A Robustness Sampler

Alan Welsh

Centre for Mathematics and its Applications
Australian National University
Errors

Suppose we want to estimate an unknown number \( \mu \) from \( n \) discrepant observations/measurements \( Y_1, \ldots, Y_n \). A modern formulation is to write \( Y_i - \mu \sim \text{independent } G \).

Simpson (1756, 1757) interpreted the differences \( Y_i - \mu \) as measurement errors, called \( G \) the “distribution of error” and used continuous distributions for \( G \).

Generally, the errors represent variability in the data, whatever the source(s), and not necessarily mistakes.

Simpson argued that the mean is a better estimator of \( \mu \) than a single “good” observation: \( \text{Var}(\bar{Y}) = \text{Var}(Y_i)/n \).
Outliers

Outliers are observations which stand out in a dataset. They can be

- **correct values** which are
  - **representative outliers**: extreme cases from the process of interest (Chambers, 1986)
  - an important different process e.g. Rayleigh found 7 out of 15 observations on the atomic weight of nitrogen in a separate group and discovered argon
  - a correct minority e.g. test results where only a few candidates get the correct answer.

- **gross errors**: spurious, erroneous, discordant, anomalous or aberrant observations which result from mistakes or blunders.

Outliers focus our thinking: What are we doing and why?
Outliers and Long-tailed distributions

For unimodal models, outliers occur in the tails of the distribution so can be modelled by long-tailed distributions.

Gaussian contamination model (Tukey, 1960): For $G = \Phi$,

$$F(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/3)$$

represents mild deviations from normality: either a slight fattening of the tails or (mild) gross errors.

Semiparametric gross error model (Huber, 1964):

$$F(x) = (1 - \epsilon)G(x) + \epsilon H(x)$$

with $\epsilon$ fixed and $H$ arbitrary.

It is conceptually useful to think of long-tails for representative outliers and contamination for gross errors but outlier type is usually a matter of conjecture (Anscombe, 1960).
Contamination
Rejection of Outliers

The presence of outliers may indicate a need for a different analysis of the outliers or of the rest of the data. A natural approach is to discard outliers and then analyse the remaining data (Boscovich, 1755):

astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability. (D. Bernoulli, 1777)

Bessel (1838) wrote that he never rejected observations.
Pierce (1852) and Chauvenet (1963) introduced explicit rejection rules.

LAD (Boscovich 1760, Laplace 1789) and the median (Laplace, 1774) were not specifically advocated for treating outliers.
Normal theory

The use of the mean was promoted by the fact that it is

- the least squares estimator (Legendre 1805)
- the MLE if the errors are normally distributed (Gauss 1809, actually using the mean to justify assuming normality) and the central limit theorem justifies the normal distribution for error (Laplace 1810) (Bernoulli had proposed using maximum likelihood to determine the weight to give to observations - after rejecting gross errors)
- “the most advantageous method”: under normality, among linear estimates, asymptotically the mean has smallest expected error (Laplace, 1811)
- BLUE (Gauss-Markov theorem): Laplace’s result holds exactly without normality (Gauss, 1823)
The Gauss-Markov Theorem

*In the model* \( y = \mu + \epsilon, \ \epsilon \sim (0, \sigma^2 I) \), *the mean is the best linear unbiased estimator (BLUE) of* \( \mu \). *If the distribution of error is Gaussian, linearity can be dropped.*

- **Parameter Choice**: the mean is not a sensible target parameter if it includes gross-errors (mean of \( F \) vs mean of \( G \))
- **Linearity = Normality**: away from normality, all linear estimators are poor. (Outlier rejection is a nonlinear estimator.)
- **Unbiasedness** is not always compelling: under contamination, some bias is inevitable.
Hypothesis testing

Neyman-Pearson (1928) theory of hypothesis testing focussed on optimal tests.

Pearson (1929) did a simulation (by hand) on the effect of beta, gamma and t errors on the level of one- and two-sample t-tests. The level was affected by t errors (mild long tails).

