BREAKDOWN OF COVARIANCE ESTIMATORS

by

Werner A. Stahel

RESEARCH REPORT No. 31

December 1981

FACHGRUPPE FUER STATISTIK
EIDGENOESSISCHE TECHNISCHE HOCHSCHULE
8092 ZUERICH

Breakdown of covariance estimators

by

Werner A. Stahel

Fachgruppe für Statistik, Eidg. Technische Hochschule

8092 Zürich

SUMMARY

The covariance-location model is considered as a model generated by a group of transformations. A new variant of Hampel's (1971) "breakdown point" is defined which disqualifies some clearly undesirable non-invariant estimators with high ordinary breakdown point. For estimators which are invariant with respect to the transformations generating the model, the two concepts coincide. Finally a new class of invariant covariance estimators with breakdown point 1/2 is given. The idea is to exclude all observations that have too extreme a projection on any direction of the observation space.

1. INTRODUCTION

Consider the problem of estimating the (variance-)covariance matrix and location vector of a p-dimensional normal distribution. We restrict our attention to estimators which may be written as functionals on the space of probability distributions. The value of such an estimator for a given sample is the value of the functional at the empirical distribution. Most of the commonly used estimators may be written as or approximated by a functional, and are consistant for its value at the distribution of the observations (compare Huber, 1981).

The breakdown point is the measure of an important aspect of robustness of estimators. Roughly speaking, it is determined as follows: mix a model distribution with an arbitrary distribution - the "contamination" - in proportions $(1-\varepsilon):\varepsilon$. Check whether you can move the value of the estimator (functional) for that mixture beyond any bound just by varying the contamination. The breakdown point is supremum over all ε for which this is not possible.

The classical estimators for the parameters of the p-dimensional normal distribution have breakdown point zero, which indicates their complete lack of robustness.

Maronna (1976) and Huber (1977) proposed a class of M-estimators (for a general definition, see Huber, 1981, Section 8.4). They noticed that their breakdown properties were still unsatisfactory for higher dimensions of the observation space: the breakdown point is less than one over the dimension (Maronna 1976, Huber 1981, Stahel 1981).

Siegel (1979) proposed a method, named "repeated medians", which yields an estimator with breakdown 1/2 in quite general paramtric models. Unfortunately, it is not directly applicable here since every element of the covariance matrix is estimated separately, and procedures of that kind need not to produce positive definite estimated matrices.

The aim of this paper is to introduce a new class of estimators with breakdown point 1/2 (section 4). Before doing so, I would like to discuss the notion of breakdown (section 3) in the context of a general class of models, which may be called the "models generated by transformations" (section 2).

2. MODELS AND INVARIANCE

The three best known statistical models, namely the location-scale, regression, and covariance-location models, may be subsumed under the following concept: Let

$$\mathbb{B} = \{B_x(\cdot;\theta) \mid \theta\varepsilon\Theta \subset \mathbb{R}^q\}$$

be a parametrized group of transformations $\mathbb{X}\to\mathbb{X}\subset\mathbb{R}^p$, and let P^o be any fixed distribution in $\mathbb{X}$. Then the distributions $\{P(\cdot;\theta) \mid \theta\varepsilon\Theta\}$, determined by

$$Z \sim P^o \Rightarrow B_x(Z ; \theta) \sim P(\cdot;\theta)$$

define a parametric model if $\theta \mapsto P(\cdot;\theta)$ is one-one (compare Fraser, 1968).

The example discussed in this paper is the covariance-location model: Let $\theta = [\mu , \Sigma]$, $\mu\varepsilon \mathbb{R}^p$, Σ a positive definite symmetric matrix of order p , and

$$B_x(x;\mu , \Sigma) = B\cdot x+\mu ,$$

where $B\cdot B^T = \Sigma$ is Cholesky's factorization of Σ (B is lower triangular with positive diagonal). P^o shall be any radial symmetric distribution, that is, the projection $X/|X|$ (for $|X|\neq 0$) of X onto the unit hypersphere shall be uniformly distributed, independently of the radius $|X|$, under P^o . If P^o is the (p-variate) standard normal, then $P(\cdot;\mu , \Sigma)$ is the normal distribution with location μ and variance-covariance matrix Σ . (We could use any other identified factorization $\Sigma = B \cdot B^T$ instead of Cholesky's.) On purpose, singular covariance matrices are excluded from the model by definition.

