[1]

Breakdown and Groups II [2]

By P. L. Davies and U. Gather

University of Duisburg-Essen, University of Dortmund

This extends the work of Davies and Gather (2004).

1. Introduction The notion of breakdown point was introduced by Hampel (1968, 1971) and has since played an important rôle in the theory and practice of robust statistics. In Davies and Gather (2004) it was argued that the success of the concept is connected to the existence of a group of transformations on the sample space and the linking of breakdown and equivariance. For example the highest breakdown point of any translation equivariant functional on the real line is 1/2 whereas without equivariance considerations the highest breakdown point is the trivial upper bound of 1. The situation considered in Davies and Gather (2004) requires the existence of "banned" parameter values such as ∞ in the case of translation and 0 and ∞ in the case of scale. In the discussion of Davies and Gather (2004) Tyler pointed out that there are situations where there are no banned parameter values but that one may nevertheless wish to have some concept of breakdown. The immediate example is that of directional data (see Mardia (1972)) where there is no banned direction but a concept of breakdown might prove useful. It may seem that

[1]Received??

AMS 2000 subject classifications. Primary 62G07; secondary 65D10, 62G20.

Key words and phrases. equivariance, breakdown point, robust statistics

[2]Research supported in part by Sonderforschungsbereich 475, University of Dortmund.

breakdown can be defined here as the smallest amount of contamination required to cause the direction to differ from the original direction by 180°. A further example explicitly mentioned by Tyler is that of principal components where breakdown could be said to occur when the first principal direction is orthogonal to the first principal direction for the non-contaminated sample. To be more mathematical, consider the unit circle S and a direction functional T on the set $\mathcal{P}$ of probability distributions over S. We take $\mathcal{P}$ to be equipped with a metric d and define the breakdown point by

$$\varepsilon^*(T, P, d) = \inf\{\varepsilon > 0 : |T(P) - T(Q)| = \pi \text{ for some } Q \text{ with } d(P, Q) < \varepsilon\} \quad (1.1)$$

where we have measured angles in radians. The problem with this definition is that there may exist a sequence $Q_j, j = 1, \ldots N$ with say $d(P, Q_j) < j\varepsilon/N$ and $T(Q_j) = T(P) + j\pi/N$. In other words we can move from P to Q in small steps Q_j and such that at each stage the value of $T(Q_j)$ is perfectly reasonable for the distribution Q_j. An explicit example is given in the rejoinder of Davies and Gather (2004) in the case of correlation. If breakdown is defined in terms of banned parameter values then such a construction is not possible. A definition of breakdown which includes directional data has been given by He and Simpson (1992). They consider gross error neighbourhoods and define breakdown to be the smallest neighbourhood within which all parameter values are attainable. We feel however the definition of He and Simpson as well as that of (1.1) refer to properties of a functional better described in terms of lack of continuity rather than in terms of breakdown. Nevertheless there is a situation which we think can be described by breakdown, namely when it is not possible to define the functional in a consistent manner. The obvious example is that of the mean $T_m(P) = \int x\, dP(x)$ for distributions P on $\mathbb{R}$ which is only defined for distributions P satisfying $\int |x|\, dP(x) < \infty$. This contrasts with the median

which can be defined in a unique manner for all distributions on the real line and still be affine equivariant.

2. Breakdown and invariant distributions

2.1. *Definition of breakdown* We use the notation of Davies and Gather (2004). Let $\mathcal{X}$ be a sample space and $\mathcal{G}$ a group of measurable transformations $g : \mathcal{X} \to \mathcal{X}$ with identity ι. We denote the set of all probability distribution P on $\mathcal{X}$ by $\mathcal{P}$ which is equipped with a metric d. Let Θ be some parameter space equipped with a group structure $\mathcal{H}_{\mathcal{G}}$, induced by $\mathcal{G}$, consisting of elements $h_g, g \in \mathcal{G}$, for which $h_{g_1} \circ h_{g_2} = h_{g_1 \circ g_2}$. A functional $T : \mathcal{P}_T \to \Theta$ with $\mathcal{P}_t \subset \mathcal{P}$ is called equivariant if the following hold:

(a) $\mathcal{P}_T$ is closed under all $g \in \mathcal{G}$,

(b) T is well defined on $\mathcal{P}_T$,

(c) $T(P^g) = h_g(T(P))$ for all $P \in \mathcal{P}_T$ and $g \in \mathcal{G}$.

