L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Robustness: Where are we now? [1]

Peter J. Huber

University of Bayreuth, Germany

Abstract: The development of selected robustness concepts since their inception in the 1960's is sketched and their current status is reviewed.

Key words: Linear models, estimation, design, regression.

AMS subject classification: 62F35, 62F12, 62A99.

He looked into the water and saw that it was made up of a thousand thousand thousand and one different currents, each one a different colour, weaving in and out of one another like a liquid tapestry of breathtaking complexity; and Iff explained that these were the Streams of Story, that each coloured strand represented and contained a single tale.
(Salman Rushdie, *Haroun and the Sea of Stories*, 1990)

1 The first ten years

A good case can be made that modern robustness begins in 1960, with the papers by J.W. Tukey on sampling from contaminated distributions, and by F.J. Anscombe on the rejection of outliers. Tukey's paper drew attention to the dramatic effects of seemingly negligible deviations from the model, and it made effective use of asymptotics in combination with the gross error model. Anscombe introduced a seminal insurance idea: sacrifice some performance at the model in order to insure against ill effects caused by deviations from it. Most of the basic ideas, concepts and methods of robustness were invented in quick succession during the following years and

[1]First published in *Student* (1995), Vol.1, No.2, 75-86. Reproduced by permission of the Presses Académiques Neuchâtel, Switzerland.

were in place by about 1970.

In 1964 I recast Tukey's general setup into an asymptotic minimax framework and was able to solve it. Important points were: the insistence on finite, but small deviations, the formal recognition that a large error occurring with low probability is a small deviation, and the switch from the then prevalent criterion of relative efficiency to absolute performance. At the same time, I introduced the notion of maximum likelihood type or M-estimators.

Hampel (1968) added the formal definition of qualitative robustness (continuity under a suitable weak topology), infinitesimal robustness in form of the influence function (von Mises derivative of a statistical functional) and the notion of breakdown point.

A nagging early worry had been the possible irrelevancy of asymptotic approaches: conceptually, 1% gross errors in samples of 1000 are entirely different from the same error rate in samples of 5. This worry was laid to rest by Huber (1965, 1968): both for tests and for estimates, there are qualitatively identical and even quantitatively similar exact finite sample minimax robustness results.

The end of the decade saw the first steps of an extension of asymptotic robustness theory beyond location to more general parametric models, namely the introduction of shrinking neighborhoods by C. Huber-Carol (1970), as well as the first attempt at studentizing (Huber, 1970).

The basic methodology for Monte Carlo studies of robust estimators was established in Princeton 1970/71, see Andrews et al. (1972). That study more or less finished off the problem of robustly estimating a single location parameter in samples of size 10 or larger, opening the way for more general multi-parameter problems.

In this paper I shall pick a few of the more interesting robustness ideas and follow the strands of their stories to the present. What has happened to the early strands since their invention? What important new strands have begun in the 1970's and 1980's? I shall avoid technicalities and instead give a reference to a recent survey or monograph, where feasible. Completeness is not intended; for a recent "far from complete" survey of research directions in robust statistics, with more than 500 references, see Stahel's article in Stahel and Weisberg (1991, Part II, p. 243).

2 Influence functions and pseudo-values

Among the basic robustness concepts, influence functions have become a standard tool, especially after the comprehensive treatment by Hampel et al. (1986). Also the trick to robustize classical procedures through the

use of pseudo-values is becoming common knowledge, even though it has received only scant coverage in the literature. This trick is that one calculates robust fitted values $\widehat{y}_i$ by iteratively applying the classical procedure to the pseudo-values $y_i^* = \widehat{y}_i + r_i^*$ instead of y_i. Here, the pseudo-residual $r_i^* = \psi(r_i)$ is obtained by cutting down the current residual $r_i = y_i - \widehat{y}_i$ with the help of a function ψ proportional to the desired influence function (i.e. with the ψ-function defining an M-estimate). For examples see in particular Bickel (1976, p. 167), Huber (1979), and Kleiner, Martin and Thomson (1979). If ψ is chosen equal rather than merely proportional to the influence function, the classical formulas, when applied to the pseudo-residuals r_i^* instead of the residuals, yield asymptotically correct error estimates for ANOVA and other purposes (Huber 1981, p. 197).

There have been some very interesting extensions of influence function ideas to time series (Künsch, 1984).

3 Breakdown and outlier detection

For a long time, the breakdown point had been a step-child of the robustness literature. The paper by Donoho and Huber (1983) was specifically written to give it more visibility. Recently, I have begun to wonder whether it has given it too much, the suddenly fashionable emphasis on high breakdown point procedures has become counter-productive. One of the most striking examples of the usefulness of the concept can be found in Hampel (1985): the combined performance of outlier rejection followed by the sample mean as an estimate of location is essentially determined by the breakdown of the outlier detection procedure.

