Confidence distributions based on M-estimators

Sam-Erik Walker
Dept. of Math., Univ. of Oslo, 13 May 2015
Outline

1. M-estimators
2. Proper scoring rules
3. Asymptotics
4. Robustness
5. Numerical example
6. Concluding remarks
M-estimators

• Forms a broad class of estimators obtained as the minimum (or maximum) of sums of functions of the data, and was proposed in 1964 by Peter J. Huber for generalizing MLE

\[ \rho\text{-type: } \hat{\theta} = \operatorname{arg\,min}_\theta \sum_{i=1}^{n} \rho(x_i, \theta) \]

\[ \psi\text{-type: } \sum_{i=1}^{n} \psi(x_i, \hat{\theta}) = \sum_{i=1}^{n} \nabla_\theta \rho(x_i, \hat{\theta}) = 0 \]

• Examples of M-estimators:
  • Least squares estimators
  • MLE
  • Huber’s M-estimator
  • Estimators derived from proper scoring rules
Huber’s M-estimator

\[ \sum_{i=1}^{n} \psi_{k} \left( \frac{x_i - \hat{\theta}_n}{\hat{\sigma}_n} \right) = 0 \]

where \( \psi_{k} (x) = \begin{cases} +k & x \geq +k \\ x & -k \leq x \leq +k \\ -k & x \leq -k \end{cases} \)

• Here \( k \) is a tuning constant determining the degree of robustness
• If \( k \to \infty \) then \( \hat{\theta}_n \) is the mean; If \( k \to 0 \) then \( \hat{\theta}_n \) is the median
• Huber’s motivation was that unrestricted \( \psi \) -functions have undesirable properties, being unstable to outliers
• \( \hat{\sigma}_n \) is usually a robust estimate of scale, e.g., \( \text{MAD} = \text{median} \left( |x_i - \text{median}(x_i)| \right) \)
• Robust start value for location: Median
Proper scoring rules

- A scoring rule is a real function \( S(x, F) \) where \( x \) is a sample from some outcome space, and \( F \) is a model distribution over this outcome space.

- Define \( S(G, F) = E_G \{S(X, F)\} \) where \( X \sim G \) the data distribution.

- The scoring rule is said to be proper iff \( S(G, F) \geq S(G, G) \) for all \( F \).

- It is said to be strictly proper iff \( S(G, F) > S(G, G) \) for all \( F \neq G \).

- If \( S \) is proper then \( cS + h(x) \) is also proper if \( c > 0 \), \( h(x) \) arbitrary function.

\[
D(G, F) = S(G, F) - S(G, G)
\]
Proper scoring rules: Inference

• If the model distribution is \textit{parametric}, i.e., \( F = F_\theta \), and if the unknown data distribution \( G \) is approximated by \( \hat{G}_n(x) = \sum_{i=1}^{n} I(x_i \leq x) / n \), we obtain the following \textit{empirical proper scoring rule estimator} of \( \theta \)

\[
\hat{\theta}_n = \arg \min_{\theta} D \left( \hat{G}_n, F_\theta \right) = \arg \min_{\theta} \left\{ S \left( \hat{G}_n, F_\theta \right) - S \left( \hat{G}_n, \hat{G}_n \right) \right\} = \arg \min_{\theta} S \left( \hat{G}_n, F_\theta \right) \\
= \arg \min_{\theta} E_{\hat{G}_n} \left\{ S \left( X, F_\theta \right) \right\} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} S \left( x_i, F_\theta \right) = \arg \min_{\theta} \sum_{i=1}^{n} S \left( x_i, F_\theta \right)
\]
Proper scoring rules: Inference

• Thus, $s$ plays the role of the $\rho$-function in M-estimation
• Further, if $s$ is differentiable w.r.t. $\theta$, we have

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} S(x_i, F_\theta) \iff \sum_{i=1}^{n} s(x_i, F_{\hat{\theta}_n}) = 0$$

i.e., the optimal parameters represent the solution of the so-called scoring rule estimating equation(s), $p$ equations if there are $p$ parameters, thus $s$ plays the role of the $\psi$-function in M-estimation
Proper scoring rules: Examples

• The most prominent proper scoring rule is the *logarithmic score* (Good, 1952)

\[ S(x, F_\theta) = -\log f(x, \theta) \]

leading empirically to MLE

• The associated divergence is the *Kullback-Leibler* divergence, and the associated entropy is the *Shannon* entropy

\[ D(G, F_\theta) = S(G, F_\theta) - S(G, G) = \int \{\log g(x) - \log f(x, \theta)\} g(x)dx \]
Proper scoring rules: Examples