Some tests (like the t-test) have approximate robustness of level (validity = significant results are meaningful but non-significant ones are not) but usually lack robustness of power (efficiency = both kinds of results are meaningful).

These and similar results of Pearson (1931), Geary (1936) and Gayen (1950) encouraged the development of nonparametric tests (e.g. Wilcoxon 1945) designed to maintain there level.
But ...

Tests for equality of variance do not even have robustness of validity!

When I became aware that the nominal probability of type-I error for the standard test of the equality of variances of two populations is invalidated by nonnormality ... I found little consolation in the optimal properties someone once established for that test. (Scheffé, 1942)

Box (1953) argued the need for robustness: procedures which are not seriously invalidated in either level or power by violation of the assumptions.
Back to estimation

Anscombe (1960): Outlier rejection rules should not be studied as tests (rejection rates) but for their effect on the efficiency of estimators. He introduced the famous insurance analogy: a (small) loss of efficiency when there are no outliers is a premium paid for protection (increased efficiency) when there are.

Tukey (1960) considered a different problem: How do estimators optimised under a model (such as the Gaussian) perform under other distributions? He explored the ARE of estimators as a function of $\epsilon$ in the Gaussian contamination model and showed that small indistinguishable perturbations can have large effects on the efficiency of estimators.

Although Fisher (1922) had unintentionally demonstrated the lack of robustness of the sample variance, this was not understood at the time.
Summary: Origins of Robustness

Gross errors in data: these had obviously serious consequences.

Sensitivity: small changes can have serious effects.

Philosophy: exact models cannot be justified either theoretically or empirically.

Theory: stability properties (under perturbation) are interesting and important.
Goals of Robustness

Explore sensitivity and develop methods which perform well when the model holds and are insensitive to

- Deviations from assumed distributional shape
  - Gross errors
  - Small deviations (approximate shape)
  - Rounding and grouping

- Dependence

Formal theories concentrate on deviations from distributional shape (usually normality) and particularly on outliers which destroy inferences but can be relatively easily dealt with.

Correlation is probably most important but is very difficult to deal with other than by modelling and accommodating.
Tools

**Breakdown point**: the largest fraction of a sample that can be moved arbitrarily without moving the estimator to the boundary of the parameter space.

‘Breakdown’ can occur before reaching the boundary so the definition is too extreme. In practice, “arbitrarily” is often restricted, undermining the intent of the definition.

**Sensitivity curve**

\[ SC_{n-1}(x) = n \{ T_n(Y_1, \ldots, Y_{n-1}, x) - T_{n-1}(Y_1, \ldots, Y_{n-1}) \} \]

For the mean and variance, we obtain

\[ SC_{n-1}(x) = x - \bar{Y}_{n-1} \quad SC_{n-1}(x) = \frac{(n - 1)}{n} (x - \bar{Y}_{n-1})^2 - \hat{\sigma}_{n-1}^2. \]

Sensitivity curves can be computed numerically.
Influence Functions

$IF(x; F)$, the first derivative of the statistic $T$ at $F$, where $x$ plays the role of a co-ordinate in the infinite dimensional space of distributions.

The IF “describes the (approximate and standardised) effect of an additional observation in any point $x$ on a statistic $T$ given a (large) sample with distribution $F$” so the sensitivity curve is a finite sample version of the IF.

The IF allows linear Taylor series approximation to $T$ so we can learn about $T$ in a neighbourhood of $F$ by studying $T$ at $F$. (For larger deviations, we can study the IF at other distributions in the neighbourhood). The size of the neighbourhood over which the approximation works is determined by the breakdown point.

