METRON (2021) 79:127–135
https://doi.org/10.1007/s40300-020-00185-3

The main contributions of robust statistics to statistical science and a new challenge

Elvezio Ronchetti[1]

Received: 21 April 2020 / Accepted: 8 August 2020 / Published online: 18 August 2020

Abstract
In the first part of the paper, we trace the development of robust statistics through its main contributions which have penetrated mainstream statistics. The goal of this paper is neither to provide a full overview of robust statistics, nor to make a complete list of its tools and methods, but to focus on basic concepts that have become standard ideas and tools in modern statistics. In the second part we focus on the particular challenge provided by high-dimensional statistics and discuss how robustness ideas can be used and adapted to this situation.

Keywords Data science · Statistical models · Neighborhoods of a model · Game theory · Minimax approach · Huber function · Statistical functionals · Gâteaux derivative · Fréchet derivative · Influence function · Breakdown point · M-estimators · Generalized Method of Moments · Generalized Estimating Equations · High-dimensional statistics · Saturated models · Penalized methods · Oracle properties

1 Introduction

Robust statistics deals with deviations from ideal models and their dangers for corresponding inference procedures. Its primary goal is the development of procedures which are still reliable and reasonably efficient under small deviations from the model, i.e. when the underlying distribution lies in a neighborhood of the assumed model. Therefore, one can view robust statistics as an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality.

If we consider the seminal papers [18,24,51] as the beginning of the systematic development of the theory and applications of robust statistics, the field is about 60 years old. During the period 1978 − 2017 according to the Current Index to Statistics, we can find about 8000 papers about robust methods in core journals in statistics and related fields, 2000 of which in the last period 2008 − 2017, i.e. 200 per year. Many more papers have been and are published in journals in applications fields. *Metron* has had its share during its 100 years of existence.

✉ Elvezio Ronchetti
Elvezio.Ronchetti@unige.ch

[1] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, 1211 Geneva, Switzerland

This shows not only that the field is still active, but more importantly that it has penetrated mainstream statistics. In order to evaluate its impact to the general theory and practice of statistics, we do not provide an extensive review of the field, but we focus on basic ideas, concepts, and tools developed early which are the backbone of robust statistics and have become standard tools in modern statistics and have had an important impact in its development.

The paper is organized as follows. In Sect. 2 we list and discuss the main contributions of robust statistics which have penetrated mainstream statistics and have become standard ideas and tools in modern statistics. Section 3 is devoted to the particular challenge provided by high-dimensional statistics and discusses the role of robust statistics in this situation. In the last section we draw some conclusions.

2 Main contributions of robust statistics

In this section we focus specifically on some key ideas developed in the framework of robust statistics and analyze their impact on modern statistics and data science. This is not a full review of robust statistics, but rather a list of basic ideas which originated within the robustness literature and have become standard ideas in modern statistics.

A large and rich literature on robust statistics has been developed in the past decades. An account of the basic general theory can be found in the classical books [27] (2nd edition [28]), [20,38]. Additional general books include, [40,42,44,48,56] Ch. 5, [10,17,23,30], and the quantile regression approach in [31]. A recent review is provided in [5].

2.1 Models as approximations

It is a basic tenet of science that models are only approximations to reality. However, perhaps because of the great success of statistical theory and practice starting from Fisher and continuing in the forties and the fifties, the implications of the sometimes stringent assumptions underlying the derivation of optimal statistical procedures have been somewhat neglected.

Tukey's seminal paper [51] opened the eyes of the statistical community about the dramatic loss of efficiency of optimal procedures in the presence of small deviations from the assumed stochastic model. Of course, good data analysts had been aware in the past of this danger, but Tukey's paper called for a systematic and theoretical investigation of this problem with the goal to develop procedures which are robust against such deviations. Perhaps this aspect is becoming even more important nowadays with the flourishing of (new) procedures and tools to analyze complex data.

2.2 Data analysis

Robust methods provide often multiple solutions to a given statistical (data-analysis) problem. For instance and at the very least, the data analyst has to decide how much robustness and efficiency s(he) would like to impose on a given procedure. This opens the door to possible multiple analyses of a statistical (data analysis) problem, a point among many others, stressed by Tukey in [52], a path-breaking paper on the future of data analysis. Almost 60 years later, this is an important issue in the present discussion about the role of data science; for a general discussion, see [11]. Incidentally, Tukey's paper was unique also in its form: it was much longer (67 pages) than a typical paper published in the *Annals of Mathematical Statistics* and it contained almost no explicit mathematical development.

