THE HISTORICAL DEVELOPMENT OF ROBUST STATISTICS

Elvezio Ronchetti
University of Geneva, Switzerland
Elvezio.Ronchetti@metri.unige.ch

We focus on the historical development of robust statistics by highlighting its contributions to the general development of statistical theory and applications. The basic robustness concepts and tools should be included in a natural way both in undergraduate and graduate statistics and econometrics curricula. We argue that this is more effective than treating robust statistics as a special (advanced) topic and we illustrate this point by means of an example drawn from economics and finance.

INTRODUCTION

Classical statistics and econometrics are based on parametric models. Typically, assumptions are made on the structural and the stochastic parts of the model and optimal procedures are derived under these assumptions. Standard examples are least squares estimators in linear models and their extensions, maximum likelihood estimators, and the corresponding likelihood-based tests. Many classical statistic and econometric procedures are well-known for not being robust, because their results may depend crucially on the exact stochastic assumptions and on the properties of a few observations in the sample. These procedures are optimal when the assumed model is exactly satisfied, but they are biased and/or inefficient when small deviations from the model are present. The results obtained by classical procedures can therefore be misleading on real data applications.

The theory of robust statistics deals with deviations from the assumptions on the model and is concerned with the construction of statistical procedures which are still reliable and reasonably efficient in a neighborhood of the model; see the books by Huber (1981), Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Maronna, Martin, and Yohai (2006), and Dell'Aquila and Ronchetti (2006) for an overview. Therefore it can be viewed as a statistical theory dealing with approximate parametric models and a bridge between the Fisherian parametric approach and the full nonparametric approach. It is a reasonable compromise between the rigidity of a strict parametric model and the potential difficulties of interpretation of a full nonparametric analysis.

Robust statistics is now some 40 years old. Indeed one can consider Tukey (1960), Huber (1964), and Hampel (1968) the fundamental papers which laid the foundations of modern robust statistics. Research is still active: a quick search in the Current Index of Statistics lists 1617 papers on robust statistics between 1987 and 2001 in statistics journals and related fields. Many more can be found in journals in application fields. Arguably the impact of a field in statistics can be measured by the amount of fundamental ideas, concepts and techniques which have become standard tools in modern statistical analysis. In the next section we review and discuss some of the basic ideas developed in robust statistics which have become standard concepts and tools in modern statistics. We then draw some implications related to teaching (robust) statistics. Finally we illustrate some of these points with an example from economics and finance.

MAIN CONTRIBUTIONS OF ROBUST STATISTICS TO MODERN STATISTICS

Here is a list of main ideas, concepts, and tools which robust statistics contributed to modern statistics. We focus only on those basic ideas which were developed in robust statistics but which are nowadays general tools in modern statistics.

a) Models are only approximations to reality

Of course, this is a standard statement in science, but robust statistics helped to stress and quantify this point. Starting with Tukey (1960), it demonstrated the dramatic loss of efficiency of optimal procedures in the presence of tiny deviations from the assumed stochastic model. This opened up the door to search for better alternatives and for multiple tools for data analysis.

b) Multiple analyses and solutions of a data-analysis problem

This point, among many others, was put forward by Tukey (1962) in his path-breaking paper on the future of data analysis. Robust statistics contributed to develop the idea that multiple tools are necessary to analyze real data and that real problems can have multiple solutions.

c) The minimax approach

This approach borrowed from game theory was Huber's (1964) elegant solution of the robustness problem, viewed as a game between Nature which chooses a distribution of the data in a neighbourhood of the model and the statistician who chooses an estimator in a given class. The payoff is the asymptotic variance of the estimator at a given distribution. Sometimes minimax solutions can be pessimistic, but it turned out that this wasn't the case here. The resulting estimator, Huber's estimator, became the basic building block of any robust procedure and is a basic tool beyond robust statistics.

d) Statistical functionals (and expansions); Gateaux and Fréchet differentiability

Statistical functionals had already been considered by von Mises (1947) but Hampel's (1968, 1974) approach recast the robustness problem in the language of functional analysis (continuity, differentiability, etc.). In particular the influence function (the Gateaux derivative of a functional) became the most important single heuristic tool to analyze the stability of statistical procedures and to develop new robust procedures. Its many links with the classical statistical theory (linear term in the asymptotic expansion of an estimator, basic tool to compute the asymptotic variance of an estimator etc.) and with other important ideas such as the sensitivity curve and the jacknife, make the influence function an important concept in modern statistics. Moreover, statistical functionals played an important role later in the development of the bootstrap and nonparametric techniques.

e) M-estimators (and estimating equations)

Huber's (1964) M-estimators represent a very flexible and general class of estimators which played an important role in the development of robust statistics and in the construction of robust procedures. However, this idea is much more general and is an important building block in many different fields including, for instance, longitudinal data, econometrics, and biostatistics.

f) The breakdown point

The breakdown point introduced by Hampel (1968, 1971) is a measure of global stability for a statistical functional and as such is a typical robustness measure. However, the quest for high breakdown point estimators in the field of robust statistics has pushed the development, among other things, of general computational techniques and resampling algorithms which can be used in more general settings; see Rousseeuw (1984), Rousseeuw and Leroy (1987).

