IMS Lecture Notes–Monograph Series
Optimality: The Third Erich L. Lehmann Symposium
Vol. 57 (2009) 31–46

DOI: 10.1214/09-LNMS5705

On the Non-Optimality of Optimal Procedures

Peter J. Huber[1]

Abstract: This paper discusses some subtle, and largely overlooked, differences between conceptual and mathematical optimization goals in statistics, and illustrates them by examples.

Contents

1. Introduction

First, we shall identify those parts of statistics that rely in a crucial fashion on optimization. The most conspicuous among them are: classical mathematical statistics, decision theory, and Bayesian statistics.

Classical mathematical statistics was created by R. A. Fisher [9], in a paper concerned with estimation, and by J. Neyman and E. S. Pearson [32], in a paper concerned with testing. It was brought to completion by E. L. Lehmann in his lecture notes (1949, 1950); those notes later grew into two books [30, 31]. Around the same time when Lehmann produced his lecture notes, A. Wald [38] expanded the scope of mathematical statistics by creating statistical decision theory.

[1]POB 198, 7250 Klosters, Switzerland, e-mail: peterj.huber@bluewin.ch
Keywords and phrases: optimality, superefficiency, optimal robustness, breakdown point, optimal design, Bayesian robustness.

The central concerns of classical mathematical statistics were *efficiency* in estimation (i.e. minimum variance), and *power* in testing problems, both being optimality concerns. Decision theory confirmed the central interest in optimality, but shifted the emphasis to *admissibility* and *minimaxity*.

A heavy slant towards optimality, of a different origin, holds also for Bayesian statistics. For a given model, consisting of a prior distribution and a family of conditional distributions, the Bayes formula by definition gives the "best" procedure; it is admissible in decision theoretic terminology.

The above-mentioned three areas of statistics appear to be the only ones where optimality is central to the theory. Elsewhere, optimality seems to provide mere icing on the cake. Note that the papers of Fisher and of Neyman-Pearson imprinted subsequent generations of statisticians with an (often uncritical) love of optimality. By 1960, as a young mathematical statistician you would not dare submit a new procedure to a reputable journal, unless you could prove some optimality property. (Later, there was a reversal, and too many statistical algorithms may have slipped through the editorial gates with enthusiastic but inadequately substantiated claims.)

2. On Optimization and Models

Mathematical optimization always operates on some model. Models are simplified approximations to the truth; the hope is that optimality at the model translates into approximate optimality at the true situation. In the sciences, the main purpose of models is different: they are to assist our conceptual understanding, and to help with communication of ideas.

In traditional statistics there is no methodology for assessing the *adequacy* of a model. At best, traditional statistics can reject a model through a goodness-of-fit test — and Bayesian statistics cannot even do that. A non-rejected model is not necessarily adequate, and even more embarrassing, a rejected model sometimes may provide a perfectly adequate approximation.

3. Classical Mathematical Statistics and Decision Theory

Classical mathematical statistics provides a clean theory under very restrictive assumptions, such as restricting the underlying models to exponential families, or the procedures to unbiasedness or invariance.

Decision theory clarified the classical views and reduced the dependence on restrictions. It also opened new areas, in particular optimal design theory (Kiefer [26]). But on the whole, decision theory was less successful than originally hoped. The two principal success stories are the Stein estimates (James and Stein [25]), relating to admissibility, and robustness (Huber [16]), relating to minimaxity.

4. Tukey's 1962 Paper

In his long 1962 paper "The Future of Data Analysis", while ostensibly talking about his personal predilections, Tukey actually redefined the field of statistics. Tukey's central theme was his emphasis on *judgment* (Section 7). At the same time, he played down the importance of mathematical rigor and optimality (Sections 5 and 6). Possibly the most important issue worked out in his long and multifaceted paper was that there is more to theoretical statistics than mathematical

statistics. This reminds one of Clausewitz [3, 4], who castigated the theorists of military strategy of his time because they "considered only factors that could be mathematically calculated".

In his paper, Tukey eschewed models. Why? Perhaps because in traditional statistics models erroneously are considered as *substitutes* for the truth, rather than as *simplified approximations*. Note in particular his quote of Martin Wilk at the end of Section 4: "The hallmark of good science is that it uses models and 'theory' but never believes them".

Tukey of course was not the first to question the role of models and of optimality. Statistical methods based on ranks and nonparametrics had become popular precisely because they avoided dependence on uncertain models and were valid under weaker assumptions, even if they lacked the flexibility and wide applicability of the parametric approaches.

But the problems with models and optimality go deeper. They have less to do with the idealized models *per se*, but more with the procedures optimized for them.

5. Pitfalls of Optimality

There are four basic pitfalls, into which mathematically optimal procedures can be trapped:

(i) the Fuzzy Concepts Syndrome:
sloppy translation of concepts into mathematics,
(ii) the Straitjacket Syndrome:
overly restrictive side conditions,
(iii) the Scapegoat Syndrome:
confuse the model with the truth,
(iv) the Souped-Up Car Syndrome:
optimize speed and produce a delicate gas-guzzler.