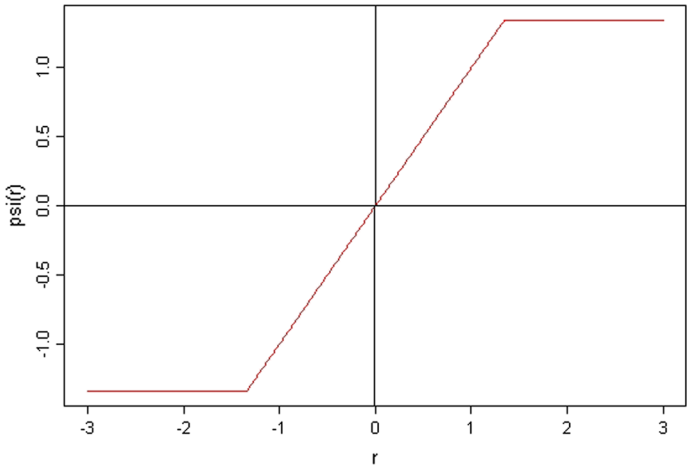

Fig. 1 The Huber function (c = 1.34)

Sometimes the possibility to provide different analyses to a given data-analytic problem is viewed as a negative point. Notice however, that the seemly uniqueness and optimality of a classical statistical procedure such as the least squares estimator in the linear model, is often obtained by paying a high price either in terms of stringent stochastic assumptions (e.g. normality of the errors) or by heavy restrictions on the class of admissible procedures (e.g restriction to *linear* estimators) as stated by the Gauss–Markov theorem.

2.3 The minimax approach

[24] was a seminal paper and contains several important contributions. Among others, Huber provided an elegant game theoretic solution in the location model, by formalizing the robustness problem as a game between the Nature, which chooses a distribution G in the neighborhood $\mathcal{F}_\epsilon(F)$ of the model F (see (2)) and the Statistician, who chooses an estimator for the location parameter in the class $\{\psi\}$ of $M-$estimators (see (1)), where the payoff is the asymptotic variance $V(\psi, G)$ of the estimator. This game has a saddlepoint $(\tilde{G}, \tilde{\psi})$, where $\tilde{\psi}$ is the Maximum Likelihood Estimator under the least favorable distribution $\tilde{G}$, i.e. the distribution minimizing the Fisher information in the neighborhood. Therefore, there exists a *minimax estimator* $\tilde{\psi}$, which solves the problem

$$\inf_{\psi} \sup_{G \in \mathcal{F}_\epsilon} V(\psi, G),$$

i.e. it minimizes the worst possible asymptotic variance of an $M-$estimator in the neighborhood $\mathcal{F}_\epsilon(F)$.

When F is the normal distribution, $\tilde{\psi}$ is the so-called *Huber function*

$$\psi_c(r) = \begin{cases} r & |r| \le c \\ c \cdot sign(r) & |r| > c, \end{cases}$$

shown in Fig. 1 and the corresponding $M-$estimator is the Huber estimator.

This formalization of the problem through minimax theory was further exploited in [25] to formalize robust testing, with an elegant interpretation in the framework of capacities and upper and lower probabilities; see [29].

2.4 Statistical functionals

Statistical functionals play a central role in Hampel's approach to robustness; see [18,19]. The basic idea is to view statistical procedures as functionals of an underlying distribution G and study their behavior in a neighborhood of a model distribution F.

Derivatives of functionals, such as Gâteaux and Fréchet derivatives, are used to linearize a functional by means of the first term of a von Mises expansion ([54]) and to describe its local stability. In particular, the influence function (the Gâteaux derivative in direction of a point mass) is a key tool to investigate the robustness properties of a statistical procedure and to construct new robust methods. Its boundedness is crucial to achieve local robustness. The importance of the influence function goes beyond its role in robust statistics. For instance, it has a strong connection with the jackknife and it appears as linearization of any asymptotically normal estimator and therefore in its asymptotic variance. These ideas have been extended to semiparametric and nonparametric models; see [8]. Statistical functionals are key concepts in modern statistics, e.g. in nonparametric statistic and in the bootstrap and other resampling methods.

2.5 *M*-estimators

M-estimators are solutions of estimating equations ([16]) defined at the population level by orthogonality or moment conditions

$$E_F[\psi(X;\beta)] = 0. \tag{1}$$

Huber ([24,26,27]) defined M-estimators as the building blocks to construct new robust estimators and investigated in detail their statistical properties. Noteworthy is his proof in [26] of the consistency and asymptotic normality of multivariate M-estimators under very weak assumptions. In this context appears the so-called sandwich estimator of the asymptotic covariance matrix of an M-estimator; see [13,26,57].