4 Studentizing

Whenever we have an estimate, we ought to provide an indication of its statistical accuracy, say by giving a 95% confidence interval. This is not particularly difficult if the number of observations is very large, so that the estimate is asymptotically normal with an accurately estimable standard error, or also in one-parameter problems without nuisance parameter, where the finite sample theory of Huber (1968) can be applied.

Otherwise, we end up with a tricky problem of studentization. To my knowledge, there has not been much progress beyond the admittedly unsatisfactory initial paper of Huber (1970). There are not only many open questions with regard to this crucially important problem, it is even open what questions one should ask! A sketch of the principal issues follows.

In the classical normal case, it follows from sufficiency of $(\overline{x}, s)$ and an

invariance argument that such a confidence interval must take the form

$$(\overline{x} - k_n s/\sqrt{n},\ \overline{x} + k_n s/\sqrt{n})$$

with k_n depending on the sample size, but not on the sample itself. In a well-behaved robust version, a confidence interval might take the analogous form

$$(T - K_n S/\sqrt{n},\ T + K_n S/\sqrt{n})$$

where T is an asymptotically normal robust location estimate and S is the location invariant, scale equivariant, Fisher consistent functional estimating the asymptotic standard deviation of $\sqrt{n}T$, applied to the empirical distribution. In the case of M-estimates for example, we might use

$$S(F)^2 = \frac{S_0^2 \int \psi^2 dF}{(\int \psi' dF)^2}$$

where the argument of ψ, ψ' is $y = (x - T)/S_0$, i.e. a robustly centered and scaled version of the observations, say with $S_0 = MAD$. If we are interested in 95% confidence intervals, K_n must approach $\Phi^{-1}(0.975) \approx 1.96$ for large n. But K_n might depend on the sample configuration in a non-trivial, translation and scale invariant way: since we do not have a sufficient statistic, we might want to condition on the configuration of the sample in an as yet undetermined way.

While the distribution of $\sqrt{n}T$ typically approaches the normal, it will do so much faster in the central region than in the tails, and the extreme tails will depend rather uncontrollably on details of the unknown distribution of the observations. The distribution of S suffers from similar problems, but here it is the low end which matters. The question is: what confidence levels make sense and are reasonably stable for what sample sizes? For example, given a particular level of contamination and a particular estimate, is $n = 10$ good enough to derive accurate and robust 99% confidence intervals, or do we have to be content with 95% or 90%? I anticipate that such questions can (and will) be settled with the help of small sample asymptotics, assisted perhaps by configural polysampling (below).

5 Shrinking neighborhoods

Direct theoretical attacks on finite neighborhoods work only if the problem is location or scale invariant. But for large samples, most point estimation problems begin to resemble location problems, so it is possible to derive quite general asymptotically robust tests and estimates by letting those neighborhoods shrink at a rate $n^{-1/2}$ with increasing sample size. This idea

was first exploited by C. Huber-Carol (1970), followed by Rieder, Beran, Millar and others. The final word on this approach is contained in Rieder's book (1994).

The principal motivation clearly is technical: shrinking leads to a manageable asymptotic theory. But there is also a philosophical justification: since the standard goodness-of-fit tests are just able to detect deviations of the order $O(n^{-1/2})$, it makes sense to put the border zone between small and large deviations at $O(n^{-1/2})$. Larger deviations should be taken care of by diagnostics and modelling, smaller ones are difficult to detect and should be covered (in the insurance sense!) by robustness.

This does not mean that our data samples are supposed to get cleaner if they get larger. But the shrinkage of neighborhood faces us with a dilemma, namely a choice between the alternatives:

- improve the model; or
- improve the data; or
- stop sampling.

Note that adaptive estimation is *not* among the viable alternatives. The problem is not one of reducing statistical variability, but one of avoiding bias, and the ancient Emperor-of-China paradox applies (you can get a fantastically accurate estimate of the height of the emperor by averaging the guesses of 600 million Chinese, most of whom never saw the emperor!).

The asymptotic theory of shrinking neighborhoods is, in essence, a theory of infinitesimal robustness and suffers from the same conceptual drawback as approaches based on the influence function: infinitesimal robustness (bounded influence) does not automatically imply robustness. The crucial point is that in any practical application we have a fixed, finite sample size, and we need to know whether we are inside the range of n and ε for which asymptotic theory yields a decent approximation. This range may be difficult to determine, but the breakdown point often is computable and may be a useful indicator.

