Stat. Meth. & Appl. (2007) 15:271–293
DOI 10.1007/s10260-006-0034-4

ORIGINAL ARTICLE

A survey of robust statistics

Stephan Morgenthaler

Accepted: 8 November 2006 / Published online: 3 January 2007

Abstract We argue that robust statistics has multiple goals, which are not always aligned. Robust thinking grew out of data analysis and the realisation that empirical evidence is at times supported merely by one or a few observations. The paper examines the outgrowth from this criticism of the statistical method over the last few decades.

1 A brief history

Providing sound methods for the analysis of data gathered through experimentation or observation is the goal of robust statistics. Robustness of statistical methods in the sense of insensitivity to grossly wrong measurements is probably as old as the experimental approach to science. In the course of experimentation, things can go wrong. A catalyst is used at too high a concentration, a value is jotted down wrongly, instead of brand A, brand B is used for some of the experiments. The possibilities are endless. If we act as if nothing of the sort could ever happen, the resulting statistical summaries will not be sound. On the other hand, ad-hoc robust procedures are often easy to construct. They are obtained by cleaning the data before statistical treatment. The first generation of robustness, therefore, dealt mainly with outlier nomination, formal outlier rejection tools, and graphical methods that could help in spotting outliers. This

This research was supported in part by the Swiss National Science Foundation.

S. Morgenthaler (✉)
Institute of Mathematics, Ecole Polytechnique fédérale de Lausanne, Lausanne, Switzerland
e-mail: stephan.morgenthaler@epfl.ch

S. Morgenthaler
EPFL FSB IMA; Station 8, 1015 Lausanne, Switzerland

may seem an easy task, but is in reality not so. For complex data types encountered in multiple regressions or multivariate estimation, finding the outliers can be overwhelming.

An essential ingredient of all data is their inconclusive nature. Knowledge gained through data and experimentation is on some level uncertain and fluid. Dating back to Gauss and the least squares method, statisticians have argued in favor of particular statistical methods on the basis of some optimality property. The usual argument was "my method is more efficient than any other". Conveniently forgotten in this type of argumentation was the fact that efficiency was defined with respect to quite strong assumptions about the data. Within the emerging theoretical thinking about statistics, some started to question the validity of deducing the method to be used from the assumed model. Around the same time, more formal robustness demands appeared. At first, this touched mostly on the validity of tests (see for example Pearson, 1931) and was studied with the help of concrete alternatives to the base model. How did the test that was calibrated for a particular model behave under alternative specifications? In particular, how robust was it to non-normality (Box, 1953). This question can equally well be applied to estimators. What happens to the efficiency of an estimator if the data distribution does not conform to the model under which the estimator is optimal? Tukey (1960) discussed a telling example in which very low frequency events could utterly destroy the average performance of so-called optimal estimators. Even in situations where data sampled from a perturbed distribution could hardly be distinguished from data sampled from the model distribution, the seemingly best method was severely hampered.

In parallel with these ideas, distribution-free tests and estimators were being developed. These seem at first blush a good answer to the critique of the over-reliance on a particular model. The analysis of the methods derived from distribution-free procedures has led to a better understanding of the mechanics of typical robust procedures. The median, which can be derived from the sign test, offers an exemplary case study. It is a nonlinear function of the data, focuses its attention only on the central part of the sample, and does not change its value drastically even when up to half of the observations are faulty. Those are properties a robust estimator should possess. However, its asymptotic efficiency at the normal model is a lowly 64%. But, then, stepping from a particular model to a completely unspecified distribution is maybe too radical a change.

Few books in statistics have been as original and thought-provoking as Peter J. Huber's Robust Statistics (1981). It pulled together the various vague ideas, case studies and emerging theories published previously and proposed a coherent theory for the second generation robustness. In Huber's view, robustness starts with a traditional parametric model and then proceeds by considering broadened assumptions under which the estimation and testing of the parameter still made sense. Instead of assuming a perfect normal model for the measurement errors, under which the mean is the most efficient summary, he asked what would happen if the data were roughly normal, but not exactly so. He thus broadened the circumstances under which the performance of the statistical summary would be judged. In such a context the old style "most efficient

statistical method" no longer exists. Its place is taken by a compromise method, whose efficiency remains high over the whole set of broadened circumstances. One can make this mathematically precise by considering for each statistical method its worst performance and then searching for the method with the best worst-case performance. It turns out that the least squares method, which is so popular among users of statistics, is particularly vulnerable under the best worst-case criterion. This follows from the fact that least squares estimators are linear functions of the observations and thus any perturbation in any of the observations is linearly transmitted to the estimators. This in turn leads to a catastrophic worst-case performance.

In some cases the minimax problem posed by Huber has a simple solution in the form of a least-favorable model, close by the original model one started out with. The best worst-case performance is then achieved by the optimal estimator for this new model. This solution depends of course on the extent to which the original model is broadened. Robust methods bridge the gap between narrowly optimal classical methods and the distribution-free methods. In Huber's world, the user must specify the size of the neighborhood. The estimation of the mean μ of a normal distribution is an illuminating example of this theory. Broadening this model can be done in various ways, for example by considering mixtures of $1-\epsilon$ parts of the normal with ϵ parts of an arbitrary contamination. If the contamination is symmetric about μ, the new parameter in the enlarged problem is the median of the distribution, and the inference problem remains well-posed. For $\epsilon = 0$, Huber's estimator is the arithmetic mean and for $\epsilon = 0.5$ it is the median. For intermediate values it turns out to be similar to a trimmed mean with a trimming rate that is monotone increasing with ϵ. More precisely, the solution corresponds to a modified least squares criterion in which the quadratic is replaced by a function with linear growth in the tail and quadratic behavior in the middle. The corresponding least-favorable model retains an exact normal form in the middle, but the tails are modified to a log-linear rather than a log-quadratic form. In this simple location model, robust methods correspond quite closely to the procedures that were earlier developed for data containing outliers and thus, robustness essentially turns out to offer protection against unanticipated heavy tails. For restricted parameter spaces additional phenomena may occur. In estimating scale for example there is a danger of implosion, which means seriously underestimating the scale. Because the deviations from the model that matter most are heavy tails, one can construct robust models directly. The estimators derived in traditional ways from robust models already have robustness properties and no additional fiddling is necessary. Methods of this kind are described in (Morgenthaler and Tukey, 1991 and Lange et al., 1989).

Other important contributions to robustness that emerged out of the idea of considering neighborhoods of models are the use of functional analytic tools to derive robustness indicators such as the influence function and related optimality theories (Hampel, 1974, Hampel et al., 1986). Such tools are based on the change in the performance of statistical methods under slight modifications of the original model. The influence function in particular has had a big impact. Its study suggests that robustness can be achieved by modifying the traditional

likelihood score function through truncation. The maximal bias under arbitrary contaminations of size ϵ is another basic quantity. It leads to the notion of breakdown (Hodges, 1967, Hampel, 1974, and Rousseeuw, 1984) and bias robustness (see for example Donoho and Liu, 1988 or Martin,Yohai and Zamar, 1989).

