Some thoughts about classification

Some thoughts about classification

by

Frank Hampel
E-Mail: hampel@stat.math.ethz.ch

Research Report No. 102
January 2002

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

Paper to be presented at the IFCS 2002 Cracow, Poland 16-19 July 2002

Some thoughts about classification

Frank Hampel

Seminar for Statistics, ETH Zurich, Switzerland
E-Mail: hampel@stat.math.ethz.ch

Abstract

The paper contains some general remarks on the high art of data analysis, some philosophical thoughts about classification, a partial review of outliers and robustness from the point of view of applications, including a discussion of the problem of model choice, and a review of several aspects of robust estimation of covariance matrices, including the pragmatic choice of a weight function based on empirical and theoretical evidence. Several sections contain new (or at least original) ideas: There are some proposals for incorporating robustness into Bayesian practice and theory, including weighted log likelihoods and Bayes' theorem for weighted data. Some small ideas refer to artificial classification in a continuum, to a "robust" (Prohorov-type) metric for high-dimensional data, and to the use of multiple minimum spanning trees. A promising but difficult research idea for clustering on the real line, based on a new smoothing method, concludes the paper.

1 Introduction

When I was asked to give a keynote lecture at this grand conference, I first hesitated and thought it might be some kind of misunderstanding. After all, I had never properly done research in classification or clustering, my research in robustness theory (which might and should be a general background tool at this conference) lies decades back, I did not follow up the literature, and my (intensive) practical work in data analysis is long past. On the other hand, my present work on the foundations of statistics can clarify many concepts and controversies and can reconcile frequentists and Bayesians on a higher level, but apart from the concepts (and a few simple examples) it is not yet applicable.

But then I thought that many researchers, both in theory and applications, are so engulfed in what they are presently doing, and busy reading the most modern literature of their respective narrow fields, that it might occasionally be a good idea for them to hear a talk "from the outside", which gives them a bird eye's view of statistics, which reminds them of the roots of some of the things they are doing (in the hope of often still badly needed clarification), and which exposes them to a number of fresh ideas, some of which may become a useful tool or may lead to fruitful further research.

In this spirit, I am starting with some general philosophical thoughts about classification, and then I recall some basic ideas, tools and results in robustness theory, which are still often unknown, distortedly known, misunderstood or ignored, but which should be in the tool box of every practical statistician, including every pragmatic Bayesian. An outline of how the results of robustness theory might be incorporated into a more canonical Bayesian theory (which may mean an extension of canonical Bayesian theory) follows, including the use of weighted likelihoods and Bayes' theorem for weighted data.

I apologize if some of my proposals, here and later in the paper, are already (unknown to me) somewhere in the literature; if they are not new, they are at least original, and I think it is better an idea is published twice than it is not published at all.

After mentioning the future potential of the use of upper and lower (instead of only ordinary) probabilities in statistics, I return to details of robust covariance matrices and associated weight functions. Several small subsections on more or less tentative ideas follow. The section on subdividing a continuum may be mathematically simple, but logically interesting (leading to a nontransitive definition of "potentially equal"). The robust metric introduced is of potential use anywhere in multivariate statistics where a metric occurs; its practical usefulness has to be tried out. The idea of superimposing several minimum spanning trees for interactive exploratory data analysis has its roots in some practical experiences, but again it has to be tried out in order to learn about its practical value. The last proposal is more an extensive research program: to use the existing beginnings of a "truly smooth" (neither biased nor locally wiggly), but very complicated smoothing method (an improved counterpart of splines, in a way) for developing tests for clusters, modes, and mixture components for data on the real line.

Before we go into the details of the next sections, we should remember the high standards of good data analysis (which are missing in mathematical theories). For example, in regression (and elsewhere), we should be "fitting equations to data" (Daniel and Wood 1980), and not "fitting data to equations" (or preconceived models), as is frequently done. Daniel (cf. also the highly recommendable book by Daniel 1976 on ANOVA) shows that in order to be an excellent data analyst, it suffices to know very little mathematics, but to have a lot of common sense, including a deep intuitive understanding of simple mathematical formulae, a down-to-earth interpretation of the data, and a clear and thorough understanding of the practical problem to be solved.

This does not mean that no high-power mathematics is needed in statistics; for example, the very basic problem of the violation of the independence assumption of supposedly i.i.d. data (even measurement data in the hard sciences), of which first-class statisticians such as Newcomb, K. Pearson, "Student" and Jeffreys were clearly aware, could only be modeled and attacked successfully after Kolmogorov invented abstractly, and Mandelbrot introduced into parts of statistics, the mathematically highly demanding and "pathological" increment processes of self-similar processes; but the consequences for practical statistics (cf. Hampel et al. 1986, Ch. 8.1; Kuensch et al. 1993) are in full agreement with what C. Daniel and undoubtedly other first-rate practitioners knew qualitatively already from experience. (This topic of lack of independence is not discussed here further.)

Unfortunately, what is very often missing (even in theoretical discussions), is the interpretation of mathematical results in the light of the problems of practical data analysis (see Hampel 1998a). Often there is a deep gap, and a lot of misuse of the results of mathematical statistics. What should also be more emphasized is the iterative going back and forth between the statistical analysis and the background knowledge from the subject matter science; it is often nonsense to require that a purely statistical analysis be complete in itself.

An early source for openminded and thorough data analysis (not blinded by any mathematical theory), which should be compulsory reading for every data analyst, is "Student" (1927). There are also many excellent data analyses in Jeffreys (1939). For more recent examples of creative and innovative data analyses (besides C. Daniel), see the writings by J.W. Tukey, F. Mosteller, G.E.P. Box and others. There are also some condensed but deep data analyses in Hampel (1987), including germs for new methods in design of experiment. An interesting example of modern thinking about data analysis can be found

in Davies (1995).

2 Some philosophical thoughts

Classification, the separation and naming of appearances, is one of the most basic cultural activities of humanity; it is a fundament for our science and civilization. Thus, astronomy started with naming the stars and constellations and separating out the planets (in the old sense). Modern geology began with the statistical separation (by the number of surviving fossils) and naming (from Eocene to Pleistocene) of consecutive geological layers by Lyell. Modern biology began with Linné's system of nomenclature, which basically is still used today.

Names of human persons, gods (or God), spirits, and animals often had a mythical importance and power attached to them (we still find remnants of this). Nowadays, names are more and more replaced (at least in part) by numbers (such as passport number, social security number), a trend enhanced by the computer. But even the simple step from "one" to "two" needs the separation of the world into (at least) two entities.