Let $\mathbb{A}$ be a group of transformations $A_x : \mathbb{X} \to \mathbb{X}$ and let A_p be the transformation of distributions induced by A_x . A model is said to be invariant under $\mathbb{A}$ if $A_x(X)$ has a model distribution whenever X has, that is, if

$$\forall A_x \epsilon \mathbb{A} \, , \ \forall \theta \epsilon \Theta \ \exists \ \bar{\theta} \epsilon \Theta \quad \text{with}$$

$$A_p(P(\cdot;\theta)) = P(\cdot;\bar{\theta}) \ .$$

If a model is generated by a group $\mathbb{B}$ of transformations, it is invariant under $\mathbb{B}$. The covariance-location model is invariant under the group $\mathbb{A}$ of all regular linear transformations ("affinely invariant") , and $\mathbb{A}$ is larger than the generating group $\mathbb{B}$.

An estimator may show a corresponding invariance property:

$$T(P) = \theta \Rightarrow T(A_p(P)) = \bar{\theta} \ .$$

Estimators which are invariant under the generating group $\mathbb{B}$ may be defined without loss of generality by the conditions under which P^o corresponds to the estimated value:

$$T(P) = \theta \Leftrightarrow T(B_p^{-1}(P;\theta)) = \theta^o$$

$(\theta^o$ determined by $B_x(\cdot;\theta^o) =$ identity) .

3. BREAKDOWN POINT

Definition. The "(gross error) breakdown point" of an estimator (functional) T at a distribution P is

$$\mathrm{Bge}(P,T) = \sup\{\varepsilon \leq 1 \mid \exists\, K(\varepsilon) \text{ kompakt}, \subsetneq \Theta \text{ with } (Q \in U(P;\varepsilon) \Rightarrow T(Q) \in K(\varepsilon))\} ,$$

where $U(P;\varepsilon)$ is the gross error 'neighbourhood' of P :

$$U(P;\varepsilon) = \{Q \mid Q=(1-\varepsilon)\cdot P+\varepsilon\cdot Q' , \quad Q' \text{ any distribution}\}$$

The simplest way of discussing the breakdown of an estimator consists in finding a sequence $[Q_n]$ in a 'neighbourhood' $U(P;\varepsilon)$ for which $T(Q_n)$ "diverges to the edge" of Θ , that is $T(Q_n) = \theta_n$ with

$$\forall K \subsetneq \Theta , \quad K \text{ kompakt}, \quad \exists\, n_K \text{ with } (n \geq n_K) \Rightarrow \theta_n \notin K) .$$

If such a sequence exists for a given ε , then $\mathrm{Bge}(P,T) \leq \varepsilon$. - Hampel (1971) coined the term "breakdown point" somewhat differently. His notion is conceptually more satisfactory but, on the other hand, more difficult to handle.

Huber (1981) and Stahel (1981) give the most general form of an affinely invariant M-estimator in the covariance-location model. They also show with the above reasoning that the breakdown point of such estimators is typically not greater than one over the dimension p of the observation space. (Maronna, 1976, showed

$\text{Bge} \leq (p+1)^{-1}$ for a special class of such estimators; despite these results there are, for special P, M-estimators with Bge approaching 1/2).