In the three decades since the laying of the foundations of robustness, the theory of has been further consolidated. All the strands we mentioned above, from outlier identification through nonparametric and semiparametric modeling, have been generalized and expanded. The robustness indicator used when constructing methods, be it a high breakdown point, a limited susceptibility to bias, a low gross error sensitivity or a minimax performance over a set of alternative models is crucial. Because none of the possibilities is intrinsically preferable and all choices are defensible, the availability of robust methods has been greatly enhanced. Some of the books and articles describing such enhancements are as follows: Maronna (1976), Kleiner et al. (1979), Rousseeuw and Leroy (1987), Tyler (1987), Morgenthaler and Tukey (1991), Kent and Tyler (1991), Maronna et al. (1992), Rousseeuw and Croux (1993), Davies (1993, 1995), van Aelst and Rousseeuw (2000).

Robust inference poses computational challenges. It is often necessary to solve a high-dimensional quasi-convex optimization problem or even explore a search space to minimize some quite arbitrary criterion function. The work horse of robust statistics has been local linearization and the iteratively reweighted least squares algorithm. It has the advantage of being quite easily added to existing statistical packages, but does not always converge to the global optimum. For high-dimensional problems stochastic optimizers (simulated annealing or even guided random selection of potential solutions) have been proposed and shown their usefulness.

2 The current state of robustness

Statistics and with it robustness are undergoing big changes of emphasis, mainly for two reasons, both driven by the increase in cheap computing power. More complex techniques that require sophisticated numerical simulation are available to today's users. At the same time, the models being fitted to data have also become much more intricate, which raises the question of the meaning of robustness in this changed context?

In the location and scale problems that were the main inspiration for second generation robustness, the unusualness or outlyingness of observations can be defined with respect to other observations taken under identical circumstances. In more complex models, starting with linear regression, this is not true. It is no longer possible to identify outliers by internal comparison alone. Instead outlier nomination must to some degree be based on assuming the truth of an underlying model. This contradicts the aims of robustness.

A possible resolution of this challenge consists in generalizing the basic robustness philosophy by taking parts of the postulated model at face value (for example the linear regression structure), while broadening the assumptions on

other parts. One may, for example, broaden the conditions on the distribution of the random errors, or on the correlation structure of the errors, or on the additivity of the errors, or choose still other aspects. It is then possible to derive estimators that are robust with regard to certain types of deviations from the original regression model. If one restricts attention to deviations in the error distribution, modifying the residual sum of squares is a natural choice. The modification proposed by Huber consists in taking a quadratic function of the residuals for small values and a linear function for large absolute residuals. This method has, however, infinite gross error sensitivity and zero breakdown point because contaminations at remote points in the predictor space may exert a large pull on the estimated regression coefficients. Should this be a cause of worry?

This result brings into focus a conditionality issue in robustness. The first generation methods relied on least squares applied to cleaned data. For data sets without apparent outlier, no cleaning was needed and the raw data were used. This seems quite reasonable. Whether or not trimming or downweighting need to be applied to an observation can be decided after looking at the data and need not be decided beforehand. Even though it is not explicitly built into the method, second generation robust techniques do in fact act conditionally on the data, at least to some extent. Huber's location estimator will be exactly equal to the mean for well-behaved data. Similarly, whether remote points in the predictor space are present should be checked in the actual data. Why should we worry about possible contamination at such points if there are none present in the data? For protection against heavy tailed error distributions one can, for example, formulate an extended model containing various error distributions with various tail indices. The most efficient estimator in the extended model averages over the estimates corresponding to the various tail indices (Morgenthaler and Tukey, 1991). Such an estimator is explicitly conditional on the observed data configuration.

Another possible solution to the challenge discussed above consists in searching for the simplest accurate description of a large portion of the data. This point of view is similar to hoping that after outlier removal, good fits to the remaining observations can be obtained. In a prediction problem, for example, a model that fits 80% of the observations reasonably well with a linear equation and a single predictor might well be considered preferable to one that manages to increase this proportion to 100%, but at the cost of including three additional predictors in a nonlinear fashion. The 20% of the observations not accounted for by the simple model are sometimes called outliers, but this is a misleading term in our context. Statistical methods that are robust in this sense provide alternative and often surprising fits. In principle, quite distinct robust answers may well be equally good and may all provide insights. High-breakdown regression estimates (Rousseeuw, 1984, Davies, 1990) and forward search procedures (Atkinson and Riani, 2000, Atkinson Riani and Cerioli, 2004) provide examples of such methodology. As a bonus, these methods offer effective outlier identification. By this we mean faulty experimental measurements, hidden in a series of experimental runs. In a regression context, for example, it is hard to

find these observations, even if the predictors to be used are known. Looking for large residuals from robust fits is the most reliable identification technique.

The aim of providing robust alternatives to traditional multivariate methods has played a large part in recent developments. Again, the meaning of the term outlier is not as clear-cut as one might like. Imagine an experiment, in which during each run several variables are measured. Each measurement can potentially go wrong and produce faulty data. If this model is used, then outlyingness must be defined with regard to the marginal distribution of each variable separately. If with chance ϵ and independently from the other measurements any variable is measured wrongly and if a total of d variables are observed, then the chance that a whole experimental run produces no faulty data in any of the variables is $(1-\epsilon)^d$. For $\epsilon = 20\%$, the probability of a clean record is equal to 33% when $d = 5$ and a tiny 1.1% when $d = 20$. If we deal with such outliers by identification and removal, we have to deal with a data matrix with around a fraction of ϵ of the values missing at random. Here, the measured variables play a privileged role and univariate analyses, variable by variable, make sense. In multivariate methods, one often considers linear combinations of the original variables, in particular orthogonal projections. If we resort to such methods before data cleaning or downweighting, we will mix outliers into almost all projected data points.

On the other hand and in addition, mistakes could happen in preparing an experimental run. For example, the treatment dose prescribed by the protocol was not correctly apportioned, the temperature was set to high, or some other glitch. All of the variables affected by this mistake are then faulty. Here, outlyingness has to be defined in a truly multivariate context, that is, an outlier may remain invisible if we look for it variable-wise. Univariate robust methods can again be used but it is not enough to consider the privileged marginal projections. A search over the sphere $\mathcal{S}^{d-1}$, that is, over all possible directions must be undertaken (Stahel, 1981). Alternatively, one may compute a Mahalanobis-like distance of each data point from the center of the distribution and thus rank and weigh the observations according to their potential outlyingness (Maronna, 1976). The outcomes are similar to the regression situation discussed above. It turns out that procedures based on internal weighing are quite easily fooled if one is allowed to add outliers in arbitrary ways. Sturdier methods can be gotten by searching for a good fit to a part of the data while ignoring the remaining part.