The IF of the variance of an estimator is called the change of variance function.
Requirements for robustness

• Good efficiency (small variance) under the model
• Non-zero breakdown point
• B-robustness: bounded IF (finite gross error sensitivity
  = \sup_x |IF(x; G)|)
• Small local shift sensitivity =
  \sup_{x \neq y} |IF(x; G) - IF(y; G)| / |x - y|
• V-robustness: bounded from above CVF (finite change-of-variance sensitivity
  = \sup_x CVF(x; G) / \text{Var}(\hat{T}))
Robust skate board
M-estimators

A very flexible class of estimators can be defined by estimating equations

\[ 0 = \sum_{i=1}^{n} \psi(y_i, \theta). \]

The influence function is

\[ IF(x; F) = \frac{\psi(x, \theta_0)}{E_F \psi'(Y, \theta_0)} \]

i.e. proportional to the estimating function \( \psi \). Standard arguments yield the asymptotic approximation

\[ \hat{\theta} \sim N(\theta_0, \frac{E_F \psi(Y, \theta_0)^2}{n\{E_F \psi'(Y, \theta_0)\}^2}). \]

The change-of-variance function can be computed from the approximate variance.
Outlier rejection revisited

Diagnostics

Data Cleaning
Fit the model, use diagnostic methods to find outliers, discard these if necessary and refit the model.

- outliers may be more difficult to find with classical methods
- classical methods often fail to find multiple outliers
- cleaned data are not really normal because of detection errors
- the inference ignores the effect of cleaning
- best rejection methods are inferior to best robust methods because of smooth weighting
The pretest approach

To apply a two sample t-test to compare two means

• test each sample for normality

• test for equality of variances

if neither hypothesis is rejected, do the t-test.

What do we do if a hypothesis is rejected? Rejection may be due to unimportant deviations and is more likely in large samples. The test for equality of variances is notoriously nonrobust.

It is impossible to prove precise null hypotheses and rejection is more likely in large samples.

Type I errors (false acceptance) affect the process.

This ignores other assumptions like independence.
Supermodels
Embed the model in a larger (parametric) supermodel (e.g. Gaussian in Pearson system, student t, skew t etc) and proceed as if the supermodel holds exactly.

- supermodels with few extra parameters are not very flexible
- belief in an exact supermodel is no more tenable than belief in the original model
- finding a supermodel to fit may require too much effort trying to model features which are not necessarily of interest, important and about which there is not much information.

Supermodels which include long-tailed distributions generate robust estimators but it is important not to assume the supermodel holds exactly.
Estimators vs Models

Believing in the model or being robust is reflected in the form of the variance used for the estimator.

The asymptotic variance of an M-estimator is the sandwich variance

\[
\text{Var}(\hat{\theta}) \approx \frac{A_F(\theta_0)}{n B_F(\theta_0)^2}
\]

where \( A_F(\theta) = E_F \psi(Y_i, \theta_0)^2 \) and \( B_F(\theta) = E_F \psi'(Y_i, \theta_0) \).

If \( \psi \) is the derivative of a log-likelihood \( \Psi \) and the model \( \Psi \) holds, then \( A_\Psi(\theta_0) = B_\Psi(\theta_0) \) so the variance is the inverse Fisher information \( 1/n B_\Psi(\theta_0) \).
Variance estimation

True believers: $n^{-1}B_n(\hat{\theta})^{-1}$ or $\hat{B}_n(\hat{\theta})^{-1} = \{\sum_{i=1}^{n} \psi'(y_i, \hat{\theta})\}^{-1}$ corresponding to expected and observed information.

Nonparametric robustniks: expected $n^{-1}B_n(\hat{\theta})^{-1}A_n(\hat{\theta})B_n(\hat{\theta})^{-1}$ or observed $\hat{B}_n(\hat{\theta})^{-1}\hat{A}_n(\hat{\theta})\hat{B}_n(\hat{\theta})^{-1}$ sandwich variance

Hardcore robustniks: The estimate of variance under the Gaussian model $V_\Phi(\hat{\theta}) = n^{-1}B_\Phi(\hat{\theta})^{-1}A_\Phi(\hat{\theta})B_\Phi(\hat{\theta})^{-1}$ or extrapolating it using the change of variance function $V_\Phi(\hat{\theta}) \exp\{\epsilon \sup CVF(x; \Phi, \hat{\theta})/V_\Phi(\hat{\theta})\}$.