These pitfalls affect distinct, very different aspects of statistics, namely: (i) concepts, (ii) procedures, (iii) models, and (iv) target functions. The list of course is not exhaustive. The pitfalls shall be discussed with the help of ten examples:

Classical:

(1) superefficiency
(2) unbiasedness, equivariance
(3) efficiency at the model

Robustness:

(4) asymptotics for finite ε
(5) finite n, finite ε
(6) asymptotics for infinitesimal ε
(7) optimal breakdown point

Design:

(8) optimal designs
(9) regression design and breakdown

Bayesian statistics:

(10) Bayesian robustness

6. Examples from Classical Statistics

The three "classical" examples (1)-(3) neatly illustrate the first three pitfalls.

6.1. The Fuzzy Concepts Syndrome

Problems caused by the Fuzzy Concepts Syndrome mostly are relics from earlier development stages of statistical theories. In a conference on Directions for Mathematical Statistics, I had argued (Huber [22]): "In statistics as well as in any other field of applied mathematics [...] one can usually distinguish (at least) three phases in the development of a problem. In Phase One, there is a vague awareness of an area of open problems, one develops *ad hoc* solutions to poorly posed questions, and one gropes for the proper concepts. In Phase Two, the right concepts are found [...]. In Phase Three, the theory begins to have a life of its own, [...] and its boundaries of validity are explored by leading it *ad absurdum*; in short, it is squeezed dry." In the 1970s there had been widespread anxiety about the future of mathematical statistics. As a deeper reason for this anxiety I had proposed the diagnosis that too many of the then current activities belonged to the later stages of Phase Three.

In the groping phase, somewhat reckless heuristics can be beneficial. The concepts inevitably are fuzzy, and correspondingly, they are translated into mathematics in a sloppy fashion. But recklessness, fuzziness and sloppiness should be cut down at the latest at the beginning of the squeezing phase (the "consolidation phase", in Erich Lehmann's terminology). Though, in the later phases it is tempting to concentrate on the mathematical formalism and to neglect a re-examination of its conceptual origins. And admittedly, even in the mathematical formalism, any attempts to eliminate sloppiness in its entirety will lead to an admirable, but non-teachable theory, as already Whitehead and Russell with their monumental *Principia Mathematica* [39] have demonstrated.

In mathematical statistics, asymptotics is exceptionally prone to sloppiness. Details notoriously are not adequately elaborated. Indeed, the expression "asymptotic theory" itself is used misleadingly. In standard mathematical usage asymptotic theory ordinarily is concerned with asymptotic expansions. Statistics knows such expansions too (e.g. Edgeworth expansions), but mostly, "asymptotic theory" denotes what more properly should be called "limiting theory". A few examples follow.

- Remainder terms? With asymptotic expansions, the first neglected term gives an indication of the size of the error. In statistics, asymptotic results hardly ever are complemented by remainder terms, however crude. That is, whatever the actual sample size is, we never know whether an asymptotic result is applicable.
- What kind of asymptotics is appropriate? In regression, for example, we have n observations and p parameters. Should the asymptotics be for p fixed, $n \to \infty$, or for $p/n \to 0$, or for what?
- Order of quantifiers and limits? Usually, one settles on an order that makes proofs easy.

Example 1. Perhaps the most illuminating case of the Fuzzy Concepts Syndrome has to do with superefficiency. There is a famous pathological example due to Hodges (see LeCam [28]). Assume that the observations $(x_1, \ldots, x_n)$ are i.i.d. nor-

mal $\mathcal{N}(\theta, 1)$. Estimate θ by

$$\begin{aligned} T_n &= \bar{x}, \quad \text{if } |\bar{x}| \geq n^{-1/4}, \\ T_n &= \bar{x}/2, \text{ if } |\bar{x}| < n^{-1/4}. \end{aligned}$$

Then T_n is consistent for all θ, with asymptotic variance n^{-1} for $\theta \neq 0$, but $\frac{1}{4}n^{-1}$ for $\theta = 0$. That is, the estimate T_n is efficient everywhere, but superefficient at 0. See Lehmann ([31], p. 405–408) for a discussion of various responses to the unpleasantness caused by Hodges' example.

Informally, asymptotic efficiency means that in large samples the variance of the estimate approaches the information bound, and this for all θ. Everyday language is notoriously ambiguous about the order of the quantifiers. For example we may spell out asymptotic efficiency as:

$$(\forall \varepsilon > 0)\ (\forall \theta)\ (\exists n_0)\ (\forall n > n_0)\ \{T_n \text{ is } \varepsilon\text{-efficent}\}, \tag{1}$$

where we define ε-efficiency by, say,

$$\{T_n \text{ is } \varepsilon\text{-efficent}\} \ = \ \left\{E_\theta\left(n(T_n - \theta)^2\right) < 1/I(\theta) + \varepsilon\right\}. \tag{2}$$

But then, for any fixed n, T_n might be arbitrarily bad for some θ. Therefore, since we do not know the true value of θ, we never will know whether an estimate satisfying (1) is any good, however large n is. In other words: while the definition of asymptotic efficiency may be technically in order, it is conceptually inacceptable.

