \tag{4.2}$$

and hence $G_1 = \{g : g(x) = Ax + a, \det(A) \neq 1\}$. We have $\Delta(P) = \sup\{P(B) : B$ is a hyperplane of dimension $\leq k - 1\}$ and Theorem 3.1 is now Theorem 3.2 of Davies (1993).

4.3. *Regression functionals and the translation group.* $\mathcal{X}$ is now $(k+1)$-dimensional Euclidean space $\mathbb{R}^k \times \mathbb{R}$, where the first k components define the design points and the last component is the corresponding value of y. The group G consists of all transformations

$$g\big((x^t, y)^t\big) = (x^t, y + x^t a)^t, \qquad (x^t, y)^t \in \mathbb{R}^k \times \mathbb{R}, \tag{4.3}$$

with $a \in \mathbb{R}^k$. The space Θ is $\mathbb{R}^k$ and a functional $T : \mathcal{P} \to \Theta$ is equivariant with respect to the group if $T(P^g) = T(P) - a$. The arguments go through as in Section 4.2 and the result is Theorem 3.1 of Davies (1993).

4.4. *Time series and realizable linear filters.* We denote the space of doubly infinite series of complex numbers by $\mathbb{C}^{\mathbb{Z}}$ and define

$$\mathcal{X} = \mathcal{X}_\delta = \left\{ x \in \mathbb{C}^{\mathbb{Z}} : \sum_{j=0}^{\infty} |x_{n-j}|(1+\delta)^{-j} < \infty \text{ for all } n \in \mathbb{Z} \right\} \tag{4.4}$$

for some $\delta > 0$ and equip $\mathcal{X}$ with the usual Borel σ-algebra. Define the group $\tilde{G}$ by

$$\tilde{G} = \left\{ \tilde{g} : \tilde{g} : \Gamma_{1+\varepsilon} \to \mathbb{C}, \text{ analytic and bounded with } \inf_{z \in \Gamma_{1+\varepsilon}} |\tilde{g}(z)| > 0 \right\}, \tag{4.5}$$

where Γ_r denotes the open disc in $\mathbb{C}$ of radius r and $\varepsilon > \delta$. Each such $\tilde{g} \in \tilde{G}$ has a power series expansion $\tilde{g}(z) = \sum_{j=0}^{\infty} g_j z^j$ and defines a linear filter g on $\mathcal{X}$,

$$(g(x))_n = \sum_{j=0}^{\infty} x_{n-j} g_j, \qquad n \in \mathbb{Z}. \tag{4.6}$$

The linear filters g form the group G. The parameter space Θ is the space of finite distribution functions F on $(-\pi, \pi]$. For $F \in \Theta$ and $g \in G$ we define $h_g(F)$ by

$$h_g(F) = F_g \qquad \text{where } dF_g(\lambda) = \big|g\big(\exp(i\lambda)\big)\big|^2 \, dF(\lambda). \tag{4.7}$$

Finally, the pseudometric D on Θ is defined by

$$D(F_1, F_2) = \begin{cases} \displaystyle\int_{-\pi}^{\pi} \left| \log\left(\frac{dF_1}{dF_2} \right) \right| d\lambda, & F_1 \asymp F_2, \\ \infty, & \text{otherwise}, \end{cases} \tag{4.8}$$

where $F_1 \asymp F_2$ means that the two measures are absolutely continuous with respect to each other. The conditions placed on the group G imply that

$$\inf_{\lambda \in (-\pi, \pi]} \big|g\big(\exp(i\lambda)\big)\big| > 0, \qquad dF_g/dF = \big|g\big(\exp(i\lambda)\big)\big|^2$$

and

$$D\big(F, h_g(F)\big) = 2 \int_{-\pi}^{\pi} \big|\log\big(g\big(\exp(i\lambda)\big)\big)\big| \, d\lambda$$

for any F in Θ and $g \in G$. This implies

$$D\big(F, h_{g^n}(F)\big) = 2n \int_{-\pi}^{\pi} \big|\log\big(g\big(\exp(i\lambda)\big)\big)\big|\, d\lambda$$

and hence

$$\lim_{n\to\infty} n \int_{-\pi}^{\pi} \big|\log\big(g\big(\exp(i\lambda)\big)\big)\big|\, d\lambda = \infty$$

unless $|g(\exp(i\lambda))| = 1$, $-\pi < \lambda \le \pi$. This, however, would imply $g(z) = z$ and so we see that $G_1 \neq \varnothing$. Theorem 3.1 gives

$$\varepsilon^*(T, P, d, D) \le \big(1 - \Delta(P)\big)/2.$$

In the present situation the definition (3.3) of $\Delta(P)$ reduces to

$$\Delta(P) = \sup\left\{ P(B) : B = \left\{ x : x_n = \sum_{j=0}^{\infty} x_{n-j} g_j, n \in \mathbb{Z} \right\}, g \in G_1 \right\}, \tag{4.9}$$

which is effectively the maximum probability that x is deterministic. If P is a stationary Gaussian measure with spectral distribution F whose absolutely continuous part has density f_{ac}, then the Szegö (1920) alternative is $\Delta(P) = 0$ or 1 according to whether

$$\int_{-\pi}^{\pi} \log\big(f_{ac}(\lambda)\big)\, d\lambda > \text{or} = -\infty.$$

4.5. *The Michaelis–Menten model.* The Michaelis–Menten model may be parameterized as

$$y = \frac{ax}{cx + 1/a} + \varepsilon, \qquad a, c, x \in \mathbb{R}_+ = (0, \infty) \tag{4.10}$$

with $\theta = (a, c)$. $\mathcal{X}$ is $\mathbb{R}_+ \times \mathbb{R}$ and the elements g of G are defined by $g((x, y)) = (\alpha x, y)$ with $\alpha > 0$. The elements h_g of the induced group are given by $h_g(\theta) = (a/\sqrt{\alpha}, c/\sqrt{\alpha})$. We take the metric D to be given by

$$D(\theta_1, \theta_2) = |a_1 - a_2| + |a_1^{-1} - a_2^{-1}| + |c_1 - c_2|.$$

As $g((x, y)) = (x, y)$ only for $g = \iota$ we see that $G_1 \neq \varnothing$ and that $\Delta(P) = 0$. This implies a highest finite-sample breakdown point of $\lfloor (n+1)/2 \rfloor / n$, which is clearly attainable. Extensions to the real linear fractional group are possible.

4.6. *Logistic regression I.* Logistic regression is a binomial model with covariates. For the binomial distribution itself it has been shown by Ruckstuhl and Welsh (2001) that a breakdown point of 1 is attainable by functionals which are equivariant with respect to the two-element group $G = \{\iota, g\}$ where $g(x) = 1 - x$ and $h_g(p) = 1 - p$. As pointed out by Peter Rousseeuw (comment at the ICORS

2002 meeting in Vancouver), this is the natural group for the binomial distribution. The logistic regression model is

$$P(Y=1|x)=\exp(\theta_0+x^t\tilde{\theta})/\big(1+\exp(\theta_0+x^t\tilde{\theta})\big),\qquad \theta=(\theta_0,\tilde{\theta}^t)^t\in\mathbb{R}^{k+1}, \tag{4.11}$$

where $x^t=(x_1,\dots,x_k)$ are the covariates associated with the random variable Y. The sample space is $\mathcal{X}=\{0,1\}\times\mathbb{R}^k$ and the parameter space Θ is $\mathbb{R}^{k+1}$. The group G is generated by the composition of transformations of the form

$$(y,x^t)^t\to(1-y,x^t)^t, \tag{4.12}$$

$$(y,x^t)^t\to\big(y,\mathcal{A}(x)^t\big)^t, \tag{4.13}$$

where $\mathcal{A}$ is a nonsingular affine transformation $\mathcal{A}(x)=Ax+a$. The group H_G of transformations of Θ induced by G is given by

$$h_g(\theta)=-\theta,\qquad g \text{ as in (4.12)}, \tag{4.14}$$

$$h_g\big((\theta_o,\tilde{\theta}^t)^t\big)=\big(\theta_0-a^t(A^t)^{-1}\tilde{\theta},\big((A^t)^{-1}(\tilde{\theta})\big)^t\big)^t,\qquad g \text{ as in (4.13)}. \tag{4.15}$$

The metric D on Θ is taken to be the Euclidean metric. All the conditions for Theorem 3.1 are satisfied except that $G_1=\varnothing$ and indeed the constant functional $T(P)=0$ for all P is equivariant with breakdown point 1. If the constant functional is not thought to be legitimate, an alternative one is the following. For $\varepsilon>0$ we define T by

$$T(P)=\arg\min_{\theta_0,\tilde{\theta}}\int\left[\left(y-\frac{\exp(\theta_0+x^t\tilde{\theta})}{1+\exp(\theta_0+x^t\tilde{\theta})}\right)^2+\varepsilon(\theta_0+x^t\tilde{\theta})^2\right]dP(x,y). \tag{4.16}$$

The additional term is a form of regularization which prevents explosion in the case where the sets of x's with $y=1$ and with $y=0$ are separated by a hyperplane. The functional T is equivariant. Consider a data set which is such that any set of $(k+1)$-vectors $(1,x_{j_i}^t)^t$, $i=1,\dots,k+1$, is linearly independent. On denoting the empirical distribution of a replacement sample by P_n^* we note that $T(P_n^*)$ remains bounded for all replacement samples which contain at least $k+1$ of the original sample's values. The finite-sample breakdown point is therefore $1-k/n$.

4.7. *Logistic regression II.* We consider the growth model

$$Y(t)=\exp(a+bt)/\big(1+\exp(a+bt)\big)+\varepsilon(t), \tag{4.17}$$

which has an obvious equivariance structure. We define $\psi(y)$ by

$$\psi(y)=\max\big\{0,\min\{1,y\}\big\}$$

and a functional T by

$$T(P) = \arg\min_{a,b} \int \big(\psi(y) - \exp(a+bt)/\big(1+\exp(a+bt)\big)\big)^2 \, dP(y,t).$$

Given a data set $(y(t_i), t_i)$, $i = 1, \ldots, n$, we see that T will only break down if there exists a t such that $y(t_i) = 0$ for all $t_i < t$ and $y(t_i) = 1$ for all $t_i > t$ or vice versa. From this it follows that in general the finite-sample breakdown point will be $1 - 1/n$. This is much higher than the breakdown point of the LMS functional, which is about $1/2$ [see Stromberg and Ruppert (1992), Section 5].

5. Attaining the bound.

5.1. *Location functionals.* The translation equivariant L_1-functional

$$T(P) = \arg\min_{\mu} \int (\|x - \mu\| - \|x\|) \, dP(x) \tag{5.1}$$

attains the bound of $1/2$ of Section 4.1. It is not affine equivariant and attempts to prove the bound of $1/2$ for affine equivariant functionals in $\mathbb{R}^k$ with $k \geq 2$ have not been successful [Niinimaa, Oja and Tableman (1990), Lopuhaä and Rousseeuw (1991), Gordaliza (1991), Lopuhaä (1992) and Donoho and Gasko (1992)]. The proof of Theorem 3.1 also fails for the affine group as $G_1 = \varnothing$. That a bound of $1/2$ does not hold globally is shown by the example $\mathcal{X} = \mathbb{R}^2$ with point mass $1/3$ on the points $x_1 = (0, 1)$, $x_2 = (0, -1)$, $x_3 = (\eta\sqrt{3}, 0)$. More generally, in k dimensions there are samples for which $1/(k+1)$ is the maximal breakdown point. In spite of this, there are samples where a finite-sample breakdown point of $1/2$ is attainable. The construction is somewhat complicated and may be found in Davies and Gather (2002).

5.2. *Scatter functionals.* The median absolute deviation (MAD) has a finite-sample breakdown point of $\max(0, 1/2 - \Delta(P_n))$, which is less than the upper bound of Theorem 3.2. We propose a modification of the MAD which does attain the upper bound. For a probability measure P we define the interval $I(P, \lambda)$ by $I(P, \lambda) = [\text{med}(P) - \lambda, \text{med}(P) + \lambda]$ and write

$$\Delta(P, \lambda) = \max\{P(\{x\}) : x \in I(P, \lambda)\}.$$

The new scale functional MAD^* is defined by

$$\text{MAD}^*(P) = \min\big\{\lambda : P(I(P, \lambda)) \geq \big(1 + \Delta(P, \lambda)\big)/2\big\},$$

which can easily be calculated. It achieves the upper bound of Theorem 3.2. The breakdown point in terms of metrics depends on the metric used [see Huber (1981), page 110]. For the Kuiper metric based on one interval the breakdown point is $(1 - \Delta(P))/3$ [see also Davies (1993)] while for the Kuiper metric based on three intervals it is $(1 - \Delta(P))/2$ [see Davies and Gather (2002)].

6. Final remarks. As mentioned in the Introduction the definition of breakdown point should meet the following three goals: it should be simple, it should reflect the behavior of statistical functionals for finite samples and it should allow useful comparisons between statistical functionals. We examine these demands more closely for the case of a location functional in $\mathbb{R}$. The definition of breakdown point (2.4) involves a limiting operation and this is an essential part of its simplicity. If ∞ in (2.4) were replaced by some large number the simplicity would be lost. The simplification resulting from the limiting operation will only be successful if the resulting definition reflects the behavior for finite samples. The situation is analogous to the limiting operation of differentiation which reflects the behavior of the function for small but finite values. The breakdown points of $1/n$ for the mean and $1/2$ for the median do reflect their finite-sample behavior. As the median is translation equivariant and the highest breakdown point for such functionals is $1/2$, we seem to have achieved all three goals. If no restrictions were imposed on the class of allowable functionals, then breakdown points of 1 become attainable. We know of no situation not based on equivariance considerations where it can be shown that the highest breakdown point for a class of reasonable functionals is less than 1. A referee suggested the following example: estimate b in the model $E(y|x) = bx$ from $2m$ points at $x = 0$ and another m points at $x = 1$ where the conditional distribution of y given x is normal with mean zero and variance 1. The problem is to construct a consistent estimator with a breakdown point of more than $1/3$. We construct one with breakdown point 1. We give a finite-sample version. The data points are $(x_1, y_1), \ldots, (x_n, y_n)$ with empirical distribution P_n. If the x_i are all equal we put $T(P_n) = 0$. Otherwise we set

$$T(P_n) = \max\{-n, \min\{n, T_{\mathrm{LS}}(P_n)\}\}, \tag{6.1}$$

where T_{LS} is the least squares estimator through the origin. As $|T(P_n)|$ is bounded by n for any empirical distribution P_n, it has finite-sample breakdown point 1. On the other hand it is consistent. Equivariance considerations prohibit such a construction. In certain situations location functionals which are not translation equivariant may be preferred. If, for example, there is prior knowledge about the range of possible values of the location, then this can be exploited to give a breakdown point of 1. In all the situations we have considered where a breakdown point of 1 is attainable, it has proved to be quite easy to produce a perfectly sensible functional which attains or almost attains a breakdown point of 1. If this had been the case for equivariant functionals, we suspect that not so much research would have been devoted to the problem of high breakdown functionals. The breakdown point of $1/2$ for the median reflects its behavior at the following samples:

$$(1.5, 1.8, 1.3, 1.5 + \lambda, 1.8 + \lambda, 1.3 + \lambda), \tag{6.2}$$

$$(1.5, 1.8, 1.3, 1.51 + \lambda, 1.8 + \lambda, 1.3 + \lambda). \tag{6.3}$$

In both cases as λ tends to infinity the median breaks down in spite of the fact that the proof of Theorem 3.2 only covers the behavior at sample (6.2). Indeed any

translation equivariant functional will break down at sample (6.2) but it is easy to define translation equivariant functionals which do not break down at sample (6.3). Although a functional which does not break down at (6.3) may seem artificial, there are quite plausible situations where a similar phenomenon occurs. The noise may be simple white noise and the signal a very small subset of the data which lies very close to a straight line. It may well be possible to find this subset in spite of 99% of the data being noise and moreover, this may be accomplished in an equivariant manner. The behavior of the median at sample (6.3) is not explained by its translation equivariance and its breakdown point of $1/2$. The median must have some other, as yet unspecified, property beyond equivariance which makes the breakdown point of $1/2$ a good description of its behavior. Thus even in the case of equivariance the success of the concept of breakdown point would seem to be more fragile than is generally supposed. It is perhaps a case of invisible small print.

Acknowledgments. We acknowledge the work of two referees and an Associate Editor whose comments on the two versions of this paper led to a number of improvements in content and style.

REFERENCES

BECKER, C. and GATHER, U. (1999). The masking breakdown point of multivariate outlier identification rules. *J. Amer. Statist. Assoc.* **94** 947–955.

DAVIES, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21** 1843–1899.

DAVIES, P. L. and GATHER, U. (1993). The identification of multiple outliers (with discussion). *J. Amer. Statist. Assoc.* **88** 782–801.

DAVIES, P. L. and GATHER, U. (2002). Breakdown and groups. Technical Report 57, SFB 475, Univ. Dortmund.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827.

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

ELLIS, S. P. and MORGENTHALER, S. (1992). Leverage and breakdown in L_1 regression. *J. Amer. Statist. Assoc.* **87** 143–148.

GATHER, U. and HILKER, T. (1997). A note on Tyler's modification of the MAD for the Stahel–Donoho estimator. *Ann. Statist.* **25** 2024–2026.

GENTON, M. G. (2003). Breakdown-point for spatially and temporally correlated observations. In *Developments in Robust Statistics*, *International Conference on Robust Statistics 2001* (R. Dutter, P. Filzmoser, U. Gather and P. J. Rousseeuw, eds.) 148–159. Physica, Heidelberg.

GENTON, M. G. and LUCAS, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 81–94.

GORDALIZA, A. (1991). On the breakdown point of multivariate location estimators based on trimming procedures. *Statist. Probab. Lett.* **11** 387–394.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HAMPEL, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bull. Inst. Internat. Statist.* **46** (1) 375–391.

HE, X. and FUNG, W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *J. Multivariate Anal.* **72** 151–162.

HE, X. and SIMPSON, D. G. (1992). Robust direction estimation. *Ann. Statist.* **20** 351–369.

HE, X. and SIMPSON, D. G. (1993). Lower bounds for contamination bias: Globally minimax versus locally linear estimation. *Ann. Statist.* **21** 314–337.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

HUBERT, M. (1997). The breakdown value of the L_1 estimator in contingency tables. *Statist. Probab. Lett.* **33** 419–425.

LOPUHAÄ, H. P. (1992). Highly efficient estimators of multivariate location with high breakdown point. *Ann. Statist.* **20** 398–413.

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248.

LUCAS, A. (1997). Asymptotic robustness of least median of squares for autoregressions with additive outliers. *Comm. Statist. Theory Methods* **26** 2363–2380.

MA, Y. and GENTON, M. G. (2000). Highly robust estimation of the autocovariance function. *J. Time Series Anal.* **21** 663–684.

MARTIN, R. D. and JONG, J. (1977). Asymptotic properties of robust generalized M-estimates for the first order autoregressive parameter. Bell Laboratories Technical Memo, Murray Hill, NJ.

MENDES, B. V. M. (2000). Assessing the bias of maximum likelihood estimates of contaminated GARCH models. *J. Statist. Comput. Simul.* **67** 359–376.