In addition to these two outlier models, others can be imagined. Actual data most likely present a mixture of several types of challenges. Linked to the different models considered above is the question of affine equivariance. Let $\mathrm{Y} \in \mathbb{R}^{n\times d}$ be a data matrix and consider the linear transformation $\mathrm{Y}\,\mathrm{V}$ for $V \in \mathbb{R}^{d\times d}$. Affine equivariance for the variance estimate demands that after transformation it be equal to $\mathrm{V}^T\mathrm{S}\mathrm{V}$, where S is the estimate before transformation. Under an elliptical model, affine equivariance is a natural requirement, whereas in more general cases it is too restrictive.

The leading model for multivariate data consists of the elliptical distributions, which are characterized by a positive definite matrix and a spherical (radial) density. The multivariate normal distribution is a member of this class.

By construction, such models do not exhibit outlier behavior, because the whole data set will be representative of a coherent model. If we wish, we can introduce the tail index of the radial distribution into the model, but this will not affect the estimation of the variance. In order to introduce whole-case outliers, one has to mix $1 - \epsilon$ parts of one elliptic model with ϵ parts of a separate one. In simpler situations, building robust models is a successful strategy. In multivariate situations, the writer does not know of any entirely convincing proposals. The robust multivariate methods proposed in the literature are based on various requirements such as high breakdown, efficiency at the normal, or computability and are derived on heuristic grounds.

Other active research areas one should mention in an overview of robustness are stochastic process data, model or variable selection, smoothing and signal processing, and spatial statistics (for example, Genton, 2001).

3 Some remarks on the future development of robustness

The stability of the conclusions drawn from experimental and observational data remains a pressing problem. The hope that the study of robustness would give birth to an entirely new way of doing statistics, a new way that would replace the classical methods entirely, has not been fulfilled. There seem to be several reasons for this. For one, statistics as a whole has moved on and incorporated the teachings of robustness. Also, leaving standard models such as the Gaussian, the Poisson, or the Binomial behind and abandoning the fitting by likelihood, multiplies the available choices. This is reflected in the diversity of the robust methods a user can choose from. The fact that no new standards have been created has not helped the acceptance of robustness.

More subtly though, the robust school has influenced modern statistics heavily. All statisticians today are aware of the dangers of a data analysis that owes more to model assumptions than experimental facts. And most research papers contain at least an acknowledgment of the robustness aspects of the problem being studied. Because the models being used have become more realistic and accurate, the need for subsequent robustness work has diminished. Statistics has been and remains today most helpful in situations where the random noise is at a relatively high level. These are precisely the situations where reliance on a restricted probability model can do the most damage by leading to totally wrong conclusions. The detection of a tiny signal hidden among a heap of Gaussian noise, for example, might be feasible if things were exactly this way, but when exact Gaussianity is replaced by a broader set of circumstances, the problem might become unsolvable.

Robust statistics is well-suited to play a leading role in two areas: routine data analyses performed with the help of statistical packages and data mining. For a long time, the writer believed that robust methods would replace the standard least squares techniques that are preprogrammed in statistical packages. Such programs are mostly used by people untrained in statistics and thus the lack of robustness would ultimately prove to be too costly. That this has not happened

is presumably because these users believe that the superiority of least squares is a fact. The costs of using run of the mill and inadequate models is quite real however. One merely has to think of all the assertions of public health findings that cannot be confirmed by subsequent research and which have cost billions in research money. Other fields have other traditions. The data mining and statistical learning community is quite open to try alternative methods. There, the emphasis is not so much on confirming experimental findings as on finding patterns and rules that can be used in a predictive manner. Robust methods, with their ability to focus on parts of the data only, can produce surprising and useful results in this context. Robust principal components or robust classification methods, for example, are sure to have an impact in this growing field.

Discussion

Anthony C. Atkinson, Marco Riani and Andrea Cerioli
The London School of Economics, London WC2A 2AE, UK
Dipartimento di Economia, Università di Parma, Italy

We enjoyed reading Stephan Morgenthaler's magisterial survey of robust statistics; in particular we appreciate his emphasis on the link between the physical mechanism generating the recorded numbers and appropriate robust methods. This theme is announced in his first paragraph, but informs much of the subsequent discussion.

As Morgenthaler emphasises, classical statistical methods give answers of known high quality to well-formulated problems about precisely modelled data. The results of such analyses are typically single estimates, with associated standard errors, with all the data contributing to the estimates. On the other hand, some of the robust methods he describes downweight or reject some observations. But they again lead to single estimates and tests with known properties, although these properties are unfortunately not so well known amongst users of statistics as are the properties of such estimators as least squares in regression. Other robust methods mentioned in Sect. 2 seek to include all observations. For example, in regression or multivariate analysis, normal distributions can be replaced with t distributions of unknown degrees of freedom. All data are again modelled. Although the point estimates of the parameters may have good properties, estimates of standard errors and confidence intervals for predictions can be so wide as to provide highly pessimistic predictions. Whether downweighting or accommodating is to be preferred must depend on the kind of data met in the subject field. Our experience is that normal models with some form of outlier, which may be completely rejected, are more informative than very long-tailed distributions that incorporate all observations in a single model.

We are pleased by Morgenthaler's emphasis on methods that perhaps fit to only part of the data −80% was mentioned. We would like to emphasize that we see no need to state in advance the fraction to be fitted. Not only can the fraction be chosen conditionally on the data, but many fractions can be fitted.

With modern computing power we can even explore a whole series of subsets of the data.

We see the forward search, mentioned in Sect. 2, as a generalization of this approach in which the model (or models) are fitted to increasing sized subsets of the data. As the subset size increases the method of fitting moves from very robust to highly efficient likelihood methods. If the data and classical model agree, the journey from fitting a few observations to virtually all will be uneventful – parameter estimates, test statistics and residual plots will remain sensibly constant. But if we have two or more populations, one of which might be a straggle of outliers, there will come a point where the stable progression of fits is rudely interrupted. Then decisions will need to be made about clustering and about outlier rejection. Atkinson and Riani (2002) show such a progression for t tests about coefficients in regression models. On the other hand, if the model is systematically incorrect, the change in parameter estimates, test statistics and residual plots may be gradual but relentless. The presence of t-distributed errors in regression produces such plots when a normal model is fitted.

