Statistics 580

A Brief Review of Robust Statistics

Introduction

What is robustness? One point of view is concerned with the distribution of a statistic, be it a test statistic or a point estimate. If the distributional properties remain substantially unchanged under certain departures in the underlying model, then the statistic is said to have the desirable property of being robust against those departures. For example, suppose that estimation of location parameter \( \mu \) of a population distribution is of interest. Then it is well known that for a wide range of distributions, confidence limits for \( \mu \) based on the \( t \)-distribution have approximately the appropriate coverage probability. This means that the \( t \)-statistic is robust in the sense stated above. The approach followed by the Princeton Robustness study and related studies is to find procedures that are robust in this sense but also to gain a measure of efficiency of these procedures against a range of distributions. They eliminated from consideration any statistic that is optimal only under the underlying model. For example, for estimating location \( \mu \) under departures from the normal model \( \bar{x} \) is very inefficient, and thus \( \bar{x} \) was not considered in these studies.

Robust Estimation of Location

Suppose that the \( X_1, X_2, \ldots, X_n \) is an i.i.d. sample from a population with cdf \( F_{\mu, \sigma}(x) \) and density \( f_{\mu, \sigma}(x) \), where \( f \) is symmetric about some point, and \( \mu \) and \( \sigma \) are location and scale parameters, respectively. Consider the standardized density \( f \) i.e.,

\[
f_{\mu, \sigma}(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right).
\]

For the normal distribution, for example,

\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.
\]

The problem is to estimate the location parameter \( \mu \), but both \( f \) and \( \sigma \) are unknown. If \( f \) is known there is no problem because a method of estimation appropriate for that distribution could be used. While estimating \( \mu \) robustly, it is hoped to keep any loss of information from the data low, at least in comparison to other estimators. If \( T_n(x_1, \ldots, x_n) \) and \( T^*_n(X_1, \ldots, x_n) \) are two estimators for \( \mu \) and if

\[
\sqrt{n}(T_n - \mu) \xrightarrow{L} N(0, \sigma^2_f(T)) \\
\sqrt{n}(T^*_n - \mu) \xrightarrow{L} N(0, \sigma^2_f(T^*))
\]

then Asymptotic Relative Efficiency is defined as

\[
ARE(T^* : T) \equiv \frac{\sigma^2_f(T)}{\sigma^2_f(T^*)}.
\]
If this is large over a range of possible distributions, then $T^*$ is considered to be more efficient than $T$.

**Example:**

Hodges and Lehmann (1963) suggest a location estimate based on the Wilcoxon signed rank sum test. This is the median of all pairwise averages:

$$
\tilde{X} = \text{median}_{i,j} \left\{ \frac{X_i + X_j}{2} \right\}
$$

Comparing $\tilde{X}$, where $\tilde{X}$ is the sample mean, and $\check{X}$ under different distributions they found that under

- Cauchy: $\text{ARE}(\tilde{X} : \check{X}) = \infty$
- Normal: $\text{ARE}(\tilde{X} : \check{X}) = 2/\pi \approx 0.9549$
- Any Location Family: $\text{ARE}(\tilde{X} : \check{X}) = 0.864$

So they recommend $\tilde{X}$ since it is never much worse than $\check{X}$, but can be infinitely better!

**Example**

Comparing $X^* = \text{median}(X_1, \ldots, X_n)$ to the sample mean $\check{X}$ under different distributions, where $\text{ARE}(X^* : \check{X}) = \sigma^2_\check{X}(X^*)/\sigma^2_{\tilde{X}}(X^*)$:

- Normal: \[
\sigma^2_\check{X}(X^*) = \frac{1}{4(f(0))^2} = \frac{1}{4(1/\sqrt{\pi})^2} = \frac{\pi}{2}
\]
  \[
\sigma^2_{\tilde{X}}(X^*) = \text{Var}(X) = 1
\]
  \[
\text{ARE} = 2/\pi
\]

- Double Exponential: $f(x) = \frac{1}{2}e^{-|x|}$
  \[
\sigma^2_\check{X}(X^*) = \frac{1}{4(f(0))^2} = \frac{1}{4(1/2)^2} = 1
\]
  \[
\sigma^2_{\tilde{X}}(X^*) = \text{Var}(X) = 2
\]
  \[
\text{ARE} = 2
\]

Therefore, for the double exponential for example, the median is more efficient.