MÜLLER, C. H. and UHLIG, S. (2001). Estimation of variance components with high breakdown point and high efficiency. *Biometrika* **88** 353–366.

NIINIMAA, A., OJA, H. and TABLEMAN, M. (1990). The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. *Statist. Probab. Lett.* **10** 325–328.

PAPANTONI-KAZAKOS, P. (1984). Some aspects of qualitative robustness in time series. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 218–230. Springer, New York.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

ROUSSEEUW, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications* (W. Grossman, G. Pflug, I. Vincze and W. Wertz, eds.) 283–297. Reidel, Dordrecht.

RUCKSTUHL, A. F. and WELSH, A. H. (2001). Robust fitting of the binomial model. *Ann. Statist.* **29** 1117–1136.

SAKATA, S. and WHITE, H. (1995). An alternative definition of finite-sample breakdown point with applications to regression-model estimators. *J. Amer. Statist. Assoc.* **90** 1099–1106.

STAHEL, W. A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH Zürich.

STROMBERG, A. J. and RUPPERT, D. (1992). Breakdown in nonlinear regression. *J. Amer. Statist. Assoc.* **87** 991–997.

SZEGÖ, G. (1920). Beiträge zur Theorie der Toeplitzschen Formen. *Math. Z.* **6** 167–202.

TATUM, L. G. and HURVICH, C. M. (1993). High breakdown methods in time series analysis. *J. Roy. Statist. Soc. Ser. B* **55** 881–896.

't HOOFT, G. (1997). *In Search of the Ultimate Building Blocks*. Cambridge Univ. Press.
TERBECK, W. and DAVIES, P. L. (1998). Interactions and outliers in the two-way analysis of variance. *Ann. Statist.* **26** 1279–1305.
TYLER, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.* **22** 1024–1044.

FACHBEREICH 06—MATHEMATIK
UND INFORMATIK
UNIVERSITÄT DUISBURG–ESSEN
45117 ESSEN
GERMANY
E-MAIL: davies@stat-math.uni-essen.de

FACHBEREICH STATISTIK
UNIVERSITÄT DORTMUND
44221 DORTMUND
GERMANY
E-MAIL: gather@statistik.uni-dortmund.de

DISCUSSION

BY MARC G. GENTON[1] AND ANDRÉ LUCAS

Texas A&M University and Vrije Universiteit

1. Introduction. In their interesting paper Davies and Gather draw our attention to what they call the "small print" in definitions of breakdown. Working from a formal group structure and a notion of equivariance, they show by a number of examples that a definition of breakdown may be void if not accompanied by a reasonable and precise group structure. This leads them to what we would label their key remark in Section 6: "We know of no situation not based on equivariance considerations where it can be shown that the highest breakdown point for a class of reasonable functionals is less than 1."

Though we agree with their general point that one has to take care not to come up with void definitions, or put differently, to make the small print explicit, we want to draw attention to the relation of their results to an alternative definition of breakdown. In particular, we claim that a different perspective on the notion of breakdown may resolve some of the small print issues.

The definition used by Davies and Gather in their equation (2.4) is a standard one and has its roots in the domain of location and scale estimation. As we argued in Genton and Lucas (2003), it is less useful in a setting with dependent data. For example, in a simple autoregression (AR) of order 1,

$$Y_t = \theta Y_{t-1} + e_t, \qquad \theta \in (-1, 1), e_t \sim N(0, 1), \tag{1}$$

the ordinary least squares (OLS) estimator for θ is driven to zero by replacing one of the Y_t's by an arbitrarily large number. Note that the OLS estimator thus

[1]Supported in part by NSF Grant DMS-02-04297.

tends to the center rather than the edge of the parameter space. Still, most people would agree that the estimator has lost its usefulness if only one extreme outlier is added. The reason is that the estimator no longer conveys useful information on the uncontaminated data. It is this latter notion that we want to put to the fore.

2. Breakdown point for (in)dependent observations. First, we would like to acknowledge that the breakdown definition as introduced in Genton and Lucas (2003) is subject to criticism raised by Davies and Gather (personal communication). One can easily construct an example of an estimator with breakdown point of 1 that would lose its information content on the uncontaminated process upon the addition of only one outlier. This is mainly due to the lack of a limiting operation in our original definition. Therefore, for the sake of this comment we introduce the following slightly adapted and simplified version of the definition in Genton and Lucas (2003).

Let Y denote a vector containing the sample of observations, and let $\mathcal{Y}$ denote the set of allowable samples. For example, in the asymptotic case Y might be a specific AR(1) process, while $\mathcal{Y}$ is the set of all stationary AR(1) processes. In a finite sample, Y might be a specific vector in $\mathbb{R}^n$, while $\mathcal{Y}$ is equal to $\mathbb{R}^n$. Let Z_k^ζ be an additive outlier process consisting of k outliers of magnitude ζ, such that we observe $Y + Z_k^\zeta$ rather than Y. To formalize the notion of information content on the uncontaminated process, we introduce the concept of badness set, which in this case we define as

$$R^*(Z_k^\zeta, \mathcal{Y}) = \{\theta(Y + Z_k^\zeta) | Y \in \mathcal{Y}\}, \tag{2}$$

where $\theta(\cdot)$ denotes the Fisher consistent estimator functional. Let μ denote an appropriate measure for the badness set. In most cases the Lebesgue measure suffices. Then we define the breakdown point of an estimator as

$$\text{bdp} = \frac{1}{n} \min\left\{k - 1 \,\middle|\, \text{for all compact } \mathcal{Y}' \subset \mathcal{Y}: \inf_{Z_k^\zeta} \mu\big(R^*(Z_k^\zeta, \mathcal{Y}') \cap R^*(0, \mathcal{Y}')\big) = 0\right\}. \tag{3}$$

An extension to the asymptotic case is straightforward. To see how the definition works, consider the regression example in Section 6 of Davies and Gather. We have $\mathcal{Y} = \mathbb{R}^{n\times 2}$ and $R^*(0, \mathcal{Y}) = [-n, n]$. The estimator is given by $\theta(Y) = \max(-n, \min(n, \theta^{\text{OLS}}(Y)))$, with $\theta^{\text{OLS}}(Y)$ the standard OLS estimator. We set μ to the standard Lebesgue measure. By taking $k = 1$ and letting the size of the outlier (ζ) diverge, the intersection of the two badness sets in the definition becomes $\{n\}$ or $\{-n\}$, which is a singleton with Lebesgue measure zero. Therefore, the estimator has broken according to our new definition. This appears reasonable as the estimator no longer conveys information about possibly uncontaminated samples.

3. Time series. The advantages of a different perspective on breakdown become even more apparent in the time series setting. Again consider our AR(1) example from (1). In the asymptotic case, define the i.i.d. additive outlier process $Z_{p,t}^{\zeta}$ with $\mathrm{P}[Z_{p,t}^{\zeta} = \zeta] = \mathrm{P}[Z_{p,t}^{\zeta} = -\zeta] = p/2$, and $Z_{p,t}^{\zeta} = 0$ otherwise. Figure 1 presents plots of the badness set $R^*(Z_p^{\zeta}, \mathcal{Y})$ associated with three estimators of θ as a function of ζ for $p = 5\%, 25\%, 50\%$. Here $\mathcal{Y}$ is the set of all stationary AR(1) processes; see the comment in the discussion below. We set μ to the standard Lebesgue measure.

The first estimator is the OLS estimator which in the above setting yields the badness set (2) based on the explicit expression

$$\theta_{\mathrm{OLS}}(Y + Z_p^{\zeta}) = \frac{\theta}{1 + p(1-\theta^2)\zeta^2}. \tag{4}$$

Letting the size of the outliers (ζ) diverge, we see that unless $p = 0$, the estimator θ_{OLS} tends to zero and the corresponding badness set becomes $\{0\}$; see the first row of Figure 1. Therefore, the asymptotic breakdown point of the OLS estimator for the AR(1) parameter θ is 0 in the setting described above.

The second estimator is the least median of squares (LMS) estimator of θ. It yields a badness set (2) based on the expression $\theta_{\mathrm{LMS}}(Y + Z_p^{\zeta}) = \arg\min_{\tilde{\theta} \in [-1,1]} c$

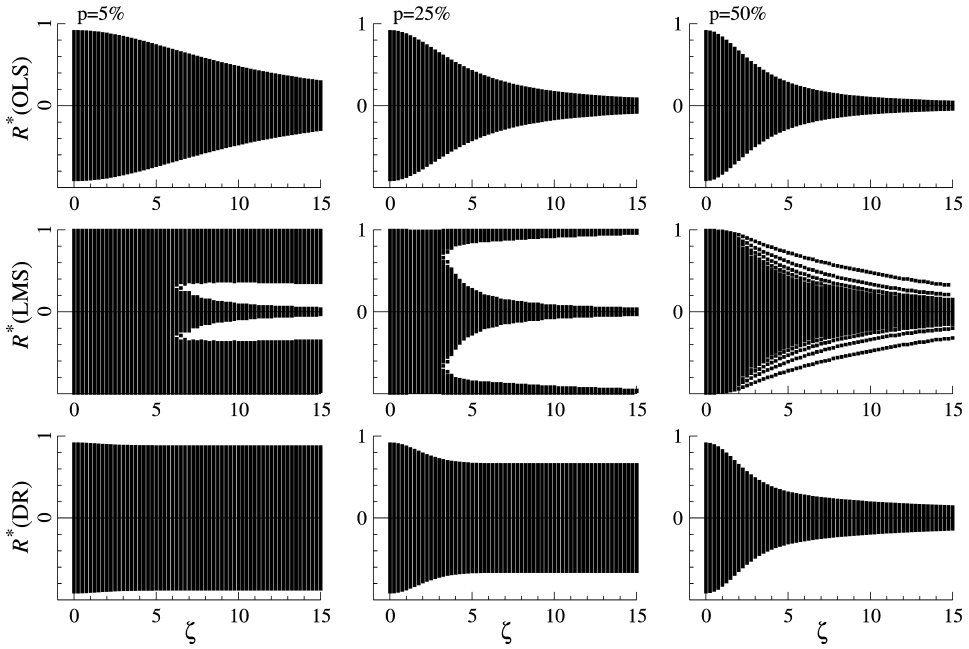

FIG. 1. *Plots of the badness set* $R^*(Z_p^{\zeta}, \mathcal{Y})$ *associated with three estimators of* θ *in the* AR(1) *as a function of* ζ *for* $p = 5\%, 25\%, 50\%$: *OLS* (top); *LMS* (middle); *DR* (bottom).

under the constraint

$$\frac{1}{2} = (1-p)^2\chi^2\left(\frac{c}{\tau^2};0\right) + p(1-p)\left[\chi^2\left(\frac{c}{\tau^2};\frac{1}{\tau^2}\zeta^2\right) + \chi^2\left(\frac{c}{\tau^2};\frac{\tilde{\theta}^2}{\tau^2}\zeta^2\right)\right] + \frac{p^2}{2}\left[\chi^2\left(\frac{c}{\tau^2};\frac{(1-\tilde{\theta})^2}{\tau^2}\zeta^2\right) + \chi^2\left(\frac{c}{\tau^2};\frac{(1+\tilde{\theta})^2}{\tau^2}\zeta^2\right)\right], \tag{5}$$

where $\tau^2 = 1 + (\theta - \tilde{\theta})^2/(1-\theta^2)$ and $\chi^2(x;\delta^2)$ denotes the cumulative distribution function evaluated at x of a chi-square random variable with noncentrality parameter δ^2. The second row of Figure 1 indicates that the badness set for $p = 5\%$ still takes a continuum of values, whereas it tends to the set $\{-1, 0, +1\}$ for $p = 25\%$. For $p = 50\%$, the badness set collapses to $\{0\}$ as ζ diverges. Therefore, letting the size of the outliers (ζ) diverge, the asymptotic breakdown point of the LMS estimator for the AR(1) parameter θ can be computed from (5) to be 22.1% in the setting described above.

The third estimator is the deepest regression (DR) estimator of θ defined by $\text{median}_t(Y_t/Y_{t-1})$. Under the additive outlier process described above, we need to consider the distribution of $(Y_t + Z_{p,t}^{\zeta})/(Y_{t-1} + Z_{p,t-1}^{\zeta})$. It yields a badness set (2) based on the expression $\theta_{\text{DR}}(Y + Z_p^{\zeta})$ given by the value c satisfying

$$\frac{1}{2} = (1-p)^2 G(c;0,0) + \frac{p(1-p)}{2}[G(c;\zeta,0) + G(c;0,\zeta) + G(c;-\zeta,0) + G(c;0,-\zeta)] + \frac{p^2}{4}[G(c;\zeta,\zeta) + G(c;\zeta,-\zeta) + G(c;-\zeta,\zeta) + G(c;-\zeta,-\zeta)], \tag{6}$$

where $G(x;a,b)$ is the cumulative distribution function evaluated at x of the ratio of two correlated normal random variables with means a and b, variances $1/(1-\theta^2)$ and correlation θ [see Hinkley (1969)]. The third row of Figure 1 indicates that the badness set still takes a continuum of values for $p = 5\%$ and $p = 25\%$, whereas it collapses to $\{0\}$ for $p = 50\%$ as ζ diverges. Thus, the asymptotic breakdown point of the DR estimator for the AR(1) parameter θ can be computed from (6) to be 50% in the setting described above.

It is interesting to note that the breakdown points of the LMS and DR estimators are markedly different for the AR(1) process above, whereas they are the same (50%) in the setting of simple regression. This indicates that our definition of breakdown allows us to distinguish between various robust estimators in the time series setting.

4. Discussion. The definition in (3) appears less dependent on a group structure than the definition used by Davies and Gather. Of course, also the definition in (3) has its limitations. For example, the definition cannot be used if

one wants to assess the breakdown of an estimator at a specific sample, that is, if $\mathcal{Y}$ is a singleton. The main drawback of conditioning on the sample is that one has to be very explicit about the region toward which the estimator breaks down, for example, to the edge of the parameter space. This may not be trivial for dependent data, as was shown in the AR(1) example for LMS. Moreover, conditioning the breakdown behavior on a specific sample may relate more to properties of the sample rather than of the estimator. The breakdown notion in (3) based on information revelation about the possible uncontaminated samples resolves this issue. That notion, however, can most easily be operationalized if there is a continuum of possible samples, which suffices for most cases studied in the literature.

A second possible limitation of (3) is that the user has to be explicit about the set $\mathcal{Y}$ of possible samples (or processes) Y. For example, if we consider stationary AR(1) processes in the asymptotic setting, the (asymptotic) breakdown point of the OLS estimator is 0. If, however, we consider AR(1) processes characterized by $\theta \in [-1, 1]$, the breakdown point is 1: the OLS estimator retains information about the distinction between stationary processes and processes with θ arbitrarily close to 1. In that sense the estimator does not break down, while it has broken down if one only wants to distinguish between alternative stationary processes; see Figure 1.

Finally, the definition in (3) is not very explicit about the measure μ. As mentioned, the Lebesgue measure suffices in most cases of practical interest. Despite the fact that empirical data have finite precision, one can work under the assumption that Y lies in a continuum to derive the breakdown properties of the estimator. The properties derived are usually also relevant for a setting with finite precision data. We do not exclude, however, that examples can be constructed where the Lebesgue measure is inappropriate. For example, the parameter space may be discrete and finite. In such cases, alternative measures μ must be used. Additionally, the restriction that the measure of the intersection of badness sets is zero may have to be replaced by something more complicated, like an infimum of $\inf_{Z_k^\zeta} \mu(R^*(Z_k^\zeta, \mathcal{Y}') \cap R^*(0, \mathcal{Y}'))$ over k.

The ideas and cautionary remarks in the paper of Davies and Gather are important and relevant. Effectively, they promote that breakdown is only a useful notion for "sensible" estimators and argue that equivariance is the crucial notion here. We argued that they mainly build on a restricted notion of breakdown. The focus of future research should be put on developing alternative definitions of breakdown that are less susceptible to the criticisms raised by Davies and Gather. The definition in (3) is such an attempt and tries to formalize the phenomena illustrated in Figure 1. In finite samples it is still susceptible to counterexamples, for example, $\theta(Y) = \max(-n, \min(n, \theta^{\mathrm{OLS}}(Y))) + 2(\mathrm{frac}(Y_1) - 1)/n$, where $\mathrm{frac}(x)$ denotes the fractional part of x, for Davies and Gather's example in Section 6, but the examples become increasingly contrived. Moreover, in the

asymptotic setting the small print issue appears to become even smaller, especially if we limit ourselves to estimators that are consistent and satisfy some form of continuity in the observations. Further developments along these lines appear promising.

REFERENCES

GENTON, M. G. and LUCAS, A. (2003). Comprehensive definitions of breakdown-points for independent and dependent observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 81–94.

HINKLEY, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika* **56** 635–639.

DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143
USA
E-MAIL: genton@stat.tamu.edu

DEPARTMENT OF FINANCE
ECO/FIN, VRIJE UNIVERSITEIT
DE BOELELAAN 1105
1081HV AMSTERDAM
THE NETHERLANDS
E-MAIL: alucas@feweb.vu.nl

DISCUSSION

BY FRANK HAMPEL

ETH Zürich

1. Introductory remarks. It is a great pleasure for me to be invited to comment upon the nice and elegant and in parts thought-provoking paper by Davies and Gather. The authors also asked me specifically to comment upon the historical roots of the breakdown point (BP), and my thoughts about it. I shall try to do so, stressing in particular aspects and work that are not published.

2. Some thoughts with the definition of the breakdown point. In my thesis [Hampel (1968)] I developed what was later also called the "infinitesimal approach to robustness," based on one-step Taylor expansions of statistics viewed as functionals (the "influence curves" or "influence functions"), a technology which for ordinary functions has long been indispensable in engineering and the physical sciences, and also for much theoretical work. However, it was always clear to me that this technology needed to be supplemented by an indication up to what distance (from the model distribution around which the expansion takes place) the linear expansions would be numerically, or at least semiquantitatively, useful. The simplest idea that came to my mind (simplicity being a virtue, also in view of Ockham's razor) was the distance of the nearest pole of the functional (if it was unbounded); see the graphs in Hampel, Ronchetti, Rousseeuw and Stahel [(1986),

pages 42, 48, 177]. Thus, right after defining the "bias function" (without using this term) as the (more complicated) bridge between model and pole, I introduced the "break-down point" on page 27 (Chapter C.4) of my thesis and, in a slight variant (by not requiring qualitative robustness anymore and therefore treating it as a purely global concept), as "breakdown point" on page 1894 in Hampel (1971). I was, of course, clearly inspired by Hodges (1967), whose intuition went in a similar direction, and by his "tolerance of extreme values"; however, his concept is not only much more limited, it is formally not even a special case of the breakdown point. [And contrary to a claim someone spread later, the term "breakdown point" does not occur anywhere in Hodges (1967).]

My definition of the BP is asymptotic, because I believe that suitable, elegant and properly interpreted asymptotics is much more informative and more generally applicable than specific or even clumsy finite-sample definitions. However, I also believe that asymptotic results need interpretations (and often numerical checks) in finite-sample frameworks, and lack of this may even be the biggest gap separating mathematical statistics from good applications of statistics [cf. Hampel (1998)].