There are two features to which we want to call attention. One is the systematic fitting of the models to many subsets of the data, leading to a series of estimates and inferences that are informative about the structure. This is to be contrasted with the single fit typical of many robust methods. The other is that much robust work has brought with it from mathematical statistics an emphasis on algebra and the presentation of results as tables of numbers. We believe that an important contribution of the forward search is to provide informative graphical summaries of the relationship between data and the properties of the series of fitted models. For interpretation of these plots it is often helpful to have bands for the distribution of test statistics, which may have to be found by simulation. Examples for clustering are given by Atkinson, Riani and Cerioli (2006).

A criticism of much work on robustness is that there is an assumption of symmetry. In Huber's motivating example described in Sect. 1 the data are "roughly normal", which is formalized as a normal distribution with symmetrical contamination. Particularly in multivariate work, the normal distribution is a most powerful tool for the analysis of data. But often normality applies only after the data have been transformed. In our work we use the family of parametric power transformations introduced by Box and Cox to obtain symmetry and approximate normality. However, the information about transformation comes from the extreme observations; if these are downweighted or deleted by a robust estimation procedure, evidence about the correct transformation will be lost. The forward search also here provides powerful and easily visualised methods of establishing the correct transformation through forward plots of estimated parameters and score statistics for hypothesised transformations. Once the correct transformation has been found for most of the data, the estimated transformation parameter will be sensibly constant for the search until the outliers, if any, enter. A difficulty is that outliers on one transformed scale may not appear as such under another transformation. Therefore several searches for different transformations may be needed to establish the distinction between

observations that are outlying and those that are influential for the transformation parameter. Once the data have been satisfactorily transformed, robust and or forward procedures can be applied.

A final comment is that Morgenthaler discusses the analysis of data, but does not touch on the design of the experiments that lead to that data. In regression experiments many optimum design criteria, such as D-optimality (see, for example, Atkinson, Donev and Tobias, 2007) lead to observations at points with equal leverages, so that the problems of outliers at very high leverage points are avoided. The experimental design literature uses the phrase "robust design" in two ways, neither of which relate to concerns in Morgenthaler's article. One, associated with the name of the Japanese engineer Taguchi, has to do with the design of products that are robust against poor or incorrect conditions of use. The other has to do with the construction of experimental designs that are robust against assumptions about, for example, models and parameter values. There is rather less work on the design of experiments with robust analysis in mind, but see Müller (1997). Perhaps the hope is that a correctly designed experiment will obviate the need for a robust analysis!

Reference

1. Atkinson AC, Donev AN, Tobias RD (2007) Optimum experimental designs, with SAS Oxford University Press, Oxford
2. Atkinson AC, Riani M (2002) Forward search added variable *t* tests and the effect of masked outliers on model selection. Biometrika 89:939–946
3. Atkinson AC, Riani M, Cerioli A (2006) Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani S, Cerioli A, Riani M, Vichi M. (eds), Data, analysis, classification and the forward search. Springer, Berlin Heidelbery New York, pp. 163–171
4. Müller CH (1997) Robust planning and analysis of experiments. Lecture Notes in Statistics, Vol. 124. Springer, Berlin Heidelberg New York

Christophe Croux and Peter Filzmoser

K.U.Leuven, Belgium
Vienna University of Technology, Austria

First of all we would like to thank Professor Morgenthaler for sharing with us his views on Robust Statistics. In this short note we give some further comments on the field of robust statistics, as a complement to Professor Morgenthaler's paper.

When we give talks on our research on robust statistics, it is often asked: "what precisely is an outlier?" An answer that we give is: "an outlier is an observation which is unlikely to have been generated by the imposed model." Such an answer implies that, for example in a regression setting, an observation may be an outlier with respect to a linear regression model, but not necessarily anymore with respect to a quadratic regression model. Even in a univariate

setting, identification of outliers relies on an implicit of explicit formulation of an underlying model (most often the normal model). If observations are taken under identical circumstances, all of them from a Cauchy distribution, then it might happen that some observations are far away from the majority of the other data points, but there is no reason to call them outliers, at least not if we take the Cauchy as our model distribution. Note that an outlier is not always a highly influential observation; influential observations are defined as having a high impact on the value of an estimator computed from the sample. Influential observations are defined with respect to an estimator, and not with respect to a model. If a model is postulated, the definition of an outlier can be made precise. If one believes that a postulated model is always a simplification of reality, as most statisticians do, then in fact none of the observations will follow exactly the model, leading to 100% of outliers, most of them moderate. Then an appropriate formulation is to state that the true distribution is "close" to the model distribution, where closeness is defined in terms of a certain metric. Such an approach was in fact already taken in the early work of Peter Huber.

Sometimes researches in robust statistics say "an outlier is an observation that behaves differently from the large majority of the other points." However, whether an observation behaves different from another can only be verified having an underlying model in mind. But from a data-analytic viewpoint the above statement can still make sense: the idea is that the selection of an appropriate model should be done using the large majority of the observations, for example 75% of them. Such an approach will result in more simple, and hence easier-to-interpret, models. Then we have an identified model, and one can detect outliers with respect to this model. The main difficulty is then to find this 75% of observations which are revealing a simple model structure.

The issue of robustness of estimators with respect to outliers is well studied, and robust rules (i.e. not subject to the masking effect) for the identification of outliers in several settings, including multivariate and non-linear models, are available. Our feeling is that more attention needs to be given to other types of model deviations. Indeed, if we consider a linear regression model, then outliers, or heavy tailed error distributions, are only one possible source of model deviation. But it is also possible to have autocorrelated or heteroscedastic error terms, which may bias the statistical inference when not accounted for. Robust corrections for these types of model deviations need to be proposed. Another issue is the linearity assumption, which may only be approximately valid. Bringing robustness ideas into non-parametric or semi-parametric methods can break new grounds here. There has been research on robust non-parametric methods, but we feel that there is room for further developments here.

As he wrote in his paper, Professor Morgenthaler believed at a certain moment that robust methods would replace standard least squares techniques. The fact that this did not happen up to now (we give no prediction for the future) might have two more reasons: (a) robustness ideas are lacking in a standard statistical training, and (b) software for robustly analyzing data is not available or too complicated for practical use. Concerning (a) we believe that already basic courses on statistics should consider the aspect of robustness. People have to

be aware that not only the statistical properties of an estimator are important, but also the behavior of the estimators under deviations from the ideal model have to be encountered. Without going into much detail, the principles of robust methods can easily be explained graphically, and don't need an advanced mathematical treatment. Moreover, by demonstrating that for "clean" data the robust method gives approximately the same answer as the classical method, the advantages of robust methods should become immediately clear. Analyzing real data examples (rather than showing the effect of robust estimation on simulated data) will again contribute to underline the necessity of robust methods.