On the other hand it is easy to construct coordinate dependent estimators that achieve a higher breakdown point: eliminate first the observations showing up as univariate outliers in any of the p coordinates and calculate any covariance-location estimate from the rest. Instead of being a preferable procedure, such a rule shows a weakness of the previously defined concept of breakdown point in the covariance-location model:

Most frequently, this model is serving to describe dependence relations between variables or helps detecting so-called "multivariate" outliers. Therefore, if the majority of the points representing a sample happen to be near a hyperplane, it is highly desirable that this structure be detected. This idea leads to the following concept:

Definition. The "(gross error) breakdown point at the edge" for an estimator T and a given model is

$$\text{Bge}^*(T) = \sup\{\varepsilon \leq 1 \mid \text{ if } \theta_n \text{ diverges to the edge and } Q_n \varepsilon U(P(\cdot;\theta_n);\varepsilon)\ , \text{ then } T(Q_n) \text{ diverges to the edge}\}\ .$$

In the context of the covariance-location model we require the following: if $[Q_n]$ is a sequence of distributions, each of which is contained in an ε-'neighbourhood' of a model distribution

$P(\cdot;\mu_n,\Sigma_n)$, and if Σ_n tends to a singular matrix, then the estimated covariance matrix should also tend to a singular matrix.

For invariant estimators in a model generated by transformations, this new kind of breakdown coincides with the first one. For a precise statement, it is useful to note that the group structure of the transformations $B_x[\cdot;\theta]$ induces an operation in Θ : Let $\odot$ be defined by $\theta = \theta'\odot\theta'' \Leftrightarrow B_x[\cdot;\theta] = B_x[B_x[\cdot;\theta''];\theta']$.

θ^o shall correspond to the identity, and θ^- to the inverse $(\theta^-\odot\theta=\theta^o)$.

Proposition. In a model generated by a group of transformations let the operation $\odot$ have the following property: for all K , $K' \subsetneq \Theta$ compact, there is a $K'' \subsetneq \Theta$, compact such that $\{\theta\odot\theta' | \theta\in K, \theta'\in K'\} \subset K''$ and $\{\theta^- | \theta\in K\} \subset K''$.

Then, for any estimator invariant under the generating transformations, the breakdown point at the edge coincides with the ordinary breakdown point.

Proof. If $Q_n \in U(P(\cdot;\theta_n);\varepsilon)$, then $\bar{Q}_n = B_p(Q_n;\theta_n^-) \in U(P^0;\varepsilon)$ and

$$T(Q_n) = \theta_n \odot T(\bar{Q}_n) \quad .$$

If θ_n diverges whereas $T(Q_n)$ stays within a compact proper subset K of Θ , then $T(\bar{Q}_n)$ diverges, so that $Bge \leq Bge^*$. On the other hand, let $Q_n \in U(P^o;\varepsilon)$ such that $T(Q_n)$ diverges. Then

$\theta_n = T(Q_n)^{-}$ diverges whereas $T(B_p(Q_n;\theta_n)) = \theta^o$ does not. Therefore Bge* $\leq$ Bge . ///

In the covariance-location model, the operation is given by

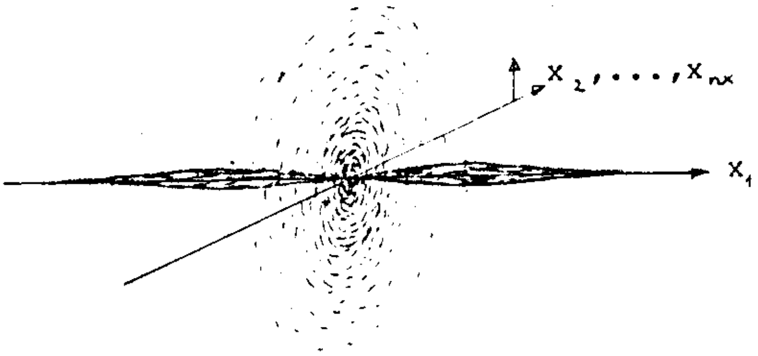
$$[\mu,\Sigma] \odot [\mu',\Sigma']$$

$$= [B\cdot\mu'+\mu,\ B\cdot\Sigma'\cdot B^T]\ ,\ (B\cdot B^T=\Sigma)\ ,$$

and the condition of the proposition is easily verified since Σ-matrices of "K-sets" have determinants bounded away from 0 and ∞ .