In a deep philosophical sense, the world is basically One. This attitude is explored in detail in the old Indian philosophy of Advaita Vedanta, and is also found in mystics of probably all religions, including the Christian one (for example in Meister Eckhart). But as they all emphasize, this "Oneness" cannot appropriately be described by our language, only circumscribed and approximated from the outside, as it were, because language means words, and words automatically mean separation. "Every word a lie." - However, outside rare mystical experiences, we all live on the other side of the "veil of Maya", in a dualistic world with differences and with language, which reflects these differences. It is interesting to note that in the Bible one of the first tasks God gives to Adam (still in Paradise) is to name and thereby separate all the animals and birds. This can be seen as the beginning of classification.

In modern research on classification, especially in cluster analysis, it may be useful to remember occasionally that, in a sense, all classification is more or less arbitrary, that its boundaries are fuzzy (even those between life and death, as modern medicine had to acknowledge), and that we probably should ask more often inhowmuch (not: whether or not) a given classification scheme matches our observations. For example, in biology the only halfways reasonably definable classification concept (limiting arbitrariness) is the species concept; but even on this level, there is a permanent merging and splitting going on (as I am noting for birds and orchids). And what about a species like the Common Ringed Plover (Charadrius hiaticula), a small shorebird, which breeds from Baffin Land and Greenland through northern Europe, Asia and North America, gradually changing in what is called a Cline, until it meets its relatives again on Baffin Land, but now it looks and behaves like a different species (Ch. semipalmatus)? (If one wants to cling to the species concept, a split has to be made at an arbitrary line - for convenience, the Bering Sea was chosen.)

In many situations, there are "natural" groups, separated from each other (apart, perhaps, from a few "hybrids"), to be described by the classification scheme. Even there, much may depend on the degree of refinement of the differentiation, and the overall structure may be more like a complicated tree, with branches at different "heights" (as in the dendrograms of hierarchical clustering). It is also possible that another than the "best-fitting" structure is required by outside reasons, most often a linear sequence for listing for convenience "canonically", for example, all birds of the world, although every specialist is aware of a better fitting tree structure (which also is changing over time), with a few

small "loose" (hard to localize) branches.

In other situations, there is a continuum to be separated artificially, for convenience, into several classes. The continuum may be on a qualitative scale (e.g., hardness) or described by 1 (e.g., brightness of stars), 2 (e.g., off-the-peg clothing) or more quantitative variables; and the values of these variables may be determined more or less accurately. For example, many interpretations of a biplot or correspondence analysis by sociologists, psychologists etc. would fit in here. There are also diseases measured on a quantitative scale (e.g., blood particle counts); and while usually a "norm" region is interpreted as meaning "healthy" and all other values as meaning "sick", one might ask whether for some purposes it might not be better to define 3 or more classes, such as "clearly healthy", "clearly sick", and some region of increasing suspicion in between. (It is interesting to note that on a hard technical and empirical level, methods for dealing with outliers which include a smooth transition zone have proven to be clearly superior to all methods which use "hard", sudden "rejection of outliers", cf. Hampel 1985).

In more difficult situations, nothing or not much is known about the type of structure, which is to be inferred solely or mainly from the data. For example, for a long time it was not known whether double stars and star clusters (such as the Pleiades) also belong together physically (some do, some do partially).

Probably many clusters as found in cluster analysis can be described as a scatter around a zero-dimensional point; but some clusters may be scatters around a linear or curved one-dimensional structure, or around a higher-dimensional structure. It can be seen as a higher level task of cluster analysis to find not only points belonging together, but also the dimension and structure describing the clusters. (Some techniques needed may be based upon factor analysis. It should also be noted that solutions need not be unique: a zero-dimensional cluster with a highly elongated scatter ellipsoid may look the same as a linear one-dimensional cluster, and so on; and the dimension might even vary within the cluster.)

3 A brief review of outliers and gross errors

A problem with which all practical statisticians, be they frequentists or Bayesians, have to deal, is that of outliers and gross errors. Outliers - or more generally, outlying substructures - are data which are "far away" from the bulk of the data, or more precisely, which do not fit the pattern suggested by the majority of the data. "Outlier" is an ill-defined concept, with only artificial borders within a fuzzy transitional zone; nevertheless it is a useful concept, as long as one keeps its very limitations in mind. (As in other situations with classification, there are clear cases of class A, there are clear cases of class B, and there are other cases which do not fit in clearly.)

Outliers are not the same as gross errors. Gross errors are cases where "something went wrong"; for example, equipment failure, transmission errors, false categories in a structured design, or mistaken measurement of a member of a different population. The percentage of gross errors differs of course with the quality of the data; but it may still be surprising to some that even for routine data (i.e. data not taken with very special care) in the exact sciences, 1-10% gross errors appear to be the rule rather than the exception (cf., e.g., Hampel et al. 1986, Ch. 1.4). For medical data, about 10% and even about 25% gross errors have been cited. And even after careful checking, to the satisfaction of the authors, 15 data sets in the behavioral sciences still contained around 1% and up to 4% (and only in one case zero) gross errors (Rosenthal 1978). Thus, it is clear that we have to live with them. And since a single, sufficiently distant gross error can completely spoil

a normal theory (or least squares) analysis, we also have to cope with them.

Outliers may be gross errors; but they may also come from a genuinely longtailed distribution (e.g., the effects on one of the four levels of the famous hierarchical random effects model in Bennett 1954 are best modeled by a Cauchy among all t-distributions); or they may come from a different distribution and (though formally still a "gross error") may give a valuable - if not the most valuable - unsuspected new information.

Clearly, known gross errors should be given zero weight in the analysis of the bulk of the data; but they should be looked at separately and not be discarded immediately, in order to see what can be learned from them. If an outlier is known or assumed to be a proper observation, it is a common mistake to give it full weight in, for example, a normal theory analysis. This is a costly mistake because the outlier clearly shows that the model of the normal distribution is wrong, and that the true model is longer-tailed; and the best (e.g., maximum likelihood and Bayes) methods for longtailed distributions give very low influence (cf. the influence function in, e.g., Hampel et al. 1986) and hence very low weight to outlying points. (The only case where a full weight is justified is a truly nonparametric situation (e.g., when a mean or total is required no matter what the distribution), with its associated severe difficulties.) Thus, fortunately, the treatment of outliers of different types in connection with the bulk of the data is qualitatively, and can be made exactly, the same in most situations; it is only their interpretation which differs.