Regarding the development of statistical software there is definitely still work left for the robustness community. Easy-to-use software should be made available, and it should be made easy to use. This means that input and output of robust routines should have about the same structure as that for the classical procedures. This will allow a simple use of the tools without having much effort to first study all different parameter settings for the robust procedures. Among other goals, these requirements have been defined in the project "Robust Statistics and R", see http://www.statistik.tuwien.ac.at/rsr/. As the name already indicates, robust statistical methods should be implemented in the statistical software package R (http://www.r-project.org). The fact that R is freely available gives hope that robust methods will be routinely used in the future. Besides robust procedures for analyzing data, the availability of interactive graphical tools for exploratory data analysis as well as tools for robust diagnostics are important. These allow a much deeper insight into multivariate data structures, and provide important additional information to summary tables.

Laurie Davies and Ursula Gather

University of Duisburg-Essen, Germany, and Technical University of Eindhoven, Netherlands University of Dortmund, Germany

We congratulate Stephan Morgenthaler on this fine survey of robust statistics. He has given us a broad and well-chosen overview of the field, written in a light tone, which makes the paper well readable and accessible to researchers from other disciplines. In such an overview it is not possible to delve deeply into all aspects of robustness and we would like to comment on only two topics where we think a more detailed although still not complete discussion may be of advantage. One concerns the problem of estimation and the other is related to outlier detection.

1. When discussing Huber's minimax problem Morgenthaler writes 'If the contamination is symmetric about μ, the new parameter in the enlarged problem is the median of the distribution, and the inference problem remains well-posed'. This raises the question of the role of symmetry in robust statistics. If the analysis is always conducted within the realm of parametric models then there is the question as to which parametric model to use but the problem of which parameters to estimate does not arise. In robust statistics we

allow full neighbourhoods of models and as a result we leave the world of parametric models. Instead of asking which parameter we wish to estimate we must ask the question as to which functional we wish to estimate. In the one-dimensional location problem all affine equivariant functionals will estimate the centre of symmetry for any symmetric distribution. In the case of the normal distribution this is μ and remains μ if only symmetric contamination is allowed. All well-defined affine equivariant functionals estimate μ in this situation and not only the median. If we allow a full topological neighbourhood of the normal distribution then this will contain non-symmetric distributions and affine equivariance will no longer suffice to give a unique answer. The question as what we now wish to estimate can no longer be answered within statistics. It is useful to distinguish two uses of the word 'estimate' which are almost always conflated in statistics. We formulate the problem in the terms of analytical chemistry. Given a sample of water we wish to estimate the amount of mercury present. This is in the primary meaning of the word 'estimate'. In statistics the data are often analysed by assuming a parametric model and then 'estimating' the parameters of the model. If the data were indeed generated under the model then this use of the word 'estimate' is acceptable. If however the data were not generated under the model we are trying to 'estimate' something whose existence is, to put it mildly, in doubt. The existence of mercury molecules in the water sample on the other hand is not in doubt. In practice, if a parametric model is used then we choose some parameter or function of the parameters and identify, provisionally and speculatively, the quantity of mercury in the water sample with the value of this parameter or with some function of the parameters. If we use a robust approach and if the basic model is symmetric then we must choose one of the many possible functionals and accept its values for non-symmetric data. This is typically the case for data from analytical chemistry. The question then is, which functional do we choose. The answer can only be decided upon by the practical results which, without being completely conclusive, do at least allow a comparison of sorts between the different functionals. In interlaboratory tests laboratories have to analyse water samples which have been prepared and the amount of contamination is known, at least to some extent. This knowledge is not complete due to evaporation and interaction with other chemicals but it does offer some indication. If we have to choose between the median and the (asymmetric) elimination of outliers followed by the mean then the latter functional would seem to be better in the empirical sense. Of course what we are judging here is a procedure rather than a given functional although the distinction in the present case is not important. A discussion of the role of symmetry and the analysis of water samples can be found in Tukey (1993) (Sect. 13: S is for symmetrical models and Symmetrical challenges, Sect. 15: 15 X is for eXamples, chemical or other) available from `www.stat-math.uni-essen.de/~davies/tukey.html` .

2. When discussing outlier detection, Morgenthaler writes 'In more complex models . . . it is no longer possible to identify outliers by internal comparison alone. Instead outlier nomination must to some degree be based on assuming

the truth of an underlying model. This contradicts the aims of robustness.' We point out that from an epistemological perspective there is just no other choice; an observation – even in the univariate case – can only be outlying with respect to some target model. It is not necessary that this be the truth. In practice the concept of an outlier cannot be very precise but in order to compare different identification rules one needs a definition to make outlier detection amenable to a formal analysis. The following is based on Davies and Gather (1993). For the normal distribution $N(\mu, \sigma^2)$ and $\alpha \in (0,1)$ the α-outlier region is defined by

$$out(\alpha, N(\mu, \sigma^2)) \;=\; \{x \in \mathbb{R}\colon\, |x - \mu| > \sigma\, z_{1-\alpha/2}\}, \tag{1}$$

which is just the union of the lower and the upper $\alpha/2$-tail regions. The extension to other distributions is clear. Each point located in the outlier region is called an α-outlier. This definition of an outlier refers indeed only to its position in relation to some model for the good data. No assumptions are made concerning the distribution of these outliers or the mechanism by which they are generated.

One can then formulate the task of outlier nomination for the normal distribution as follows: For a given sample $\boldsymbol{x}_n = (x_1, \ldots, x_n)$ which contains at least $[n/2] + 1$ i.i.d. observations distributed according to $N(\mu, \sigma^2)$, we have to find all those x_i that are located in $out(\alpha, N(\mu, \sigma^2))$. The parameters μ, σ^2 may be unknown and then have to be estimated in a robust manner to avoid masking and swamping effects. The level α can be chosen to be dependent on the sample size. The advantage of this approach is that it allows complete control of the outliers. The number and positions are at the disposal of the statistician. Morgenthaler mentions Tukey's demonstration that optimal estimators can quickly lose their optimality in the presence of hardly detectable perturbations. Tukey used a mixture model for this

$$(1 - \varepsilon)N(0,1) + \varepsilon N(0,9).$$

We note that he did not use this model to generate outliers. However it is unfortunately the case that mixture models of the form

$$(1 - \varepsilon)N(0,1) + \varepsilon H$$

are often used for outlier generation. Two weaknesses of such models for this purpose are that the data remain i.i.d. and that one cannot control the number of outliers in a given sample which is then itself a random variable.

References

1. Davies PL, Gather U (1993) The identification of multiple outliers (with discussion). J Am Stat Assoc 88:782–801

2. Tukey JW (1993) Issues relevant to an honest account of data-based inference, partially in the light of Laurie Davies's paper. Princeton University, Princeton

Ricardo A. Maronna and Víctor J. Yohai

University of La Plata and C.I.C.P.B.A., Argentina
University of Buenos Aires and CONICET Argentina

We congratulate Prof. Morgenthaler for his survey on the current status of robust methodology. This survey has stimulated us to address some points which we consider relevant for further developments in robustness.