Since I consider the finite-sample interpretations of an asymptotic definition (even different ones under different circumstances) an integral part of the properly interpreted definition, I never felt the need to introduce a general explicit definition of a finite-sample breakdown point. In fact, different needs require different definitions. In my eyes, the BP should be a flexible tool adapted to the requirement of specific problems (see also below).

Informal finite-sample BPs have been used in Andrews et al. (1972), and, for example, in many of my papers, starting with Hampel (1973). Often, the lower (or else upper) gross-error finite-sample BP is sufficient. But Grize (1978) showed the need for the "total-variation BP" in a specific situation concerning correlations. A standard reference is Donoho and Huber (1983). But in the background remains the fact that the BP is originally defined with the Prohorov distance. Very often we can forget this somewhat awkward distance and simplify; but whenever it is needed, we have to be ready to dig it out again.

The use of the Prohorov distance needs some explanation, also in view of the paper under discussion. Many good mathematical statisticians strive for the greatest generality, without regarding the practical implications. In some way, this is legitimate (and even required by the mathematical side of statistics). But I rather try to find the specific concepts most suitable for the problem at hand. Thus, as I explained elsewhere [e.g., in Hampel (1968)], I find it necessary to use the weak (formerly weak*) topology for general robustness problems, which is metrized by the Prohorov (former spelling Prokhorov) distance, which in turn has a nice interpretation in terms of the model deviations occurring in real life. [For more technical details, see Huber (1981).] This does not preclude the possibility of simplifying in specific situations. For example, Huber's (1964) gross-error model is theoretically too narrow in scope, but it captures the most important deviations

from the idealized model, and the solutions found are also useful and good in the more general situation [cf. Hampel (1992)].

Note that there is nothing about equivariance in my definition. If it is available, it simplifies life often tremendously and allows a nice mathematical theory with beautiful theorems; but I do not consider it an intrinsically necessary part of a general statistical theory (cf. also Fisher's view on his general theory of estimation).

It may be considered ironic in view of the present discussion paper that in my definition of the BP (with the compact proper subset of the parameter space), I explicitly thought of correlation statistics as examples, where there is no equivariance at all. As the authors correctly observe, such BPs have not become popular at all (so far), giving some credit to their stress on group structures. Compare also below.

3. Some further developments. The above idea of combining linear extrapolation with the BP was very successful in the cases tried [cf. Hampel et al. (1986), Subsection 1.3e, in particular Table 1 on page 50, which reproduces Huber's (1964) Table I—and thus his minimax results—very accurately; and Hampel (1983), page 214, which reproduces some of the Monte Carlo results in Andrews et al. (1972)]. As a rule of thumb, under mild conditions the linear extrapolations seem to be very accurate up to BP/4, and still numerically quite usable even somewhat beyond BP/2.

Another for me quite surprising success was the explanation of the (partly unsuspectedly bad) empirical behavior of various rejection rules just by means of the BP [see Hampel (1985)].

For regression I introduced the conditional BP given the design in Hampel [(1975), page 379] (implicitly and condensed because of the page limit imposed). It is more intricate, but also more informative (once the design is fixed or the data are in) than the unconditional BP [which was mostly used later on, except, e.g., in Hampel et al. (1986), page 328, unfortunately without stressing the difference between the two BPs].

Some definitions of variants of the BP, adapted to specific ANOVA-type problems, have been given by Hampel (1987), by Mili, Phaniraj and Rousseeuw (1990) and by Ruckstuhl (1995); see also Stahel, Ruckstuhl, Senn and Dressler (1994).

The BP seemed to be trivial, with BP $= 50\%$ easily possible in the models considered, until Maronna (1976) essentially showed the upper bound to be $= 1/\text{dimension}$ for "nice" equivariant estimators in multivariate and multiple regression situations. Much effort has since been put into keeping the equivariance and reaching BP $= 50\%$ with "pathological" estimators [the first prototype having been the "shordth" or "minimum median deviation" method in Hampel (1975), page 380, later popularized under the name "least median of squares"]. But from a practical point of view, I find it more reasonable to give up exact equivariance.

Gross errors are often partly in single coordinates and are not equivariant, even if the ideal model is.

For general nonlinear models, equivariance may not be attainable at all, but it may make perfect sense to look at the ("a posteriori") BP at and in a neighborhood of the fitted model. Compare also the highly condensed first sentence of 3.3 on page 380 in Hampel (1975), valid also under nonequivariance (and suggesting nice quantitative theorems under equivariance).

4. The thesis by Grize. In his unpublished Diplomarbeit, Yves-Laurent Grize (1978) made a thorough and deep investigation of the breakdown properties and influence functions of correlation measures, notably of the Kendall (K), Spearman (S) and quadrant (Q) rank correlations. He noted that the BP actually depends on the model F, and that also the specification of the "distance" may make a difference. For some F's, $\mathrm{BP(K)} = \mathrm{BP(S)} = 1$, while for others $\mathrm{BP(K)} = (3/2)\mathrm{BP(S)} < 1$, and in again another situation $\mathrm{BP(K)} = 0.29$, $\mathrm{BP(Q)} = 0.25$ and $\mathrm{BP(S)} = 0.21$. Grize showed that for correlations, the gross-error BP is often not suitable, and he used the better-fitting total-variation BP instead. He briefly also discussed the possibility of the (much more complicated) Prohorov-distance BP, and of ranks (by gross mistakes) outside the range from 1 to n. It appears that often Kendall's rank correlation is considerably more robust than Spearman's (and that there are some meaningful numbers and results to be taken out for statistical practice), but a lot depends on the precise specification of the situation.

My first reaction was disappointment. The results were just not as simple and beautiful as we then were used to in robust statistics. But the thesis is a valuable piece of work, and I regret very much that by some unfortunate circumstances it never found its way into the printed literature. Perhaps the time was not yet ripe for it. It seems that in recent years, some fragments of it are being rediscovered [cf. Bin Abdullah (1990) and Dehon and Croux (2003)], partly with seemingly contradictory results ("BP small" vs. "$\mathrm{BP} = 1$"), which may be due to insufficient care for the fine details (which really matter here). As the recent interest in the (formerly "too complicated") "bias function" (cf. above) shows, it may well be that in the near future "complicated" BPs without a natural equivariance structure will become more popular.

5. The small print. It seems to me that in the regression-through-0 example of the discussion paper, there is the same play with asymptotics (concerning both BP and consistency) going on which Fisher [(1956); cf. also Hampel (1998)] complained about when he defended his definition of consistency against Neyman's. In the case of the two location samples, I guess that the unnamed estimator not breaking down under the specific model and alternatives of (6.3) has a BP of 0.005—if the Prohorov distance is taken into account, which in such situations cannot be neglected. Thus, it really seems to be a case of small print that has been forgotten.

Acknowledgment. The author is grateful to U. Gather for digging out some references.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*: *Survey and Advances*. Princeton Univ. Press.

BIN ABDULLAH, M. (1990). On a robust correlation coefficient. *The Statistician* **39** 455–460.

DEHON, C. and CROUX, C. (2003). Maxbias curves for correlation estimators. Presentation at the International Conference on Robust Statistics, Univ. Antwerp, Belgium.

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

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, London.

GRIZE, Y. L. (1978). Robustheitseigenschaften von Korrelationsschätzungen. Diplomarbeit, Swiss Federal Institute of Technology (ETH), Zürich.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

HAMPEL, F. R. (1973). Robust estimation: A condensed partial survey. *Z. Wahrsch. Verw. Gebiete* **27** 87–104.

HAMPEL, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bull. Inst. Internat. Statist.* **46** (1) 375–391.

HAMPEL, F. R. (1983). The robustness of some nonparametric procedures. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.) 209–238. Wadsworth, Belmont, CA.

HAMPEL, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics* **27** 95–107.

HAMPEL, F. R. (1987). Some problems in statistics. In *Proc. First World Congress of the Bernoulli Society* (Y. Prohorov and V. V. Sazonov, eds.) **2** 253–256. VNU Science Press, Utrecht.

HAMPEL, F. R. (1992). Introduction to Huber (1964): Robust estimation of a location parameter. In *Breakthroughs in Statistics* **2**. *Methodology and Distribution* (S. Kotz and N. L. Johnson, eds.) 479–491. Springer, New York.

HAMPEL, F. R. (1998). Is statistics too difficult? *Canad. J. Statist.* **26** 497–513.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York.

HODGES, J. L., JR. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 163–186. Univ. California Press, Berkeley.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67.

MILI, L., PHANIRAJ, V. and ROUSSEEUW, P. J. (1990). High breakdown point estimation in electric power systems. In *Proc. 1990 IEEE International Symposium on Circuits and Systems* **3** 1843–1846. IEEE, New York.

RUCKSTUHL, A. F. (1995). Analysis of the T_2 emission spectrum by robust estimation techniques. Ph.D. dissertation no. 11170, Swiss Federal Institute of Technology (ETH), Zürich.

STAHEL, W. A., RUCKSTUHL, A. F., SENN, P. and DRESSLER, K. (1994). Robust estimation in the analysis of complex molecular spectra. *J. Amer. Statist. Assoc.* **89** 788–795.

SEMINAR FÜR STATISTIK
ETH ZÜRICH
CH-8092 ZÜRICH
SWITZERLAND
E-MAIL: hampel@stat.math.ethz.ch

DISCUSSION

BY XUMING HE

University of Illinois at Urbana–Champaign

The notion of breakdown point has influenced the robust statistics literature for over three decades. Professors Davies and Gather make a convincing argument that the common understanding of a high breakdown point is intimately connected to a group structure in the sample space. I would like to applaud the authors for a fine piece of work to formalize the connection. Inspired by their work, I would like to offer my opinions on the nature and the future of the breakdown point.

A high breakdown point is usually considered to be a virtue of a statistical procedure, because such a procedure is less affected by least favorable configurations of data contamination. Thus arises a natural question of how high the breakdown point can be in a given problem. For location equivariant functionals, $1/2$ is a tight upper bound on the breakdown point. In more structured problems and in more general settings, the defintion of a breakdown is less straightforward. Stromberg and Ruppert (1992) considered nonlinear regression. He and Simpson (1992) provided a definition of breakdown for general parameter spaces which might be compact. Instead of citing more work on breakdown in specific settings, I would emphasize that it is the simplicity and intuition that has made the breakdown point a popular measure of global robustness. Intuitively speaking, the breakdown point is the smallest fraction of data contamination that could make an estimator or test statistic totally uninformative or unusable.

The point I would like to make is that it is better to remember the spirit, not the letter, of any definition of the breakdown point. To illustrate this point, let us use the following definition of a finite-sample breakdown point.

Given a sample X_n of size n, the breakdown point of $T_n(X_n)$ is

$$\varepsilon_n^* = \min\left\{m/n : \sup_{X_{n,m}^*} d\big(T_n(X_{n,m}^*), T_n(X_n)\big) = \infty\right\},$$

where $X^*_{n,m}$ is obtained by replacing m out of n points in X_n by arbitrary values, and $d(a,b) \in (0,\infty)$ is some distance measure between a and b. Let us assume here that $d(a,b)$ can take arbitrarily large values. If T is a location estimator, one often takes $d(a,b) = |a-b|$. If T is a scale estimator, one may take $d(a,b) = |\log(a/b)|$. For most location, scale and regression estimators, the breakdown points do not depend very much on the initial sample X_n, but this is not always the case. Because the breakdown point is defined at each sample, we can easily modify any estimator T_n so that it will not break down at all according to this definition. For example, we can take $T^*_n(X_n) = \max\{-n, \min\{n, T_n(X_n)\}\}$ for a location estimator, and it will never be unbounded at any contamination.

Such a construction, however, violates the spirit of a high breakdown estimator, albeit it is mathematically legitimate. For a location estimator this problem can be eliminated by imposing location equivariance. In a general setting it is not clear what can be done. I use this example to stress that we should not try to exploit the mathematics of a statistical concept without a clear sense of purpose.

When someone claims to have found an estimator with breakdown point equal to 1, my first reaction tends to be that it might not be an appropriate use of the notion. Understanding and imposing a group equivariance structure on the estimator certainly helps, but it cannot eliminate inappropriate use of breakdown. In some problems (e.g., logistic regression) the group structure that can be identified might be very limited.

The notion of breakdown for a test statistic does not always carry the same implications as for an estimator. Davies and Gather discuss in their treatment of logistic regression whether the parameter value of 0 should be considered as a breakdown. I agree with the authors that the value of 0 plays no special role for an estimator. To study the breakdown of a statistical test, the value 0 often plays a special role. I refer to He, Simpson and Portnoy (1990) for more detail, but simply point out the obvious that one cannot judge the appropriateness of a breakdown definition without further specifics.

Take, for example, the classification tree. It is reasonable to say that a procedure breaks down if the classification rule becomes no better than a random guess. However, it is not obvious at all how to construct a tree with the highest possible breakdown point. Other notions of breakdown are also possible here.

I hope that the breakdown point will remain as a simple and intuitive concept. Maybe it falls into the same category as "outlier," where some degree of vagueness would win over more users. When every statistician starts to talk about his or her own notion of a breakdown point, I think that we have made it.

Having made my main point, I would like to use this opportunity to offer my thoughts on some of the controversial issues surrounding the breakdown point.

1. Is the breakdown point a conservative measure of robustness? Yes, it is by definition. But as long as we know what it is doing, it is not bad to be conservative. 2. Are there good reasons to aim for the highest possible breakdown point? Usually not. I am not voicing objections to research on the highest possible breakdown

point procedures; such research can offer insights. In choosing a good statistical procedure, we have to balance breakdown with other measures of quality. 3. Some say that high breakdown estimators are usually locally unstable. There is some truth to this, but again, one has to strike a balance between breakdown and local stability. This statement would be as true or untrue as "efficient estimators usually have low breakdown points." 4. High breakdown point estimators are usually too hard to compute. It is easy to propose a difficult-to-compute high breakdown estimator, but advances in methodological research and in computing power are already making more and more high breakdown procedures practical. Obviously I like the fact that SAS procedures based on high breakdown method are being added.

Finally, what role will the notion of breakdown point play in the future? I am not good at predicting the future, but I hope that it will be in every statistician's mind in evaluating the quality of a statistical procedure. It is in our best interest to keep it as simple and intuitive as possible so that it will be understood and appreciated by every statistician (plus more). In addition to research papers such as this one under discussion, I hope to see educational papers, too, that will be accessible by a broader audience. If I use the NSF jargon, I hope to see both scientific merit and broader impacts. I think that we will get there if we all try.

REFERENCES

HE, X. and SIMPSON, D. G. (1992). Robust direction estimation. *Ann. Statist.* **20** 351–369.

HE, X., SIMPSON, D. G. and PORTNOY, S. (1990). Breakdown robustness of tests. *J. Amer. Statist. Assoc.* **85** 446–452.

STROMBERG, A. J. and RUPPERT, D. (1992). Breakdown in nonlinear regression. *J. Amer. Statist. Assoc.* **87** 991–997.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT
URBANA–CHAMPAIGN
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: x-he@uiuc.edu

DISCUSSION

BY HANNU OJA

University of Jyväskylä

1. Breakdown, equivariance and invariance. The authors are to be congratulated for their excellent paper, which nicely clarifies the role of equivariance in

finding upper bounds for the breakdown points of functionals. The breakdown point approach, with upper bounds showing how far one can go, has achieved great success in the univariate and multivariate location, scale, scatter and regression estimation problems. The authors justifiably argue that this is due to the fact that the acceptable, well-behaved estimates in these contexts have natural equivariance properties. In constructing reasonable estimates and test statistics, one therefore considers statistics satisfying certain conditions (invariance, equivariance, unbiasedness, consistency, etc.). If there are no restrictions, the upper bound is one as the breakdown point (using the common definition) of a "stupid" constant functional, for example, is one.

The paper is clearly written with several illustrative examples. The constructive proof of the main Theorem 3.1 illustrates how one can concretely break down an equivariant estimate:

1. Pick a transformation g corresponding to the set with the supremum probability mass in (3.3).
2. Apply the transformation g or g^{-1} repeatedly to contaminate a (random) half of the data outside the set with the supremum probability mass.

In the one-sample location problem with sample size $n = 2k$, for example, the translation equivariance of a location estimate $T(x_1, \ldots, x_n)$ means that

$$T(x_1 + c, \ldots, x_k + c, x_{k+1}, \ldots, x_n) - T(x_1, \ldots, x_k, x_{k+1} - c, \ldots, x_n - c) = c$$

and, consequently, the estimate can be broken either by repeatedly shifting the first half of the data by $+c$ or by repeatedly shifting the second half of the data by $-c$.

The theory thus yields upper bounds for the breakdown points of the affine equivariant univariate location and scale functionals but does not say anything about *affine invariant* skewness and kurtosis statistics, for example. Clearly the invariant classical skewness statistic

$$b_1 = \frac{((1/n)\sum(x_i - \bar{x})^3)^2}{((1/n)\sum(x_i - \bar{x})^2)^3}$$

does not break down with one outlying observation, although affine equivariant second and third central moments both do. For a single outlier going to infinity, the central moments move "beyond all bounds" but b_1 converges to a constant $(n-2)^2/(n-1)$. As this limit is not data dependent at all, the contaminated statistic b_1 does not convey any information on the original data points. Is this a breakdown? Another strange example is the estimation problem for the parameters of the linear predictor $\theta_0 + \theta' x$ in the generalized linear model. Again, for $n = 2k$, an equivariant estimate of θ satisfies

$$\hat{\theta}\left(\begin{pmatrix} c \cdot x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} c \cdot x_k \\ y_k \end{pmatrix}, \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}\right)$$
$$= \frac{1}{c}\hat{\theta}\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \begin{pmatrix} (1/c) \cdot x_{k+1} \\ y_{k+1} \end{pmatrix}, \ldots, \begin{pmatrix} (1/c) \cdot x_n \\ y_n \end{pmatrix}\right)$$

and the estimate can be moved beyond all bounds *or* to zero by repeatedly multiplying half of the data by c or by $1/c$. The estimate then seems to become uninformative. Is something wrong with the definitions of the breakdown and the breakdown point? What do we really mean when we say that a breakdown occurs?

2. When does the breakdown occur? Since the early notions by Hampel (1971), the concept of breakdown point has been widely discussed and further developed by several contributors. For considering and comparing different approaches we adopt the following notation. Let $X = (x_1, \ldots, x_n)$ be an original "true" sample of size n lying in the sample space $\mathcal{X}$. The statistic ("estimate") considered is denoted by $T(X)$ with possible values in $\mathcal{T} = \{T(X) : X \in \mathcal{X}\} \subset \mathbb{R}^p$. We say that a point t is interior to $\mathcal{T}$ if it belongs to $\mathcal{T}$ and there is a neighborhood of t which contains only points of $\mathcal{T}$. A point $t \in \mathbb{R}^p$ is exterior to $\mathcal{T}$ if it does not belong to $\mathcal{T}$, and if there exists a neighborhood of t which contains no points of $\mathcal{T}$. Finally, t is called a boundary point of $\mathcal{T}$ if t is neither interior nor exterior to $\mathcal{T}$. Often $\mathcal{T} = \mathbb{R}^p$ and then there are no boundary points.