Pragmatic compromisers: Estimate nonparametrically those terms that are non-zero in the Gaussian case - effectively proceed nonparametrically as if there were symmetric errors.

This is a bit like survey analysis: It all comes down to how you estimate the variance!!!!
Weighting and getting SEs correct

We can write an estimating equation for a location estimator as

\[ 0 = \sum_{i=1}^{n} w_i(\mu)(Y_i - \mu), \quad w_i(\mu) = \psi\{(Y_i - \mu)/\sigma\}/\{(Y_i - \mu)/\sigma\} \]

and estimate \( \mu \) from the sequence of iteratively reweighted means

\[ \hat{\mu} = \{\sum_{i=1}^{n} w_i(\mu)\}^{-1} \sum_{i=1}^{n} w_i(\mu)Y_i. \]

Weighted least squares theory suggests that the variance is

\[ \text{Var}(y_i) \frac{\sum_{i=1}^{n} w_i(\mu)^2}{\left\{\sum_{i=1}^{n} w_i(\mu)\right\}^2} \neq \sigma^2 \frac{\mathbb{E}\psi\{(Y_i - \mu)/\sigma\}^2}{\left[\mathbb{E}\psi'(Y_i - \mu)/\sigma\right]^2}. \]
Huber’s (1964) Minimax Approach

Under the gross error location model, choose $\psi$ so that the asymptotic variance of $\hat{\mu}$ defined by

$$0 = \sum_{i=1}^{n} \psi\{(y_i - \mu)/\sigma\}$$

is as small as possible at the worst distribution in the gross error model.

The solution is the Huber M-estimator $\psi(x) = \max\{-k, \min(k, x)\}$ which is generated by the least favourable distribution with density proportional to $\exp(-x^2/2)$ for $|x| \leq k$ and $\exp(-k|x| + k^2/2)$ otherwise.

Huber also treated scale and proposed methods for simultaneously estimating location and scale.
Discussion

The contamination proportion $\epsilon$ has to be known but results are not very sensitive over the main range.

The gross error model is richer than supermodels. It omits gaps, troughs, rounding but allows for the most important deviations.

Scale is needed to determine the distance at which treatment occurs but scale estimation changes the neighbourhood because those with the same estimated scale (rather than scale) are considered.

Bias is a problem: Huber assumed symmetry to eliminate bias.

The approach does not generalize to arbitrary models.
Hampel’s (1968) Infinitesimal Approach

Minimise the asymptotic variance subject to a bound on the maximum value of the absolute influence function (gross error sensitivity).

For the normal location problem, this leads to the Huber estimator again.

Adding the side condition that the influence function vanishes outside a region leads for the normal location problem to the redescending hyperbolic tangent estimators.

The neighbourhoods are not infinitesimal - the IF is used as an approximation over finite neighbourhoods.
More on Bias

The uncertainty in the model means the parameters are not uniquely defined.

Although this inherent and unavoidable indeterminacy is small if $\epsilon$ is small and is rather irrelevant for practical purposes, it poses awkward problems for the theory... Huber (1964)

The consequences are seen by writing

$$n^{1/2} \{ \hat{T} - T(G) \} = n^{1/2} \{ T(F) - T(G) \} + n^{1/2} \{ \hat{T} - T(F) \}$$

and noting that we somehow need to make $T(F) = T(G)$.