On closer inspection we conclude that the order of quantifiers in (1) does not correspond to our intuitive concept of asymptotic efficiency. An improved version is obtained by interchanging the second and third quantifiers:

$$(\forall \varepsilon > 0)\ (\exists n_0)\ (\forall \theta)\ (\forall n > n_0)\ \{T_n \text{ is } \varepsilon\text{-efficent}\}. \tag{3}$$

It turns out that this version excludes superefficiency.

But version (3) still is negligent. Conceptually, unbounded loss functions are unsatisfactory. Technically, the awkward fact is that for very long-tailed distributions, the expectation in (2) may fail to be finite for all n and all "reasonable" estimators (i.e. for all estimators T_n whose value is contained in the convex hull of the observations, cf. Huber [20], p. 1047), while the limiting distribution exists and has a finite variance. To obtain a definition of asymptotic efficiency working more generally we might rewrite (3) to

$$(\forall c > 0)(\forall \varepsilon > 0)(\exists n_0)(\forall \theta)(\forall n > n_0)\left\{E_\theta\left(\left(\left[\sqrt{n}(T_n - \theta)\right]_{-c}^{+c}\right)^2\right) < 1/I(\theta) + \varepsilon\right\}. \tag{4}$$

Here, $[x]_a^b = max(a, \min(b, x))$. Of course, (4) is not yet the final word; for example, we might want to replace the global uniform bound by a local one.

In my opinion Hodges' example should not be considered as a local improvement of the standard estimate, comparable to the James-Stein estimate, but rather as an ingenious spotlight on a conceptual inadequacy of the traditional formalization of asymptotic efficiency. This interpretation is not new. In particular, the crucial technical result, namely that one-sided locally uniform bounds suffice to prevent superefficiency, had been published in an abstract more than 40 years ago (Huber [18]). But I never had found a congenial outlet for the philosophical side of the result; it took the present symposium to provide one.

6.2. The Straitjacket Syndrome

Example 2. Classical examples of the Straitjacket Syndrome, that is of overly restrictive side conditions on the procedures, are well known and do not need a detailed discussion. One is furnished by unbiasedness: unbiased estimates may not exist, or they may be nonsensical, cf. Lehmann ([31], p. 114). Other examples occur with invariance (more properly: equivariance): equivariant estimates may be inadmissible (Stein estimation).

6.3. The Scapegoat Syndrome

This subsection is concerned with excessive reliance on idealized models. The word "scapegoat" refers to the pious belief that the gods of statistics will accept the model as a substitute for the real thing.

As statisticians, we should always remember that models are simplified approximations to the truth, not the truth itself. Sometimes they are not even that, namely when they are chosen for ease of handling rather than for adequacy of representation; typical examples are conjugate priors in Bayesian statistics. The following eye-opening example gave rise to robustness theory.

Example 3. In 1914, Eddington had advocated the use of mean absolute deviations, against root-mean-square (RMS) deviations, as estimates of scale. Fisher [8] objected and showed that for normal errors RMS deviations are 12% more efficient. Tukey [36] then pointed out that for the contaminated normal error model

$$F(x) = (1-\varepsilon)\Phi(x/\sigma) + \varepsilon\Phi(x/(3\sigma)) \tag{5}$$

mean absolute deviations are more efficient for all $0.002 < \varepsilon < 0.5$.

The unfortunate fact is that errors in real data typically are better approximated by a contamination model with a contamination rate ("gross error rate") in the range $0.01 < \varepsilon < 0.1$, than by the normal model.

The main lesson to be learnt from the Eddington–Fisher–Tukey example is that the standard normal error model may be quite accurate, especially in the center of the distribution. The problem is that the tail behavior of real data, to which the traditional estimates are highly sensitive, usually is rather indeterminate and difficult to model. The mistake of Fisher (and others) had been to treat the standard model as the exact truth.

We note a few conclusions from such examples:

- Optimality results put in evidence what can (and what cannot) be achieved in an *ideal* world.
- Notoriously, optimal procedures are unstable under *small deviations* from the ideal situation.
- The task thus is to find procedures that achieve *near optimality* under the ideal situation, but that are more *stable* under small deviation.

In 1964, I had begun to implement a program suggested by this under the name of robustness. The guiding ideas were:

– Keep the optimality criterion (asymptotic variance, ...).
– Formalize small deviations (ε-contamination, ...).
– Find best sub-optimal procedures (best in a minimax sense).