**Example**

Tukey(1960) considers a random sample $X_1, \ldots, X_n$ from the contaminated Normal distribution

$$
X \sim N(0, k^2) \quad \text{with probability} \quad 1 - \epsilon
$$
$$
X \sim N(0, 9k^2) \quad \text{with probability} \quad \epsilon
$$

Two estimators, $d_n = \frac{1}{n} \sum |X_i|$ and $s_n = \sqrt{\frac{1}{n} \sum X_i^2}$ for $k$ are compared. The $\text{ARE}(s_n : d_n)$ is
L-Estimators

This class of estimators consists of linear combinations of order statistics. Suppose $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a distribution $F$. Let

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

denote the ordered sample. Then L-estimator of location is defined as

$$T_n = \sum_{j=1}^{n} a_j X_{(j)} \quad , \quad \sum_{j=1}^{n} a_j = 1$$

Since $f$ is assumed to be symmetric, it is also required to have

$$a_{n-j+1} = a_j$$

for all $j$. This guarantees unbiasedness of $T_n$, provided $E(T_n)$ exists.
Examples:

mean: choose \( a_j = 1/n \) for all \( j \)

trimmed mean: \((\alpha - \text{trimmed mean})\)
delete \([n\alpha]\) points from both ends of ordered data and average the rest.

\[
T_n(\alpha) = \sum_{j=m+1}^{n-m} \frac{X(j)}{n-2m}
\]

where \( m = [n\alpha] \), and \([\cdot]\) is the truncation function. It is generally recommended that \( .1 \leq \alpha \leq .2 \). 50\% trimming corresponds to the median.

Winsorized mean: similar to the \( \alpha \)-trimmed mean but instead of deleting extreme points they are replaced by a common value.

\[
T_n(\alpha) = \frac{m (X(m+1) + X(n-m)) + \sum_{j=m+1}^{n-m} X(j)}{n}
\]

where \( m \) is defined above.

Gastwirth proposal: \( T_n = .3X_{(m+1)} + .4\bar{X} + .3X_{(n-m)} \) where \( m = [n/3] \), and \( \bar{X} = \text{median} (X_1, \ldots, X_n) \).

Trimean: \( T_n = .25X_{m+1} + .5\bar{X} + .25X_{n-m} \) where \( m = [n/4] \), \( \bar{X} = \text{median} (X_1, \ldots, X_n) \).

**M-Estimators**

In looking at L-estimators carefully, it is clear that extreme observations are either completely deleted (given zero weights) or downweighted in some fashion. The M-estimators are generalizations of maximum likelihood estimators of location parameter in a specified distribution (hence, the “M”), which automatically downweights large residuals. Consider a location-scale density.

\[
f_{\mu,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)
\]

as before. For a random sample \( x_1, \ldots, x_n \), the likelihood function is

\[
L(\mu, \sigma) = \Pi_{j=1}^{n} \frac{1}{\sigma} f\left(\frac{x_j - \mu}{\sigma}\right)
\]

and

\[
\ell(\mu, \sigma) = \log L(\mu, \sigma) = -n \log \sigma + \sum_{j=1}^{n} \log f\left(\frac{x_j - \mu}{\sigma}\right).
\]
Minimizing $\ell$ with respect to $\mu$ and $\sigma$,

$$
\frac{\partial \ell}{\partial \mu} = \sum_{j=1}^{n} \frac{\partial}{\partial \mu} \frac{f(\frac{x_j - \mu}{\sigma})}{f(\frac{x_j - \mu}{\sigma})} = -\frac{1}{\sigma} \sum_{j=1}^{n} \frac{f'(\frac{x_j - \mu}{\sigma})}{f(\frac{x_j - \mu}{\sigma})}
$$

where $f'(u) = \frac{\partial f(u)}{\partial u}$, and

$$
\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{\sigma^2} \sum_{j=1}^{n} (x_j - \mu) \frac{f'(\frac{x_j - \mu}{\sigma})}{f(\frac{x_j - \mu}{\sigma})}
$$