4 The question of model choice

If, as it occurs frequently in classification and cluster analysis, a model of multivariate normality is tentatively entertained and there appear to be outliers which are not gross errors, many scientists, both Bayesians and frequentists, are inclined to change the model, by adding more parameters and including longtailed distributions. I do not think this is the best idea. It prevents the worst — and in that sense it is defensible —, but it generally leads to rather mediocre procedures, as is shown by a number of 3-parameter and adaptive procedures in the Princeton Monte Carlo study of location estimates (Andrews et al. 1972), including those named after Hogg, L.J. Savage, Takeuchi, Jaeckel, and others. Such models cannot claim either to be the exact "true model"; they are more complicated, mathematically less nice and harder to interpret; they either lose efficiency by switching between simple models, or they try to estimate ill-determined parameters and thus are in danger of doing overfitting (which may be a partial explanation for their surprisingly mediocre performance); and they contradict one of the deepest principles of experienced data analysis (C. Daniel, in his Berkeley lectures 1968): use (and first search for) the simplest model reasonably possible, even if it is "significantly wrong"(!), because it is more useful, more reliable and better generalizable (C.D.'s term) than a more complicated one. (If a complicated model is not too much ad hoc and can be interpreted, or if it is even called for by the background knowledge of the problem, and if simultaneously the sample size is large enough, then such a complicated model may be the simplest one reasonably possible.)

Thus, in general I'd suggest to keep the simple model of normality (or whatever the most basic model is, such as normality after a simple transformation(!), or exponentiality), but to treat it explicitly as only an approximation to reality. This is done in robustness theory, the stability theory of statistical procedures (cf. Huber 1981 and Hampel et al. 1986, especially Ch. 1; for a recent overview, including also the problem of violation of the independence assumption, which is not treated in this talk, cf. also Hampel 2002). The theory allows at the same time for outliers, gross errors, longtailed (and shorttailed)

distributions, small visible distortions of the model distribution and the imperceptible deviations from the ideal model distribution which still can have large effects on the efficiency (cf., e.g., Tukey 1960). There has been a rich technology built up, with concepts such as M-estimators (or estimating equations; slight but powerful generalizations of maximum likelihood estimators), influence curve or influence function IF (describing the local effect of each data point or potential observation on whatever is being estimated; closely related to the jackknife and various forms of sensitivity curves, to perturbation theory and sensitivity analysis, and, more mathematically, to derivatives of functionals and - highly useful - Taylor expansions), and breakdown point BP (a global robustness concept, essentially the smallest fraction of arbitrary gross errors or outliers which can make a procedure totally unreliable). On this general basis (which holds for every "reasonable" parametric model!), a specific technology was developed, and checked by Monte Carlo studies, for how to deal with approximately normal data with outliers in practice, leading to 2- and 3-part redescending M-estimators, Tukey's biweight, and other estimators of the same type (cf. Andrews et al. 1972, Hampel 1985, Hampel 1997). (Huber-estimators and related ones are only a first, though important and necessary step; in the same way, in which Huber-estimators sacrifice a few percent efficiency under the strict normal distribution for being arbitrarily much better than least squares under more realistic alternatives, good redescenders sacrifice a few percent under Huber's least favorable distribution, for being at least 10-20% more efficient than Huber-estimators under more realistic distributions which include outliers. But Huber was right in warning against the careless and thoughtless use of redescenders, without sufficient understanding about the problem of multiple solutions.)

Bayesians seem to have problems with the idea of an approximate model, which does not fit into their traditional paradigm. Some suggestions for them are given below (Sec. 5).

Some decades ago, many statisticians switched to nonparametric procedures "because normality is not true"; now, for the same reason, some are switching to models with upper and lower probabilities. Either method may be perfectly appropriate in a given situation, but they are not, or at least not fully justified by the reason given. There is no need to "pour out the baby with the bath water" and abandon the model (e.g. normality) entirely.

We can try to summarize a fairly simple treatment of data from an approximate model (which is still sufficiently general for most purposes):

(i) give central ("clearly good") observations weight one, and distant outliers (which in classification and cluster analysis may already belong to another population) weight zero;

(ii) decrease the weight continuously, and not too quickly, in the "zone of doubt" from one to zero.

The quantitative details are not so important, but it is essential that the weights decrease rather smoothly and not too quickly, especially in the range where they are still high. Otherwise inefficiency and instability may result. (Cf. Hampel 1997.)

For some more details on which redescending methods to use in practice, see also Hampel (1980).

It should be stressed that not only "clear outliers", but also "doubtful outliers" should be looked at separately (and interpreted, if possible); in connection with other knowledge, they may give valuable information (cf. the example in Hampel 1987, Sec. 3, p. 101ff).

With the downweighting, we are not doing a "test whether the value is an outlier". If one really wants to know how "unlikely" a value is, it is not appropriate to use the

normal distribution as basis; both Bayesians and frequentists would have to estimate the longtailedness of the true distribution of the "good" data (which would lead to very different results); but most good scientists will prefer practical expertise to a formal rejection rule (for the interpretation of outliers. For assessing the structure of the bulk of the data, "doubtful outliers" should still be kept with their partial weight).

If for nonstatistical reasons (e.g. because of simplicity) a "classification" (separation) of the data into only two "classes" ("good" and "bad") is required (although this is highly artificial), then rejection rules based on the median deviation or MAD (the median of the absolute deviations from the sample median) work best, and even they lose at least about 10-20% efficiency unnecessarily, compared with good "smooth redescenders". Subjective rejection (if well done) also prevents the worst, but it has been shown to lose also about 10-20% efficiency unnecessarily (Relles and Rogers 1977, Hampel et al. 1986, Ch. 1.4). It has often been said that the whole methodology of rejection of outliers is faulty in principle; but in addition most rejection rules proposed in the literature (cf., e.g., Barnett and Lewis 1994) are only mediocre to bad. Even the frequently used largest Studentized residual ("Grubbs's rule") is only mediocre, with a breakdown point around 10%, and works only for fairly high-quality data. See Hampel 1985 for details. Not even the interquartile range can fully replace the MAD (cf. Andrews et al. 1972, Ch.7E3).

A few words might be added on the question of efficiency. Many statisticians have objected to robust procedures for the normal model because robust procedures "lose efficiency". But this argument is nonsensical. These statisticians lose not a few, but dozens percent efficiency with normal theory methods, without even noticing it, the misleading and false suggestions or claims by Cox and Hinkley (1968), Stigler (1977) and the Gauss-Markov theorem (to cite a few sources) notwithstanding (cf., e.g., Hampel 1998a). On the other hand, a few dozen percent efficiency loss may often be bearable in practice, as long as the analysis is good otherwise (with respect to systematic errors, model choice, etc). It may also be stressed that the arithmetic mean is "much less nonrobust" than the empirical variance (for which the efficiency is in fact not rarely close to zero), so the mean without extreme outliers may well be used with rather modest efficiency standards. (With high standards, if trying to have just a few percent avoidable efficiency loss, one needs the utmost of robustness technology.) In the topics of this conference, simple and flexible tools seem often more important than the last refinements, hence details about how to "optimize" robust procedures will at most be alluded to.