We next construct a contaminated sample. Let $S = (s_1, \ldots, s_n)$ be a vector of zeros and ones indicating the contamination, and $Y = (y_1, \ldots, y_n)$, also in $\mathcal{X}$, a sample of "outliers." The contaminated sample then consists of the observations $(1 - s_i)x_i + s_i y_i$, $i = 1, \ldots, n$. The number of outlying or alien observations is accordingly $s = \sum s_i$. The contaminated value of the estimate is then

$$T(X, Y, S) = T\big((1 - s_1)x_1 + s_1 y_1, \ldots, (1 - s_n)x_n + s_n y_n\big).$$

The breakdown of the estimate is most often defined as follows.

DEFINITION 1. T breaks down at X with s outliers if there exist a sequence (Y_m) in $\mathcal{X}$ and S with $\sum s_i = s$ such that $\|T(X, Y_m, S)\| \to \infty$.

The breakdown point then gives the smallest fraction of outliers (s/n) that suffices to "*drive the estimate beyond all bounds*." According to this definition, a constant estimate $(T(X) = t_0)$ can never be broken down. Note also that the scale estimate and scatter matrix estimate are usually thought to break down also if they converge to a boundary point of $\mathcal{T}$ (scale estimate converges to zero and the smallest eigenvalue of the scatter matrix estimate converges to zero). In the paper this is taken care of with a suitably chosen pseudometric; see, for example, Section 4.2. Another possibility is to give a new definition:

DEFINITION 2. T breaks down at X with s outliers if there exist a sequence (Y_m) in $\mathcal{X}$ and S with $\sum s_i = s$ such that either (i) $\|T(X, Y_m, S)\| \to \infty$ or (ii) $T(X, Y_m, S) \to t_0$ where t_0 is a boundary point of $\mathcal{T}$.

If the constant functional $T(X) = t_0$ does not depend on X, $\mathcal{T} = \{t_0\}$ and t_0 is also a boundary point. Therefore T breaks down for all S. Note that the boundary

point t_0 in the definition may depend on X, however. In the simple regression example suggested by the referee and analyzed in Section 6, the statistic $T(P_n)$ has values in $\mathcal{T} = [-n, n]$ and it breaks down (in the sense of Definition 2) if $\sum s_i = 1$.

Genton and Lucas (2003) take a different viewpoint and argue that a crucial property of an estimator $T(X, Y, S)$ is that it takes different values for different values of $X \in \mathcal{X}$ and that the breakdown occurs if this property is lost. In this spirit one can say that:

DEFINITION 3. T breaks down with s outliers if there exist a sequence (Y_m) in $\mathcal{X}$ and S with $\sum s_i = s$ such that either (i) $\|T(X, Y_m, S)\| \to \infty$, for all $X \in \mathcal{X}$, or (ii) $T(X, Y_m, S) \to t_0 \in \mathbb{R}^p$, for all $X \in \mathcal{X}$.

In this definition, it is remarkable that the interior or boundary point t_0 is not allowed to depend on X at all. This definition solves the problem with the classical skewness statistic; b_1 can be made to break down with a single extreme outlier. I wonder whether the techniques and results in the paper by Davies and Gather could be expanded to cover this definition also. Genton and Lucas (2003) seem in fact to be still more permissive and say that $T(X, Y, S)$ breaks down if

$$\mathcal{T} \cap \left\{ \lim_m T(X, Y_m, S) : X \in \mathcal{X} \right\}$$

collapses to a finite set; an empty set and a singleton $\{t_0\}$ are then special cases. Given a continuum of values of X, one expects a continuum of possible values of the estimate. In the linear predictor estimation problem this definition implies that the breakdown point of an equivariant estimate of θ is at most one half.

All the approaches described above work with the worst possible scenario represented by a strategically chosen sequence of the sets of outlying observations (Y_m). In practice, the observed contaminated value of the estimate $T(X, Y, S)$ is in $\mathcal{T}$, however, and not a boundary point, and one can ask whether the estimate still conveys useful information about the true data cloud or not. Then, instead of speculating about the sequences (Y_m), one may consider the set of possible values of $T(X, Y, S)$ for all choices of $Y \in \mathcal{X}$. With (at most) s outliers, the set of possible values of $T(X, Y, S)$ is

$$\mathcal{T}_s(X) := \left\{ T(X, Y, S) : Y \in \mathcal{X}, \sum s_i = s \right\}.$$

Then clearly

$$\{T(X)\} = \mathcal{T}_0(X) \subset \mathcal{T}_1(X) \subset \mathcal{T}_2(X) \subset \cdots \subset \mathcal{T}$$

and the value of the estimate is totally determined by s outliers if $\mathcal{T}_s(X) = \mathcal{T}$. More generally, we can define that:

DEFINITION 4. T breaks down with s outliers if the set $\mathcal{T}_s(X)$ does not depend on X.

Note that if T is affine equivariant/invariant, then also $\mathcal{T}_s(X)$ is affine equivariant/invariant. Assume next that the observed value of $T(X, Y, S)$ is t. If we knew the maximum number of outliers in the data set, say s, but S and Y are unknown, the observed event $\mathcal{T}_s(X) \ni t$ may still be informative. In the univariate location case with $n = 2k - 1$ and $\mathcal{X} = \mathbb{R}^n$, $\mathcal{T}_s(X) = \mathbb{R}$ for the sample mean if $s > 0$. But for $\mathcal{X} = [0, \infty)^n$, for example, the breakdown point of the mean is one as $\mathcal{T}_{s-1}(X) \ni t \Longleftrightarrow x_{(1)} \le n \cdot t$. For the sample median, the event

$$\mathcal{T}_s(X) \ni t \quad \Longleftrightarrow \quad x_{(k-s)} \le t \le x_{(k+s)}, \qquad s = 0, \ldots, k,$$

is clearly data dependent and therefore carries information about the data cloud.

REFERENCES

GENTON, M. G. and LUCAS, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **85** 81–94.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF JYVÄSKYLÄ
P.O. BOX 35
FIN 40351 JYVÄSKYLÄ
FINLAND
E-MAIL: ojahannu@maths.jyu.fi
URL: www.maths.jyu.fi/~ojahannu

DISCUSSION

BY PETER J. ROUSSEEUW

Universiteit Antwerpen

1. General comments. The interesting paper of Davies and Gather (henceforth [DG]) pulls together results on upper bounds on the breakdown value of translation equivariant location estimators [Donoho (1982)], regression estimators [Rousseeuw (1984)] and affine equivariant scatter estimators [Davies (1987)] into a single framework of group equivariance. I can only agree with them on the important role of the latter notion in obtaining nontrivial bounds. [By the way, I prefer the term breakdown *value* myself because it is not a point, and the term "value"

captures both its dimension (one) and its orientation (we aim for higher, not lower values).]

The theory in [DG] is formulated for estimators that are uniquely defined, but it seems to work just as well in the general case. Then $T(P)$ is a set, and we can follow the implicit convention of saying that it breaks down when any member of $T(P)$ does. We only need to redefine $D(T(P), T(Q))$ in (2.4) as a supremum over all pairs of members of $T(P)$ and $T(Q)$.

The new applications of the theory are fascinating, for example, to the Michaelis–Menten model (with a nontrivial bound) and logistic regression (without one). I am less convinced by the fragility argument illustrated by the difference between the contaminated samples (6.2) and (6.3). It is true that this proof of the upper bound only covers (6.2), but in some sense that is enough since breakdown is a worst-case concept and the bound is not specific for the median but for all translation estimators. But anyway, the behavior of the median at (6.3) can be derived from that at (6.2) by a variety of other properties that it possesses. For instance, its monotonicity property alone suffices. Or we can use the property that when you move the observations over distances of at most δ, the median changes by at most δ. This holds for any $\delta > 0$, and is a Lipschitz property for the metric on samples defined as

$$d\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) = \min_{\pi \in S_n} \max_i |x_i - y_{\pi(i)}|, \tag{1}$$

where S_n is the set of all permutations on $\{1, \ldots, n\}$. Note that (1) is equal to $\max_i |x_{i:n} - y_{i:n}|$ [see Rousseeuw and Leroy (1987), pages 127–128]. People who compute maxbias curves always use the properties of the actual estimator. Perhaps we should not expect much more elegant results for contaminated samples that break down the estimator than for those that yield a finite bias.

2. The maximal breakdown value of affine equivariant location estimators. From here on I will focus on the open problem in Section 5.2 of [DG]. It has been known since Donoho (1982) that the finite-sample breakdown value (fsbv) of translation equivariant estimators of location is at most $\lfloor (n+1)/2 \rfloor / n$ and that this bound is sharp. The bound obviously holds also for affine equivariant location estimators, but it may not be sharp for them. In one dimension ($k = 1$) it is, but for $k \geq 2$ this has been an open problem for over 20 years. During that time many affine location estimators were constructed with an fsbv of $\lfloor (n-k+1)/2 \rfloor / n$, such as the MVE and MCD of Rousseeuw (1984), location S-estimators [Davies (1987), Rousseeuw and Leroy (1987)] and a modification of the Stahel–Donoho estimator [Tyler (1994), Gather and Hilker (1997)]. Since $\lfloor (n-k+1)/2 \rfloor / n$ is known to be the sharp upper bound for affine equivariant scatter estimators [Davies (1987)], it has seemed plausible that it could also be the upper bound for affine location. Over the years there have been several attempts to attain the upper bound $\lfloor (n+1)/2 \rfloor / n$, but as far as I know none has succeeded. [The result in Zuo

(2004) does not count because it uses a weaker version of the fsbv which requires that all the contaminating points coincide.]

Let us consider any data set $X = \{x_1, \ldots, x_n\} \subset \mathbf{R}^k$ (from here on always $k \geq 2$ and $n > k$) which is in general position (GP). By GP we mean that no more than k data points lie on any affine hyperplane. This holds a.s. when sampling from an absolutely continuous distribution. The convex hull $\mathrm{conv}(X)$ is then a polytope with faces that contain exactly k data points. [In $\mathbf{R}^3$ the faces are two-dimensional, and in general they are $(k-1)$-dimensional.] Note that $\mathrm{conv}(X)$ can be stretched arbitrarily by replacing even a single point of X by an outlier. Since we are studying very robust estimators T, it is natural to require that T should not lie on the boundary of $\mathrm{conv}(X)$ or outside of $\mathrm{conv}(X)$. A slightly more general formulation of this requirement is the following condition:

(C_h) Let $X = \{x_1, \ldots, x_n\} \subset \mathbf{R}^k$ be in general position, with $n > k \geq 2$. Let u be a direction such that the inner products $y_i = u'x_i$ satisfy $y_1 = \cdots = y_h < y_{h+1} \leq \cdots \leq y_n$ (after renumbering) for the specified number h, with $1 \leq h \leq k$. Then there exists an $\alpha > 0$ (which depends only on k and the $y_1, \ldots, y_n$) such that $u'T(X) \geq y_h + \alpha$.

The typical case is to take $h = k$. For any face of $\mathrm{conv}(X)$ we can take the orthogonal direction u pointing to the inside of $\mathrm{conv}(X)$, so Condition (C_k) says that T cannot lie on or arbitrarily close to the boundary of $\mathrm{conv}(X)$ or outside of it. [Note that $\mathrm{conv}(X)$ is the intersection of halfspaces containing X and having a face of $\mathrm{conv}(X)$ on their boundary.] For $h < k$ the condition becomes somewhat weaker; for example, Condition (C_1) only says that T cannot come arbitrarily close to a vertex of $\mathrm{conv}(X)$ or lie outside of $\mathrm{conv}(X)$.

Condition (C_k) is intuitive for a robust estimator. For instance, Condition (C_k) holds for all estimators that can be written as a weighted mean $(\sum_i w_i x_i)/(\sum_i w_i)$ where $0 \leq w_i \leq 1$ and at least $k+1$ of the w_i equal 1 [it suffices to put $\alpha = (y_{k+1} - y_k)/n$]. This encompasses, for example, trimmed means and the minimum covariance determinant estimator (MCD). Moreover, a robust estimator would typically be expected to have a reasonably large Tukey depth, for example,

$$\mathrm{depth}(T, X) \geq k + 1 \tag{2}$$

(at least for large enough n, when there are many depth contours). Condition (2) implies Condition (C_k) and is another way of saying that T should not be in the outskirts of the data cloud.

THEOREM 1. *Consider a data set $X = \{x_1, \ldots, x_n\} \subset \mathbf{R}^k$ in general position with $n > k$. Let T be an affine equivariant location estimator satisfying Condition* (C_h) *with $1 \leq h \leq k$. Then* $\mathrm{fsbv}(T, X) \leq \lfloor (n - h + 1)/2 \rfloor / n$.

PROOF. Put $\hat{\theta} := T(X) \in \mathbf{R}^k$. Since X is in GP, $\mathrm{conv}(X \cup \{\hat{\theta}\})$ has at least one face not containing $\hat{\theta}$. Take an h-subset S of the k data points on this

face. Then there exists an affine hyperplane L which contains S such that both $\hat{\theta}$ and $X \setminus S$ lie strictly on the same side of L. Assume w.l.o.g. that $0 \in L$. Denote the unit normal vector to L in the direction of $X \setminus S$ as e_1 and take an orthonormal basis $\{e_2, \ldots, e_k\}$ of L. After renumbering, the $x_{i1} := e_1' x_i$ satisfy $0 = x_{1,1} = \cdots = x_{h,1} < x_{h+1,1} \leq \cdots \leq x_{n,1}$, hence $\hat{\theta}_1 \geq \alpha > 0$ by Condition (C_h).

Let us assume that T cannot be broken down by replacing any m-subset B of X, where $m = \lfloor (n-h+1)/2 \rfloor$, by an arbitrary m-set B' yielding the contaminated data set $X' := (X \setminus B) \cup B'$. This means that there exists a finite radius M such that for any contaminated data set X' of this type it holds that $T(X') \in B(\hat{\theta}, M) := \{x \in \mathbf{R}^k; \|x - \hat{\theta}\| \leq M\}$.

We will now construct a linear transformation which leaves S invariant and moves $X \setminus S$ as well as $\hat{\theta}$. For this we consider the "shear transform" g_γ given by the nonsingular matrix

$$\left[\begin{array}{cc|c} 1 & 0 & 0 \\ \gamma & 1 & 0 \\ \hline 0 & 0 & I_{k-2} \end{array}\right] \tag{3}$$

relative to the basis $\{e_1, \ldots, e_k\}$, for $\gamma \in \mathbf{R}$. We note that $g_\gamma(e_j) = e_j$ for all $j \neq 1$, hence $g_\gamma(x_i) = x_i$ for $i = 1, \ldots, h$, but at the same time $g_\gamma(e_1) = e_1 + \gamma e_2$. Denoting $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$, we find $g_\gamma(\hat{\theta}) = (\hat{\theta}_1, \hat{\theta}_2 + \gamma\hat{\theta}_1, \hat{\theta}_3, \ldots, \hat{\theta}_k)^T$ with $\hat{\theta}_1 > 0$, hence $\|g_\gamma(\hat{\theta}) - \hat{\theta}\| = |\gamma|\hat{\theta}_1$ goes to infinity for increasing γ. Analogously, the image of any data point x_i with $i = h+1, \ldots, n$ is of the form $g_\gamma(x_i) = x_i + \gamma x_{i1} e_2$, so all $g_\gamma(x_i)$ move in the direction of e_2 and $(g_\gamma(x_i))_1 = x_{i1}$. Each point travels a distance $\|g_\gamma(x_i) - x_i\| = |\gamma||x_{i1}| \geq |\gamma||x_{h+1,1}|$.

Let us partition $X \setminus S$ into two sets A and B with $|B| = m = \lfloor (n-h+1)/2 \rfloor$ and $|A| = n - h - |B|$. (If $n - h$ is even, we find $|A| = |B|$, whereas for odd $n - h$ we have $|A| = |B| - 1$.) We will replace B by $B_\gamma := g_\gamma(B)$ yielding the contaminated data set $X'_\gamma := S \cup A \cup B_\gamma$. Note that X'_γ is in GP for all but a finite number of γ values. Put $\Gamma = \{\gamma; X'_\gamma \text{ is in GP}\}$. For all $\gamma \in \Gamma$ it holds that $T(X'_\gamma) \in H := \{z \in \mathbf{R}^k; z_1 \geq \alpha\}$ by Condition (C_h).

For any γ the image of $B(\hat{\theta}, M)$ through g_γ is an ellipsoid with center $g_\gamma(\theta)$. For a large enough $\gamma \in \Gamma$ it holds that $B(\hat{\theta}, M) \cap g_\gamma(B(\hat{\theta}, M)) \cap H = \varnothing$. We know that $T(X'_\gamma) \in B(\hat{\theta}, M)$ by assumption. On the other hand, we can also write $X'_\gamma = g_\gamma(S \cup A_{-\gamma} \cup B)$, which implies $T(X'_\gamma) \in g_\gamma(B(\hat{\theta}, M))$. Since $T(X'_\gamma) \in H$ it follows that $T(X'_\gamma) \in B(\hat{\theta}, M) \cap g_\gamma(B(\hat{\theta}, M)) \cap H = \varnothing$. This contradiction proves the desired upper bound on fsbv. □

In the typical case where $h = k$, Theorem 1 yields the upper bound $\lfloor (n-k+1)/2 \rfloor / n$ which has been attained. This says that any affine location estimator T with a higher fsbv must be somewhat strange in the sense of not satisfying Condition (C_k), so T can be arbitrarily close to the boundary of $\text{conv}(X)$ or

even lie outside it. Any T which were to attain the translation equivariant bound $\lfloor (n+1)/2 \rfloor / n$ cannot even satisfy Condition (C_1), so at times it must be arbitrarily close to a vertex of $\text{conv}(X)$ or lie outside it. It is counterintuitive that an estimator with maximal fsbv would have such a low Tukey depth (at most 1).

So far the only published result with higher fsbv than $\lfloor (n-k+1)/2 \rfloor / n$ is the projection median (PM) of Zuo (2003), which attains $\lfloor (n-k+2)/2 \rfloor / n$ by using a univariate scale estimator MAD_{k-1} in its definition. By Theorem 1, this estimator cannot satisfy Condition (C_k). Here is a bivariate counterexample (which can be extended to $\mathbf{R}^k$). Start with the data points $z_1 = (0, \delta)$ and $z_2 = (0, -\delta)$ for some $\delta > 0$. Add m points (x_i, y_i) with x_i equispaced between 10 and 20 and $y_i = x_i + \delta u_i$ where the noise u_i is such that these points are in GP. Add another m points with the same x_i but with $-y_i$. Then the $n = 2m + 2$ points of Z are in GP for all but finitely many δ. When $\delta \to 0$, the outlyingness $\text{Out}(0, 0)$ tends to the outlyingness of 0 relative to $\{0, 0, x_1, x_1, \ldots, x_m, x_m\}$; hence for any $0 < \delta < 1$ we have $\text{Out}(0, 0) < M$ for some $M < \infty$. We will prove that for any $\varepsilon > 0$ there is a $\delta_0 > 0$ such that $\delta < \delta_0$ implies $\|\text{PM}(Z)\| < \varepsilon$. By projecting in the direction orthogonal to $y = -x$ we see that MAD_1 tends to 0, so for small enough δ all points (not necessarily data points) in $\mathbf{R}^2$ lying farther than $\varepsilon/\sqrt{2}$ away from the line $y = -x$ have $\text{Out} > M$. The same holds for points farther than $\varepsilon/\sqrt{2}$ from $y = x$. Therefore $\text{PM} \to (0, 0)$; hence α in Condition (C_k) is zero.