This leads to the following definition of breakdown

$$\varepsilon^*(T, P, d) = \inf\{\varepsilon > 0 : d(P, Q) < \varepsilon \text{ for some } Q \notin \mathcal{P}_T\} \tag{2.2}$$

with of course $\varepsilon^*(T, P, d) = 0$ if $P \notin \mathcal{P}_T$. We note that this concept of breakdown does not require a topology on the parameter space Θ. As an example we consider the mean T_m which, as mentioned at the end of the last section, is defined only for distributions with a finite absolute first moment. As any neighbourhood of any distribution P on the real line contains distributions with an infinite absolute first moment it follows that the mean has a breakdown point of zero. In contrast the median can be well-defined for all distributions on the real line and consequently has a breakdown point of one in the sense of (2.2) above.

Suppose that there exist distributions P with $P^g = P$ for some g with $h_g \neq h_\iota$. We denote the set of all such distributions by $\mathcal{P}_{ginv}$. If T is equivariant and $g \in \mathcal{P}_{ginv}$ we have $T(P) = T(P^g) = h_g(T(P))$ which is not possible as $h_g \neq h_\iota$. We see that

$$\mathcal{P}_{ginv} \subset \mathcal{P} \setminus \mathcal{P}_T \tag{2.3}$$

for every equivariant functional T. This implies

$$\varepsilon^*(T, P, d) \leq \inf\{\varepsilon > 0 : d(P, Q) < \varepsilon \text{ for some } Q \in \mathcal{P}_{ginv}\}. \tag{2.4}$$

We note that (2.4) gives an upper bound for the breakdown point which is the same for all equivariant functionals T.

2.2. *Finite sub-groups* Suppose $\mathcal{G}$ contains a finite sub-group $\mathcal{G}_k$ of order $k \geq 2$ so that $g^k = \iota$ for all $g \in \mathcal{G}_k$. For any distribution P we set

$$P_k = \frac{1}{k} \sum_{j=0}^{k-1} P^{g^j}. \tag{2.5}$$

Then $P_k^g = P_k$ so that $P_k \in \mathcal{P}_{ginv}$ and hence

$$\varepsilon^*(T, P, d) \leq d(P, P_k).$$

If the metric d satisfies (2.1) and (2.2) of Davies and Gather (2004) we have

$$d(P, P_k) \leq \frac{k-1}{k}\ d(P, \tilde{P}_k) \leq \frac{k-1}{k}$$

where

$$\tilde{P}_k = \frac{1}{k-1} \sum_{j=1}^{k-1} P^{g^j}.$$

Examples of sample spaces and groups $\mathcal{G}$ with subgroups of order $k = 2$ are the unit circle and the unit sphere. In both cases the maximum breakdown point of any direction functional is $1/2$.

2.3. *Total variation and finite sample breakdown points* Although the above results can be extended to finite sample breakdown points this may not always make sense. In particular if P is an empirical measure then the breakdown point measured using the total variation metric $d = d_{tv}$ may be reduced from $1/2$ to $1/n$ by the smallest of alterations in the values of the data. The same applies to the finite sample breakdown points. This is the only example we know where the use of a metric which allows for minor alterations in the values of the data points leads to a completely different breakdown point.

REFERENCES

Davies, P. L. and Gather, U. (2004). Breakdown and groups (with discussion). *Annals of Statistics* to appear.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph. D. thesis, Dept. Statistics, Univ. California, Berkeley.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods (with discussion). *Proceedings of the 40th Session of the ISI*, Vol. **46**, Book 1, 375–391.

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

Mardia, K. V. (1972). *Statistics of directional data.* Academic Press, London, New York.