5 Some suggestions for Bayesians

5.1 General remarks

Bayesians seem to have problems with robustness, especially with robustness against deviations from the parametric model (there is now quite a bit of work concerning certain aspects of robustness against changes of the prior distribution). The most common way out in practice still seems to be the replacement of the original parametric model, such as normality, by another, more complicated ad hoc model. These models are, strictly speaking, as unrealistic as the original model; if (as is frequently the case) they are chosen with good intuition, they do work for a full neighborhood of the original model, but this can only be proven by robustness theory. For an old numerical data set, see Dempster (1975) and the subsequent discussion, where the results (apart from the formalism) are virtually identical with those of "classical" robustness theory.

It may also be noted that a sequence of Bayes estimators (indexed by n, for a fixed

prior) can often also be viewed as a sequence of functionals, a different one for each n. In order to apply asymptotic theory for a Bayes estimator for a given n, one ought to use the functional for that fixed n (which the sequence of Bayes estimators "crosses" on its smooth but nondirect way to infinity), go to infinity with it (and NOT with the Bayes estimators) and extrapolate back from infinity onto this fixed n. This allows also to transfer concepts such as influence function and breakdown point onto Bayes estimators for each given n in a meaningful and informative way.

5.2 Robustified likelihood function

Some Bayesians may want to cling to their original model and to an unmodified likelihood function, yet be somewhat robust. For them I offer the following tentative suggestion. All they have to do is to replace the most extreme observations by pseudo-observations, which behave like data from the ideal model and do not contain dangerous outliers. It is well known that transition to ranks loses amazingly little information, and it is also known that the estimator derived from the Fisher-Yates-Terry-Hoeffding-van der Waerden test or normal scores test is robust and asymptotically fully efficient under the model of normality, although already in small neighborhoods of the normal, it is empirically surpassed by the Hodges-Lehmann-estimator derived from the Wilcoxon test, as predicted by its lower breakdown point (Hampel 1983). The basic idea is now to determine iteratively a central region with unchanged data, and replace the most extreme ones by something related to the normal scores.

More precisely, let (under the ideal model) for $i \leq n$ the X_i be $i.i.d. \sim N(\theta, 1)$, and assume the X_i are already ordered. Let t be a trial value for $\hat{\theta}$ (starting with the median), then keep all X_i with $|X_i - t| \leq c$ (with $c = 2$ or perhaps $= 1.5$). Let X_k be the largest $X \leq t + c$; let $\Phi(c) = d$, where Φ is the cumulative standard normal distribution; then, for all $j > k$, replace X_j by something like $t + \Phi^{-1}(aj + b)$ with $a = (1 - d)/(n - k)$ and $b = 1 - (n + 1)(1 - d)/(n - k)$ (for variants, see below). That is, replace the upper tail by an "ideal" normal sample. Do analogously for the lower tail. Call the new pseudo-observations, depending on t (including the central ones) Y_i. Compute the new $t = \Sigma Y_i/n$ and iterate until convergence. This means, solve the implicit equation $\Sigma(Y_i(t) - t) = 0$.

The new pseudosample would contain about the same information as the original one if that came from an exact normal distribution, but it does not contain dangerous outliers anymore. Its likelihood can be plugged into Bayes' theorem in the usual way.

Some variants for the argument of the inverse cumulative normal are the following. One may try to put some more "natural" variability into the tail pseudovalues; this can be achieved by replacing j by $j + \epsilon$, where the ϵ's are independent uniform on $[-1/2, 1/2]$. One may also try to get a smoother transition between $X_k (= Y_k)$ and Y_{k+1}, for example by putting $Y_{k+1} = X_{k+1}$ and starting the (perhaps approximately) linear scale in the argument of Φ^{-1} from there.

A related approach would be the following. Start with a trial value t and an $\epsilon > 0$, such as $\epsilon = 1/(n + 1)$, and replace each $X_i - t$ by $\Phi^{-1}((1 - \epsilon)\Phi(X_i - t))$. Compute the mean as the correction of the old t and iterate until convergence. Instead of Φ, it is even easier to use the cumulative logistic distribution. This procedure can also be restricted to the tails of the sample, as above; it contains a bit more information from the original sample and is also somewhat robust.

If t is a good robust starting value, the first step of the iterations (in all these cases) may already suffice (cf. the experiences in Andrews et al. (1972) and Bickel's (1975) proofs of asymptotic equivalence).

5.3 Bayes' theorem for weighted data

One of the most useful and important descriptions and tools in robust statistics is down-weighting of outlying and nearly outlying points. The weights are to be determined, e.g., by the influence function of the M-estimator used, and thus are determined randomly by the whole sample. (M-estimators are basic for being able to reject outliers smoothly.) In general, the weights change with the parameter component being estimated, and the class of multivariate estimates obtained with "random weighting" (one weight only for all parameters in each data point, as in "iteratively reweighted least squares") is not the same as that obtained by "random transformations" (e.g., iterative Huberizing), cf. Hampel (1978). For multiple regression (with fixed x-variables), they are the same, but not in multivariate analysis. The class of all M-estimators is even more general. Nevertheless, for reasons of simplicity, we may often restrict ourselves to estimators based on "random weighting", as has been implicitly done in some sections above.

Since the weights are random weights, frequentists have to worry about their variability (which caused some complications in the seventies when such estimators were first discussed). I believe Bayesians (and adherents of the likelihood school) have a great advantage here, because they believe in the strong likelihood principle. For them, it does not matter how the weights were obtained; all they have to accept is that somehow these weights measure the internal consistency of a data point with the rest of the sample (independently of any prior belief), and then they can treat them as fixed weights.

Since the log likelihood ratios $\log(l_x(\theta_1)/l_x(\theta_2))$ measure the evidence of one parameter value over another one, and since the evidence can be weakened, even down to zero, by multiplication with the weight attached to it, it appears very natural to define the joint likelihood of a sample of i.i.d. observations, weighted by weights w_i ($0 \le w_i \le 1$), as $\Pi l_{x_i}^{w_i}(\theta)$, and the log likelihood as $\Sigma w_i \log l_{x_i}(\theta)$. ($\Sigma w_i$ can be taken as the effective sample size.) With apriori-distribution $\alpha(\theta)$, we thus obtain Bayes' theorem for weighted data:

$$\alpha(\theta|x_1,\ldots,x_n) = \alpha(\theta)\cdot\Pi l_{x_i}^{w_i}(\theta)\ /\ [\int \alpha(\theta')\Pi l_{x_i}^{w_i}(\theta')d\theta']$$

This formula, together with just some reasonable weights derived from robustness considerations, allows Bayesians to make inference which is robust against deviations of the distribution of the data from the assumed model.