To remove this difficulty, we can

- (choose the target $T$) define the parameters by arbitrary functionals of the distribution (e.g. by the expectation, or the median of $F$); makes estimation an nonparametric problem
- (choose the estimator $\hat{T}$) define the parameter to be estimated
by the asymptotic value of the estimator as $n \to \infty$; this leads to a single $T$

- (restrict the distributions $F$) restrict attention to distributions over which the parameters are uniquely defined; works for pure location but not in general and we don’t have control over $F$

- (restrict the class of distributions asymptotically; Huber-Carol, 1970, Bickel, 1978, Rieder, 1978) make the class of distributions shrink to $G$ as $n \to \infty$; this makes robustness a finite sample problem but models are usually poorer as $n$ increases and shrinkage is to a distribution other than the ideal one

- (Hampel, 1968) estimate the correct parameter under the ideal model and limit the unavoidable bias under other models.
Comparing neighbourhoods
‘Good’ points are downweighted

Robust methods downweight observations in the tails whether they are outliers or not, gross errors or correct observations.

If we adopt a model with a longer tailed distribution, then such observations will be downweighted in location estimation.

The philosophy is that observations in the tails are downweighted for safety (insurance) but, even if these are all completely valid observations, the estimator is still sensible and reasonably efficient.
**Trimming and Winsorizing**

Order the data $X_{(1)} \leq \ldots \leq X_{(n)}$.

**Trimmed Mean**: remove the smallest and largest $\alpha n$ observations and take the mean of the rest (Anonymous 1821).

**Winsorized Mean**: replace the smallest $\alpha n$ observations by $X_{(\alpha n+1)}$, replace the largest $\alpha n$ observations by $X_{(n-\alpha n)}$ and take the mean of the modified sample.

What are the estimators doing to individual observations? Does the trimmed mean reject (give zero weight) to extreme observations? Does the Winsorized mean give the extreme observations the weight of the nearest retained observation?

The data ordering makes these questions difficult to answer but the answers are given by the influence functions.
Both IFs are constant for $x < F^{-1}(\alpha)$, linear for $F^{-1}(\alpha) \leq x \leq F^{-1}(1 - \alpha)$ and constant for $x > F^{-1}(1 - \alpha)$.

The trimmed mean has a continuous IF and it Winsorizes rather than rejects extreme observations. It is like the Huber M-estimator. The Winsorized mean has jumps at the quantiles: it Winsorizes but not to the quantile values.

Ironically: The sample variance of the Winsorized sample is the variance of the trimmed mean not of the Winsorized mean. The variance of the Winsorized mean is actually difficult to estimate (contrary to what is often claimed).

The IFs obviously depend on F: they are asymmetric for asymmetric distributions. (Also true for M-estimators through scale.) It makes sense to allow asymmetric trimming.
Winsorization

Trimming
Searls’ Winsorized Mean

Searls (1966) considered estimating the mean $EY$ by

$$\hat{W}_K = n^{-1} \sum_{i=1}^{n} \min(Y_i, K).$$

This estimator truncates large observations but requires no ordering. It (metrically) Winsorises the estimator rather than the estimating equation.

Huber M-estimator = metrically Winsorized mean should be metrically Winsorized estimating equation = metrically trimmed mean!

Notice that the target is fixed as the mean - this only makes sense if the outliers are correct observations. The issue is no longer robustness but efficiency, making the problem nonparametric.
Searls showed that there exists a $K$ such that

$$MSE(\hat{W}_K) = \{E \min(Y, K) - EY\}^2 + n^{-1} Var\{\min(Y, K)\} < MSE(\bar{Y}).$$

The optimal $K$ satisfies

$$0 = n^{-1}(n - 1)E \min(Y, K) - EY + n^{-1}K$$

so it depends on $F$ and has to be estimated.

This estimator has been discussed by Ernst (1980) and then in ABS by Gross (1985), Gross et al (1986), Kokic and Bell (1994) and Clark (1996).

Estimation of $K$ is done by plugging in an estimate of $F$. ABS uses information from past surveys and actually treats $F$ as known. If the same data is used, the estimator becomes an adaptive estimator. In this case, the IF involves that of the sample mean so the estimator is not robust. It is also not clear whether adaptation is achieved or not.
Regression problems

Complicated by outliers in covariate space and the fact that covariates are conditioned on and not modelled.