Define

$$
\psi(u) = -\frac{f'(u)}{f(u)}
$$

and note that if $f$ is symmetric, $\psi$ is an odd function. Setting partial derivatives equal to zero yields the system of maximum likelihood equations:

$$
\sum_{j=1}^{n} \psi \left( \frac{x_j - \mu}{\sigma} \right) = 0
$$

$$
\sum_{j=1}^{n} \left[ \left( \frac{x_j - \mu}{\sigma} \right) \psi \left( \frac{x_j - \mu}{\sigma} \right) - 1 \right] = 0.
$$

**Example:** Normal Distribution $N(\mu, \sigma^2)$

$$
f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}
$$

Thus

$$
f'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot (-x)
$$

and

$$
\psi(x) = -\frac{f'(x)}{f(x)} = +x.
$$
Maximum likelihood equations are

\[ \frac{1}{\sigma} \sum_{j=1}^{n} (x_j - \mu) = 0 \]

and

\[ \frac{1}{\sigma^2} \sum_{j=1}^{n} [(x_j - \mu)^2 - 1] = 0 \]

giving

\[ \hat{\mu} = \bar{x} \text{ and } \hat{\sigma} = \sqrt{\frac{\sum(x_j - \bar{x})^2}{n}} \]

**Example:** Double Exponential

\[ f(x) = \frac{1}{2} e^{-|x|} \]

Thus

\[ f'(x) = \begin{cases} \frac{1}{2} e^{-|x|} & , x < 0 \\ -\frac{1}{2} e^{-|x|} & , x > 0 \\ \text{arbitrary} & , x = 0 \end{cases} \]

and

\[ \psi(x) = -\frac{f'(x)}{f(x)} = \text{sgn}(x) \]

The maximum likelihood equation for location becomes \( \sum \text{sgn} (x_j - \mu) = 0 \) implying that the estimate of \( \mu \) is that value of \( x_j \) that allocates \( n/2 + \) signs and \( n/2 - \) signs to \( x_j - \mu \). This of course, is \( \hat{\mu} = \text{median} (x_1, \ldots, x_n) \).
General Definition of the M-Estimator

Let $\rho$ be a real-valued function on $\mathbb{R}$. Then consider the class of estimators of location which minimizes a function of scaled deviations of the observation from the estimate, i.e. minimizes

$$\sum_{j=1}^{n} \rho \left( \frac{x_j - T_n}{s_n} \right)$$

with respect to $T_n$ or, equivalently solve

$$\sum_{j=1}^{n} \psi \left( \frac{x_j - T_n}{s_n} \right) = 0.$$

Notice that if $\rho$ corresponds to $-\log f$, solving the above problems yield the m.l.e. of $\mu$ with respect to the density $f$. In M-estimation, $s_n$ is often taken to be a “robust” estimate of scale such as

- normalized interquartile range (IQR): $(Q3 - Q1)/1.35$
- normalized median of absolute deviations (MAD): $\text{median} \{|x_j - \tilde{x}|\}/.6754$

Sometimes the location estimator does not depend on $s_n$. But the basic problem in M-estimation is how to select $\psi$ (or $\rho$)? Obviously, $\rho$ and $\psi$ could be selected to correspond to some density. For example, if Cauchy density

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

is used, assuming unit scale for now.

$$\psi(x) = \frac{2x}{1 + x^2}.$$

A way to choose a $\psi$-function is seen by expressing M-estimators as a weighted mean:

Let $T_n$ be obtained by solving

$$\sum_{j=1}^{n} \psi \left( \frac{x_j - T_n}{s_n} \right) = 0.$$
Then \( T_n \) can be written as
\[
T_n = \frac{\sum w_j x_j}{\sum w_j}
\]
where
\[
w_j = \frac{\psi \left( \frac{x_j - T_n}{s_n} \right)}{\left| \frac{x_j - T_n}{s_n} \right|}
\]
For symmetric distributions, \( w_j \) are non-negative since \( \psi \) is an odd function. So to choose \( \psi \), one could use the weighted mean formulation to decide how much weight to be given to observation depending upon how “far out” they are. The weighted mean formulation also suggests a convenient method for computing M-estimators from a sample. This is called iteratively reweighted means scheme:

1. Start with \( T_0 \) (say, the median)
2. Compute \( s_n \) and \( w_j \) from \( T_0 \)
3. Compute new \( T \) as the weighted mean
4. Go to Step 2 if not converged.