In most cases, the w_i are $\equiv 1$ for the "good" data - which have to exist for defining a model structure. However, sometimes we may discount the whole data batch for outside reasons, for example, if we doubt its overall quality, and then we may multiply all weights by a factor < 1, so that we obtain $\max w_i < 1$.

It is clear that one needs "enough" data for a purely internal "consistency check" (at least 3, preferably many more, in the case of the normal with unknown mean and variance). However, if also the prior is used for checking the reasonableness of the data, even a single observation can be downweighted. This gives an automatic solution to the problem of obvious discrepancy between prior and the data (more or less keeping the prior), although in some cases we would rather trust the data than the prior. (We could still treat the data as a more or less internally consistent group of "outliers" compared with our prior belief, which describe a "new" population of data sources.)

Bayesians who are not quite so pragmatic as to accept weights from some "outside" source such as common sense or robustness theory, may try to develop a purely Bayesian internal check for consistency of the data. (Perhaps this has been done already.) However, if they are sufficiently pragmatic to still care a little bit about frequentist properties, they

should try to develop this check in alignment with the empirical and theoretical results of robustness theory.

5.4 Upper and lower probabilities

An extension of Bayesian theory which may help to model the state of our knowledge more flexibly and appropriately (and which can also precisely model the "state of total ignorance" with respect to a parameter), is the incorporation of upper and lower probabilities. This has been done by professed Bayesians such as Dempster (1967, 1968) and Good (cf. e.g., Good 1983 and the literature therein); it has been worked out by Shafer (1976), Smets and others to the so-called Dempster-Shafer belief function theory; and it is also in the background of the work by Berger (1984) and others on robustness against changes of the apriori-distribution. Upper and lower probabilities are more appropriate to describe uncertain and ambiguous knowledge and hence for statistical inference, while Bayesian proper probabilities are in general only suitable for decisions (if this distinction is being made). A unifying statistical theory which combines the aleatory probabilities of the Neyman-Pearson theory with the epistemic probabilities of the Bayesians, using (new) frequentist epistemic probabilities as a bridge which contain Fisher's fiducial probabilities in a corrected form, is outlined in Hampel (1998b, 2000). This theory is not yet suitable for general practical use, but it clarifies already many concepts, up to almost complete symmetry between the (extended) frequentist and the Bayesian approach.

I do not know how much has already been done, but I can imagine the (sensible) introduction of upper and lower probabilities into classification and cluster analysis to become very fruitful.

6 Some remarks on robust covariance matrices

6.1 General remarks

A basic and central tool in much of multivariate analysis are covariance matrices. It is clear that they can suffer much from outliers and gross errors (remember that an empirical variance is even "much more nonrobust" than an arithmetic mean); and unfortunately, their robustification encounters difficulties because of the "curse of dimension" ("too many directions in a high-dimensional space"). Robust M-estimators of location and scatter are an elegant and valuable tool for low dimensions; but it came as a shock to the "robustniks" community when Maronna (1976) (and later others) proved that under weak conditions the breakdown point for all equivariant M-estimators and in fact for all "smooth" and "reasonable" equivariant estimators is $\leq 1/p$, where p is the dimension of the data. (Hence, starting with dimension around 10, the reliability of these methods is too low.)

As a reaction, "high breakdown point" estimators, such as the Stahel-Donoho estimator (Stahel 1981a,b; Donoho 1982), the minimum covariance determinant (MCD) estimator and others were developed for covariance matrices (and similarly for multiple regression, the "least median of squares" estimator going in fact back to Hampel 1975), which theoretically keep a breakdown point 1/2 for all dimensions, but which have to be approximated by a cumbersome computer search "in all directions" of a high-dimensional space.

While I think these methods are a legitimate and even fascinating topic for theoretical research, I do not think they should be recommended for general routine practical use. (There are always problems where even the most crazy looking methods invented in mathematical theory may be most appropriate. Thus, I recall a scientist who had the

problem of discovering a main effect analysis of variance structure in his data, which contained outliers, without knowing the experimental conditions (that is, the levels of the effects); but this is precisely the problem addressed by the sequence of estimators beyond the "shordth" (Hampel 1975, p. 380), the "shordth" being also the basis for the "least median deviation" or "least median of squares" estimator, which is also briefly discussed for general nonlinear models in that paper.)

In particular, the condition of equivariance has been overly stressed (for mathematical convenience?). Even though an affine equivariant "ideal" model is often appropriate, this is not true for the gross errors, which tend to occur quite often in single coordinates, thus breaking this structure. It may often be more appropriate first to downweight and eliminate the worst outliers in single coordinates, and then to "robustify" the rest in an affine equivariant way (thus still achieving "approximate" affine equivariance). Because of the "curse of dimension", it may well be that no simple routine method would be fully appropriate under all circumstances; but in any case, I would think that a box of flexible and easily understandable tools combined with interactive analysis and interpretation of the data might be quite useful in real practice.

6.2 Some aspects of weight functions

Given a sample of p-dimensional vectors X_i, we may start by first "robustifying" in single coordinates. In each coordinate, the median med is the most robust (in many ways) estimate of location, and the median deviation or median absolute deviation MAD (Hampel 1968, 1974; Andrews et al. 1972) is the most robust estimate of scale. Using them, we can center and scale the whole sample (by linear transformations) to robust location zero and robust dispersion diagonal one. Given this, we can already downweight and reject clear outliers in single coordinates, using the weights ($w(x) = \psi(x)/x$) corresponding to one of the "smoothly redescending M-estimators" (defined by $\int \psi((x-T)/MAD)dF(x) = 0$), such as 25A or biweight (cf. Section 4). - By the way, if deemed desirable, med and MAD can also be replaced by more efficient though similarly robust estimators.