The robustness notion I had adopted corresponds to Tukey's 1960 version. Though, this is not the unique interpretation of robustness occurring in the literature. In the 1970's, under Tukey's influence, there was a semantic shift, adopted by many, namely that the purpose of robustness was to provide procedures with a strong performance for a *widest possible* selection of heavy-tailed distribution.

But I still prefer the original 1960 version. In particular, I hold that robustness should be classified with *parametric* procedures, and that *local stability* in a neighborhood of the parametric model is the basic, overriding requirement.

7. Problems with Optimality in Robustness

Robustness had been designed to safeguard against pitfalls of optimal procedures. But optimal robustness is vulnerable to the very same pitfalls, and there are even a few new variants. The conceptual problem mentioned below in Example 4, and its solution described in Example 5, both have received less resonance in the robustness literature than they would have deserved. While the influence function without doubt is the most useful heuristic tool of robustness, one ought to be aware that optimality results based on it are no better than heuristic (Example 6).

7.1. Asymptotic Robustness for Finite $\varepsilon > 0$

Example 4. In the decision theoretic formalization of my 1964 paper I had imposed an unpleasant restriction on Nature by allowing only symmetric contaminations. The reason for this was that asymmetric contamination causes a bias term of the order $O(1)$. Asymptotically, this bias then would overpower the random variability of the estimates (which typically is of the order $O(n^{-1/2})$). Automatically, this would have led to the relatively inefficient sample median as the asymptotically optimal estimate. On the other hand, for the sample sizes and contamination rates of practical interest, the random variability usually is more important. Simultaneously, the symmetry assumption had permitted to extend the parameterization to the entire ε-neighborhood and thereby had made it possible to maintain a standard point-estimation approach.

The assumption of exact symmetry is repugnant, it violates the very spirit of robustness. Though, restrictions on the distributions are much less serious straitjackets than restrictions on the procedures (such as unbiasedness). The reason is that after performing optimization under symmetry restrictions, one merely has to check that the resulting asymptotically "optimal" estimate remains nearly optimal under more realistic asymmetric contaminations, see Huber ([23], Section 4.9).

Curiously, people have worried (and still continue to worry!) much more about the symmetry straitjacket than about a conceptually much more serious problem. That problem is that 1% contamination has entirely different effects in samples of size 10 or 1000. Thus, asymptotic optimality theory need not be relevant at all for modest sample sizes and contamination rates, where the expected number of contaminants is small and may fall below 1. Fortunately, this question could be settled through an exact finite sample theory – see the following example. This theory also put to rest the problem of asymmetric contamination.

7.2. Finite Sample Robustness for Finite $\varepsilon > 0$

Example 5. To resolve the just mentioned conceptual problem, one needs a finite sample robustness theory valid for finite $\varepsilon > 0$. Rigorous such theories were devel-

oped early on, see Huber [17] for tests and Huber [19] for estimation. The latter covers the same ground as the original asymptotic robustness theory, namely single parameter equivariant robust estimation. Gratifyingly, it leads to procedures that are qualitatively and even quantitatively comparable to the M-estimators obtained with the asymptotic approach.

This finite sample approach to robustness does not make any symmetry assumptions and thus also avoids the other objections that have been raised against asymptotic robustness theory. In particular, by aiming not for point estimates, but for minimax interval estimates, it bypasses the parameterization and asymmetry problems. Despite its conceptual importance, the finite sample theory has attained much less visibility than its asymptotic and infinitesimal cousins. I suspect the reason is that the approach through an unconventional version of interval estimates did not fit into established patterns of thought. In the following I shall sketch the main ideas and results; for technical details see Huber [19].

Just as in the original asymptotic theory, we consider the one-parameter location problem and assume that the error distribution is contained in an ε-neigborhood of the standard normal distribution. The optimally robust finite sample estimator turns out to be an M-estimate T defined by

$$\sum \psi(x_i - T) = 0, \tag{6}$$

where $\psi(x) = [x]_{-k}^{k} = \max(-k, \min(k, x))$ for some $k > 0$. But instead of minimizing the maximal asymptotic variance, this estimator is optimal in the sense that it minimizes the value α for which one can guarantee

$$P\{T < \theta - a\} \leq \alpha, \quad P\{T > \theta + a\} \leq \alpha \tag{7}$$

for all θ and all distributions in the ε-neighborhood.

We have three free parameters, n, ε and a. Interestingly, the characteristic parameter k of the ψ-function depends only on ε and a, but not on the sample size n. In (7), instead of minimizing α for fixed a, we might alternatively minimize a for fixed α. The asymptotic theory can be linked to these exact finite sample optimality results in several different fashions. In particular, if we let $n \to \infty$, but keep both α and k fixed, then a and ε of the optimally robust estimates go to 0 at the rate $O(n^{-1/2})$. Conceptually, ε-neighborhoods shrinking at a rate $O(n^{-1/2})$ make eminent sense, since the standard goodness-of-fit tests are just able to detect deviations of this order. Larger deviations should be taken care of by diagnostics and modeling, while smaller ones are difficult to detect and should be covered (in the insurance sense) by robustness.