1 Are global extrema best?

Many high breakdown point estimates are defined as the absolute minimum of a functional. For example, MM and S regression estimates, and also S estimates of multivariate location and dispersion are thus defined. The absolute minimum is usually unattainable, except when the number of parameters is one or two. The usual procedure is to start from a (possibly inefficient) estimate with high breakdown point, and iterate to attain a local minimum of the functional. However, besides the breakdown point it is necessary to consider the contamination bias of an estimate. It turns out that the best approach when bias is taken into account is to use as starting value an estimate with the lowest available bias. The local minimum thus obtained has, compared to the global one, the same breakdown point and the same asymptotic efficiency, *but* a lower asymptotic bias!.

The first two facts can be proved theoretically. We have no general proof of the third, but have demonstrated it for regression and for multivariate analysis in (Maronna et al. 2006, Sects. 5.9 and 6.8). Thus, surprising as it may seem, even if the global minimum were revealed to us, we would do better by iterating from a good starting point.

2 Elementwise contamination

Consider an $n \times p$-dataset. In general, robust multivariate procedures are resistant when the proportion of "atypical rows" is small (in any event, smaller than 1/2), and cannot cope with missing values. There are however situations where it is conceivable that each of the np data values may be contaminated. For instance, in computer vision, where each row represents an image and each column a pixel (De la Torre and Black, 2001); or in the analysis of microarray data, where each row represents a case and each column a gene. Here p is very large. Assume that each element is altered at random with probability ε. Then the probability π that a row contains at least a "bad" element is 1-$(1-\varepsilon)^p$, and this can be very large for large p even if ε is small. For example, with $p = 100$

and $\varepsilon = 0.01$ we have $\pi \approx 0.63$, so that the majority of rows would contain atypical values. Missing values are also frequent. In fact, even if the proportion of missing values is small, it may happen that most rows and columns contain some missing value, so that one cannot simply keep only complete rows or columns. The usual procedures cannot cope with this type of data. It is necessary to work "elementwise", which implies giving up any kind of equivariance. This model has been studied by van Aelst et al. (2006).

The problem that seems most amenable to attack is principal components. Here one can exploit the equivalence between principal components and approximation of a matrix by another of low rank. Croux et al. (2003) have proposed an approach based on "alternating regressions" using weighted L_1 regression, which is resistant to cellwise contamination and can be adapted to handle missing values. We are currently working on an approach to this problem based on the minimization of an M-scale of residuals.

We expect this approach to be more efficient, more robust, and computationally faster than former ones.

3 Fast hybrid estimates

The first algorithms for computing high breakdown point estimates for linear regression or covariance matrix were based on pure subsampling. A remarkable improvement with respect to these algorithms are the so called "fast algorithms" introduced by Rousseeuw and Van Driessen (1999).

The modification proposed by Rousseeuw and Van Driessen is to improve each of the candidates obtained by subsampling by means of a small number of "concentration steps" such that each of them decreases the loss function. Doing this, in most of the cases one or two steps are enough to transform a poor candidate into a good one. Rousseeuw and Van Driessen (1999) propose concentration steps for the minimum covariance determinant (MCD) estimate; Rousseeuw and Van Driessen (2006) do the same for the least trimmed squares (LTS) estimate; and Salibian-Barrera and Yohai (2006) propose an algorithm for regression S-estimates. The use of concentration steps reduces dramatically the computing time required high breakdown estimates, which allows the application of these procedures to very large data sets.

One common property of the concentration steps mentioned above is that they guarantee the reduction in the value of the corresponding loss function. However in many cases, for example for computing the minimum volume ellipsoid (MVE) estimate, there is not a natural concentration step with this property. However in Maronna et al. (2006) it is shown that if the MVE is computed by applying to each candidate obtained by subsampling the concentration step of the MCD, the results notably improve. The reason is that even if we cannot guarantee that always the MCD concentration step produces a reduction in the loss function of the MVE, this generally occurs when it is applied to a bad candidate. This "hybrid estimate" is much more robust than the MVE based on pure subsampling and has similar computing times.

It should be noted that a similar approach for the LMS estimate was already proposed by Stromberg (1993).

Reference

1. Croux C, Filzmoser P, Pison G, Rousseeuw PJ (2003) Fitting multiplicative models by robust alternating regressions. Stat Compu 13:23–36
2. De la Torre F, Black MJ (2001) Robust principal components analysis for computer vision. In: Proceeding international conference on computer vision, 2001. http://citeseer.ist.psu.edu/torre01robust.html
3. Gabriel KR, Zamir S (1979) Lower-rank approximation of matrices by least squares with any choice of weights. Technometrics, 21:489–498
4. Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York
5. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
6. Rousseeuw PJ, Van Driessen K (2006) Computing LTS regression for large data sets. Data Mining Know Discovery 12:29–45
7. Salibian-Barrera M, Yohai VJ (2006) A fast algorithm for S-regression estimates. J Comput Graph Stat 15:414–427
8. Stromberg AJ (1993) Computation of high breakdown nonlinear regression parameters. J Am Stat Assoc 88:237–244
9. Van Aelst S, Alqallaf F, Yohai VJ, Zamar RH (2006) A model for contamination in multivariate data. Submitted for publication

Hannu Oja and Frank Critchley

Tampere School of Public Health, University of Tampere, Finland
The Open University, United Kingdom

It is a pleasure to congratulate the author on an excellent survey of robust statistics, its history, state-of-play and future prospects.

The classical tools for robustness (such as the breakdown point and influence functions) have been developed for simple sampling designs, and new developments may be needed for more complex models and designs (e.g. the concept of the partial influence function for the several samples case). In the following, we mainly focus on approaches to the analysis of multivariate data.

In model-based developments one usually assumes that the data come from a distribution which is, in some sense, close or similar to the nominal distribution (often a normal distribution). The strategy is often to embed the nominal parametric model in a larger nonparametric or semi-parametric model with the same set of parameters of interest (but accepting additional disturbance in the observed values). Robust estimation and testing procedures for the parameters are then hoped to be almost optimal in the nominal model, but still valid in the larger model. Huber extended the normal model by considering a mixture of a normal distribution and an arbitrary symmetric distribution (with the same

symmetry centre). Again, the multivariate normal model is often extended to a semi-parametric model of elliptically symmetric distributions. Then the d-variate observations y_i are generated by $y_i = \Lambda z_i + \mu$ with d-variate z_i having a standardized spherically symmetric distribution.The parameters in the semi-parametric model of elliptical distributions, the location vector μ and scatter matrix $\Sigma = \Lambda\Lambda'$, then retain their interpretation. In this extension, it is both natural and entirely feasible to restrict attention to estimates/tests which are suitably equivariant/invariant under affine transformation.