A problem is to get good robust estimates of the correlations. There are a number of proposals for the correlation of each pair of variables. One possibility are rank correlations, such as Spearman's or Kendall's rank correlation. Their (moderately good) robustness properties, including their (somewhat intricate) breakdown properties, were first explored by Grize (1978). The most extreme correlation in some ways, the quadrant correlation, does often seem to lose too much information; but a less extreme class obtained by "smooth limiting" (cf. Huber 1981, Ch. 8.3), transforming the standardized data by a monotone ψ-function and taking the (biased) correlations of the transformed values, may be better applicable and leads to a positive semidefinite covariance matrix. Otherwise, one of the nicest proposals in this context (highly robust starting points) is probably the application of the idea in Gnanadesikan (1977, Ch. 5.2, Formula (70); there are many more proposals in this chapter): let Y_{ki} and Y_{kj} be the standardized *i-th* and *j-th* coordinates of the X_k; then define $r_{ij} = ((MAD_k\{Y_{ki} + Y_{kj}\})^2 - (MAD_k\{Y_{ki} - Y_{kj}\})^2)/((MAD_k\{Y_{ki} + Y_{kj}\})^2 + (MAD_k\{Y_{ki} - Y_{kj}\})^2)$. The main disadvantage is that if these pairwise highly robust correlations are put together, the resulting covariance matrix is (in general) not positive semidefinite.

There are other possibilities (cf. also the proposals in Hampel 1975; one claim there is wrong, cf. Hampel et al. 1986, p. 431). But for proceeding pragmatically and simply, we may tentatively just multiply all the weights ($= \psi(x)/x$, with ψ defining, e.g., 25A, cf. Andrews et al. 1972) of the single coordinates together to give a fixed weight for each

data point (eliminating all clear outliers in single coordinates). Then we may iteratively compute weighted covariance matrices, with these fixed weights multiplied with iterative weights derived from the "robust Mahalanobis distances" in each step. The final multiplied weights can be used in an ordinary Bayesian analysis of the weighted sample (cf. Section 5.3 above), as if the weights had been given apriori. Frequentists still should at least approximately acknowledge the effects of the weights being only estimated, e.g. by using the ψ-function formulas of M-estimators.

If the first weight function is $\equiv 1$ in a large central region, the resulting procedure is even approximately affine equivariant, if all data are "good". The first weight function can be seen as giving a partial "precleaning" of the data.

For the choice of the weight function, compare also the general remarks towards the end of Section 4. If a quick descent is required for outside reasons, e.g. in order to separate nearby clusters, then the form of the tanh-estimator should be used or approximated, the "gross-error sensitivity" can and probably should be decreased, and the advantage of a lower "rejection point" is to be weighted against the danger of local instability (for concepts and arguments, cf. e.g. Hampel et al. 1986 and Hampel 1997).

Even more specifically: If r is the standardized distance of a point from the center, or the Mahalanobis distance given by a weighted covariance matrix during the iterations, we may put $w(r) \equiv 1$ for $r \leq c$ and $w(r) \equiv 0$ for $r > z$, with $z > c$ (preferably $z > 3c$). (Some default values might be $c = 2$ or 2.5 and $z = 8$ for one-dimensional data, but there is a whole range of smooth redescenders, e.g. from 25A to 12A, see Andrews et al. 1972.) In between c and z, the most primitive idea would be to go down linearly (which, by the way, is not the same as the linear descent of ψ in HMD or in the third part of the three-part redescenders). However, it is better first to go down more slowly, as in a quadratic with derivative 0 at c (and only later linearly). A popular form of weight function is Tukey's biweight (Beaton and Tukey 1974); however, it redescends in one region almost as quickly as it ascends, and this could cause problems with very special data sets (Hampel et al. 1986, p. 408f).

The "curse of dimension" shows itself when we consider the choices of c and z for higher dimensions p. If c is constant with respect to p, then a smaller and smaller percentage of data will receive weight 1. If c (and z) is increasing with p (e.g. such that this percentage stays constant), then in each direction it becomes harder and harder to downweight and reject data points; but this is to be expected, because there are "so many" directions in which data and hence also outliers could be.

Covariance matrices can be used to characterize classes of clusters. It appears natural to allow data points to have positive weights in two (or more) classes or clusters. (This happens, somewhat similarly, also in "fuzzy clustering", cf., e.g., the FANNY method in Kaufman and Rousseeuw 1990.) This is often almost a necessity in a first round (for safety and efficiency reasons), and as long as only inference is considered, it may just describe the fact that the point fits into two (or more) populations. But when a decision is required (even if it be artificial), then each point may be put into the population where it has the largest weight. Depending on circumstances, we may even use rather quickly redescending weight functions in a later round (up to "hard rejection"), in order to separate the populations better.

The foregoing discussion may be naive in some ways, in trying to simplify the highly complex situation with robust covariance matrices to some essentials (in fact, some problems with the speed of convergence have already been noted), and there is room for many improvements and refinements. But that the basic ideas actually do work very nicely, has been shown by Hennig (1998, 2001, Hennig and Christlieb 2002) in his work on fixed point

clusters, starting iterative weighting more than once from small homogeneous subsets, where he gives both theoretical proofs of properties and successful empirical results of the procedures.

7 Artificial classes in a continuum

7.1 The problem in the one-dimensional case

Assume measurements or counts Y are being made in a finite interval on R (or Z), and that they can be modeled as an ideal value or parameter X (unknown) plus a random term ϵ, whose distribution is known (or can be estimated). Assume we want to subdivide the interval of X-values into finitely many classes for convenience (for example, in order to distinguish and name different degrees of severeness of a disease, when the severeness can be approximately measured), but because of the random fluctuation ϵ, there will be misclassifications when Y is used for the unknown X. How do we get a maximum number of classes (and hence maximum information from the classification) while keeping the probabilities of misclassification small in a suitable way?

First, it is clear that for X near a boundary of two classes, the probability of misclassification is at least near 50%, at least for misclassification into a neighboring class, if not also other classes. All we can reasonably require is that the probability of misclassification into a nonneighboring class is small, say, at most 2α (e.g., = 10% or 5%) for all nonneighboring classes. But then we can solve the problem by successive computation of the boundaries.

If we have a location or shift model (with restricted parameter range), we can start with a whole class for the parameter on the lower boundary of the range of X, take the upper α-point of the distribution of Y under this parameter as the next class boundary point, the upper α-point under this next point as the following one, and so on. If we have a general continuous one-parameter model, we have to check each time whether the previous point is below the lower α-point of the distribution under the present boundary value, otherwise we have to shift the latter upwards until this condition is fulfilled. (The reason for this slight complication is obviously that the variance or distribution of ϵ may change with the parameter.)

In a discrete case, say, the binomial distribution, we try to classify the possible discrete values in such a way that 2 observations in nonneighboring classes most likely do not come from the same parameter value. This means, we try to find out how many parameter groups and observation groups we can "half-"distinguish (in the sense that only misclassifications into neighboring groups are allowed to be likely). It will be interesting to compute how many classes are possible, for example, for $n = 20$ compared with $n = 10$ in the binomial case.