The cause of outliers is ambiguous but affects how we react to them: outliers in $X$ may be due to gross errors in $X$, priceless observations, or accurate but useless observations (outside the range of validity of the model). Outliers in $X$ and $y$ may be due to gross errors in $X$ or $y$, or model misspecification.

Generalized M-estimators include additional weights to downweight outliers in $X$.

M-estimators have low breakdown point. We can pursue high breakdown. (Huber not small deviations: Rouseeuw either high or low - not much in between.) (i) General position requirement (any $p$ rows give a unique parameter determination) ignores collapse to a lower dimension and excludes designed covariates. Good design with
built in redundancy is the way to go. (ii) High breakdown estimators are unstable (nonconvex minimisation).

Equivariance properties complicate achieving high breakdown: regression ($\beta(y + Xv, X) = \beta(y, X) + v$ for any $v$: this enables us to translate the data), scale ($\beta(cy, X) = c\beta(y, X)$ for any $c > 0$: this enables us to change the scale of measurement) and affine equivariance ($\beta(y, XA) = A^{-1}\beta(y, X)$ for any nonsingular $A$: this enables us to make linear transformations of $X$ without changing the fitted values).

... dealing with high positional influence requires what-if analyses and human judgement rather than a blind, automated robustness approach (Huber, 2009)
Finite population robustness

Design-based robustness

- Hulliger (1991) treated the Horvitz-Thompson estimator under pps sampling as a regression estimator and then applied robust regression to produce

\[ T_{HT} = \bar{x}_U \hat{\beta}_{HT}, \quad \hat{\beta}_{HT} = \arg\min \sum_{s} \frac{1}{\pi_i} \left( \frac{y_i - \beta x_i}{x_i^{1/2}} \right)^2. \]

- Beaumont, Haziza and Ruiz (2009)?

This runs into difficulties of robustness when conditioning. Kokic and Bell postulated an underlying distribution for population values but they averaged over all possible sample selections. (Clark, 1996)

Model-based robustness = robust prediction

Gross errors and unique values are called nonrepresentative outliers; other outliers (which can be expected to occur in the nonsample part
of the population as well) are representative outliers. (Chambers, 1986) Outliers are often asymmetric.

Under the model,

$$EY_i = \beta x_i \quad \text{Var}(Y_i) = \sigma^2 v(x_i),$$

Chambers proposed

$$\hat{T} = \sum_s Y_i + \hat{\beta} \sum_r x_i + \sum_s \frac{x_i/\sigma v(x_i)^{1/2}}{\sum_s x_j^2/\sigma^2 v(x_j)} \phi\left(\frac{y_i - \beta x_i}{\sigma v(x_i)^{1/2}}\right) \sum_r x_j.$$  

Here $\phi$ is different from the $\psi$ function used to estimate $\beta$. Its role is to put some of the outliers back, to explicitly acknowledge that similar outliers occur in the nonsample part. (If they don’t then drop the third term.)

Predicting representative outliers is more difficult than just doing robust estimation.
Conditional and Bayesian Robustness
Robustness is about sensitivity to changes but conditioning fixes things so disallows changes. This is like trying to buy insurance retrospectively - only after we know we needed it. (relevant to FPP)

Huber: the model is under the statistician’s control, the true underlying state is not. Bayesians don’t worry about what is not under their control.

Using supermodels leads to infinite recursion.

Sensitivity analysis is a practical approach but note the importance of specifying the range (small changes tend to show small sensitivity; larger ones show larger sensitivity) and the lack of connection to robustness (if data contains outliers, small changes in the model can have big effects so sensitivity is informative about the presence of outliers but not about the choice of model).

The separation of the procedure from the model is problematic. In a location and scale problem, we need to deal with both problems
simultaneously and cannot think about them separately. This makes model robustness elusive.

Pragmatic approach: use least favourable models, flat priors and discount the tails of the posterior (because the first two effectively lengthen its tails).