The proposition evokes an argument for looking at invariant estimators with a high (ordinary) breakdown point. Clearly, the coordinate dependent procedure mentioned above does not improve the breakdown point at the edge, but the question whether there are other non-invariant estimators with a high breakdown point at the edge remains open.

It is quite interesting to look at the situation leading to a breakdown of the affinely invariant M-estimators for $\varepsilon = 1/p$. Let P^* be the projection of P^o onto a hyperplane, and let $\tilde{P}$ be the symmetrized distribution of $|X|$, $X \sim P^*$. Mix a fraction $(1-1/p)$ of P^* on any hyperplane with a fraction $1/p$ of $\tilde{P}$ on a perpendicular to it, just like the figure shows



If Q_n approaches this distribution (while $P(\cdot;\theta_n)$ goes to P^*), the estimated covariance matrix will usually approach a multiple of the identity matrix instead of a singular one.

4. A COVARIANCE-LOCATION ESTIMATOR WITH BREAKDOWN POINT 1/2

While univariate outliers in one of the coordinates were eliminated in the procedure mentioned above, we now exclude all observations sticking out in any projection:

Definition. For each direction $d \in \mathbb{R}^p$, $|d| = 1$, let L_d and S_d be a (one-dimensional) location and scale estimator (functional) of the distribution of the projection $d^T \cdot X$, $X \sim P$, respectively (with the corresponding invariance properties), and

$$R(x,P) = \sup_d \{ |d^T \cdot x - L_d| / S_d \} .$$

Then the "projection estimator" corresponding to L, S and a weight function $W: \mathbb{R}^+ \to \mathbb{R}$ is defined as the ordinary weighted covariance-location estimator with weights $W(R(X,P)^2)$:

$$\hat{\mu}(P) = \int W(R(x,P)^2) \cdot x \; P(dx) \; / \int W(R(x,P)^2) \; P(dx)$$

$$\hat{\Sigma}(P) = c \cdot \int W(R(x,P)^2) \cdot (x-\hat{\mu}) \cdot (x-\hat{\mu})^T \; P(dx) \; / \int W(R(x,P)^2) \; P(dx) ,$$

where c is a fixed constant used to achieve Fisher-consistency at the normal distribution.

The invariance of such estimators follows from the definition of $R(x,P)$.

Proposition. Suppose that

(i) L and S have breakdown point $1/2$ at the projection of P^o on a straight line;

(ii) W is poitive and bounded, $0 \leq W(r^2) \leq w^*$, and there are $r_1 > 0$, r_2 and $a > 0$ such that
$W(r^2) \geq a$ for $r \leq r_1$ and $W(r^2) = 0$ for $r \geq r_2$;

(iii) P^o has a positive density.

Then, for $\varepsilon < 1/2$, $\hat{\mu}(Q)$ and $\hat{\Sigma}(Q)$ are bounded for $Q \varepsilon U(P;\varepsilon)$, P a model distribution. If μ is known and L_d is set to 0 , then the breakdown point for the covariance part $\hat{\Sigma}$ is $1/2$.

For the combined estimator, I could not prove that $\hat{\Sigma}$ may not approach a singular matrix; therefore I cannot state that the breakdown is $1/2$.

Proof. It suffices to consider P^o for the model distribution. Let $\varepsilon < 1/2$. Because of (i), there is a b_2 such that $R(x) \geq r_2$ if $|x| > b_2$, regardless of the contamination. $W(R(x,Q)^2)$ may not vanish almost surely (P^o) because of (ii) and (iii), whence the estimates are well defined and bounded. - It remains to show that the smallest eigenvalue of $\hat{\Sigma}$ is bounded away from 0 for $\mu=0$. But since S_d has this property, there is a b_1 with $W(R(x,Q)^2) > a$ for $|x| < b_1$ and all $Q \varepsilon U(P^o,\varepsilon)$, and by (ii),

$$\begin{aligned} d^T \cdot \hat{\Sigma} \cdot d &= \int (d^T \cdot x)^2 \cdot W(R(x,Q)^2)\ Q(dx) \ / \int W(R(x,Q)^2)\ Q(dx) \\ &\geq \int_{|x| \leq b_1} (d^T \cdot x)^2\ W(R(x,Q)^2)\ Q(dx)\ /\ w^* \\ &\geq (1-\varepsilon) \cdot (a/c) \cdot \int_{|x| \leq b_1} (d^T \cdot x)^2\ P^o(dx)\ .\ /// \end{aligned}$$