A different, recent, semiparametric extension of the multivariate normal distribution is the independent component (IC) model where the y_i are now generated by $y_i = \Lambda z_i + \mu$ with (in a certain way) standardized z_i having independent components. In the IC model, location vector and scatter matrix estimates may be defined in a different way and new tools for robustness studies may be needed: the regular location and scatter estimates (M-estimates, S-estimates, etc.) are built for the elliptical model only and do not work here (in the estimation of Λ). The scatter matrices, for example, should have the so-called independence property.

Another parametric extension of the multivariate normal model is a multivariate skew-normal distribution (with new parameters to describe the skewness of the distribution), which can be further broadened to the semiparametric model of multivariate skew-elliptical distributions. These models may be used for the robust analysis of the data if the data are perturbed by an unknown selective sampling procedure (and not by independent individual outliers). It is now unclear whether the affine equivariant/invariant property is still feasible. Also, the traditional definition of the breakdown point is not natural in this context.

Still one more extension of the normal model with i.i.d. observations is to allow dependence between observations. In particular, the observations may be clustered in a known or unknown way (e.g. students in schoolclasses, repeated measures on the same individual). When several levels of hierarchy are present (school/class/student), the data are called multilevel or hierarchical data. The hierarchy then yields covariance structures between measurements which should then be taken into account in the analysis. Here we may wish to have methods which are resistant to disturbances in the dependence structure.

As our last example, we mention the data with small sample size n but high dimension d which also poses new challenges in the development of robust methods. In the analysis of gene expression data, for example, one is hunting outlying genes (variables) not outlying individuals. But what do we mean by outlying variables?

Daniel Peña

Department of Statistics, Universidad Carlos III of Madrid, Spain

This paper presents a insightful Survey on Robust Methods and I want to congratulate the author for this well-written paper. The field is so broad that

it is not possible to cover all its faces in a few pages but the author has been successful in pointing out some key issues in the use of robust procedures in the past and nowadays. I would like to add to the list of good references in robustness presented in Sect. 1 the excellent new book by Maronna, Martin and Yohai (2006), which contains an outstanding presentation of the robust methods and their applications.

I think that the field of robustness is now reaching maturity. In the first stage, in the 1960s and 1970s, where the main ideas were developed, the struggle to establish the field was concentrated in differentiating it as much as possible from the previous outlier detection methods. They were judged not only as inefficient, which was true in may cases, but as wrongly oriented. The lack of recognition to the contributions from other approaches in order to emphasize their own has been a youth disease in the development of all new paradigms and, for instance, Bayesian statistics had a similar problem in its fight for survive against frequentist procedures. A key idea of the robustness approach that was strongly emphasized was that you should not delete outliers, but downweight suspicious observations in the estimation procedure. Of course if you downweight too much in order to have a high breakdown point you loose efficiency under normality and, as indicated in the paper, this trade-off has important implications because some of the most popular robust estimation methods can be very inefficient under the central model. In fact it has been shown that we can have a highly robust and efficient procedure if the final estimate is computed by least squares on some cleaned data set. Thus, we may be in part coming back to the main idea in the 1960s of outlier detection, although this outlier identification is made by using much better and powerful detection tools developed in the robust field.

I expect that in the future there will be a change in focus in the robust field from robustness to large measurement errors to robustness to the assumptions of the model, including the functional form of the model itself. This may lead to a new formalized field of model diagnostics, a key part of the scientific model building process. The new theoretical developments may or may not be based on minimax ideas, as in the Huber/Hampel paradigm, which although allowing for beautiful mathematical developments, can be criticized from many directions. This new theory should incorporate the existing exploratory diagnostic methods into a broader theory on building robust models for complex heterogeneous data. For instance, in many industrial experimentation it is assumed that the data are independent, when in fact they are correlated, and using a robust approach does not provide any protection in these cases. More important, the standard Huber/Hampel approach is based on assuming that the majority of the data follow the assumed model, and this may be very naive in many cases. Theoretical deployments to expand the field in these directions will be important.

For this reason although I agree with the final conclusion of the paper that the existing robust methods are well suited for routine analysis of large data sets, I am not so sure they can be useful for data mining and knowledge discovery. In the exploratory analysis of large data sets, as in data mining problems, we have to

be prepared for all sorts of heterogeneity, including not having a central model but a convex mixture of models. For instance, instead of a unique regression line we may have four different regression lines depending on the values of a set of explanatory variables. Also, it may happen that none of these clusters include more than 40% of the data. In this type of situations, which are very common in practice, the present paradigm of robustness is not useful, as the concepts of breakdown point or influence curve are not relevant. The approach to robustness developed by Huber and Hampel, data generated by a parametric model with some unknown contaminated distribution, was in the 1970s an important generalization from the single parametric model approach that was the usual approach in Statistics. However, since then the field of Statistics has had an impressive growth in the last 30 years, and much more sophisticated problems are now considered. The developments in nonparametric statistics and local modelling approaches, mixture models, heteroskedastic models, and complex Bayesian statistics models estimated by Markov Chain Monte Carlo methods, have enlarged very much the kind of heterogeneity problems that statisticians are trying to solve nowadays. For instance, an alternative to assuming outliers from a contaminated distribution is to assume model heteroskedasticity and this approach has deep roots in econometrics and other fields. On the other hand, there has been a huge increase in the use of local models, that do not try to fit the whole sample, but only some neighborhood of each point, where the effect of outliers on the estimation can be more easily controlled. Also, the search for outliers in genetic problems has brought up the study of inliers, that is groups of genes with the same behavior which appear as clusters of concentrated outliers in the middle of the data. Again, the present robust procedures are not well suited for this type of applications. Finally, the robust methods have been mainly developed for models with symmetric errors and in many fields we have very skewed distributions where this idea does not fit very well.

Given all this it is not surprising that the robust procedures we have today, although very useful in many situations, have not become the standard way to analyze data. I do not think that the reason is that people are just using least squares, rather they are using more sophisticated methods in which the problem of contamination for outliers is only a small part which can be handled by different procedures. A broader view of the aim of our field may contribute to enlarge the impact of the robustness ideas in applied statistical problems.

Reference

1. Maronna R, Martin D, Yohai YJ (2006) Robust statistics. Wiley, New York

Peter J. Rousseeuw and Stefan Van Aelst

University of Antwerp, Department of Mathematics and Computer Science, Belgium.

Ghent University, Department of Applied Mathematics and Computer Science, Belgium

Stephan Morgenthaler has given a nice overview of the history and development of robust statistics. We will complement the survey with some of our viewpoints, and mention some developments in robust statistics related to our own work.