An iterative formula can be presented in the form:
\[
T_{(i+1)} = T_{(i)} + s_n \frac{\sum \psi(u_j)}{\sum \psi(u_j)}
\]
where \( u_j = \frac{x_j - T_{(i)}}{s_n} \), \( s_n = \text{med}\{x_j - T_0\} \) and \( T_0 = \text{med}\{x_j\} \). Another computational method for M-estimators is the Newton-Raphson iterative scheme, based on:
\[
T_{(i+1)} = T_{(i)} + s_{(i)} \frac{\sum \psi(u_j)}{\sum \psi'(u_j)}
\]
where \( u_j = \frac{x_j - T_{(i)}}{s_{(i)}} \), and \( s_{(i)} = \text{med}\{x_j - T_{(i)}\} \). If the Newton-Raphson algorithm is used for only a single iteration, then the resulting estimator is called a one-step M-estimator. The choice of starting value \( T_{(0)} \) is important in this case, and generally \( T_{(0)} = \text{median} \) is used.

**Tukey’s Bisquare Estimator**
\[
w_j = \begin{cases} 
1 - \left( \frac{x_j - T_n}{C s_n} \right)^2, & \text{if } \left| \frac{x_j - T_n}{s_n} \right| < 1 \\
0, & \text{otherwise}
\end{cases}
\]
\( s_n \) is usually taken to be the MAD estimate = \( \text{med}\{|x_j - T_0|\} \) and \( C \) to be 6 or 9 and iteration is continued until convergence. The associated \( \psi \)-function is
\[
\psi(u) = \begin{cases} 
u(1 - u^2)^2, & |u| < 1 \\
0, & |u| \geq 1
\end{cases}
\]
Huber Estimators

\[ \rho(x) = \begin{cases} 
  x^2/2 & , |x| \leq C \\
  C|x| - \frac{C^2}{2} & , |x| > C 
\end{cases} \]

\[ \psi(x) = \begin{cases} 
  x & , |x| \leq C \\
  C \cdot \text{sgn}(x) & , |x| > C 
\end{cases} \]

These choices correspond to maximum likelihood estimation from a density

\[ f(x) = \begin{cases} 
  c_1 e^{-\frac{1}{2}x^2} & , |x| \leq C \\
  c_1 e^{\frac{C^2}{2} - C|x|} & , |x| > C 
\end{cases} \]

that is, the density has a normal center but exponential tails.
Redescending Estimators

Hampel’s $\psi$ functions

This $\psi$-function has slope 1.0 in the middle, but the rest depends on choices for a, b, and c and gives zero weight to data points beyond C.

Andrews Sine Estimator

$$\psi(x) = \begin{cases} \sin(x/a\pi) & |x| \leq a \\ 0 & |x| > a \end{cases}$$

Hogg’s Adaptive Estimator

First determine $k = $ sample kurtosis from the sample, then use it to compute the location estimate depending on it’s value according to the following:

- $k < 1.9$ use outer mean
- $1.9 < k < 3.1$ use mean
- $3.1 < k < 4.5$ use 25% trimmed mean
- $4.5 < k$ use Gastwirth mean
Conclusions from the Princeton Study

A Monte Carlo procedure outlined in the previous section was used in the Princeton Study to compute various measures of efficiency such as small sample variances etc., of a large number of robust estimators of location. A variety of distributions were employed for generating data and several sample sizes were used. For example sample sizes of 5, 10, 20, 40 were used for Normal and Cauchy and sample sizes of 10 and 20 for most other distributions that included various degrees of freedom and forms of contamination. All distributions were long-tailed with respect to the normal distribution. The following is a summary of conclusions:

- The best were the 3-part M-estimators i.e., Hampel's proposal
  - with bends at (1.7, 3.4, 8.5) or (2.1, 4.0, 8.2)
  - and used MAD, (not adjusted to be unbiased for $\sigma$ under the normal case) scale estimator i.e., $s_n = \text{median}\{|x_j - T_n|\}$.
- Andrews' sine estimator also did well
  $$
  \psi(x) = \begin{cases} 
  \sin(x/2.1) & |x| \leq 2.1\pi \\
  0 & |x| > 2.1\pi
  \end{cases}
  $$
  - also used MAD for $s_n$
- One-step Huber estimators starting at the median did quite well
- Adaptive estimators were excellent for $n > 50$
- Tukey's Bisquare estimator was not included in the study
- The 25% trimmed mean is a good compromise if one wants to avoid more elaborate computation.