The maximization involved in the definition of $R(x,P)$ poses a serious computational problem. Heuristic considerations show that there may be many local maxima even if the global maximum is much lower than these. (This is especially true if L and S are the median and the median deviation.) Since, therefore, ordinary nonlinear programming cannot solve the problem, one could try evaluating the function to be maximized at all points of a fine "grid" on the hypersphere. To my knowledge, such a grid has not been given yet; in higher dimensions, it would need very many grid points in order to be accurately fine. (A similar problem arises in connection with the "function plot" of Andrews (1972): The proposed curve on the hypersphere does not pass sufficiently near to all points on that sphere.)

A practicable solution for finite sample sizes n seems nevertheless possible by adjusting the general idea of Siegel (1979):

Choose at random p indices $[i_k]$ from $\{1,2,\ldots,n\}$ and find d perpendicular to the hyperplane through the observations with indices $[i_k]$. Repeat this construction m times and treat the maximum over the m directions as the global maximum. (A local maximization procedure could be used to improve this preliminary solution). In the situation mentioned above, where the "good" observations lie near a hyperplane, the procedure will detect the structure if the indices $[i_k]$ select "good" observations exclusively for at least one of the m choices. The probability for this is

$$p^* \approx 1-(1-(1-\varepsilon)^p)^m$$

where ε is the proportion of contamination in the sample, if n/p is large.

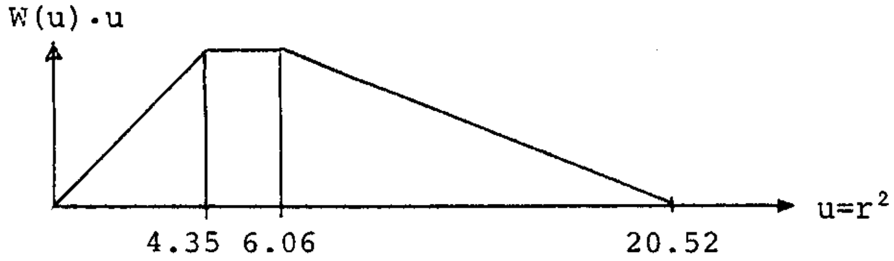
The number m of choices necessary for $p^* = 0.95$ is given below:

ε p	.Ø5	Ø.1	Ø.3	Ø.5
2	2	2	5	11
3	2	3	8	23
5	3	4	17	95
1Ø	4	7	1Ø5	3Ø67
2Ø	7	24	3753	3141252

In order to get a first impression of the feasability of the method, I run a "mini-simulation" with $p=5$ and 50 replicates. The distribution was constructed, like the figure at the end of §3 shows, as a flat disk with a stick through it:

$$x^{(i)} \sim \mathcal{N}_1(0,0.05^2) \times \mathcal{N}_4(0,I) \quad , \quad i=1,2,\ldots,35 \quad ;$$
$$\sqrt{\chi_4^2} \times \delta(0,0,0,0) \quad , \quad i=36,\ldots,50 \quad .$$