As noted in Sect. 2 of the survey, statistical models have become more intricate, but also the sample size and dimension of datasets has increased enormously in the past years, posing new challenges for robust statistics. For example, in chemometrics data the dimension often exceeds the sample size, which requires special robust techniques such as those developed by Hubert and Vanden Branden (2003). Since outlier detection in high dimensions is extremely difficult, robust dimension reduction techniques have become more important. Robust methods for principal components have been developed by e.g. Croux and Ruiz-Gazen (2005), Hubert et al. (2005), and Salibian-Barrera et al. (2005). Note also that even with today's fast computers the robust analysis of large, high-dimensional data requires time-efficient algorithms. Such algorithms have been developed for the minimum covariance determinant (MCD) estimator and least trimmed squares (LTS) regression (Rousseeuw and Van Driessen 1999, 2006) as well as for S-estimators (Salibian-Barrera and Yohai 2006).

If the dimension becomes really high, then as noted in Sect. 3 of the survey it is no longer practical to consider entire observations (cases) as regular or outlying. In huge dimensions, downweighting an entire observation entails a large loss of information since it is unlikely that *all* components of the observation are actually unreliable. Moreover, in low-quality data there may not be enough observations for which all the components are reliable. Such contamination models have been studied by Alqallaf et al. (2005). As mentioned in the survey it then becomes necessary to give up affine equivariance, which makes estimating a covariance matrix less useful. Other techniques such as the bagplot (Rousseeuw et al. 1999) may then be more appropriate to describe dispersion and shape.

In Sect. 3 of the survey it is suggested that today's more complex models have reduced the need for robustness. We agree in many instances. But in many other situations, such as process monitoring, the goal is to detect deviations from the overall pattern. Complex models may not pick up changes in the process because their flexibility allows them to accomodate the deviating data too easily. In such cases, robustly fitting a simpler (more rigid) model to the main pattern/trend is desirable; see Rousseeuw and Van Driessen (1999) for an example of monitoring a production line in a TV set factory. Also in computer vision, an important task is to robustly pick up the signal in the majority of the data (see e.g. Meer et al. 1991). More generally, even with complex models there can be outliers (e.g. mistakes) that do not follow the main pattern. The more complex the model, the more difficult it becomes to detect outliers by (graphical) diagnostic tools, and hence robust estimation methods are still needed.

As noted in the survey, robust statistics started by focusing on a parametric model and constructing estimators and tests that still make sense under deviations from that model. However, robustness has been shown to be useful also in nonparametric settings. For instance, depth medians (for a survey, see Liu et al 1999) and deepest regression (Rousseeuw and Hubert 1999, Van Aelst et al. 2002) are nonparametric methods with positive breakdown value.

Another aspect that was not touched upon in the survey is the need for robust standard errors and confidence regions to accompany a robust estimation method. These are typically derived from the asymptotic distribution of the estimator. However, often this asymptotic distribution is only known for the central (normal) model, making inference impossible for contaminated data. An alternative, nonparametric approach is the use of a time-efficient and robust bootstrap procedure (Salibian-Barrera and Zamar, 2002). The fast robust bootstrap procedure has been further developed in Van Aelst and Willems (2005) and Salibian-Barrera et al. (2005).

Rejoinder

Stephan Morgenthaler

I would like to thank all the discussants for their valuable contribution. Their commentaries provide clarification of central concepts and issues, or descriptions of new developments that were not touched on in my piece, or insightful background information and views.

Several of the discussants comment on the need for data fragmentation, which combines clustering with model fitting. The idea behind this concept is the following: when representing a selected subset of the data, relatively simple descriptions can be sufficient, whereas models for the whole database might be hard or impossible to construct. Thus one may find, for example, that some outcome variable in a database of clinical cases can simply be explained as a linear regression of one or two input variables. But this only holds within clusters of cases, that is, the coefficients of the regression equation vary from one subsample to another. While this is related to the separation between "good" and "bad" data that is sometimes attempted in robust statistics, data fragmentation is in my opinion often better described as an example of flexible data representation. As is pointed out in the discussion, it is quite possible that none of the data fragments is very large by itself. It is also possible in this context that the fragments overlap and that some cases belong to none of the fragments.

The concept of outlier – the "bad" data mentioned above – is a sore point for robust statisticians. Robust statistics has been criticises in some circles as propagating sanitized data analysis, where the analyser was free to delete some of the observations in order to reach the desired conclusion. This caricature is dangerous nonsense. First, gross outliers were not invented by robust statisticians, they occur even in high quality surveys and experiments with astounding frequency. Second, to invalidate classical methods no gross outliers in the data

are required. A basic goal of robust methods is to provide protection against small, unnoticable deviations from the model. Applying such methods allows the data analyst to focus much more effectively on the data that are well fitted by the model and as a consequence helps in identifying gross outliers.

Several of the pieces bring up the issue of symmetry and comment on its significance. I think this is to some extent prompted by another critique of robust statistics that has persisted and left a mark. The usual formulation of this critique says that robust statistics is based on the assumption of symmetry. In my survey I may have been guilty of overly emphasizing symmetry and I am grateful to the discussants for correcting this impression. The importance of bringing robustness to multivariate statistics and further developing this area is also brought up by several of the discussants. Finally, modern robustness cannot be separated from computational statistics. As several of the discussants show, the inverse is also true. Robust statistics has contributed important ideas to computational statistics, in particular new ideas in optimization. To all of these points I can only add that I fully agree.

Reference

1. Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH (2005) A model for contamination in multivariate data. Submitted
2. Croux C, Ruiz-Gazen A (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. J Multivariate Anal 95:206–226
3. Hubert M, Vanden Branden K (2003) Robust methods for partial least squares regression. J Chemometrics 17:537–549
4. Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: a new approach to robust principal components analysis. Technometrics 47:64–79
5. Liu RY, Parelius J, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Stat 27:783–840
6. Meer P, Mintz D, Rosenfeld A, Kim DY (1991) Robust regression methods in computer vision: a review. Int J Comput Vis 6:59–70
7. Rousseeuw PJ, Hubert M (1999) Regression depth. J Am Stat Assoc 94:388–402
8. Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: a bivariate boxplot. Am Stat 53:382–387
9. Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
10. Rousseeuw PJ, Van Driessen K (2006) Computing LTS regression for large data sets. Data Mining Know Discovery 12:29–45
11. Salibian-Barrera M, Zamar RH (2002) Bootstrapping robust estimates of regression. Ann Stat 30:556–582
12. Salibian-Barrera M, Van Aelst S, Willems G (2005) PCA based on multivariate MM-estimators with fast and robust bootstrap. J Am Stat Assoc (to appear)
13. Salibian-Barrera M, Yohai V (2006) A fast algorithm for S-regression estimates. J Comput Graph Stat 15:414–427
14. Van Aelst S, Rousseeuw PJ, Hubert M, Struyf A (2002) The deepest regression method. J Multivariate Anal 81:138–166
15. Van Aelst S, Willems G (2005) Multivariate regression S-estimators for robust estimation and inference. Stat Sinica 15:981–1001