Extensions and further developments of M-estimators include the Generalized Method of Moments (when $dim(\psi) > dim(\beta)$ in (1)) by Hansen ([21]), a backbone of modern econometrics because (1) are often derived from economic theory to characterize economic models, and Generalized Estimating Equations by Liang and Zeger([34]), an important technique for the analysis of longitudinal data in biostatistics.

2.6 The breakdown point

The breakdown point ([18]) is a measure of global reliability for a statistical procedure and gives the worst percentage of contamination that a procedure can tolerate before it becomes arbitrarily biased. It provides a worst-case scenario and it can be obtained by a typical back-of-the-envelope calculation.

This concept has opened up the search for procedures with high breakdown point, which allow to separate the structure encompassing the bulk (or the majority) of data from that possibly forming an important minority group. Therefore, these are useful exploratory tools

that allow to discover patterns in the data. Their development has revisited old concepts such as the depth of a data cloud ([35,43,53]) and has open up new research directions in different areas with an important impact in data analysis and computational statistics; see e.g. the forward search in [2,3].

2.7 Teaching

In view of the points mentioned above, it seems important to include basic robustness concepts both in undergraduate and graduate curricula in statistics and data science as well as in fields of applications. This is more effective and natural than treating robust statistics as a special (exotic) and advanced topic at the graduate level. The mathematical treatment can always be adapted to the level of the course and shouldn't represent an obstacle to convey the basic ideas and tools.

For instance the influence function and the breakdown point can be viewed as familiar concepts of calculus, where the former is a derivative that can be used to linearize complicated functions, whereas the latter describes a pole of a function.

3 A challenge: high-dimensional statistics

Large and complex data sets are increasingly common in science and we face the challenge to provide suitable procedures for analyzing these data and to investigate their statistical properties. In this framework (e.g. when the number of variables diverges with the sample size) deviations from the assumptions can be expected to have a larger impact on statistical procedures and robust statistics is likely to play an important role; see the discussion about stability in [58].

3.1 Robustifying penalized methods

Let us first focus on penalized methods, which have been proved useful in particular for estimation and model selection in high-dimensional problems and have been studied extensively. Good reviews on the topic are provided by [15,50], and [22] and a more detailed discussion can be found in [9]. In particular, many results concerning e.g. oracle properties are available in linear regression assuming gaussian or sub-gaussian errors.

From the experience with methods without penalization, it is intuitively clear that penalized estimators based on classical likelihoods (such as Lasso based on a square loss function in linear regression) will be affected by outlying points and will suffer robustness problems. It is then natural to try and robustify these methods by modifying their loss function. Along these lines, several authors proposed robust versions of Lasso in linear models: [1] proposed a trimmed version, [33] provided screening method based on rank correlations, [7] proposed the Lasso penalty for quantile regression, [14] extended the latter by proposing an adaptive penalized estimator, and [37] and [36] used a redescending loss. All these papers include simulation studies that indicate that these robustified versions, are indeed robust under some type of deviations from the stochastic assumptions. However, there is not much work on the theoretical characterization of robustness for these and more general methods. Some exceptions are [1,55], where the authors study the breakdown point of some penalized methods for linear models, [4], where a rigorous definition of the influence function of penalized

$M-$estimators is provided, and [49], where the theoretical properties of an adaptive version of the Huber regression estimator is investigated.

From a theoretical point of view, it is important to investigate the behavior of penalized methods not only when the errors follow the distribution of the classical model F, but when it lies in a ϵ-neighborhood of it:

$$\mathcal{F}_\epsilon(F) = \{G = (1-\epsilon)F + \epsilon H,\ H \text{ an arbitrary distribution}\}. \tag{2}$$

Under appropriate conditions for the penalized $M-$estimator (including the boundedness of its score function and its derivative), if this bias is not too large and the minimum signal is large enough, we obtain correct support recovery and bounded bias, i.e. a robust penalized $M-$estimator behaves as well as a robust oracle by providing

- sparsity: $\hat{\beta}_2 = 0$, for large n with high probability, where β_2 is the zero component of the parameter β;
- bounded bias: in ℓ_∞-norm: $\|\hat{\beta}_1 - \beta_1\|_\infty = O(n^{-\zeta}\log n + \epsilon)$, where β_1 is the non-zero component of the parameter β;

see [6]. Notice that the score function of classical penalized methods such as Lasso is unbounded. Thus, their bias in ℓ_∞-norm in the neighborhood is infinity.