Note that Theorem 1 fits in the framework of [DG] with G the affine group on $\mathbf{R}^k$. The main difference is that here we first fix a set B (our h-subset) and then a subgroup of G which keeps B invariant, whereas condition (3.3) in [DG] is over many possible B. Afterward we put $g := g_1$ [i.e., (3) with $\gamma = 1$], yielding $\Delta(P_n) = h/n$. The remainder of the proof of Theorem 3.1 in [DG] can then be retraced by noting that for any integer m it holds that $g^m = g_m$ (the shear transform with $\gamma = m$). We basically set aside h points and then apply our usual reasoning to the remaining $n - h$ points.

Also note that Condition (C_h) and Theorem 1 can be extended to situations without general position. As long as T satisfies Condition (C_h) without the GP condition (this is a stronger assumption), and X does have h points whose inner products with some u satisfy $y_1 = \cdots = y_h < y_{h+1} \le \cdots \le y_n$, the upper bound $\text{fsbv}(T, X) \le \lfloor (n-h+1)/2 \rfloor / n$ holds. In this situation it is even allowed that $h > k$ (which could not happen under GP).

Acknowledgment. I would like to thank Yijun Zuo for stimulating discussions.

REFERENCES

DAVIES, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Dept. Statistics, Harvard Univ.

GATHER, U. and HILKER, T. (1997). A note on Tyler's modification of the MAD for the Stahel–Donoho estimator. *Ann. Statist.* **25** 2024–2026.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

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

TYLER, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.* **22** 1024–1044.

ZUO, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.* **31** 1460–1490.

ZUO, Y. (2004). Projection-based affine equivariant multivariate location estimators with the best possible finite sample breakdown point. *Statist. Sinica* **14** 1199–1208.

DEPARTMENT OF MATHEMATICS
AND COMPUTER SCIENCE
UNIVERSITEIT ANTWERPEN
MIDDELHEIMLAAN 1
B-2020 ANTWERP
BELGIUM
E-MAIL: Peter.Rousseeuw@ua.ac.be

DISCUSSION

BY DAVID E. TYLER

Rutgers, The State University of New Jersey

The breakdown point has played an important role within robust statistics over the past 25–30 years. A large part of its appeal is that it is easy to explain and easy to understand. It is often interpreted as "the proportion of bad data a statistic can tolerate before becoming arbitrary or meaningless." In this paper Professors Davies and Gather give us a much needed critical look at this seemingly simple concept, and are to be commended for doing so.

Except for some pathological examples, such as a constant functional, one typically presumes that the breakdown point of a statistic or functional cannot be greater than $1/2$. A heuristic justification for this presumption follows from the simple argument "if over half the data is bad, then one cannot distinquish between the good data and the bad data." The authors nicely show that when one more carefully examines this "common sense" argument, then it only appears to be meaningful within a setting with an appropriate group structure. They further challenge the reader to give meaning to this expression outside of such a setting.

One may be able to modify the definition of the breakdown point in some creative way in order to meet the authors' challenge. For example, one could define the breakdown of a statistic to mean it can be made to go to the boundary

of all possible values of the statistics. Such a modification would then imply that a constant statistic has breakdown point 0, which is in intuitive agreement with the notion that a constant statistic conveys no information about the data. Such a modification also nullifies the counterexample given by the authors in Section 6.

The intent of this discussion, though, is not to attempt to defend the notion of breakdown outside of the group setting, which may only be possible on a case-by-case basis. Rather, being in general agreement with the arguments made by the authors, the focus of this discussion is to further examine the breakdown point concept within the group setting via two fundamental examples.

1. Robust principal component vectors. One approach to robust principal components is to perform a principal component decomposition on a robust covariance matrix rather than on the sample covariance matrix. The asymptotic distribution and influence function of the principal component roots and vectors follow readily from those of the robust covariance matrix; see, for example, Croux and Haesbroeck (2000). However, the breakdown point of the robust covariance matrix has no information regarding the principal component vectors, since the breakdown of a covariance matrix only implies that either the largest root can become arbitrarily large or the smallest root can become arbitrarily small.

To illustrate this point, let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ represent a sample in $\Re^d$ and let $S_n = \mathcal{Q}\Delta\mathcal{Q}'$ represent the spectral value decomposition of the sample covariance matrix S_n. Define $V_n = \mathcal{Q}\Delta^*\mathcal{Q}'$, where $\Delta^* = diagonal\{\lambda_1^2, \ldots, \lambda_d^2\}$ with λ_j being a high breakdown point scale statistic for the univariate sample $\{\mathbf{q}_j'\mathbf{x}_1, \ldots, \mathbf{q}_j'\mathbf{x}_n\}$ and where $\mathcal{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_d]$. That is, we simply replace the eigenvalues of the sample covariance matrix, which correspond to the variances of the sample principal component variables, with robust variances for the sample principal component variables. The resulting statistic V_n has a high breakdown point, namely the breakdown point of the univariate scale statistic used in its definition, whereas the breakdown point of S_n is zero. Both statistics, though, yield the same principal component vectors. So, using a high breakdown point estimate of the covariance matrix for principal components analysis is in itself meaningless, unless one can show some relationship between it and the breakdown of the principal component vectors.

The principal component vector associated with the largest root of the sample covariance matrix can be made arbitrarily close to any given vector by perturbing just one data point. One implicitly assumes this does not occur if a robust covariance matrix is used in place of the sample covariance matrix. Except for contrived examples like the one constructed in the previous paragraph, the proportion of contamination needed to make the largest principal component vector "arbitrary" is likely to be dependent on the separation between the largest root and the other roots of the robust covariance matrix of the uncontaminated data or distribution, as is the case with the influence function. Thus, the best possible bound on the breakdown point is likely dependent on the structure of the

uncontaminated data or distribution. As far as this discussant is aware, there are no known results which allow one to quantify this somewhat obvious conjecture, and so it is of interest to see how the results of this paper might apply.

Some meaningful notion of breakdown for a principal component vector is first needed. For the sake of this discussion, consider any robust version of the largest principal component vector, whether or not it is defined via a robust covariance matrix. One usually regards this as an orthogonally equivariant mapping from the data or distribution into the parameter space $\Theta = S^{d-1} = \{\theta \in \Re^d | \theta'\theta = 1\}$, with the parameters θ and $-\theta$ being viewed as equivalent. Alternatively, Θ can be taken to be the set of all one-dimensional subspaces. A natural metric on Θ is the absolute value of the angle between any two elements, that is, $D(\theta_1, \theta_2) = \arccos(|\theta_1'\theta_2|)$, as is used, for example, in Van Aelst and Willems (2004). This metric, however, does not satisfy condition (2.3) of the paper since the largest possible angle is π. Since Θ is compact and with no interior points, it is not possible to define a pseudometric on Θ which does satisfy (2.3). An intuitive definition of breakdown, though, can be obtained by simply replacing ∞ with π in definition (2.4). Breakdown is then naturally interpreted as the proportion of contamination needed for the largest principal component vector to become orthogonal to that obtained from the uncontaminated data.

Since condition (2.3) cannot be made to hold, the results of the paper do not apply here. If one attempts to extend the results of the paper by also replacing ∞ with π in the definition of G1 given by (3.3), then the set G1 is null. Even if G1 were not null, the crucial step in the proof of Theorem 3.1 is highly dependent on having an unbounded metric. So, it remains an open question as to what types of limits for breakdown are possible for this problem. Using the group equivariance property alone is probably not sufficient for answering this question since the principal component vector associated with any particular root has the same group equivariant property. Some further constraints on what is meant to be a largest principal component vector may be needed to obtain a meaningful bound on the breakdown point. This unsettled question is not specific to principal component vectors, but also applies to any parameter space for which no unbounded metric exists. Such parameter spaces arise naturally, for example, in the areas of directional data analysis and shape theory.

Perhaps some anomaly always arises not only outside of the group setting, but outside of the group setting with unbounded metrics. For principal components, a technicality arises in that it is possible for some data sets or distributions that the largest principal component vector can be any vector within some subspace of dimension $q > 1$, that is, as some q-dimensional subspace. For example, any reasonable definition of the largest principal component vector should be any vector at the standard multivariate normal distribution. The complete parameter space then corresponds to the set of all subspaces of $\Re^d$ rather than simply the set of all one-dimensional subspaces. The largest principal component "vector"

can still be restricted to equivariant functionals under the group of orthogonal transformations, and the metric D can be extended to this larger parameter space. For example, for bivariate distributions only $\Re^2$ is added to the parameter space and D can be extended by defining $D(\Re^2, \Re^2) = 0$ and $D(\Re^2, \theta) = \pi$. This implies that breakdown occurs when a "well-defined" vector becomes "undefined," that is, becomes $\Re^2$, or vice versa, which is in intuitive agreement with what one thinks of as breakdown. In this setting, though, there exists an orthogonally equivariant functional with breakdown point 1, namely the constant functional $T(P) = \Re^2$, although it is not consistent.

2. Redescending M-estimates of location. In Section 6 of their paper, the authors note that the meaning of the breakdown point may even be suspect in the well-studied simple univariate location problem. This motivates them to state that "even in the case of equivariance the success of the concept of breakdown point would seem to be more fragile than it is generally supposed." The intent of the discussion here is to elaborate on their remarks by examining in more detail the behavior of the M-estimates of location.

For a univariate sample $X^n = \{x_1, \ldots, x_n\}$, an M-estimate of location $T(X^n)$ can be defined as a solution to the M-estimating equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{c}\right) = 0, \tag{1}$$

for some function ψ and tuning constant $c > 0$. A well-known result is that for monotonic, bounded and odd ψ-functions, the breakdown point of T is $1/2$. For redescending ψ-functions, the breakdown point of the corresponding M-estimate is more complicated. Since redescending ψ-functions tend to result in multiple solutions to the M-estimating equations, it is more convenient to use the alternative definition of an M-estimate of location given by

$$T(X^n) = \arg\min_t \sum_{i=1}^{n} \rho\left(\frac{x_i - t}{c}\right), \tag{2}$$

for some ρ-function. If ρ is differentiable, then the solution to (2) also satisfies (1) with $\psi = \rho'$. If the function $\rho(r)$ is even, nondecreasing in $|r|$ and bounded, then when ρ is also differentiable the corresponding ψ function is odd and redescends to zero, that is, $\psi(r) \to 0$ as $|r| \to \infty$. Huber (1984) shows the finite-sample contamination breakdown point of such redescending M-estimates of location to be

$$\varepsilon^*(T; X^n) = \frac{1 - A(X^n; c)/n}{2 - A(X^n; c)/n}, \tag{3}$$

where

$$A(X^n; c) = \min_t \sum_{i=1}^{n} \rho\left(\frac{x_i - t}{c}\right), \tag{4}$$

and without loss of generality $\lim_{r\to\infty}\rho(r)=1$. As the tuning constant $c\to\infty$, the resulting M-estimate looks more like the sample mean (provided ρ is differentiable in a neighborhood of zero). Curiously, though, as the tuning constant $c:0\to\infty$ one can note that the breakdown point $\varepsilon^*(T;X^n):0\to 1/2$.

A convenient way to gain insight into this formula for the breakdown point (3) is by using the relationship between redescending M-estimates of location and kernel density estimators. The objective function for an M-estimate of univariate location with fixed scale and a kernel density estimate with a given window width, which can be expressed, respectively, as

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i-\mu}{c}\right)\quad\text{and}\quad\hat{f}(x)=\frac{1}{nc}\sum_{i=1}^{n}\kappa\left(\frac{x-x_i}{c}\right), \tag{5}$$

have a one-to-one relationship when $\kappa\propto 1-\rho$. This relationship has been noted, for example, by Chu, Glad, Godtliebsen and Marron (1998). The M-estimate of location for a given tuning constant c then corresponds to the mode of the kernel density estimate with window width c. The mode of a kernel density estimate based on the Gaussian kernel $\kappa(r)=e^{-r^2/2}/\sqrt{2\pi}$, for example, corresponds to an M-estimate of location based on $\rho(r)=1-e^{-r^2/2}$, which is referred to as Welsch's M-estimate in Splus and MATLAB. Likewise, the mode of a kernel density estimate based on the Epanechnikov kernel corresponds to the skipped-mean.

As an illustrative example, consider the graphs in Figure 1. The data set in the graphs is a simulated data set comprised of 80% standard normal data and 20% normal data with mean 5 and standard deviation 0.1. The three graphs show the kernel density estimates for this data using a Gaussian kernel and the corresponding objective function for the M-estimate of location, for increasing values of the tuning constant c. In the first graph, the principal mode of the density is centered about the 20% of tightly compacted points. The second graph corresponds to using a larger value of c, and the principal mode is located near the mean of the main 80% of the data. The last graph corresponds to using a relatively large value of c.

In the first graph, the principal mode would go off to infinity if the more compact 20% of the data were pushed off to infinity, and hence breakdown occurs. Although this M-estimate can be made arbitrary under 20% contamination, it is arguable whether the solution is meaningless. On the other hand, in the third graph, although the principal mode is essentially the sample mean, it will not break down even if the 20% were replaced by 45% and allowed to go to infinity since eventually it will fall outside the window width and not impact the central mode. This phenomenon also occurs if the M-estimate is made scale equivariant by introducing a robust scale such as the M.A.D.; see Chen and Tyler (2004).

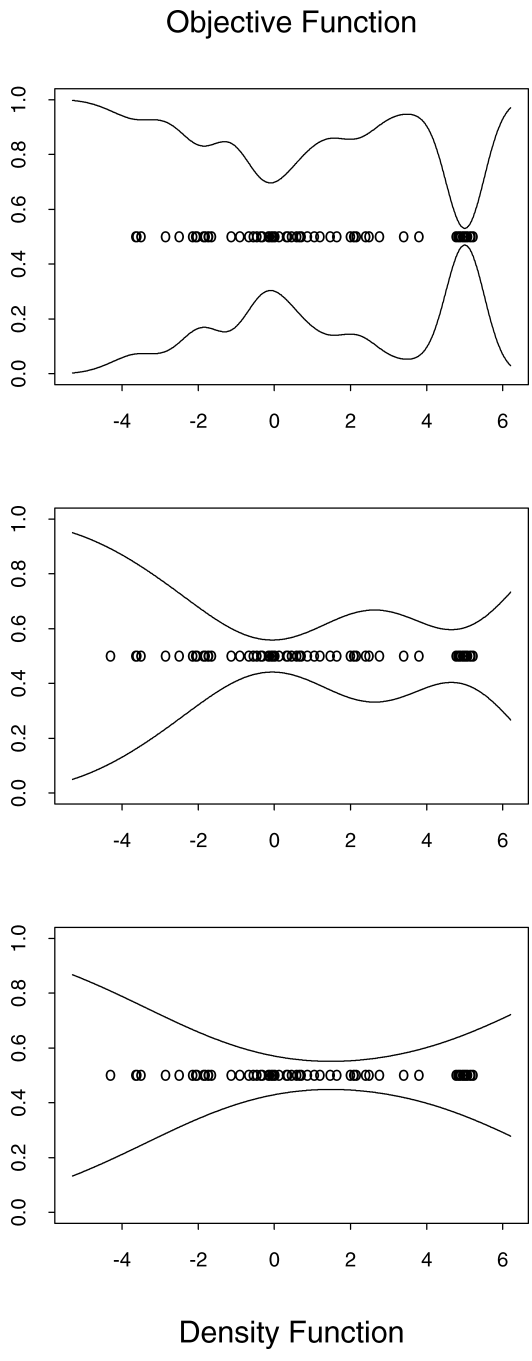

FIG. 1. *An illustration of the relationship between redescending M-estimates and kernel density estimates.*

Within the class of redescending M-estimates of location, it is arguable whether the breakdown point is more a descriptive property rather than an optimality property. A higher breakdown point redescending M-estimate is not necessarily a more desirable estimate. Note also that the nature of breakdown for a redescender differs from that of a monotonic M-estimate. The redescender can only break down if a relatively compact cluster of points goes to infinity. If the spread of the "bad" data is greater than that of the "good" data, then a redescending M-estimate cannot be broken down even if the "bad" data is in the majority, whereas such contamination would break down a monotonic M-estimate.

The above discussion helps illustrate how the simple heuristic interpretation of the breakdown point as "the proportion of bad data a statistical method can tolerate" can be misleading. It has led to some confusion in areas such as computer vision/image understanding. A relatively compact subset of the data may not be considered "bad data" for some applications but rather the data of interest. Instead, a bad data point or "outlier" may be considered a point unlike other data points, and these are the type of bad data points that one may wish to be protected against. At the 2002 ICORS conference in Vancouver, for example, the computer scientist Raymond Ng cleverly noted that a computer scientist's concept of a bad data point can be paraphrased by using the popular Sesame Street phrase "one of these things is not like the others, one of these things does not belong." If one adopts this notion, then the redescending M-estimates of location do not break down even under 99% contamination, whereas the monotonic M-estimates still have breakdown point $1/2$.

REFERENCES

CHEN, Z. and TYLER, D. E. (2004). On the finite sample breakdown points of redescending M-estimates of location. *Statist. Probab. Lett.* **69** 233–242.

CHU, C. K., GLAD, I. K., GODTLIEBSEN, F. and MARRON, J. S. (1998). Edge-preserving smoothers for image processing (with discussion). *J. Amer. Statist. Assoc.* **93** 526–556.

CROUX, C. and HAESBROECK, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika* **87** 603–618.

HUBER, P. J. (1984). Finite sample breakdown points of M- and P-estimators. *Ann. Statist.* **12** 119–126.

VAN AELST, S. and WILLEMS, G. (2004). PCA based on multivariate *MM*-estimators with fast and robust bootstrap. Preprint.

DEPARTMENT OF STATISTICS
RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
PISCATAWAY, NEW JERSEY 08855
USA
E-MAIL: dtyler@rci.rutgers.edu

REJOINDER

BY P. LAURIE DAVIES AND URSULA GATHER

University of Duisburg–Essen and Technical University Eindhoven, and University of Dortmund

We thank all the discussants for their contributions and in particular we wish to thank Hampel. The concept of breakdown point goes back to his Ph.D. thesis [Hampel (1968)] and he was the first to exhibit a high breakdown equivariant regression estimate now known as the least median of squares [Hampel (1975)], a fact which is sometimes forgotten. These two sources are the starting point of the present discussion. In his contribution Hampel gives us insight into the thoughts which led to his definition of breakdown point, intended as it was to complement the infinitesimal behavior of a functional as described by the influence function. Hampel emphasizes that equivariance considerations were not part of his definition and he had in mind correlation statistics "where there is no equivariance at all." He considers correlation in some detail and, as we disagree with him on this very topic, we give a detailed analysis of correlation statistics in our rejoinder. We hope that this will help clarify the issues involved.

1. On breakdown. The first signification of the word "breakdown" given in the Oxford Dictionary starts with the following subsignification:

"**1. a.** The act of breaking and falling down: a ruinous downfall, a collapse."