6 Design

Robustness casts a shadow on the theory of optimal designs: they lose their theoretical optimality very quickly under minor violations of linearity (Huber, 1975) or independence assumptions (Bickel and Herzberg, 1979). I am not aware of much current activity in this area, but the lesson is clear: "Naive" designs usually are more robust and better than "optimal" designs.

7 Regression

Back in 1975, the discussants of Bickel (1976) raised interesting criticisms, in particular there were complaints about the multiplicity of robust procedures, and about their computational and conceptual complexity. Bickel fended them off skillfully and convincingly.

There may have been reasons for concern then, but the situation has become worse. Most of the action in the 1980's has been on the regression front. Here is an incomplete list of robust regression estimators: L1 (going back at least to Laplace), M (Huber, 1973), GM (Mallows, 1975), with variants by Hampel, Krasker and Welsch, RM (Siegel, 1982), LMS and LTS (Rousseeuw, 1984), S (Rousseeuw and Yohai, 1984), MM (Yohai, 1987), τ (Yohai and Zamar, 1988), SRC (Simpson, Ruppert and Carroll, 1991), and no end is in sight. For an up-to-date review see Davies (1993).

Bickel would not have an easy job now, much of the "Nordic" criticism, unsubstantiated in 1975, seems to be justified now. In any engineering product, an overly rapid sequence of updates sometimes is a sign of vigorous progress, but it can also be a sign of shoddy workmanship, and often it is both. In any case, it confuses the customers and hence is counterproductive...

Robustness has been defined as insensitivity to small deviations from an idealized model. What is this model in the regression case? The classical model goes back to Gauss and assumes that the carrier X (the matrix X of the "independent" variables) is error-free. X may be systematic (as in designed experiments), or haphazard (as in most undesigned experiments), but its rows only rarely can be modelled as being a random sample from a specified multivariate model distribution. Statisticians tend to forget that the elements of X often are not observed quantities, but are derived from some model (cf. the classical non-linear problems of astronomy and geodesy giving rise to the method of least squares in the first place). In essence, each individual X corresponds to a somewhat different situation and might have to be dealt with differently. Thus, multiplicity of procedures may lie in the nature of robust regression. Curiously, most of the action seems to have been focused through tunnel vision on just one aspect: safeguard at any cost against problems caused by gross errors in a random carrier.

Over the years, I too had to defend the minimax approach to distributional robustness on many occasions. The salient points of my defense were that the least favorable situation one is safeguarding against, far from being unrealistically pessimistic, is more similar to actually observed error distributions than the normal model, that the performance loss at a true normal model is relatively small, that on the other hand the classically optimal

procedures may lose sorely if the normal model is just slightly violated, and that the hardest problems are not with extreme outliers (which are easy to detect and eliminate), but with what happens on the shoulders of the distributions. Moreover, the computation of robust M-estimates is easy and fast (the last paragraph of this section). Not a single one of these defense lines can be used with the modern "high breakdown point" regression estimates.

A typical cause for breakdown in regression are gross outliers in X; while individual such outliers are trivially easy to spot (with the help of the diagonal of the hat matrix), efficient identification of collaborative leverage groups is an open, perhaps unsolvable, diagnostic problem. However, I would advise against treating leverage groups blindly through robustness, they may hide serious design or modeling problems, and there are similar problems even with single leverage points.

The story behind an outlier among the X ("leverage point") might for example be:

- a misplaced decimal point,
- an accurate but useless observation, outside of the range of validity of the model.

If the value at this leverage point disagrees with the evidence extrapolated from the other observations, this may be because:

- the outlying observation is affected by a gross error (in X or in y),
- the other observations are affected by small systematic errors (this is more often the case than one might think),
- the model is inaccurate, so the extrapolation fails.

The existence of several, phenomenologically indistinguishable but conceptually different situations with different consequences calls for a diagnostic approach (identification of leverage points or groups), followed by alternative "what if" analyses. This contrasts sharply with simple location estimation, where the observations are exchangeable and a minimax approach is quite adequate (although one may want to follow it up with an investigation of the causes of grosser errors).

At the root of the current confusion is that hardly anybody bothers about stating all of the issues clearly: not only the estimator and a procedure for computing it must be specified, but also the situations for which it is supposed to be appropriate or inappropriate, and criteria for judging estimators and procedures. There has been a tendency to rush into print with rash claims and procedures. In particular, what is meant by the word breakdown? For many of the newer estimates there are unqual-

ified claims that their breakdown point approaches 0.5 in large samples. But such claims tacitly exclude designed situations: if the observations are equally partitioned among the corners of a simplex in d-space, no estimate whatsoever can achieve a breakdown point above $1/(2d+2)$.