3.2 Saturation in linear models

A complementary perspective between robustness and sparsity in linear models is provided by the so-called saturated regression model (or mean-shift outlier model):

$$y_i = \sum_{j=1}^{d} x_{ij}\beta_j + \gamma_i + \varepsilon_i\ ,\quad i = 1, \ldots, n$$

where $d < n$ and the γ_i are nonzero when observation i is an outlier.

It turns out that minimizing

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d} x_{ij}\beta_j - \gamma_i\right)^2 + \sum_{i=1}^{n} p_\lambda(|\gamma_i|)$$

over β and γ for a given penalty $p_\lambda(\cdot)$, we obtain an estimator of β matching the one obtained by minimizing

$$\sum_{i=1}^{n}\rho\left(y_i - \sum_{j=1}^{d} x_{ij}\beta_j\right)$$

for some loss function $\rho(\cdot)$. This is an $M-$estimator for β with score function $\psi(\cdot)$, the derivative of $\rho(\cdot)$. For instance, the Huber estimator is obtained by using the Lasso penalty.

This idea goes back to [45] (in the case of the Huber estimator), [12,39,46,47] (in the context of approximate message passing). It has been also successfully exploited by David Hendry and coauthors in the econometrics literature (Autometrics) as a variable selection tool and more recently as an outlier detection technique.

In the past few years this approach has become a popular tool in the Machine Learning community to enforce robustness in available algorithms. We believe that its connection to $M-$estimation opens the door to a beneficial cross-fertilization between the sparse modeling literature and robust statistics.

4 Conclusion

Robust statistics has contributed in an important way to the development of modern statistics by providing many ideas, concepts, and tools that are now part of mainstream statistics. There is no doubt that robustness will follow the present development of statistics and data analysis and face the same multiple challenges. A typical case is the development of robust procedures for high-dimensional and complex problems by machine learning algorithms; see the method of median-of-means by [32] and the method of robust gradient estimation by [41].

Acknowledgements The author would like to thank the editors of the special issue and the reviewing team for their useful remarks.

Funding Open access funding provided by University of Geneva.

References

1. Alfons, A., Croux, C., Gelper, S.: Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann. Appl. Stat. **7**, 226–248 (2013)
2. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer-Verlag, New York (2000)
3. Atkinson, A.C., Riani, M., Cerioli, A.: Exploring Multivariate Data with the Forward Search. Springer-Verlag, New York (2004)
4. Avella-Medina, M.: Influence functions for penalized M-estimators. Bernoulli **23**, 3778–3796 (2017)
5. Avella Medina, M., Ronchetti, E.: Robust statistics : a selective overview and new directions. WIREs Comput. Stat. **7**, 372–393 (2015)
6. Avella Medina, M., Ronchetti, E.: Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika **105**, 31–44 (2018)
7. Belloni, A., Chernozhukov, V.: ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann. Stat. **39**, 82–130 (2011)
8. Bickel, P., Klaassen, C., Ritov, Y., Wellner, J.: Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York (1998)
9. Bühlmann, P., van de Geer, S.: Statistics for High-dimensional Data: Methods. Theory and Applications. Springer, New York (2011)
10. Clarke, B.: Robustness Theory and Application. Wiley, New York (2018)
11. Donoho, D.: 50 years of data science. J. Comput. Graph. Stat. **26**, 745–766 (2017)
12. Donoho, D., Montanari, A.: High dimensional robust M-estimation: Asymptotic variance via approximate message passing. Probab. Theory Relat. Fields **166**, 935–969 (2016)
13. Eicker, F.: Limit theorems for regression with unequal and dependent errors (1967). In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 59–82
14. Fan, J., Fan, Y., Barut, E.: Adaptive robust variable selection. Ann. Stat. **42**, 324–351 (2014)
15. Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**, 101–148 (2010)
16. Godambe, V.P.: An optimum property of regular maximum likelihood estimation. Ann. Math. Stat. **31**, 1208–1211 (1960)
17. Greco, L., Farcomeni, A.: Robust Methods for Data Reduction. Chapman and Hall/CRC, Boca Raton, London, New York (2015)