2. Breakdown to points and variations. Genton and Lucas and Oja argue for the usefulness of the breakdown concept in situations not covered by the results of our paper. They claim that at least in an intuitive sense breakdown occurs if the value of a functional is driven to the boundary or to an interior point which is independent of the uncontaminated sample. A formal definition of breakdown point is given which is intended to cover such possibilities. A first version is to be found in Genton and Lucas (2003) and is referred to by Oja. It defines the breakdown point as the smallest amount of contamination which can cause the statistic to assume only a finite number of values independently of the uncontaminated observations. On the basis of this definition the arithmetic mean is claimed to have a finite-sample breakdown point of $1/n$. The argument is as follows: if the first observation of the sample is contaminated, $(\xi_1, y_2, \ldots, y_n)$, and we let ξ_1 tend to infinity, then the sample mean tends to infinity, that is, to a single value which is independent of $y_2, \ldots, y_n$. However, for any finite value of ξ_1 the arithmetic mean takes on a continuum of values on varying the uncontaminated part of the sample. The only way of reducing the arithmetic mean to a single value is to introduce the symbol ∞ as a possible value for the contamination.

The symbol ∞ is thus elevated to a real entity for data. The new definition avoids this but our reaction is similar: any definition of breakdown based on the concept of Lebesgue measure zero must be at fault. According to the new definition the functional

$$T(P_n) = \max\{-n, \min\{n, T_{\mathrm{LS}}(P_n)\}\}$$

has a breakdown point of $1/n$. We perturb it and put

$$T^*(P_n) = T(P_n) + \frac{1}{n}\int \sin(x)\, dP_n(x).$$

The set of values taken on by $T^*(P_n)$ as we vary the uncontaminated part of the sample has Lebesgue measure at least $2/n^2$ as long as not all the sample is contaminated, and the breakdown point is therefore 1. As the perturbation tends with $1/n$ to zero, T^* remains consistent and asymptotically normal at the model. Oja mentions the classical skewness statistic

$$b_1 = \frac{((1/n)\sum(x_i - \bar{x})^3)^2}{((1/n)\sum(x_i - \bar{x})^2)^3},$$

but this can be treated in the same manner by putting

$$b_1^* = b_1 + \sin(nb_1),$$

which is still invariant but does not converge.

A second criticism we made of the definition of Genton and Lucas (2003) is that any realizable functional immediately breaks down for the simple reason that it can only take on a finite number of values; all data and statistics are of finite precision. Genton and Lucas mention this in their contribution as a weakness of the new definition and so it is. No reasonable definition of breakdown can rely on the myth of a continuum of possible values for a statistic or the associated myth of infinite precision. When applying mathematics to applied problems it is important that the discrete problem can be well approximated by the continuous one. Genton and Lucas' use of infinite precision and a continuum of values and sets of Lebesgue measure zero is not of this sort. Their continuous formulations do not approximate the discrete world of statistics.

We point out further that the definitions of Genton and Lucas, and also of Oja (Definition 4), represent a complete break with the meaning of breakdown as it is used in statistics. Transferred to the statistical context the "ruinous downfall" of Section 1 is expressed in terms of distances and arbitrarily large bias. None of this is present in a concept of breakdown in terms of the number or Lebesgue measure of the set of all possible limits of contaminated samples. No mention is made of bias, that is, how far the value of the statistic can move from its value at the uncontaminated sample for a given amount of contamination. Yet it is this which has motivated robust statistics from the influence function via bias to breakdown

point. In a sense the proposal put forward by Genton and Lucas is the very opposite of this. Rather than moving arbitrarily far, the statistic has broken down if it does not move at all. It is said that it then cannot convey any information in the sample. Even this is not always the case. Consider the statistical functional T_{75} which takes on the fixed value of 75 for all data sets. This has a breakdown point of 0 according to the Genton–Lucas definition. German insurance companies are required to use life expectancies specified by law. In the case of a male they could, for example, be forced to use the functional T_{75} to estimate life expectancy in years. The effect can be felt but it is not a ruinous downfall. It would be a ruinous downfall for the German insurance companies if they had to use a value of 65 and the reason is that 65 differs from the experienced lengths of life much more than does the fixed value of 75. Here as in the usual definition of breakdown it is the discrepancy which is important.

3. Perturbations. The criticism we gave of Genton and Lucas's definitions of breakdown has wider implications. We regard robust statistics as a perturbation theory for statistics. In particular, robust statistics must concern itself with perturbations of models and data sets and, in consequence, it must be able to deal with finite precision. The perturbations involved should be realistic ones and this will in general exclude perturbations described by the gross error neighborhood, which is simply too small. Unfortunately, the idea of stability under perturbations is sometimes lost, especially in theoretical work. Suppose a theorem on the existence and uniqueness of a functional requires assumptions about the existence and differentiability of a density function. These assumptions should then not be referred to as "under weak assumptions" but rather as "under very restrictive assumptions which violate the spirit of robustness." Densities disappear under perturbations, likelihood disappears under perturbations as does the property of being a Lebesgue set of measure zero, efficiency is pathologically discontinuous, and so on. Perturbations and their consequences should be taken seriously by all who work in the area of robust statistics.

4. Affine equivariant location functionals. The example of location functionals makes use of only the translation group although it seems natural to require affine equivariance. The problem is that for the affine group we have $G_1 = \varnothing$ since if we iterate $\mathcal{A}(\theta) = A(\theta) + b$ this will in general not tend to infinity so Theorem 3.1 is not applicable. The highest breakdown point for translation equivariant functionals is $1/2$ but there are affine equivariant location functionals which are based on scatter functionals and which have a breakdown point of at least that of the scatter functional, namely $(1 - \Delta(P))/2$. The gap has not been closed but Rousseeuw gives a sufficient condition for the bound $(1 - \Delta(P))/2$ to hold. His argument makes use of the convex hull which can be seen as a form of scatter functional albeit with a low breakdown point. In Davies and Gather (2002) we showed that the bound $1/2$ is attainable at least at some empirical measures so that the gap remains.

5. Metrics on $\mathscr{P}$. We agree with Hampel's comments on the gross error neighborhood but we do not like either of the alternatives he suggests. First, total variation is not much better than the gross error neighborhood; a distribution Q lies in the ε total variation neighborhood of P if and only if $Q - P = \varepsilon(H_1 - P) - \varepsilon(H_2 - P)$ for some distributions H_1 and H_2 [see Rieder (2000), page 7]. Second, the Prohorov metric defined by

$$d_{\mathrm{pr}}(P, Q) = \inf\{\varepsilon > 0 : P(A) \le Q(A^\varepsilon) + \varepsilon\}, \tag{1}$$

where

$$A^\varepsilon = \{x : d(x, A) < \varepsilon\}, \tag{2}$$

conflates the last ε of (1) where it operates as a dimensionless probability with the ε of (2) where it represents a rounding error. We refer to Davies (1993) for a discussion of this point. Other simpler metrics are also capable of dealing with rounding errors. The Kolmogorov metric is defined by

$$d_{\mathrm{ko}}(P, Q) = \sup\{|P(I) - Q(I)| : I = (-\infty, x], x \in \mathbb{R}\}. \tag{3}$$

Let P_n be the empirical distribution of some data and P_n^* be the empirical distribution of the same data after rounding. If the rounding δ is less than the minimum gap between the unrounded observations, then $d_{\mathrm{ko}}(P_n, P_n^*) = 1/n$ assuming at least one observation to have been altered. It is sometimes argued that d_{ko} is too weak in the data analytical sense for comparing distributions. There are stronger versions which go under the name of Kuiper metrics. The Kuiper metric of order k is defined by

$$d_{\mathrm{ku},k}(P, Q) = \sup\left\{\sum_{j=1}^{k} |P(I_j) - Q(I_j)| : I_1, \ldots, I_k \text{ disjoint finite intervals}\right\}. \tag{4}$$

Kuiper metrics of order $k = 19$ are used in Davies and Kovac (2004) in the context of providing approximate densities for data. The Kolmogorov and Kuiper metrics are restricted to $\mathbb{R}$, but in higher dimensions metrics on Vapnik–Cervonenkis classes of sets retain many of their properties [see Pollard (1984)]. We refer to Davies (1993) for their use in the regression setting. The conflation of measurement error and probability in (1) can be avoided as follows. We define

$$d_{\mathrm{pk}}(P, Q) = \inf\{\varepsilon > 0 : P(I) \le Q(I^\varepsilon) + \varepsilon, \text{ for all intervals } I\}, \tag{5}$$

where I^ε denotes the interval with the same center as I but with length $|I| \exp(\varepsilon)$. All occurrences of ε in (5) are now dimensionless. The idea is not new. We refer to Davies (1992, 1993).

Hampel's second argument for the Prohorov metric is that it metricizes weak convergence but we fail to see the relevance of this. The Kolmogorov metric (3)

does not metricize weak convergence but nevertheless does have advantages over the Prohorov metric for proving central limit theorems. In particular we have

$$d_{\mathrm{ko}}(P_n, P) = O_P\big(1/\sqrt{n}\,\big) \tag{6}$$

uniformly in P. If T is a functional with a bounded influence function $I(x, T, P)$, then under appropriate regularity conditions

$$T(P_n) - T(P) = \int I(x, T, P)\, d\big(P_n(x) - P(x)\big) + o_P\big(d_{\mathrm{ko}}(P_n, P)\big), \tag{7}$$

which in the light of (6) gives us a central limit theorem for $\sqrt{n}(T(P_n) - T(P))$. The same reasoning fails for the Prohorov metric because (6) does not hold [see Kersting (1978)].

6. Metrics on Θ. We turn to the metric D on Θ which quantifies the "ruinous downfall." For location in $\mathbb{R}$ the choice $D(\theta_1, \theta_2) = |\theta_1 - \theta_2|$ seems natural but the choice $|\log(\theta_1/\theta_2)|$ for scale is not quite as obvious. It does, however, have a strong justification in that numbers often have to be standardized by division by scale. If so, a scale of zero is a "ruinous downfall." In higher dimensions breakdown in scale includes the data being concentrated on a lower-dimensional hyperplane, making it impossible to identify the influence of individual covariables. Again the word breakdown would seem appropriate. In an earlier version of our paper we considered the possibility of measuring differences in the parameter θ by differences in the corresponding distributions P_θ as in $D(\theta_1, \theta_2) = d(P_{\theta_1}, P_{\theta_2})$ for an appropriate metric d on the space of distributions, but this needs to be given more thought.

Tyler has pointed out that if the parameter space is compact, then the metric is bounded so that condition (3.1) of our paper cannot possibly be satisfied. This is true, but just as metrics on $\mathcal{P}$ are chosen for the problem, so we can choose metrics on Θ according to the problem. If breakdown is defined in terms of convergence to some parameter values such as those on the boundary, then we can choose an appropriate metric as follows. We start by considering the problem of scale in $\mathbb{R}$. The proof works by showing that if $\varepsilon > (1 - \Delta(P))/2$, then there exists an affine transformation $\mathcal{A}(x) = ax + b$ with $|a| \neq 1$ and, for any n, distributions Q_{1n} and Q_{2n} satisfying

$$d(P, Q_{1n}) < \varepsilon, \qquad d(P, Q_{2n}) < \varepsilon, \qquad T(Q_{1n}) = |a|^n T(Q_{2n}).$$

From this it follows that either

$$\liminf_{n\to\infty}\big(\min\big(T(Q_{1n}), T(Q_{2n})\big)\big) = 0 \quad \text{or} \quad \limsup_{n\to\infty}\big(\max\big(T(Q_{1n}), T(Q_{2n})\big)\big) = \infty.$$

Using this fact we can define the breakdown point by

$$\begin{aligned}\varepsilon^*(T, P, d, \{0, \infty\}) = \inf\{\varepsilon > 0 : \inf[T(Q) : d(Q, P) < \varepsilon] = 0 \\ \text{or } \sup[T(Q) : d(Q, P) < \varepsilon] = \infty\}.\end{aligned} \tag{8}$$

This definition makes no reference to a metric but two points on the boundary of the parameter space, 0 and ∞, play a special role. The metric we use in this case is $D(\theta_1, \theta_2) = |\log(\theta_1/\theta_2)|$ and, not surprisingly, the points 0 and ∞ also play a special role here. The result is that $\varepsilon^*(T, P, d, \{0, \infty\}) = \varepsilon^*(T, P, d, D)$. We see that breakdown as defined by (8) can be reformulated in terms of an appropriately chosen metric on the parameter space Θ. This remains true even if Θ is compact. Suppose Θ is equipped with a metric D^*, bounded or not, and that some parameter value θ_0 is regarded as breakdown, for example, 0 in the scale context or 1 in the correlation context. We define the metric D on Θ by

$$D_{\theta_0}(\theta_1, \theta_2) = \left| \frac{1}{D^*(\theta_1, \theta_0)} - \frac{1}{D^*(\theta_2, \theta_0)} \right|. \tag{9}$$

It follows that if we keep θ_1 constant, then $D_{\theta_0}(\theta_1, \theta_2)$ tends to infinity if and only if θ_2 tends to θ_0. If we define in analogy to (8)

$$\varepsilon^*(T, P, d, \{\theta_0\}) = \inf\{\varepsilon > 0 : \inf[D^*(\theta_0, T(Q)) : d(Q, P) < \varepsilon] = 0\}, \tag{10}$$

then clearly $\varepsilon^*(T, P, d, \{\theta_0\}) = \varepsilon^*(T, P, d, D_{\theta_0})$. If there is a set of parameter values Θ_0 which are regarded as breakdown, for example, the boundary points, we define

$$D_{\Theta_0}(\theta_1, \theta_2) = \sup\{D_{\theta_0}(\theta_1, \theta_2) : \theta_0 \in \Theta_0\} \tag{11}$$

and again we have a metric which can be used to define breakdown. We define

$$\begin{aligned} &\varepsilon^*(T, P, d, \Theta_0) \\ &\quad = \inf\{\varepsilon > 0 : \inf\{\inf[D^*(\theta_0, T(Q)) : d(Q, P) < \varepsilon] : \theta_0 \in \Theta_0\} = 0\} \end{aligned} \tag{12}$$

and it follows that $\varepsilon^*(T, P, d, \Theta_0) = \varepsilon^*(T, P, d, D_{\Theta_0})$ and also

$$\varepsilon^*(T, P, d, \Theta_0) = \inf\{\varepsilon^*(T, P, d, \{\theta_0\}) : \theta_0 \in \Theta_0\}. \tag{13}$$

Grize (1978), as we shall see below, defines breakdown as the minimum contamination such that all points in Θ_0 are reachable and not just some such point. This can be accommodated by defining

$$\begin{aligned} &\varepsilon^{**}(T, P, d, \Theta_0) \\ &\quad = \inf\{\varepsilon > 0 : \inf\{\theta \in \Theta_0 : \sup[D_{\theta_0}(T(P), T(Q)) : d(P, Q) < \varepsilon] = \infty\}\}. \end{aligned} \tag{14}$$

In contrast to (13) this definition results in

$$\varepsilon^{**}(T, P, d, \Theta_0) = \sup\{\varepsilon^*(T, P, d, \{\theta_0\}) : \theta_0 \in \Theta_0\}. \tag{15}$$

There are no doubt other variations. The conclusion is that if breakdown is defined as convergence to some set of exceptional parameter values, then this can be described by a metric as required in our theorem. It still leaves open the question as to whether such a definition of breakdown is sensible but this can only be answered on a case-to-case basis.

7. Breakdown point? Rousseeuw in his contribution argues for the use of breakdown "value" rather than "point." We do not quite understand his reasoning and while usage is never absolute, we do not see any advantage in replacing "point" by "value." Hampel mentions the analysis of variance as one situation where the term breakdown point may not be appropriate. In the simple two-way table breakdown occurs if the majority of observations in any one row or column are badly contaminated, but this is too pessimistic and gives an artificially low breakdown point. In Terbeck and Davies (1998) the breakdown or interaction *patterns* for the two-way table are characterized and it is shown how these are related to the L_1-solution and to Tukey's median polish. Other articles concerned with patterns are Ellis and Morgenthaler (1992) for L_1 regression and Kuhnt (2000) for contingency tables.

8. Affine equivariance. We agree with Hampel that affine equivariance is not always a requirement and in two or more dimensions it is more difficult to justify than in one. We made comments to this effect in our paper. Nevertheless it is not always the case that outliers are apparent in the single coordinates and to find these some sort of equivariance would seem to be required. An example of a simple data set for which it is not sufficient just to look at the coordinates is given on page 57 of Rousseeuw and Leroy (1987). Programs based on high breakdown methods are now readily available and in our opinion should be used in a routine manner [Becker and Gather (1999, 2001), Rocke (1996) and Rousseeuw and van Driessen (1999)]. The costs are negligible and the returns can be substantial.

9. Correlation. This brings us to the perhaps most important part of the discussion. Hampel argues strongly that correlation provides an example of a useful concept of breakdown which does not have an equivariance structure. We argue that he is wrong on both counts: the concept is not useful and it does have an equivariance structure, albeit a simple one. We give a detailed reply which touches on many of the points discussed so far. Grize (1978) gives two definitions of breakdown for a rank correlation functional T_{rc}. The first reads [see (14), (15)]

$$\begin{aligned}\varepsilon^{**}(T_{\mathrm{rc}}, P, d, \{-1, 1\}) = \inf\big\{\varepsilon > 0 : \sup\{T_{\mathrm{rc}}(Q) : d(P, Q) < \varepsilon\} = 1,\\ \inf\{T_{\mathrm{rc}}(Q) : d(P, Q) < \varepsilon\} = -1\big\}\end{aligned} \tag{16}$$

and the second reads

$$\begin{aligned}\varepsilon^{**}(|T_{\mathrm{rc}}|, P, d, \{0, 1\}) = \inf\big\{\varepsilon > 0 : \sup\{|T_{\mathrm{rc}}(Q)| : d(P, Q) < \varepsilon\} = 1,\\ \inf\{|T_{\mathrm{rc}}(Q)| : d(P, Q) < \varepsilon\} = 0\big\}\end{aligned} \tag{17}$$

for some appropriate metric d. For the total variation metric Grize calculates the breakdown points of Kendall's and Spearman's rank correlation for some particular distributions. We carry out a small experiment for Spearman's functional T_{sc}.