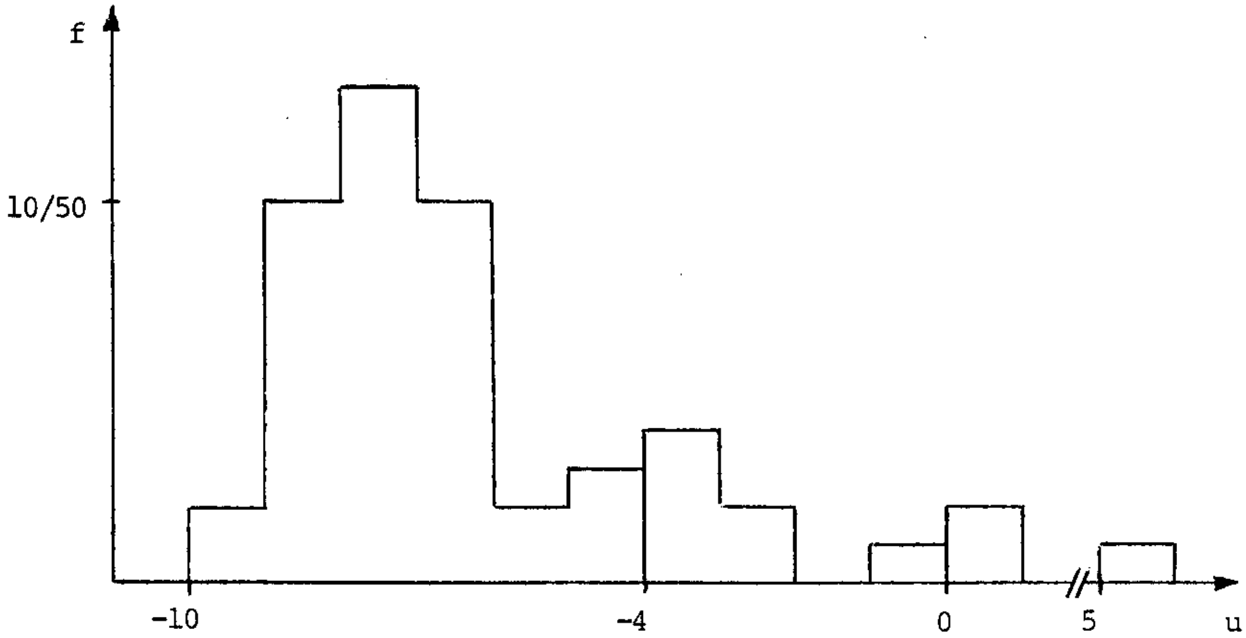
The projection estimator considered used the median and median deviation/0.6745 as L and S, and a W-function as specified by the following graph of $W(u)\cdot u$ against $u=r^2$.



(compare to χ_4^2 percentage points). No consistency constant was introduced (c=1) . The number of trials m was 17 according to the foregoing table. In one case of the 50 , none of these trials happened to select only "good" observations. The following histogram for

$u = \log_2(\Sigma_{11} / \mathrm{ave}(\Sigma_{22},\ldots,\Sigma_{55}))$

shows that this fraction was below $(1/4)^2$ - which might be seen as a success - in 40 of the 50 replicates.



Remark. In the context of regression, $Y=\beta^T \cdot X+\mathrm{Error}$, Hampel (1975, p. 380) suggests to define a very robust estimator as the β minimizing the median deviation of the residuals, which is equivalent to finding

$$\min_b \{ \operatorname{med}(|Y-b^TX-L_b|) \}$$

where L_b is the median of $Y-b^TX$. This is a problem of similar complexity as the one above (although it has to be solved only once instead of n times), and the same procedure should lead to a feasible solution.

REFERENCES

Andrews D.F. (1972) Plots of high-dimensional data; Biometrics 28, 125-36.

Fraser D.A.S. (1968) The structure of inference; Wiley, New York.

Hampel F.R. (1971) A general qualitative definition of robustness; Ann.Math.Statist. 42, 1887-96.

Hampel F.R. (1975) Beyond location parameters: Robust concepts and methods; Bull.Int.Stat.Inst. 46, 375-82.

Huber P.J. (1977) Robust covariances; in: Gupta S.S., Moore D.S. eds.: Statistical decision theory and related topics 2, 165-9.

Huber P.J. (1981) Robust statistics; Wiley, New York.

Maronna R.A. (1976) Robust M-estimators of location and scatter; Ann.Statist. 4, 51-67.

Siegel A.F. (1979) The repeated median regression algorithm; manuscript.

Stahel W. (1981) Robust estimation: Infinitesimal optimality and covariance matrix estimators (in german); Ph.D. thesis, ETH, Zuerich.