18. Hampel, F.: Contribution to the theory of robust estimation. Ph.D Thesis, University of California, Berkeley (1968)
19. Hampel, F.R.: The influence curve and its role in robust estimation. J. Am. Stat. Assoc. **69**, 383–393 (1974)
20. Hampel, F.R., Ronchetti, E., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
21. Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**, 1029–1054 (1982)
22. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
23. Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.: Robust Methods in Biostatistics. Wiley, New York (2009)
24. Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
25. Huber, P.J.: A robust version of the probability ratio test. Ann. Math. Stat. **36**, 1753–1758 (1965)
26. Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions (1967). In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L.M. LeCam, J. Neyman (Eds.), Berkeley: University of California Press, 221–233
27. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
28. Huber, P.J., Ronchetti, E.: Robust Statistics, 2nd edn. Wiley, New York (2009)
29. Huber, P.J., Strassen, V.: Minimax tests and the Neyman-Pearson Lemma for capacities. Ann. Stat. **1**, 251–263 (1973)
30. Jurecková, J., Picek, J.: Robust Statistical Methods. Chapman & Hall/CRC, New York (2006)
31. Koenker, R.: Quantile Regression. Cambridge University Press, (2005)
32. Lecué, G., Lerasle, M.: Robust machine learning by median-of-means: Theory and practice. Ann. Stat. **48**, 906–931 (2020)
33. Li, G., Peng, H., Zhang, J., Zhu, L.: Robust rank correlation based screening. Ann. Stat. **40**, 1846–1877 (2012)
34. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
35. Liu, R.Y.: On a notion of data depth based on random simplices. Ann. Stat. **18**, 405–414 (1990)
36. Loh, P.L.: Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann. Stat. **45**, 866–896 (2017)
37. Lozano, A., Meinshausen, N.: Minimum distance lasso for robust high-dimensional regression. Electron. J. Stat. **10**, 1296–1340 (2016)
38. Maronna, R.A., Martin, D.R., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)
39. McCann, L., Welsch, R.E.: Robust variable selection using least angle regression and elemental set sampling. Comput. Stat. Data Anal. **52**, 249–257 (2007)
40. Morgenthaler, S., Tukey, J.: Configural Polysampling: a Route to Practical Robustness. Wiley, New York (1991)
41. Prasad, A., Suggala, A.S., Balakrishnan, S., Ravikumar, P.: Robust estimation via robust gradient estimation. J. R. Stat. Soc. Ser. B **82**, 601–627 (2020)
42. Rieder, H.: Robust Asymptotic Statistics. Springer, New York (1994)
43. Rousseeuw, P.J., Hubert, M.: Regression depth (with discussion). J. Am. Stat. Assoc. **94**, 388–433 (1999)
44. Rousseeuw, P.J., Leroy, A.M.: Robust Regression & Outlier Detection. Wiley, New York (1987)
45. Sardy, S., Tseng, P., Bruce, A.: Robust wavelet denoising. IEEE Trans. Signal Process. **49**, 1146–1152 (2001)
46. She, Y., Chen, K.: Robust reduced-rank regression. Biometrika **104**, 633–647 (2017)
47. She, Y., Owen, A.: Outlier detection using nonconvex penalized regression. J. Am. Stat. Assoc. **106**, 626–639 (2011)
48. Staudte, R., Sheather, S.: Robust Estimation and Testing. Wiley, New York (1990)
49. Sun, Q., Zhou, W.X., Fan, J.: Adaptive Huber regression. J. Am. Stat. Assoc. **115**, 254–265 (2020)
50. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Ser. B **73**, 273–282 (2011)
51. Tukey, J.W.: A survey of sampling from contaminated distributions (1960). In: Contributions to Probability and Statistics, I. Olkin (Ed.), Stanford University Press, 448–485
52. Tukey, J.W.: The future of data analysis. Ann. Math. Stat. **33**, 1–67 (1962)
53. Tukey, J.W.: Mathematics and the picturing of data. Proc. Int. Cong. Math. **2**, 523–531 (1975)
54. von Mises, R.: On the asymptotic distribution of differentiable statistical functions. Ann. Math. Stat. **18**, 309–348 (1947)

55. Wang, X., Jiang, Y., Huang, M., Zhang, H.: Robust variable selection with exponential squared loss. J. Am. Stat. Assoc. **108**, 632–643 (2013)
56. Welsh, A.H.: Aspects of Statistical Inference. Wiley, New York (1996)
57. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica **48**, 817–838 (1980)
58. Yu, B.: Stability. Bernoulli **19**, 1484–1500 (2013)