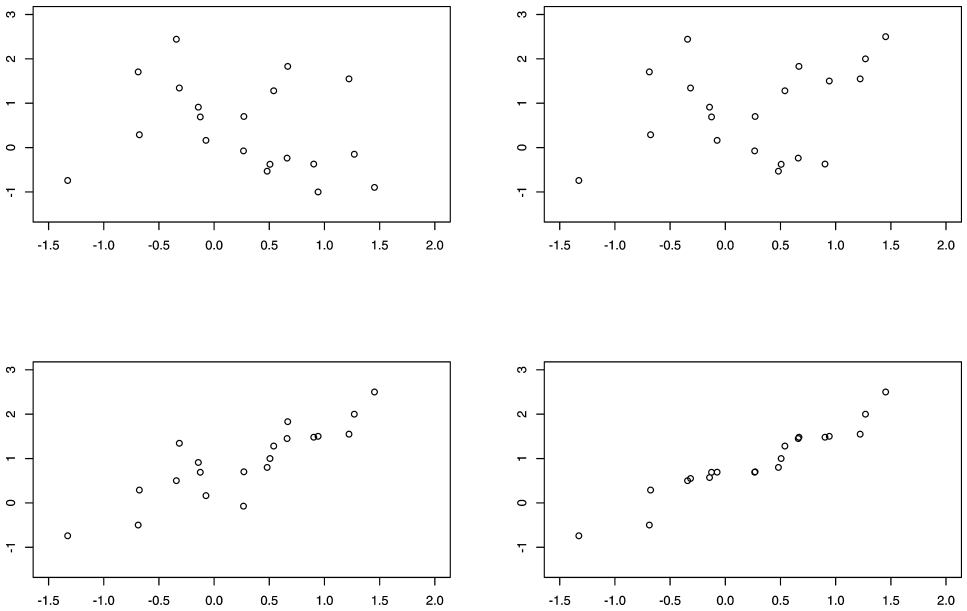

FIG. 1. *Samples differing from the initial sample* (upper left) *by* 3 *points* (upper right), 9 *points* (lower left) *and* 14 *points* (lower right) *with the rank-correlation changing from* -0.332 *to* 0.278, 0.878 *and* 1, *respectively.*

The top left panel of Figure 1 shows 20 data points with $T_{\mathrm{sc}}(P_n) = -0.332$. Initially there are various sets of six points for which $y_i = h(x_i)$ with h a nondecreasing function. We choose one and then move one of the remaining points at a time until finally after 14 moves all the points satisfy $y_i = h(x_i)$ with h nondecreasing. For the final sample the rank correlation is 1 and we have, according to (16), breakdown. The top right panel of Figure 1 shows the sample after three moves, the bottom left after nine moves. The final sample is shown in the bottom right panel. At no stage do we experience a breakdown. Each sample differs only slightly from the previous one and the values of the rank correlation are perfectly reasonable for the sample they refer to. In Hampel's terminology there is no pole. In a similar vein, Figure 2 shows a distribution considered by Grize for which he calculates the breakdown point 0.1 of Spearman's rank correlation in the sense of (17). The top left panel shows the initial distribution and the top right panel the same data after a monotone transformation. A breakdown to zero is shown in the bottom left panel and to 1 in the bottom right panel. In our opinion the bottom right panel is the only one where one would not a priori question any observation and yet this is classified as breakdown. We now play a similar game in one dimension and consider a simple standard normal sample of size 20. We consider the median and as breakdown corresponds to an arbitrarily large value of the median we start with 100. The game is now to alter the initial sample point by point until after ten moves the value of the median is at least 100. The moves are almost prescribed.

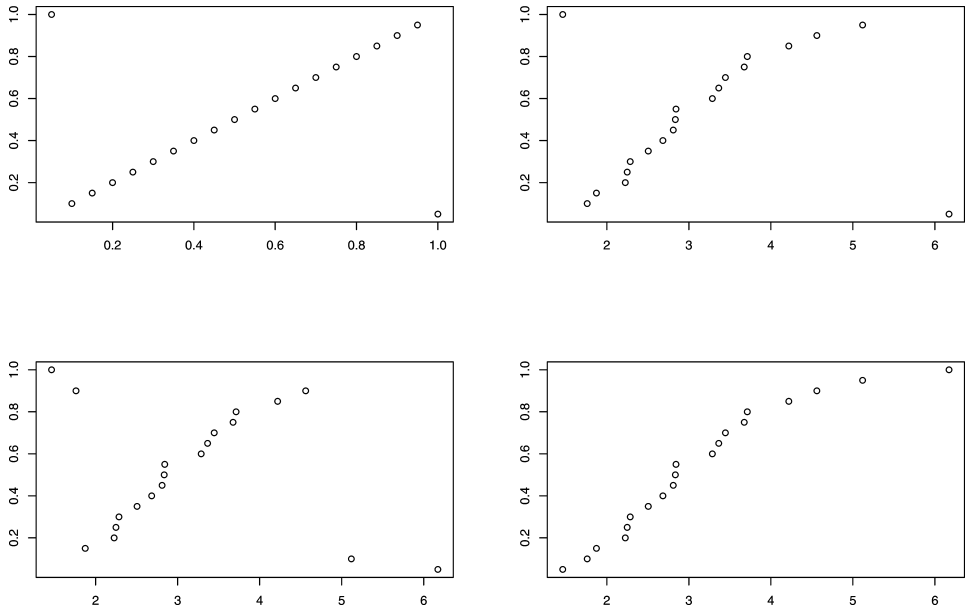

FIG. 2. *The upper left panel shows a distribution considered by Grize* (1978). *The upper right panel shows the same data after a monotone transformation. The bottom left panel shows the breakdown* [*in the sense of* (17)] *of Spearman's rank correlation to zero. The bottom right panel shows the breakdown of Spearman's rank correlation to* 1.

We choose any observation from the original sample and move it about 200 units to the right. After ten moves the median assumes a value of about 100. There is no other strategy. Even the first move alters the sample in a manner which distinguishes it immediately from the initial sample. Furthermore, when we progress from the ninth to the tenth move the median suddenly jumps from a value of about zero to one of about 100. We think this situation can be described by the word "breakdown." Moreover, it holds for any translation equivariant functional if one replaces the points carefully as in (6.2) and not as in (6.3). Any such functional must break down by the tenth move at the latest.

We now consider the usual linear correlation functional T_{lc}. For the initial data set of Figure 1 its value is -0.258. If we take any observation and move it to the point (γ, γ) and let γ tend to infinity, then T_{lc} tends to 1. In this situation it seems reasonable to use the word breakdown but perhaps discontinuity would be a better description. We analyze the problem more closely. Linear correlation can be placed into the context of our paper by introducing the following group structure. We define G_{lc} to be the group of transformations $g : \mathbb{R}^2 \to \mathbb{R}^2$ with

$$g(x, y) = (a_1 x + b_1, a_2 y + b_2) \tag{18}$$

with $a_1 a_2 \neq 0$. An equivariant functional T is one which satisfies

$$T(P^g) = \operatorname{sgn}(a_1 a_2) T(P). \tag{19}$$

Clearly the usual linear correlation functional T_{lc} is equivariant w.r.t. this group. The metric on the space of distributions is taken to be the strip metric

$$d_{\mathcal{ST}}(P, Q) = \sup\{|P(C) - Q(C)| : C \in \mathcal{ST}\}, \tag{20}$$

where $\mathcal{ST}$ denotes the set of strips C

$$C = \{(x, y) : -\delta \le ax + by + c \le \delta; a, b, c \in \mathbb{R}, \delta \in \mathbb{R}_+\}. \tag{21}$$

We note that this metric is also "correct" as it is invariant under the group G_{lc}:

$$d_{\mathcal{ST}}(P, Q) = d_{\mathcal{ST}}(P^g, Q^g), \qquad g \in G_{\mathrm{lc}}. \tag{22}$$

There is also a version of this metric which corresponds to (5) [see Davies (1993)]. To fit into the structure in our paper we also require a metric D on the parameter space $\Theta = [-1, 1]$. The precise metric is not important because of the simple nature of the equivariance given in (19). To be concrete we put in (9)

$$D^*(\theta_1, \theta_2) := |\tan(\theta_1 \pi/2) - \tan(\theta_2 \pi/2)|,$$

which is consistent with the desire to have breakdown at ± 1.

From (19) we see that the condition $G_1 \neq \varnothing$ is not satisfied and Theorem 3.1 does not provide a nontrivial upper bound. Indeed, there is an equivariant correlation functional with breakdown point 1, namely $T^o_{\mathrm{lc}} \equiv 0$, but to forestall protests we give another. For an empirical distribution P_n we define

$$T^*_{\mathrm{lc}}(P_n) = \frac{1}{N} \sum_{I, |I| \ge 3} T_{\mathrm{lc}}(I), \tag{23}$$

where I is a subset of the data containing $|I|$ observations, $N = 2^n - n - n(n-1)/2$ and $T_{\mathrm{lc}}(I)$ is, by an abuse of notation, T_{lc} evaluated at the empirical measure based on the set of observations in I. The functional T^*_{lc} is equivariant and also Fisher consistent. To calculate the breakdown point we consider an empirical measure P_n deriving from a sample of size n from a continuous distribution on $\mathbb{R}^2$ and another empirical measure Q_n. We assume that the supports of each are contained in some compact set K. The reason for these assumptions is to reduce complications due to the fact that the linear correlation coefficient as usually defined requires the existence of moments. We consider a sequence of Q_n with $\lim_{n\to\infty} T^*_{\mathrm{lc}}(Q_n) = 1$. From (23) it follows that the support of Q_n must be contained in a strip

$$C_n = \{(x, y) : -\delta_n \le a_n x + b_n y + c_n \le \delta_n\}$$

with $\lim_{n\to\infty} \delta_n = 0$. As $P_n(C_n) \le 2/n$ for sufficiently large n we have $d_{\mathcal{ST}}(P_n, Q_n) \ge 1 - 2/n$ and hence

$$\varepsilon^{**}(T^*_{\mathrm{lc}}, P_n, d_{\mathcal{ST}}, \{-1, 1\}) \ge 1 - 2/n \tag{24}$$

for this class of probability measures. We generalize this result in a manner which emulates the setting of our paper. As G_1 is empty we reformulate the definition of the functional $\Delta(P)$ of (3.3) in our paper as follows. We set

$$\Delta(P) = \sup\{P(B) : T(Q) \text{ not definable for } Q \text{ with } \operatorname{supp}(Q) \subset B\}. \tag{25}$$

For example, in the case of scale in $\mathbb{R}$ the relevant sets B are singletons and a measure concentrated on a singleton must have scale either zero or ∞ to be equivariant, both of which are excluded. If, following Grize, a linear correlation of ± 1 is defined to be breakdown, the corresponding sets are lines and this leads to

$$\Delta_+(P) = \sup\big\{P(C) : C = \{(x, y) : ax + by + c = 0\}, ab \le 0\big\}, \tag{26}$$

$$\Delta_-(P) = \sup\big\{P(C) : C = \{(x, y) : ax + by + c = 0\}, ab \ge 0\big\}, \tag{27}$$

and it follows that

$$\varepsilon^{**}(T^*_{\mathrm{lc}}, P, d_{\mathcal{TU}}, \{-1, 1\}) = 1 - \min\{\Delta_+(P), \Delta_-(P)\}. \tag{28}$$

The reasoning can be extended to rank correlation and this gives a more elegant theory as there are no problems with moments. The appropriate group is G_{rc} which consists of all transformations $g : \mathbb{R}^2 \to \mathbb{R}^2$ of the form

$$g((x, y)) = \big(\zeta(x), \eta(y)\big), \qquad \zeta, \eta, \mathbb{R} \to \mathbb{R}, \tag{29}$$

where each of ζ and η is either strictly increasing or strictly decreasing. A correlation functional T_{rc} is equivariant with respect to this group if

$$T_{\mathrm{rc}}(P^g) = \operatorname{sgn}(\zeta \circ \eta) T_{\mathrm{rc}}(P), \tag{30}$$

where $\operatorname{sgn}(\zeta) = \pm 1$ depending on whether ζ is strictly increasing or decreasing. The natural metric is the tube metric

$$d_{\mathcal{TU}}(P, Q) = \sup\{|P(C) - Q(C)| : C \in \mathcal{TU}\}, \tag{31}$$

where $\mathcal{TU}$ denotes the set of monotonic tubes C

$$C = \{(x, y) : -\delta \le h(x) + y \le \delta, h : \mathbb{R} \to \mathbb{R} \text{ strictly monotonic}, \delta \in \mathbb{R}_+\}. \tag{32}$$

The metric is "correct" in that it is invariant with respect to the group G_{rc}:

$$d_{\mathcal{TU}}(P, Q) = d_{\mathcal{TU}}(P^g, Q^g), \qquad g \in G_{\mathrm{rc}}. \tag{33}$$

As we now require correlations of ± 1 only for data points which are strictly increasing or decreasing, we define analogously to (26) and (27),

$$\Delta_+(P) = \sup\big\{P(C) : C = \{(x, y) : y = h(x)\}, h \text{ strictly increasing}\big\}, \tag{34}$$

$$\Delta_-(P) = \sup\big\{P(C) : C = \{(x, y) : y = h(x)\}, h \text{ strictly decreasing}\big\}. \tag{35}$$

From this it follows for Spearman's rank correlation functional T_{sc} that

$$\varepsilon^{**}(T_{\mathrm{sc}}, P, d_{\mathcal{TU}}, \{-1, 1\}) = 1 - \min\{\Delta_+(P), \Delta_-(P)\}. \tag{36}$$

In fact (36) holds for any functional for which $T_{\mathrm{rc}}(P) = 1$ or -1 if and only if $\Delta_+(P) = 1$ or $\Delta_-(P) = 1$, respectively, and consequently it also holds for Kendall's τ. The appearance of min in (28) and (36) is due to Grize's definition (16) of breakdown which refers to both boundary points. Usable estimates of $\Delta_+(P_n)$ are available for empirical measures P_n deriving from nonatomic i.i.d. random variables in each component; that is, the components are also independent. Let the sample be (X_i, Y_i), $i = 1, \ldots, n$, and consider the points $(X_{i1}, Y_{i1}), \ldots, (X_{ik}, Y_{ik})$ with the X_{ij}, $j = 1, \ldots, k$, in increasing order. The points lie on some curve $y = h(x)$ for a strictly increasing h if and only if the Y_{ij}, $j = 1, \ldots, k$, are also in increasing order. The probability of this is $1/k!$. There are $\binom{n}{k}$ different samples of size k and we see that the probability that at least k points lie on some curve $y = h(x)$ is at most

$$\frac{1}{k!}\binom{n}{k} \le \frac{n^k}{(k!)^2}.$$

By maximizing over k we obtain

$$\Delta_+(P_n) = O\bigl(1/\sqrt{n}\bigr)$$

and it follows from (36)

$$\varepsilon^{**}(T_{\mathrm{sc}}, P_n, d_{\mathcal{TU}}, \{-1, 1\}) \ge 1 - O\bigl(1/\sqrt{n}\bigr). \tag{37}$$

The fact that (36) also holds for Kendall's τ apparently contradicts Hampel's comments, but this is not so because it is definition (17) of breakdown to which Hampel's comments apply. To proceed we consider the problem of maximizing $\Delta_+(P)$ subject to $T_{\mathrm{rc}}(P) = 0$. For T_{sc} the answer is $\Delta_+(P) = \sqrt[3]{1/2}$, which is attained at a distribution for which the rank of x_i is i and the rank of y_i is $k + i$, $1 \le i \le n - k$, and $n - i + 1$, $k + 1 \le i \le n$, with $k = n\sqrt[3]{1/2}$. The corresponding result for Kendall's τ replaces $\sqrt[3]{1/2}$ by $\sqrt{1/2}$. If now Q is any distribution with $\Delta_+(Q) = 1$, it follows that the breakdown point [in the sense of (17)] at Q is $1 - \sqrt[3]{1/2} = 0.2063$ for Spearman's rank correlation and $1 - \sqrt{1/2} = 0.2929$ for Kendall's τ. If we now move only half the mass of $1 - \sqrt[3]{1/2}$, it is clear that we can obtain distributions Q_1 and Q_2 at which Spearman's and Kendall's rank correlations have breakdown points of $(1 - \sqrt[3]{1/2})/2$ and $(1 - \sqrt{1/2})/2$, respectively, and that these are the smallest possible breakdown points. We have not understood Hampel's claim $BP(K) = \frac{3}{2}BP(S)$ as, as far as we can see, these refer to different distributions, one with $\Delta_+(Q_1) = 0.85$ and one with $\Delta_+(Q_2) = 0.9$, but this is only a minor point. The tube metric $d_{\mathcal{TU}}$ is stronger than the strip metric $d_{\mathcal{ST}}$ but considerably weaker than the total variation metric d_{tv} used by Grize. In particular it allows for wobbling of the observations. We also note that neither metric suffers from the deficiency of the Prohorov metric of mixing dimensionless probabilities with measurement units. The class $\mathcal{ST}$ of strips has polynomial discrimination but not the class $\mathcal{TU}$

as there are arbitrarily large finite subsets of $\mathbb{R}^2$ which can be shattered by $\mathcal{T}\mathcal{U}$ [see Pollard (1984)].

Finally, we point out the differences to Theorem 3.1. If we take the usual definition of breakdown as a worst-case situation rather than Grize's definition which is a sort of best worst-case definition, then the breakdown point of Kendall's or Spearman's rank correlation is

$$\varepsilon^*(T_{\mathrm{lc}}^*, P, d_{\mathcal{T}\mathcal{U}}, \{-1, 1\}) = 1 - \Delta(P), \tag{38}$$

where

$$\Delta(P) = \max\{\Delta_+(P), \Delta_-(P)\}.$$

In the general situation we argue as follows. Let $\mathcal{P}_0$ denote the set of distributions at which T breaks down, which means that their support is contained in some exceptional subset of the sample space as in (25). Suppose $0 < \Delta(P) < 1$ and choose an exceptional subset B_0 of the sample space with $P(B_0) = \alpha$, $0 < \alpha < \Delta(P)$. If we define $Q_0(\cdot) = P(\cdot \cap B_0)/P(B_0)$ and $Q_1(\cdot) = P(\cdot \cap (\mathcal{X} \setminus B_0))/(1 - P(B_0))$, then Q_0 and Q_1 are probability measures with $P = \alpha Q_0 + (1 - \alpha)Q_1$. If the metric on $\mathcal{P}$ satisfies (2.2) of our paper, we see that $d(P, Q_0) \leq 1 - \alpha$ and this implies

$$\varepsilon^*(T, P, d, D) \leq 1 - \Delta(P). \tag{39}$$

This differs from the claim of Theorem 3.1 by the factor of $1/2$ and it is precisely the group structure which produces this factor. Because of equivariance things start going wrong before one reaches an arbitrarily small neighborhood of some point in $\mathcal{P}_0$. As Tyler mentions in his contribution, heuristic justifications for the factor of $1/2$, such as not being able to distinguish between good and bad data, are too vague. One of the challenges of this paper is to obtain the factor of $1/2$ or even some other factor without an equivariance structure.

10. Principal component vectors. Tyler argues that it may be possible to define a reasonable concept of breakdown for principal component vectors, although he recognizes that there are problems involved. The idea is that breakdown occurs if contamination results in the first principal component vector being orthogonal to the first principal component vector without contamination. This example cannot be reformulated in terms of metrics as described in Section 6, as there is no special set of parameter values Θ_0. Furthermore, it is not possible to adjust the proof of Theorem 3.1 to include this case. However, we shall now argue that it does not make sense to talk about the breakdown of the principal component vectors without reference to the corresponding eigenvalues.

Consider a two-dimensional data set for which the eigenvalues are the same. The set of first principal component vectors is now the set of all points on the unit circle with θ and $-\theta$ being identified. The smallest alteration of any observation will cause the space to collapse to a single direction, say $\theta_1 = (1, 0)$, with the

second principal component vector $\theta_2 = (0, 1)$ being orthogonal to it. It is clear that there exists an arbitrarily small perturbation of the original data set such that $\theta_1 = (0, 1)$ and $\theta_2 = (1, 0)$. In other words, there exist data sets for which arbitrarily small perturbations cause breakdown in the first principal component vector. The perturbations can be so small as to be nondetectable and any computer of finite precision or some nonoptimal numerical recipe may result in the wrong answer and be the "cause" of the breakdown. It seems to us that this situation is one which is not describable by the word "breakdown." We cannot think of any useful statistical procedure which can be made to break down by the smallest of perturbations of the data set. In practice, of course, use is not just made of the first principal component vector but of all those principal component vectors for which the eigenvalues are in some sense large. One data analytical strategy is to look at the two-dimensional plots on the first two principal component vectors, and here it is irrelevant if they are the wrong way round. The principal component vectors are defined as those directions where the variability of the data, however measured, is particularly large. The word "breakdown" can be more appropriately applied to a situation in which the large variability is the result of outliers and causes a direction of small variability to become one of large variability. It seems to us to be clear that the size of the eigenvalues will have to be taken into account. Principal component vectors do not therefore constitute a counterexample to our meta claim of no nontrivial theory of breakdown without groups. In spite of this Tyler has alerted us to the possibility of breakdown being defined in terms of a relationship between two parameter values rather than closeness to some specific parameter values. We cannot exclude the possibility of there being some perfectly reasonable concept of breakdown of this nature.