7.3. Asymptotic Robustness for Infinitesimal ε

Example 6. Parametric families more general than location and scale are beyond the scope of the above approaches to robustness. Hampel proposed to attack them via gross error sensitivity: minimize asymptotic variance at the model, subject to a bound on the influence function (see Hampel [13], and Hampel *et al.* [15]). This approach is infinitesimal in nature and stays strictly at the parametric model. In essence, it is concerned only with the limiting case $\varepsilon = 0$.

Heuristically, it combines two desirable properties of robust estimates: good efficiency at the model, and low gross error sensitivity. However, a bound on the latter *at the model* does not guarantee robustness (local stability in a neighborhood of the

model), there are counter examples with L-estimates, see Huber ([23], Section 3.5). Thus, the conceptual basis of this approach is weak. Even if it should yield robust procedures, we have no guarantee that they are approximately optimal for non-zero ε, and we thus have to pray to the statistical gods that they will accept an infinitesimal scapegoat. As a minimum, one ought to check the breakdown point of procedures constructed by this method.

There is a conceptually more satisfactory, but technically more complicated alternative approach via shrinking neighborhoods: while $n \to \infty$, let $\varepsilon \to 0$ at the rate $O(n^{-1/2})$. This particular asymptotic theory had been motivated by the finite sample approach of Example 5. It was introduced by C. Huber-Carol in her thesis [24] and later exploited by Rieder in several papers, culminating in his book [33]. The limiting results are comparable to those obtained with the infinitesimal approach, and like these, in the location parameter case they agree with those obtained in Example 4.

The principal heuristic appeal of the shrinking neighborhood approach is that in the location case it yields a sequence of estimates that have a well-defined optimality property for each n. We therefore can hope that in the general case it yields a sequence of estimates that are approximately optimal for non-zero ε. But to be honest, we have no way to check whether the heuristic arguments reliably carry beyond the location case. That is, we may run into a fifth pitfall: overly optimistic heuristics.

7.4. Optimal Breakdown Point

Hampel, at that time a student of Erich Lehmann, in his Ph.D. thesis (1968) had introduced the breakdown point by giving it an asymptotic definition. Conceptually, this may have been misleading, since the notion is most useful in small sample situations, see Donoho and Huber [6]. With large samples and high contamination rates you may have enough data to interpret the information contained in the contamination part. Therefore, rather than blindly using high breakdown point procedures, you may spend your efforts more profitably on an investigation of mixture models.

Example 7. All standard regression estimates, including the one based on least absolute deviations (the L_1-estimate, which generalizes the highly robust sample median), are sensitive to observations sitting at influential positions ("leverage points"). A single bad observation at an extreme leverage point may cause breakdown. Clearly, a higher breakdown point would be desirable. How large can it be made, and how large should it be? Via projection pursuit methods it is indeed possible to approach a breakdown point of $1/2$ in large samples, provided the data are in general position (i.e., under the idealized, uncorrupted model no p rows of the n-by-p design matrix are linearly dependent, and thus any p observations give a unique determination of θ). This is a result of considerable theoretical interest.

Unfortunately, all estimators that try to optimize the breakdown point seem to run into the Souped-up Car Syndrome. The first among them was the LMS-estimate (Hampel [14], Rousseeuw [34]).

The LMS- (Least Median of Squares) estimate of θ modifies the least squares approach by minimizing the *median* instead of the *mean* of the squared residuals:

$$\text{median}\left\{(y_i - x_i^T\theta)^2\right\}. \tag{8}$$

If the data points are in general position, its breakdown point is $([n/2]-p+2)/n \to 1/2$. But it has the following specific drawbacks:

- Its efficiency is low: the dispersion of the estimate decreases at the rate $n^{-1/3}$, instead of $n^{-1/2}$.
- Its computational complexity increases exponentially with p.
- What if the points are not in general position?

My conclusion is that an asymptotic theory for large p and n does not make much sense under such circumstances, and a small sample theory is not available.

S-estimates were introduced by Rousseeuw and Yohai [34] to overcome some of these shortcomings. They estimate θ by minimizing a suitable robust M-estimate of the scale of the residuals. Under suitable regularity conditions their breakdown point also approaches 1/2 in large samples, and they reach a high efficiency at the ideal model, with a dispersion converging at the rate $n^{-1/2}$. Unfortunately, S-estimators suffer from a serious flaw which probably cannot be removed, namely that uniqueness and continuity can only be proved under certain conditions, see Davies ([5], Section 1.6).

Moreover, Davies (*ibid.*) points out that all known high breakdown point estimators of regression are inherently unstable. Paradoxically, it thus seems that in order to achieve an optimal regression breakdown point we may have to sacrifice robustness.

8. Design Issues

8.1. Optimal Designs

Example 8. Assume that the task is to fit the best possible straight line to data originating from an exactly linear function. Then the optimal regression design puts all observations on the extreme points of the segment where observations can be made.