• The logarithmic score is a special case of a general so called *separable Bregman score* construction (Dawid, 2007)

\[
S(x, F_\theta) = -\varphi' \{ f(x, \theta) \} - \int \left[ \varphi \{ f(u, \theta) \} - f(u, \theta) \varphi' \{ f(u, \theta) \} \right] du
\]

where the defining function \( \varphi : \mathbb{R}^+ \to \mathbb{R} \) is convex and differentiable

• By setting \( \varphi(t) = t \ln t \), which is convex and differentiable, we obtain the logarithmic score
Proper scoring rules: Examples

• Another important special case of this construction is the Tsallis score obtained by setting \( \varphi(t) = t^q \) for \( q > 1 \). This yields

\[
S_q(x, F_\theta) = -q \cdot f(x, \theta)^q + (q-1) \int f(u, \theta)^q \, du
\]

\[
= \left\{-q \cdot f(x, \theta)^{q-1} + (q-1) \int f(u, \theta)^q \, du + q-1\right\} / (q-1)
\]

• Incidentally this is the same as the Basu-Harris-Hjort-Jones power divergence estimator (Basu et al., 1998) with parameter \( \alpha = q - 1 > 0 \) obtained by setting \( \varphi(t) = t \cdot t^{(\alpha)} \) where \( t^{(\alpha)} \) is the Box-Cox power transf.

• When \( q \to 1^+ (\alpha \to 0^+) \) we obtain the logarithmic score
Proper scoring rules: Examples

• Finally, an interesting so-called *local* scoring rule (which depends on $F_\theta$ only through its value at $x$, and the value of a finite number of its derivatives at $x$) is the *Hyvärinen score* (Hyvärinen, 2005)

$$S_H(x, F_\theta) = 2 \frac{\partial^2}{\partial x^2} \ln f(x, \theta) + \left| \frac{\partial}{\partial x} \ln f(x, \theta) \right|^2$$

• This has the convenient property that it can be computed without knowledge of the normalising constant of the density

• It can also be defined for multivariate distributions
Asymptotics

Theorem. Under suitable regularity conditions, e.g., Barndorff-Nielsen and Cox (1994), an M-estimator (proper scoring rule) estimator $\hat{\theta}_n$ is asymptotically normal with mean $\theta_G$ (least false parameter), and variance $V$, i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta_G) \xrightarrow{D} N(0,V)$$

where

$$V = J^{-1}KJ^{-T}; \quad J = E_G \left\{ \frac{\partial S(\theta)}{\partial \theta^T} \right\}_{\theta=\theta_G}; \quad K = E_G \left\{ s(\theta_G) s(\theta_G)^T \right\}$$

When $G = F_{\theta_0}$, the above theorem holds with $\theta_G = \theta_0$ (true parameter)
Asymptotics

• The matrix $V^{-1}$ is known as the Godambe information matrix (Godambe, 1960)

• In general for M-estimators (scoring rule estimators) $J \neq K$

• In the special case of the logarithmic score (MLE), and when the model is exact, i.e., $G = F_\theta$, we have that

$$V^{-1} = J = K$$

is the Fisher information matrix. In this case

$$V_M = J_M^{-1} K_M J_M^{-T} \geq V_{MLE} = J_{MLE}^{-1}$$
Robustness

• The *influence function* (IF) of an estimator measures the effect on it of a small contamination at a point $x$

• The supremum of the IF over the data space measures the worst influence of such contamination, so supplying a measure of gross-error sensitivity

• A desirable property for a statistical procedure is that this gross-error sensitivity is *finite*, i.e. that IF is *bounded*

• This is termed *B-robustness*
Robustness

• From the general theory of M-estimators (Huber and Ronchetti, 2009), the IF of an M- or scoring rule estimator \( \hat{\theta} \) is given by

\[
\text{IF}(x; s, G) = J^{-1} s(x, \theta_G)
\]

• Thus, if \( s(x, \theta) \) is bounded in \( x \) for each \( \theta \), then the M- or scoring rule estimator \( \hat{\theta} \) is B-robust

• The IF can also be used to evaluate the asymptotic variance of \( \hat{\theta} \) since

\[
V = E_G \left\{ \text{IF}(X; s, G) \text{IF}(X; s, G)^T \right\}
\]
Robustness: Bregman divergence estimator

• A necessary and sufficient condition for B-robustness of the Bregman estimator, where $s$ is given by

$$-s(x, \theta) = \lambda(x, \theta) - E_\theta \lambda(X, \theta) \quad \text{with} \quad \lambda(x, \theta) = \varphi''\{f(x, \theta)\} \nabla_\theta f(x, \theta)$$

is the following (Dawid et al. (2014a), Condition 5.1):