11. Fisher consistency. Hampel does not like the example of regression through the origin and neither do we. It was included as an answer to a referee as to whether it was possible to construct a Fisher consistent functional with a breakdown point of at least $1/3$. Fisher consistency seems to be the obvious candidate to replace group equivariance as a desirable property of a functional. We do not give a theorem as there are difficulties in defining what is meant by a reasonable parametric family, but a parametric family $\{P_\theta, \theta \in \Theta\}$ typically forms a very sparse subset of the set of all models. This is indicated by Figure 3 where the line represents the family of models in the space of all probability measures and the circles indicate an infinitesimal neighborhood. Fisher consistency describes the behavior of the functional only in the infinitesimal neighborhood. We are left free to define the functional elsewhere and this is what we exploit in our example. Equivariance considerations prevent this form of local definition. The orbits connect points which are far apart in the space of probability models and this prevents constructions such as the one we give.

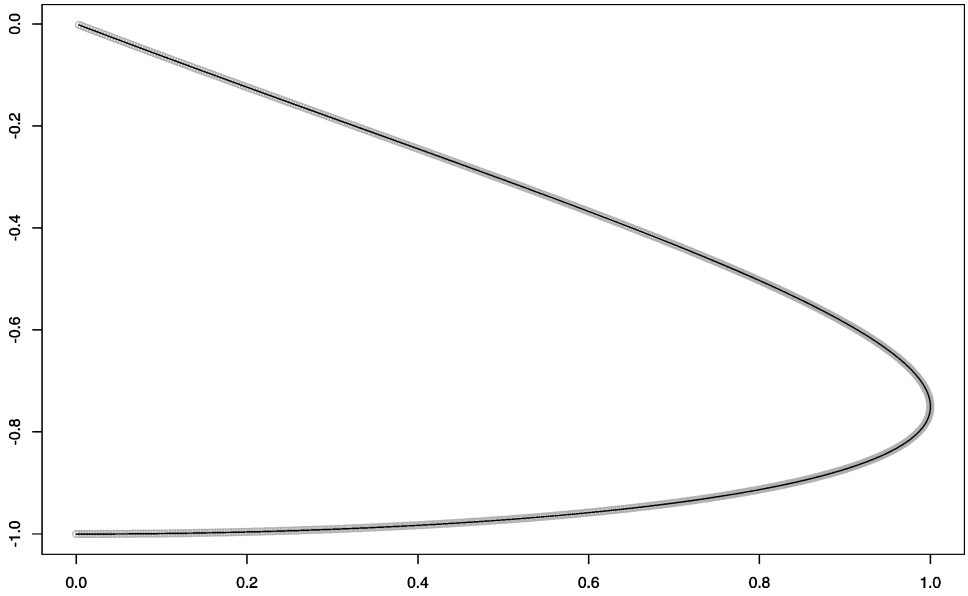

FIG. 3. *A thin parametric model and an infinitesimal neighborhood within which Fisher consistency becomes relevant.*

12. The samples (6.2) and (6.3). Hampel, Rousseeuw and Tyler all comment on the samples (6.2) and (6.3). Rousseeuw correctly remarks that one can often calculate the breakdown point of a functional directly and that such direct proofs do not rely on a repetition. Hampel says, also correctly, that the unnamed functional (there are many) must have a low breakdown and suggests that perhaps some small print is missing. What is missing is some large print explaining exactly what we intended with these two examples. Tyler saw clearly what was intended and has made some very interesting comments on (6.2) and (6.3). He also explicitly mentions the connection with the area of computer vision which was one of our motivations as we indicate below. He points out that such apparently well-understood functionals such as appropriately tuned redescending M-functionals can exhibit the same behavior. We confess to not having been aware of this and we would have chosen another example had we known. As Tyler says, under appropriate conditions redescending affine equivariant M-functionals do not break down even under 99% contamination. This is exactly the phenomenon to which we intended to bring attention.

The proof of Theorem 3.1 relies in part on exactly reproducing a portion of the data elsewhere. If there is no exact repetition as in (6.3), there will be many equivariant functionals which do not break down. One choice for sample (6.3) is

$$\bigl(T_l(\mathbf{x}_n), T_s(\mathbf{x}_n)\bigr) = \arg\min_{\mu,\sigma}\left\{\sum_{j=1}^{3} r_{(i)}(\mu,\sigma)^2\right\}, \tag{40}$$

where

$$r_i(\mu,\sigma)^2 = \min_{1\le j\le n}\left(\frac{x_j-\mu}{\sigma} - z_i\right)^2$$

and $z_1 = 1.5$, $z_2 = 1.8$ and $z_3 = 1.3$. In this connection we mention Oja's example of linear regression at the end of his Section 1. We fail to follow his argument as to why the estimate becomes uninformative. As it stands, the argument seems to make no use of the assumption $n = 2k$, in which case we can put $k = 1$ and the conclusion would seem to be that every regression equivariant functional has a breakdown point of $1/n$. If $n = 2k$ is implicitly meant, then breakdown occurs only if we cannot distinguish between the two samples. If we can distinguish between the two samples, for his example if $x_1 = x_2 = \cdots = x_k = 0$, then what is claimed as breakdown is nothing but equivariance (see Section 2 above).

At first glance the functional (40) may seem very artificial but this is not so. It is constructed to find a particular pattern in the sample, namely affine transformations of 1.5, 1.8 and 1.3. Figure 4 shows the smile of the Cheshire cat and the problem is to locate it in a sea of noise into which it gradually disappears. This is only possible as the noise does not reproduce the signal. For real examples from the area of computer vision we refer to Wang and Suter (2004). In analytical terms one can define a modified breakdown point by

$$\varepsilon^*(T, P, d, \mathcal{H}) = \inf\{\varepsilon > 0 : \sup|T(P) - T(Q)|, d(P, Q) < \varepsilon, Q \in \mathcal{H}\}, \tag{41}$$

where $\mathcal{H}$ specifies what you want to protect yourself against. If $\mathcal{H}$ does not

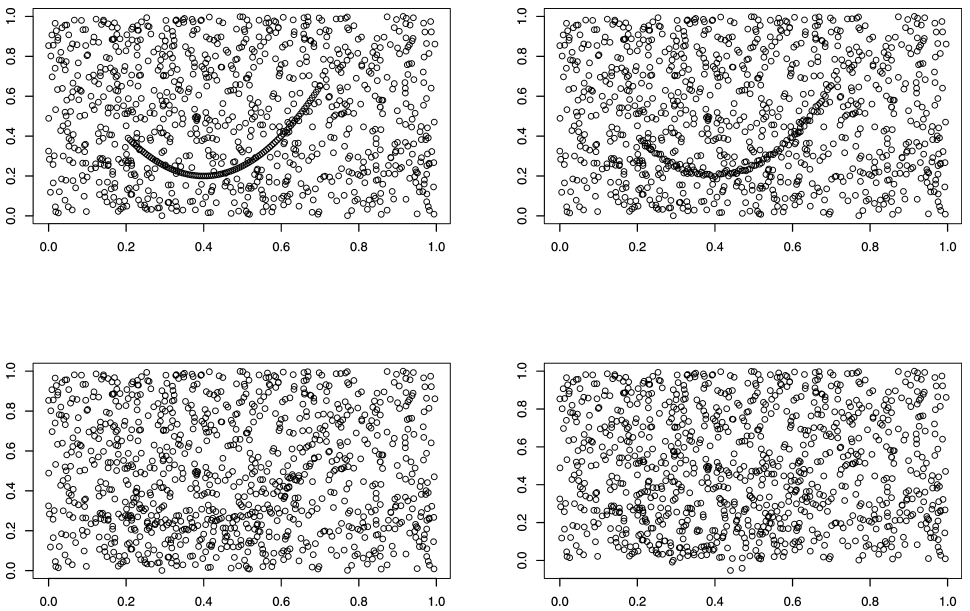

FIG. 4. *The smile of the Cheshire cat gradually disappearing as it is corrupted by noise.*

allow a repetition of the signal elsewhere, then affine equivariant functionals can attain breakdown points higher than $1/2$. Moreover, the usual high breakdown functionals are typically of no help in this situation [see Wang and Suter (2004)].

13. Nonparametric statistics. In this paper we have shown that the concept of breakdown point has been generally accepted only in situations where there is a group structure sufficiently rich to allow the calculation of a nontrivial upper bound for the breakdown point. In his contribution Hampel speculates that this could be the reason why the breakdown point for correlation coefficients has not yet been widely accepted. In spite of this and as pointed out by an Associate Editor, we have not proved a theorem to the effect that a breakdown point is only sensible when a rich group structure exists. It is difficult to imagine what such a theorem would look like. Nevertheless the paper, the contributions of the discussants and our reply do seem to indicate that it will not be easy to come to an acceptable definition of breakdown with a nontrivial upper bound without a group structure. There is perhaps another reason why some definitions of breakdown have been successful. They are defined for so-called nonparametric functionals in the sense of Bickel and Lehmann (1975a, b). One can always calculate the median of a distribution in $\mathbb{R}$ and this is not associated with a restrictive stochastic model. We wish to emphasize this as we have the impression that it is sometimes assumed that functionals are only to be applied to data which is generated by some stochastic model but with contamination. Genton and Lucas entitle a section "Breakdown point for (in)dependent observations," which suggests at least to us that they distinguish between samples which are generated by independent random variables and those which are not. The title of Genton and Lucas (2003) also tends in this direction. They write "$\mathcal{Y}$ is the set of all allowable samples" and later "$\mathcal{Y}$ is the set of all stationary AR(1) processes." We think the intention is clear. The data are generated by a stationary AR(1) process and then contaminated by the outliers. On the other hand, the only possible mathematical interpretation of "the set of all stationary AR(1) processes" is the support of the model. As the support of an AR(1) process with Gaussian innovations is $\mathbb{R}^n$, this means simply all samples, $\mathcal{Y} = \mathbb{R}^n$. Thus what at first glance seems plausible turns out to be untenable. This is the reason why the restrictions we place on data are analytical ones and not distributional ones. The median can be successfully applied to data which are very obviously dependent [see, e.g., Davies, Fried and Gather (2004)], but consider the data in Figure 5. For these data it makes no sense to artificially restrict ρ to the interval $[-1, 1]$. Rather one would point out that the data can be well approximated by an AR(1) model with $\rho = -1.25$ but not by a stationary AR(1) model. This simple message is lost if we are forced to specify a value in $[-1, 1]$. If breakdown is going to be meaningful in such a situation we suspect that it should be applied to a statistical procedure and not to the behavior of a single functional.

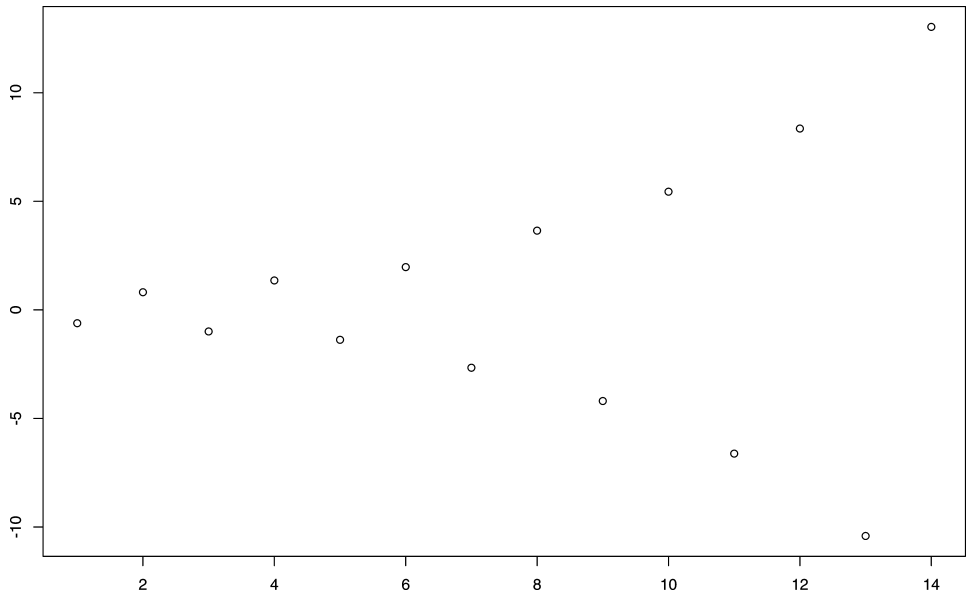

FIG. 5. *A sample of size* 14 *generated by* $X_{t+1} = -1.25X_t + 0.2Z(t)$ *with* $Z(t)$ *standard Gaussian white noise.*

14. Breakdown without groups and alternatives. We have argued above that the only generally accepted definitions of breakdown are in situations where there is a sufficiently rich group and equivariance structure. If a need is felt to extend it to other situations, we state what we think are the minimal requirements. First, the definition should be capable of being made precise. He argues that breakdown is the smallest fraction of contamination which makes a test statistic "uninformative or unusable." Later he argues that breakdown should have the same degree of vagueness as he claims to be the case with outliers. He continues that "when every statistician starts to talk about his or her own notion of a breakdown point, I think we have made it." We think there are dangers in such an attitude. Ostensive definitions of breakdown with statisticians pointing in all directions are unlikely to contribute to a general acceptance of the word. Intuition is important, but just as is the case with outliers [see Davies and Gather (1993)] much is to be gained by undertaking the attempt to give a precise definition and to investigate its consequences. This not only deepens the understanding, it also sharpens the intuition. Semantics is important and we think that any generalizations of the concept of breakdown should be such as to be recognizably referring to some common element, in particular the presence of some natural pole. Second, it is not sufficient to give a definition of breakdown and show that it gives the correct answer in some particular cases. A definition of breakdown should be subjected to some form of analysis, including its stability under perturbations. The onus is on those who propose definitions of breakdown to do this. Third, the definition

should be simple and intuitively appealing. Here we agree completely with He. If it requires more than sixty seconds to understand a definition, it is probably bad. Fourth, when calculating breakdown points use should be made of metrics which can accommodate rounding errors. Gross error neighborhoods and the total variation metric are too restrictive. Fifth, the definition should not be too restrictive and only apply to one single functional. It should apply to a whole family of functionals which offer different possibilities of quantifying the feature of interest, location, scale, correlation or nonparametric function. Sixth, there should be a class of reasonable functionals for which it makes sense to compare breakdown points. If such a definition of breakdown is not possible, there are alternatives. One is simply to compare different functionals by their continuity or bias properties, again if possible in weak metrics. For this to make sense it is not necessary that an explosion occurs. It may be that this proves more useful than trying to extend the idea of breakdown to situations for which it is not suitable.

15. Conclusion. We thank all discussants for their contributions and hope that the disagreements that are apparent have been clarified by our rejoinder. In our paper we have not proved that breakdown without equivariance is not a sensible concept. On the other hand, in all situations we are aware of in which there is no or little equivariance (made precise by our main theorem), then either (i) breakdown points of 1 are attainable or (ii) the word breakdown is inappropriate (the movement from the top right to the bottom right panel of Figure 2) or (iii) the very definition of breakdown point is inadequate. An example without these weaknesses would be interesting.

REFERENCES

BECKER, C. and GATHER, U. (1999). The masking breakdown point of multivariate outlier identification rules. *J. Amer. Statist. Assoc.* **94** 947–955.

BECKER, C. and GATHER, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Comput. Statist. Data Anal.* **36** 119–127.

BICKEL, P. J. and LEHMANN, E. L. (1975a). Descriptive statistics for nonparametric models. I. Introduction. *Ann. Statist.* **3** 1038–1044.

BICKEL, P. J. and LEHMANN, E. L. (1975b). Descriptive statistics for nonparametric models. II. Location. *Ann. Statist.* **3** 1045–1069.

DAVIES, P. L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20** 1828–1843.

DAVIES, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21** 1843–1899.

DAVIES, P. L., FRIED, R. and GATHER, U. (2004). Robust signal extraction for on-line monitoring data. Contemporary data analysis: Theory and methods. *J. Statist. Plann. Inference* **122** 65–78.

DAVIES, P. L. and GATHER, U. (1993). The identification of multiple outliers (with discussion). *J. Amer. Statist. Assoc.* **88** 782–801.

DAVIES, P. L. and GATHER, U. (2002). Breakdown and groups. Technical Report 10/2002, Sonderforschungsbereich 475, Univ. Dortmund.

DAVIES, P. L. and KOVAC, A. (2004). Densities, spectral densities and modality. *Ann. Statist.* **32** 1093–1136.

ELLIS, S. P. and MORGENTHALER, S. (1992). Leverage and breakdown in L_1 regression. *J. Amer. Statist. Assoc.* **87** 143–148.

GENTON, M. G. and LUCAS, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 81–94.

GRIZE, Y. L. (1978). Robustheitseigenschaften von Korrelationsschätzungen. Diplomarbeit, Swiss Federal Institue of Technology (ETH), Zürich.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. California, Berkeley.

HAMPEL, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bull. Inst. Internat. Statist.* **46** (1) 375–391.

KERSTING, G. D. (1978). Die Geschwindigkeit der Glivenko–Cantelli–Konvergenz gemessen in der Prohorov-Metrik. *Math. Z.* **163** 65–102.

KUHNT, S. (2000). Ausreißeridentifikation im Loglinearen Poissonmodell für Kontingenztafeln unter Einbeziehung robuster Schätzer. Ph.D. dissertation, Univ. Dortmund.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RIEDER, H. (2000). Neighborhoods as nuisance parameters? Robustness vs. semiparametrics. Technical Report 25/2000, SFB 373, Humboldt Univ., Berlin.

ROCKE, D. M. (1996). Robustness properties of *S*-estimators of multivariate location and shape in high dimension. *Ann. Statist.* **24** 1327–1345.

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

ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

TERBECK, W. and DAVIES, P. L. (1998). Interactions and outliers in the two-way analysis of variance. *Ann. Statist.* **26** 1279–1305.

WANG, H. and SUTER, D. (2004). Robust adaptive-scale parametric model estimation for computer vision. *IEEE Trans. Pattern Analysis and Machine Intelligence* **26** 1459–1474.

FACHBEREICH 06—MATHEMATIK
UND INFORMATIK
UNIVERSITÄT DUISBURG–ESSEN
45117 ESSEN
GERMANY
E-MAIL: davies@stat-math.uni-essen.de

FACHBEREICH STATISTIK
UNIVERSITÄT DORTMUND
44221 DORTMUND
GERMANY
E-MAIL: gather@statistik.uni-dortmund.de