7.2 The two-dimensional continuous case

The two-dimensional case can occur, for example, when the results of a PCA, biplot or a correspondence analysis based on a small sample (with appreciable random variability of the points) are to be classified into as many "half"-distinguishable groups as possible.

In the two-dimensional case, we have the additional problem (and flexibility) of the shape of the classes. Assume we have a circular distribution of Y with constant error distribution. A naive idea would be to divide the plane up into squares of a suitable size, by intersecting two orthogonal sets of equidistant parallels. But then each square would

have 8 neighboring squares with which it could easily be confused. A better idea might be to shift every second row by half a square-length; then each square has only 6 neighbors (though with a smaller distance of the nearest non-neighbor). Still better appears to be a filling of the plane with regular hexagons, because presumably each hexagon (if of equal size as a square) is better isolated (farther away) from its non-neighbors, reducing the error rate, or equivalently the hexagons can be made smaller (to be checked by computations).

The basic considerations (not the mathematical details) of this section may be of some value for the recent tendencies of strong subdivision of the Larus argentatus/Larus cachinnans complex (Herring gull complex) in ornithology. I do not think a geographical subdivision in which nonneighbors cannot be distinguished anymore (with reasonable certainty) would make much sense.

8 Some ideas concerning clustering

8.1 A robust metric

The following idea can be applied anywhere in multivariate analysis where a metric is used and it is feared that gross errors might occur in single coordinates. Instead of searching for the outliers separately, one simply "almost ignores" them by using a metric which is hardly affected by a few outliers.

The idea is inspired by the Prohorov distance between two probability distributions (Prohorov 1956), which has found a nice interpretation in robustness theory. We consider the same basic idea, but in a different manner. Let X_i and X_j be two p-dimensional vectors with coordinates X_{ik} and X_{jk}. Starting with the basic metric $d(X_i, X_j) = \max_k |X_{ik} - X_{jk}|$, we define the "robustified metric" $d^*(X_i, X_j) = \inf\{\epsilon : |X_{ik} - X_{jk}| \leq \epsilon$ except for a fraction ϵ of coordinates $\}$ or $d^*(X_i, X_j) = \inf\{\epsilon : \#k$ with $|X_{ik} - X_{jk}| > \epsilon$ is $\leq \epsilon p\}$. This implies that data vectors with $1, 2, \ldots$ clear outliers in single coordinates have at least distance $1/p$, $2/p, \ldots$ (also depending on the distances of the other coordinates), but not arbitrarily close to ∞; for pairs of vectors with $d < 1/p$ (without outliers), d and d^* agree. The incongruence of the two ϵ's in the definition, one describing a distance and the other describing a fraction of outliers, is also an advantage: we can scale the basic metric arbitrarily and thus decide in what distance we want to treat a pair of coordinates as containing an outlier, by putting that distance $= 1/p$ by multiplication of the basic metric with a constant. (By the way, this rescaling is also sometimes incorporated into the definition of the Prohorov distance.)

This robust metric also partly answers a remark by W.J.J. Rey (2001, orally), who correctly noted that the naive use of the breakdown point for whole data vectors in high-dimensional multivariate statistics is inappropriate, because for large p almost every data vector will contain some outlying coordinates. The problem was already noted, together with other problems, in Tukey's talk 1970 on robust regression in the Princeton robustness seminar (cf. Hampel 1997, p. 137).

8.2 Triple minimum spanning trees

Minimum spanning trees (MST) are a fast and valuable tool for getting a first hold of the (possibly complicated) structure of a data set in high dimensions. (There can and should be first many considerations concerning the choice of the metric, but we skip these here.) The MST describes a simple network covering all points and giving some information on closest neighbors of a point. But my impression is that it does not give enough information

about the neighborhood of any point. In order to get a first feeling for a local structure, we need some more neighboring points, but not too many, otherwise we would get lost again.

Therefore I suggest to try the tentative idea of an overlay of a few, preferably perhaps 3, MST. The basic MST is very fast to compute, so the others should not take much longer. The idea for the second MST is to determine an MST, but leaving out all connections already in the first tree. Similarly, the third MST must not contain any connections used in the first two MST.

The hope is that these trees together give a first indication of local clustering, local dimension, structural peculiarities, and so on, while still allowing to "look at the data" (with their tree connections). They might perhaps serve as a tool in exploratory analysis of high-dimensional data sets.

8.3 One-dimensional clusters (and modes) via a new smoothing method

Assume a (moderately large) number of points on the real line are given. Are they scattered "randomly", or do they form certain clusters? (As an example, take the observation dates of a migrating bird species - with perhaps several populations - on the time axis. Are they distributed "uniformly" (or unimodally) over the whole migration period, or are there several distinct modes which should be reproducible with future observations, and which might belong to different populations or subspecies?) A related problem is that of discovering several modes (or even "hidden modes", from different subpopulations) in a histogram from possibly heterogeneous data.

There are many, many smoothing methods for one-dimensional data. However, to my limited knowledge, they all may contain local "wiggles" which are not justified by the data, or they may contain a strong bias, deviating more globally from the data structure. In trying to formalize Tukey's ideal of a "freehand smooth", which avoids either fault, I developed the ingredients of a new smoothing methodology which minimizes the number of points of inflection in the curve and, in principle, all its derivatives. A "necessary" point of inflection describes a true feature of the data. The ingredients were (partly) put together in the thesis by Mächler (1989; 1995a and b) with a lot of sophistication, and resulting in a first, preliminary, but working computer program.

The program gives a solution essentially for each number (and approximate location) of points of inflection (and hence modes and antimodes in the derivative) given. A further, probably very demanding continuation of this work would be to develop tests (presumably by cumbersome simulations) for the necessity of a point of inflection (to test whether the "smooth" with it, or a pair of them, is significantly better than the "smooth" without). These tests could then be applied to the empirical cumulative distribution function of the observed points (which does not suffer from the disadvantages of the histogram), in order to find out which "apparent" clusters appear to be real.

Acknowledgments: I am grateful to C. Hennig, M. Mächler, G. Shafer and W. Stahel for several valuable discussions and comments.

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 University Press, Princeton, N.J.

BARNETT, V., and LEWIS, T. (1994): *Outliers in Statistical Data*. Wiley, New York. Earlier editions: 1978, 1984.

BEATON, A.B., and TUKEY, J.W. (1974): The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data. *Technometrics*, *16*, 2, 147–185, with Discussion –192.