However, a possibly more realistic version of this task is to fit the best straight line to an *approximately* linear function. In either case, one would want to make something like the expectation of the integrated mean square error as small as possible. Of course, usually one does not know much about the small deviations from linearity one might have to cope with (and does not care about them, so long as they are small). Already Box and Draper ([2], p. 622) had recognized the crux of the situation and had pointed out: "The optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if *variance were ignored completely* and the experiment designed so as to *minimize bias alone*."

In other words, the "naive" design, which distributes the observations evenly over the accessible segment, in such a situation should be very nearly optimal, since it minimizes the integrated squared bias of the fit. Apart from that, it has an advantage over the optimal design since its redundancy allows to check the linearity assumption.

These aspects have been made precise in a minimax sense by Huber [21]. The most surprising fact emerging from this study was: there is a range where the deviation from linearity is slight enough to stay below statistical detectability, yet large enough so that the "naive" design will outperform the "optimal" design and give a better linear approximation to the true function. Even though the effects are much less dramatic than in Example 3, we evidently have run here into another example of the Scapegoat Syndrome.

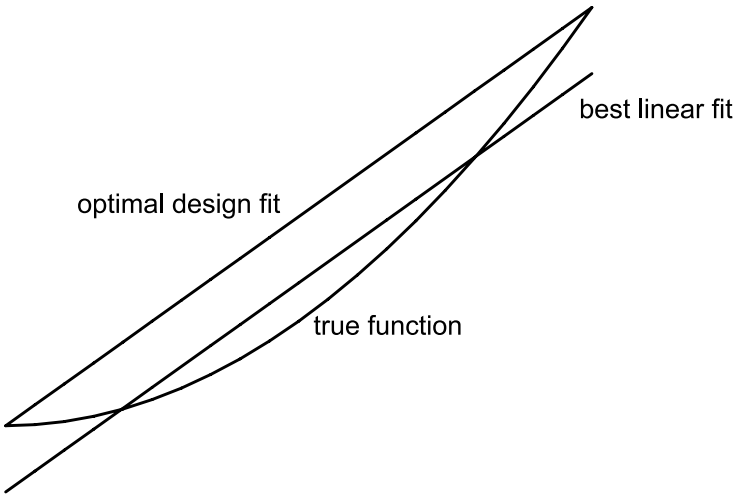

FIG 1. *Optimal design fit (based on the theoretically optimal design) and best linear fit (minimizing the integrated squared error), to a not-quite linear function. Observational errors are neglected in this figure.*

8.2. Regression Design and Breakdown

Example 9. In higher dimensions, generalizing the preceding example, an optimal linear regression design would place an equal number of m observations onto each of the $(p+1)$ corners of a p-dimensional simplex. Technically, such a design is optimal, but again, it lacks redundance.

For such a design the best possible breakdown point is

$$\varepsilon^* = \lceil m/2 \rceil / (m(p+1)) \approx 1/(2(p+1)). \tag{9}$$

This breakdown point is attained by the L_1-estimate (calculate the median at each corner). The so-called high-breakdown point LMS- and S-estimates cannot do any better.

But already an arbitrarily small jittering of the design points will bring them into general position. Then the breakdown point of LMS and S is close to $1/2$. How can this happen?

On closer inspection we see that the high breakdown point of LMS- and S-estimates is achieved by extrapolation: at each corner, you put more faith in the value extrapolated from the mp observations clustering near the far-away other p corners, than in the m local values. The fitted hyperplane thus not only loses efficiency, but becomes sensitive to small errors affecting a majority of the observations, such as rounding.

The conclusion is that high breakdown point regression is not necessarily robust. We have a clear case of the Souped-up Car Syndrome: both extremes, optimal design and optimal breakdown point, lead to estimates with undesirable properties, and a compromise is called for. A quantitative, *design-dependent* theory of robust regression would seem to be needed. The customary assumption underlying all high breakdown point regression theories in the published literature, namely that the regression carrier is a random sample from a suitable multi-dimensional continuous distribution, in my opinion is much too narrowly conceived.

9. Bayesian Statistics

Example 10. What is Bayesian robustness? Bayesian statistcs has a built-in problem with the Scapegoat Syndrome, that is, with over-reliance on the model; this problem becomes acute in connection with robustness. By definition, Bayes procedures are optimal for the chosen model, consisting of a prior $\alpha(\theta)$ and a family of conditional densities $f(x, \theta)$. Instability, and conversely robustness, thus are properties of the *model*. This was emphasized in 1978 by George Box in an illuminating, facetious but profound oral interchange with John Tukey at an ARO meeting on Robustness in Statistics (Launer and Wilkinson, [27]). Box maintained that, after all, *he* had invented the term (see Box [1]), and that he could define it as he pleased, and that in his opinion robustness was to be achieved by choosing a proper model, not by tampering with the data (by trimming or Winsorizing) as Tukey was wont to do. He did not elaborate on how to choose such a model.