It is one thing to design a theoretical algorithm whose purpose is to prove, for example, that a breakdown point $1/2$ can be attained in principle, and quite another thing to design a practical version that can be used not merely on small, but also on medium sized regression problems, with a 2000-by-50 matrix or so. This last requirement would seem to exclude all of the recently proposed robust regression estimates.

Some comments on the much maligned "plain vanilla" regression M-estimates are in order. The M-estimate approach is not a panacea (is there such a thing in statistics?), but it is easy to understand, practicable, and considerably safer than classical least squares. While I had had regression problems in mind from the very beginning (cf. Huber 1964, p.74), I had not dared to go into print until 1973, when I believed to understand the asymptotic behavior of M-estimates of regression. A necessary regularity condition for consistency and asymptotic normality of the fitted values is that the maximum diagonal element h of the hat matrix $H = X(X^TX)^{-1}X^T$ tends to 0. While watching out for large values of h does not enforce a high breakdown point, it at least may prevent a typical cause of breakdown. Moreover, with M-estimates of regression, the computing effort for large matrices typically is less than twice what is needed for calculating the ordinary least squares solution. Both calculations are dominated by the computation of a QR or SVD decomposition of the X matrix, which takes $O(np^2)$ operations for an (n,p)-matrix. Since the result of that decomposition can be re-used, the iterative computation of the M-estimate, using pseudo-values, takes $O(np)$ per iteration with fewer than 10 iterations on average.

8 Multivariate problems

Classically, "regression" problems with errors in the independent variables are solved by fitting a least squares hyperplane, that is, by solving a principal component problem and picking the plane belonging to the smallest eigenvalue. It can be argued by analogy that regression problems with potential gross errors in the carrier should be attacked through some version of robust principal components. Thus, multivariate problems, in particular of the principal component type, deserve added attention.

In 1976, Maronna showed that all M-estimates of location and scatter in d dimensions have a breakdown point $\varepsilon^* \leq 1/(d+1)$. In higher dimensions

this is shockingly low, but then Stahel (1981) and Donoho (1982) independently found estimates based on projection pursuit ideas with a breakdown point approaching 1/2 in large samples. The bad news is that with all currently known algorithms the effort for computing those estimates increases exponentially with d. We might say they break down by failing to give a timely answer!

9 Some persistent misunderstandings

Robust methods have persistently been misclassified and pooled with nonparametric and distribution-free methods. They are part of traditional parametric statistics, the only difference is that they do not assume that the parametric model is exactly true.

In part the robustness community itself is to blame for this misunderstanding. In the mid-1970's adaptive estimates - attempting to achieve asymptotic efficiency at all well-behaved error distributions - were thought by many to be the ultimate robust estimates. Then Klaassen (1980) proved a disturbing result on the lack of stability of adaptive estimates. My current conjecture is that an estimate cannot be simultaneously adaptive in a neighborhood of the model and qualitatively robust at the model.

Also the currently fashionable (over-)emphasis of high breakdown points transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but the potential presence of high contamination usually indicates a mixture model and calls for diagnostics. A thoughtless application of robust procedures might only hide the underlying problem.

There seems to be some confusion between the respective roles of diagnostics and robustness. The purpose of robustness is to safeguard against deviations from the assumptions, in particular against those that are near or below the limits of detectability. The purpose of diagnostics is to find and identify deviations from the assumptions.

The term "robust" was coined by Bayesians (Box and Andersen, 1955). It is puzzling that Bayesian statistics never managed to assimilate the modern robustness concept, but remained stuck with antiquated parametric supermodels. *Ad hoc* supermodels and priors chosen by intuition (personal beliefs) or convenience (conjugate priors) do not guarantee robustness, for that one needs some theory. Compare already Hampel (1973).

10 Future directions: small sample problems?

At present, the most interesting and at the same time most promising new methods and open problems have to do with small sample situations.

The Princeton robustness study (Andrews et al., 1972) remains one of the best designed comparative Monte Carlo studies in statistics. Among other things we learned from it how difficult it is to compare the behavior of estimates across entire families of distributions, since small systematic differences of performance easily are swamped by larger random sampling errors. A very sophisticated sampling method thereafter proposed by Tukey is based on the remark that a given sample can occur under any strictly positive density, but it will do so with different probabilities. Thus it must be possible to compare those performances very efficiently by re-using the same sample configurations with different weights when forming Monte Carlo averages. A comprehensive account of this approach is given by Morgenthaler and Tukey (1991).