BENNETT, C.A. 1954: Effect of measurement error in chemical process control. *Industrial Quality Control*, *11*, 17–20.

BERGER, J.O. (1984): The robust Bayesian viewpoint. In: J.B. Kadane (Ed.): *Robustness of Bayesian Analyses*. Elsevier Science, Amsterdam.

BICKEL, P.J. (1975): One-step Huber estimates in the linear model. *J. Amer. Statist. Ass.*, *70*, 428–434.

COX, D.R., and HINKLEY, D.V. (1968): A note on the efficiency of least-squares estimates. *J. R. Statist. Soc. B* , *30*, 284–289.

DANIEL, C. (1976): *Applications of Statistics to Industrial Experimentation*. Wiley, New York.

DANIEL, C., and WOOD, F.S. (1980): *Fitting Equations to Data*. Wiley, New York. Second edition.

DAVIES, P.L. (1995): Data Features. *Statistica Neerlandica*, *49*, 185–245.

DEMPSTER, A.P. (1967): Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, *38*, 325–339.

DEMPSTER, A.P. (1968): A generalization of Bayesian inference. *J. Roy. Statist. Soc.*, *B 30*, 205–245.

DEMPSTER, A.P. (1975): A subjectivist look at robustness. *Bull. Internat. Statist. Inst.*, *46*, Book 1, 349–374.

DONOHO, D.L. (1982): *Breakdown properties of multivariate location estimators.* Ph.D. qualifying paper, Department of Statistics, Harvard University, Cambridge, Mass.

GNANADESIKAN, R. (1977): *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

GOOD, I.J. (1983): *Good Thinking; The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis.

GRIZE, Y.L. (1978): *Robustheitseigenschaften von Korrelationsschätzungen*. Diplomarbeit, Seminar für Statistik, ETH Zürich.

HAMPEL, F. (1968): *Contributions to the theory of robust estimation*. Ph.D. thesis, University of California, Berkeley.

HAMPEL, F. (1974): The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, *69*, 383–393.

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

HAMPEL, F. (1978): Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Invited paper ASA/IMS Meeting. Amer. Statist. Assoc.* Proc. Statistical Computing Section, ASA, Washington, D.C., 59–64.

HAMPEL, F. (1980): Robuste Schätzungen: Ein anwendungsorientierter Überblick. *Biometrical J. 22*, 3–21.

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

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

HAMPEL, F. (1987): Design, modelling, and analysis of some biological data sets. In: C.L. Mallows (Ed.): *Design, Data, and Analysis, by some friends of Cuthbert Daniel*. Wiley, New York, 93–128.

HAMPEL, F. (1997): Some additional notes on the "Princeton Robustness Year". In: D.R. Brillinger, L.T. Fernholz and S. Morgenthaler (Eds.): *The Practice of Data Analysis: Essays in Honor of John W. Tukey*. Princeton University Press, Princeton, 133–153.

HAMPEL, F. (1998a): Is statistics too difficult? *Canad. J. Statist.*, *26*, 3, 497–513.

HAMPEL, F. (1998b): On the foundations of statistics: A frequentist approach. In: Manuela Souto de Miranda and Isabel Pereira (Eds.): *Estatística: a diversidade na unidade*. Edições Salamandra, Lda., Lisboa, Portugal, 77–97.

HAMPEL, F. (2000): An outline of a unifying statistical theory. Gert de Cooman, Terrence L. Fine and Teddy Seidenfield (Eds.): *ISIPTA'01 Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications*. Cornell University, June 26-29, 2001. Shaker Publishing BV, Maastricht, Netherlands (2000), 205–212.

HAMPEL, F. (2002): Robust Inference. In: Abdel H. El-Shaarawi and Walter W. Piegorsch (Eds.): *Encyclopedia of Environmetrics*, *3*, 1865–1885.

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.

HENNIG, C. (1998) Clustering and outlier identification: Fixed Point Clusters. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Berlin, 37–42.

HENNIG, C. (2001) Clusters, Outliers, and Regression: Fixed Point Clusters. *J. Multivariate Anal.* Submitted.

HENNIG, C., and CHRISTLIEB N. (2002): Validating visual clusters in large data sets: Fixed point clusters of spectral features. *Computational Statistics and Data Analysis*, to appear.

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

JEFFREYS, H. (1939): *Theory of Probability*. Clarendon Press, Oxford. Later editions: 1948, 1961, 1983.

KÜNSCH, H.R., BERAN, J., and HAMPEL F.R. (1993): Contrasts under long-range correlations. *Ann. Statist.*, *21* 2, 943–964.

KAUFMAN, L., and ROUSSEEUW, P.J. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

MÄCHLER, M.B. (1989): *Parametric' Smoothing Quality in Nonparametric Regression: Shape Control by Penalizing Inflection Points*. Ph. D. thesis, no 8920, ETH Zurich, Switzerland.

MÄCHLER, M.B. (1995a): Estimating Distributions with a Fixed Number of Modes. In: H. Rieder (Ed.): *Robust Statistics, Data Analysis, and Computer Intensive Methods – Workshop in honor of Peter J. Huber, on his 60th birthday*. Springer, Berlin, Lecture Notes in Statistics, Volume 109, 267–276.

MÄCHLER, M.B. (1995b): Variational Solution of Penalized Likelihood Problems and Smooth Curve Estimation. *The Annals of Statistics*. *23*, 1496–1517.

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

PROHOROV, Y.V. (1956): Convergence of random processes and limit theorems in probability theory. *Theor. Prob. Appl.*, *1*, 157–214.

RELLES, D.A., and ROGERS, W.H. (1977): Statisticians are fairly robust estimators of location. *J. Amer. Statist. Assoc.*, *72*, 107–111.

ROSENTHAL, R. (1978): How often are our numbers wrong? *American Psychologist*, *33*, 11, 1005–1008.

SHAFER, G. (1976): *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N. J.

STAHEL, W. (1981a): *Robust estimation: Infinitesimal optimality and covariance matrix estimators (in German)* Ph. D. thesis, no 6881, ETH Zurich, Switzerland.

STAHEL, W. (1981b): *Breakdown of covariance estimators*. Research Report 31, ETH Zurich, Switzerland.

STIGLER, S.M. (1977): Do robust estimators work on real data? *Ann. Statist.*, *6*, 1055–1098.

"STUDENT" (1927): Errors of routine analysis. *Biometrika*, *19*, 151–164.

TUKEY, J.W. (1960): A survey of sampling from contaminated distributions. In: I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann (Eds.): *Contributions to Probability and Statistics*. Stanford University Press, 448–485.