The philosophical problem of Bayesian statistics is that it is congenitally unable to separate the model, the underlying true situation, and the statistical procedure. It acts as if the model were exactly true, and it then uses the corresponding optimal procedure. A fundamentalist Bayesian, for whom probabilities exist only in the mind, will not be able to see that there is a problem of the Scapegoat type; it takes a pragmatist like George Box to be aware of it.

I shall now attempt to sketch a way around this Bayesian Scapegoat Sydrome. The central question is: what is a *robust model*? *Ad hoc* parametric supermodels, which sometimes are advertised as *the* Bayesian approach to robustness, do not guarantee robustness. There are no reliable guidelines to select such models, and the resulting procedures may suffer from instabilities.

If we proceed pragmatically, then, as a minimum requirement, the statistical conclusions from the model ought to be insensitive to occasional outliers. Sensitivity studies *à la* Berger, that is: admit that the specifications are inaccurate and find the range of implied conclusions (see Wolpert [40], p. 212), may reveal the presence of outliers: if there are outliers, small changes in the tails of the model $f(x, \theta)$ can produce large effects. Also, they may reveal conflicts between the prior and the observational evidence: if the observational evidence points to a θ far from the center of the prior, small changes in the tails of the latter can produce large effects. Thus, if a sensitivity analysis shows that the range of implied conclusions is narrow, any model in the uncertainty range will do. If not, we better choose a robust model. But then, why not choose a robust model right away? Unfortunately, sensitivity studies do not help us find a robust model.

The following is a proposal for an informal portmanteau definition of robustness, covering both Bayesian and non-Bayesian statistics:

> *Uncertain parts of the evidence should never have overriding influence on the final conclusions.*

This is supposed to apply not only to questionable data (outliers), but also to uncertainties in the model densities $f(x, \theta)$ and to uncertainties in the prior $\alpha(\theta)$, and even to vagueness in the specification of the target loss function.

How to implement such a loose definition? The first two of the above four requirements are interconnected and tricky to separate: insensitivity to dubious data features (gross errors), and insensitivity to uncertain model specifications. I claim that the following implementation should do the job for both aspects: Choose a model $f(x, \theta)$ *within the uncertainty range,* such that the conclusions are insensitive to gross errors. This has to be made precise.

The mode of the posterior density solves

$$\alpha'(\theta)/\alpha(\theta) + \sum f'(x_i, \theta)/f(x_i, \theta) = 0, \tag{10}$$

where the prime denotes the derivative with respect to θ. For a flat prior, the mode of the posterior coincides with the maximum likelihood estimate.

As Freedman [10] has expressed it, there is a "striking and mysterious fact", namely that asymptotically, Bayes and M.L. estimates behave similarly: they not only have the same asymptotic distribution, but if the true underlying distribution belongs to the parametric family, the Bayesian posterior distribution, centered at the M.L. estimate and scaled by $n^{-1/2}$, is asymptotically normal and coincides with the asymptotic distribution of the M.L. estimate, centered at the true θ and also scaled by $n^{-1/2}$. See also LeCam [29]; the result apparently goes back to Bernstein and von Mises.

Thus, if we are willing to adopt the infinitesimal approach via gross error sensitivity (see Example 6), asymptotic robustness ideas should carry over from non-Bayesian M-estimates. Though, Hampel's approach through gross error sensitivity does not apply without some caveats, since it does not automatically lead to ψ-functions that are logarithmic derivatives of probability densities (which is a necessary side condition in the Bayes context — another example of a straitjacket). Finite ε-neighborhoods need somewhat more work. Assume that the M- and Bayes estimates both are calculated on the basis of the least favorable density (instead of the unknown true underlying distribution, which is supposed to lie anywhere in the given ε-neighborhood). Then, the M- and Bayes estimates still have the same asymptotically normal distribution, but the equivalence with the asymptotic posterior is lost. Though, in the one-dimensional location case it can be shown that the asymptotic variance of the posterior then lies between the asymptotic variance of the M-estimate and the upper bound for that variance obtained from the least favorable distribution (see [23], 2nd edition, Section 15.5). — As an amusing aside on the subject of pitfalls, I might mention that the usual applications in econometrics of one the formulas relevant in this context (the so-called "sandwich formula") go so far beyond its original intention that they deserve an honorable mention in the category of overly optimistic heuristics, see Freedman [11].

In short, the heuristic conclusion, deriving from hindsight based on non-Bayesian robustness, thus is that f'/f ought to be bounded. In (10) the prior acts very much like a distinguished additional observation. Thus, in analogous fashion, also α'/α ought to be bounded. In both cases, the bounds should be chosen as small as feasible. Ordinarily, these bounds are minimized by the least informative distributions, with Fisher information used as measure of information. Thus, a possible optimization goal can be expressed:

> *A most robust Bayesian model can be found by choosing α and f to be least informative within their respective (objective or subjective) uncertainty ranges.*

For all practical purposes this is the same recipe as the one applying to the non-Bayesian case. But like there, it is difficult to implement once one wants to go beyond the location case. And if it is adopted overly literally, we might even get trapped in one the pitfalls of optimality.