IMPLICATIONS FOR TEACHING

In view of the discussion above, it seems important to include basic robustness concepts both in undergraduate and graduate curricula in statistics as well in fields of applications; see the example below. This is more effective and natural than treating robust statistics as a special (exotic) and advanced topic. Our experience shows that the mathematical treatment can be adapted to the level of the course and is in no way an obstacle to convey the basic ideas and tools.

EXAMPLE

Financial models are often estimated and tested with methodologies that do not explicitly control for the effects of small distributional deviations from the assumptions; see Knez and Ready (1997). However, because of the intrinsic complexity of financial markets and the richness of financial phenomena, we may realistically believe that some deviations from the assumptions will almost always be present when using a financial model in empirical finance. It seems therefore natural to treat financial models as approximate descriptions of the financial reality and to work with statistical procedures that can deal with some amount of "abnormal" observations and identify them systematically. In some cases, it is precisely a detailed analysis of the identified abnormal observations that will offer new insights and suggestions on the kind of features that a more accurate model should be able to fit.

Implicitly, we argue that while estimating a financial model it is important to verify first, if the majority of the data is consistent with the assumed model. If this is not the case, a more complex model can be introduced. This seems particularly meaningful in the context of empirical financial modelling, where parameter estimates and the model selected are often the input for the pricing and hedging of financial instruments. In practice, one would like to ensure that the choice of a model used to price and hedge a financial instrument is driven by the features of the majority of the observed data rather than by single data points or some particular historical period.

An illustration is provided by the re-analysis of the empirical evidence concerning a well-known class of one factor models for the short rate process, the CKLS (see Chan, Karolyi, Longstaff, and Sanders, 1992) and some recent extensions using a new statistical methodology based on robust statistics; see Dell'Aquila, Ronchetti, and Trojani (2003) where the complete analysis can be found. In particular it is demonstrated how a robust analysis can provide new insight and important implications from the point of view of financial modelling.

More specifically, CKLS nests several linear-drift one factor models for the short rate process. The main result in the financial literature is that the more appropriate models for the US short interest rates over the period 1964 to 1989 are those that allow the conditional volatility of short interest rate changes to be highly dependent on the level of the short rate. Extensions to more complex models and to Euro-mark data produce similar results.

A simple sensitivity analysis performed by moving one single observation (out of 307) in a very small range compatible with the observed volatility, shows that the results obtained by classical econometric techniques on the original sample can be completely reversed. This is an extreme example of nonrobustness which affects in a drastic way the implications from the financial point of view. On the other hand, when testing the CKLS models with a robust methodology, we find that they are all clearly misspecified and we identify a clustering of influential observations in the 1979-1982 subperiod, a time span that is well-known to coincide with a temporary change in the monetary policy of the Federal Reserve. This clustering of influential observations does not disappear when we introduce a non-linearity in the drift and allow for a parameter shift during the 1979-1982 period. This anomalous clustering of influential points may suggest a change of structure over this period, rather than the existence of a set of isolated outliers. This confirms similar results in the literature obtained by more complex and unstable models (such as switching models). It is important to stress, however, that the robust methodology leads to this conclusion by a single analysis on the original data set and without changing the basic ideal financial model.

REFERENCES

Chan, K. G., Karolyi, G. A., Longstaff, F. A. and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term rate. *Journal of Finance*, 47, 1209-1227.

Dell'Aquila, R. and Ronchetti, E. (2006, to appear). *Robust Statistics and Econometrics With Economic and Financial Applications*. New York: Wiley.

Dell'Aquila, R., Ronchetti, E. and Troiani, F. (2003). Robust GMM analysis of models for the short rate process. *Journal of Empirical Finance*, 10, 373-397.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation.* PhD Thesis, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42, 1887-1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

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

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics,* 35, 73-101.

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

Knez, P. J. and Ready, M. J. (1997). On the robustness of size and book--to--market in cross--sectional regressions. *Journal of Finance*, 52, 1355-1382.

Maronna R., Martin, R. D., and Yohai, V. (2006, to appear). *Robust Statistics: Theory, Computation and Methods*. New York: Wiley.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to Probability and Statistics*, (pp. 448-485). Palo Alto: Stanford University Press.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33, 1-67.

Von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18, 309-348.