On the other hand, it was noted by Hampel that a variant of the saddle point method can give fantastically accurate asymptotic approximations down to very small sample sizes (occasionally down to $n = 1$!).

I hope that these approaches in the near future will permit to close several open problems in the area of small samples: studentization, confidence intervals, testing. Embarrassingly, the robustification of the statistics of two-way tables still is wide open. Typically there are so few degrees of freedom per cell that ordinary asymptotic approaches are out of the question, maybe some version of small sample asymptotics may help here too.

References

[1] 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: Princeton Univ. Press.

[2] Anscombe, F.J. (1960). Rejection of outliers. *Technometrics* **2**, 123-147.

[3] Bickel, P.J. (1976). Another look at robustness: A review of reviews and some new developments. *Scand. J. Statist.* **3**, 145-168.

[4] Bickel, P.J., and Herzberg, A.M. (1979). Robustness of design against autocorrelation in time. *Ann. Statist.* **7**, 77-95.

[5] Box, G.E.P., and Andersen, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departure from assumption. *J. R. Statist. Soc.* B **17**, 1-34.

[6] Davies, P.L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21**, 1843-1899.

[7] Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Harvard University, Cambridge, MA.

[8] Donoho, D.L., and Huber, P.J. (1973). The notion of breakdown point.

In *A Festschrift for Erich L. Lehmann*, Eds. P. J. Bickel, K.A. Doksum and J.L. Hodges. Wadsworth, Belmont, CA.

[9] Hampel, F.R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Univ. of Calif., Berkeley.

[10] Hampel, F.R. (1973). Robust estimation: A condensed partial survey. *Z. Wahr. verw. Geb.* **27**, 87-104.

[11] Hampel, F.R. (1985). The breakdown point of the mean combined with some rejection rules. *Technometrics* **27**, 95-107.

[12] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence.* New York: Wiley.

[13] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

[14] Huber, P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36**, 1753-1758.

[15] Huber, P.J. (1968). Robust confidence limits. *Z. Wahr. verw. Geb.* **10**, 269-278.

[16] Huber, P.J. (1970). Studentizing robust estimates. In *Nonparametric Techniques in Statistical Inference*, Ed. M. L. Puri. Cambridge: Cambridge University Press.

[17] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799-821.

[18] Huber, P.J. (1975). Robustness and designs. In *A Survey of Statistical Design and Linear Models*, Ed. J. N. Srivastava. Amsterdam: North Holland.

[19] Huber, P.J. (1979). Robust smoothing. In *Proc. ARO Workshop on Robustness in Statistics*, Eds. R. N. Launer and G. N. Wilkinson. New York: Academic Press.

[20] Huber, P.J. (1981). *Robust Statistics.* New York: Wiley.

[21] Huber-Carol, C. (1970). Etude asymptotique de tests robustes. Ph.D. thesis, ETH, Zürich.

[22] Klaassen, C. (1980). Statistical performance of location estimators. Ph.D. thesis, Mathematisch Centrum, Amsterdam.

[23] Kleiner, B., Martin, R.D., and Thomson, D.J. (1979). Robust estimation of power spectra. *J. R. Statist. Soc.* B **41**, 313-351.

[24] Künsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843-863.

[25] Mallows, C.L. (1975). On some topics in robustness. Tech. Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

[26] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51-67.

[27] Morgenthaler, S. and Tukey, J.W. (1991). *Configural Polysampling.* New York: Wiley.

[28] Rieder, H. (1994). *Robust Asymptotic Statistics.* Berlin: Springer.

[29] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79**, 871-880.

[30] Rousseeuw, P.J., and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, Eds. J. Franke, W.Härdle and R.D. Martin. Lecture Notes in Statistics 26. New York: Springer.

[31] Siegel, A.F. (1982). Robust regression using repeated medians. *Biometrika* **69**, 242-244.

[32] Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Am. Statist. Assoc.* **87**, 439-450.

[33] Stahel, W.A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH Zürich.

[34] Stahel, W., and Weisberg, S. (Eds.) (1991). *Directions in Robust Statistics and Diagnostics* Part I and II. The IMA Volumes in Mathematics and its Applications, Vols. 33-34. New York: Springer.

[35] Student (1927). Errors of routine analysis. *Biometrika* **19**, 151-164.

[36] Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, Ed. I. Olkin. Stanford: Stanford Univ. Press.

[37] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15**, 642-656.

[38] Yohai, V.J., and Zamar, R.H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *J. Am. Statist. Assoc.* **83**, 406-413.