10. Concluding Remarks

In the 1970s statistical theory clearly had been in the third, consolidation or "squeezing" phase of the development cycle. At present, we seem to have entered a

new cycle and seem to be in the middle of a new "groping" phase, trying to get conceptual and theoretical handles on challenging problems involving extremely large and complexly structured data sets.

I hope that this time the laxness of the groping phase will be eliminated in time, and will not be cemented into place during the consolidation phase. Perhaps it may help to keep in mind the following aphorisms on optimality and optimization. They are not new, they are re-iterating sentiments already expressed by Tukey in 1962. Those sentiments subsequently had been studiously ignored by most statisticians. I hope that this time they will fare better.

- Optimality *results* are important: they show what can (and what cannot) be achieved under ideal conditions, and in particular they show whether a given procedure still has worthwhile potential for improvement.
- Optimal *procedures* as a rule are too dangerous to be used in untempered form.
- Beware of sloppy asymptotics.
- Never confuse the idealized model with the truth.
- Do not optimize one aspect to the detriment of others.
- There are no clear-cut rules on how the tempering of optimal procedures should be done — compromises are involved, and one must rely on human judgment. But if one insists on a mathematical approach, minimizing Fisher information within a subjective uncertainty range often will do a good job, both for Bayesians and non-Bayesians.

References

[1] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335.
[2] Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654.
[3] Clausewitz, C. von (1832). *Vom Kriege*, 19th ed. (1991). Dümmler Verlag, Bonn.
[4] Clausewitz, C. von (1984). *On War*. Edited and translated by M. Howard and P. Paret. Princeton Univ. Press, Princeton, NJ.
[5] Davies, P. L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21** 1843–1899.
[6] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges eds.). Wadsworth, Belmont, CA.
[7] Eddington, A. S. (1914). *Stellar Movements and the Structure of the Universe*. Macmillan, London.
[8] Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and the mean square error. *Monthly Not. Roy. Astron. Soc.* **80** 758–770.
[9] Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philos. Trans. Roy. Soc. London, Ser. A* **222** 309–368.
[10] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.
[11] Freedman, D. A. (2006). On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors". *Amer. Statist.* **60** 209–302.

[12] Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis. Univ. California, Berkeley.
[13] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **62** 1179–1186.
[14] Hampel, F. R. (1975), Beyond location parameters: Robust concepts and methods. Proc. 40th Session I. S. I., Warsaw 1975. *Bull. Int. Statist. Inst.* **46** 375–382.
[15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence*. Wiley, New York.
[16] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.
[17] Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36** 1753–1758.
[18] Huber, P. J. (1966). Strict efficiency excludes superefficiency, (Abstract). *Ann. Math. Statist.* **37** 1425.
[19] Huber, P. J. (1968). Robust confidence limits. *Z. Wahrsch. Verw. Gebiete* **10** 269–278.
[20] Huber, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.* **43** 1041–1067.
[21] Huber, P. J. (1975a). Robustness and designs. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.). North Holland, Amsterdam.
[22] Huber, P. J. (1975b). Application vs. abstraction: The selling out of mathematical statistics. *Suppl. Adv. Appl. Prob.* **7** 84–89.
[23] Huber, P. J. (2009). *Robust Statistics*, 2nd ed. Wiley, New York.
[24] Huber-Carol, C. (1970). Etude asymptotique de tests robustes. Ph.D. thesis. Eidgen, Technische Hochschule, Zürich.
[25] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **I** 311–319.
[26] Kiefer, J. (1959). Optimum experimental designs. *J. Roy. Statist. Soc. Ser. B* **21** 272–319.
[27] Launer, R. L. and Wilkinson, G. N. (Eds.) (1979). *Proc. ARO Workshop on Robustness in Statistics, April 11–12, 1978*. Academic Press, New York.
[28] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Statist.* **1** 277–330.
[29] LeCam, L. (1957). Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.* **3** 37–98.
[30] Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
[31] Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.
[32] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London, Ser. A* **231** 289–337.
[33] Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, Berlin.
[34] Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.
[35] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-Estimators. In *Robust and Nonlinear Time Series Analysis* (J. Franke, W. Härdle and R. D. Martin, eds.). *Lecture Notes in Statistics* **26**. Springer, New York.
[36] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics* (I. Olkin, ed.). Stanford Univ. Press, Stanford, CA.

[37] Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33** 1–67.
[38] Wald, A. (1950). *Statistical Decision Functions*. Wiley, New York.
[39] Whitehead, A. N. and Russell, B. (1910–13). *Principia Mathematica* **3**. Cambridge Univ. Press.
[40] Wolpert, R. L. (2004). A conversation with James O. Berger. *Statist. Sci.* **19** 205–218.