For all $\theta$, $\lambda(x, \theta)$ is a bounded function of $x$

• This condition combines properties of the Bregman divergence generating function $\varphi$, and the form of the model $f(x, \theta)$
Robustness: Bregman divergence estimator

- A sufficient condition for B-robustness of the Bregman estimator is

Dawid et al. (2014a), Condition 5.2:
(i) The Bregman gauge $\varphi''$ is locally bounded (on all intervals $(0,M)$)
(ii) Both $f(x,\theta)$ and $\nabla_\theta f(x,\theta)$ are bounded in $x$, for each $\theta$

- Tsallis score: $\varphi(t) = t^q; \varphi''(t) \propto t^{q-2}$ locally bounded for $q \geq 2$
- Log score: $\varphi(t) = t \ln t; \varphi''(t) = 1/t$ not locally bounded (on $(0,M)$)
Numerical example

• We have tested the performance of the Tsallis score/BHHJ power divergence estimator for estimating the mean and scale (sdev) for a Normal distribution using uncontaminated and contaminated data.

• We draw 100 samples from a N(0,1) (uncontaminated case) and from a contaminated distribution 0.95*N(0,1) + 0.05*N(3,1).

• Repeating this 1000 times and look at the statistics of estimates.

• We compare Tsallis/BHHJ estimates for q = 1.2, 1.5 and 2.0 with MLE.
Numerical example (cont.)

• We use the asymptotic normality of this particular M-estimator to construct coverage distributions and curves using

\[ cc(\theta) = \left| 1 - 2\Phi\left( \frac{\theta - \hat{\theta}_n}{\hat{\sigma}_n} \right) \right| \]

where \( \hat{\sigma}_n = \sqrt{\text{diag}(\hat{V}_n)} \) for \( i = 1, 2 \), with \( \hat{V}_n = \frac{1}{n} \hat{J}_n^{-1}\hat{K}_n\hat{J}_n^{-T} \).

• Note that we use the “sandwich” matrix also for MLE for contaminated data since the model is not exact!
Uncontaminated case: $N(0,1)$

Location $\text{n} = 100$  $\#\text{sim} = 1000$  Scale
Contaminated case: $0.95 \times N(0,1) + 0.05 \times N(3,1)$

Location: $n = 100$, #sim = 1000, Scale
0.95*N(0,1) + 0.05*N(3,1): Confidence curves for location parameter

Parameter
Probability
Tsallis q
1
1.2
1.5
2

n = 100  sim = 1-9
0.95*N(0,1) + 0.05*N(3,1): Confidence curves for scale parameter

Parameter
Probability
Tsallis q
1
1.2
1.5
2

n = 100  sim = 1-9
$$0.95 \times N(0,1) + 0.05 \times N(3,1)$$

Coverage probabilities in % for location = 0

<table>
<thead>
<tr>
<th>Cov. Prob. %</th>
<th>MLE</th>
<th>Tsallis q = 1.2</th>
<th>Tsallis q = 1.5</th>
<th>Tsallis q = 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>21.4</td>
<td>40.7</td>
<td>49.7</td>
<td>50.5</td>
</tr>
<tr>
<td>75</td>
<td>42.9</td>
<td>63.5</td>
<td>73.3</td>
<td>76.2</td>
</tr>
<tr>
<td>90</td>
<td>67.4</td>
<td>82.6</td>
<td>88.3</td>
<td>90.1</td>
</tr>
<tr>
<td>95</td>
<td>78.5</td>
<td>90.1</td>
<td>93.3</td>
<td>93.3</td>
</tr>
<tr>
<td>99</td>
<td>93.1</td>
<td>97.1</td>
<td>98.3</td>
<td>98.7</td>
</tr>
</tbody>
</table>

\(n = 100\)  \#sim = 1000
0.95*N(0,1) + 0.05*N(3,1)
Coverage probabilities in % for scale = 1

<table>
<thead>
<tr>
<th>Prob. %</th>
<th>MLE</th>
<th>Tsallis q = 1.2</th>
<th>Tsallis q = 1.5</th>
<th>Tsallis q = 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>5.4</td>
<td>27.2</td>
<td>42.6</td>
<td>49.4</td>
</tr>
<tr>
<td>75</td>
<td>17.2</td>
<td>48.9</td>
<td>70.5</td>
<td>73.2</td>
</tr>
<tr>
<td>90</td>
<td>42.1</td>
<td>75.2</td>
<td>87.6</td>
<td>89.5</td>
</tr>
<tr>
<td>95</td>
<td>61.7</td>
<td>87.1</td>
<td>94.1</td>
<td>95.3</td>
</tr>
<tr>
<td>99</td>
<td>90.7</td>
<td>97.4</td>
<td>98.8</td>
<td>99.2</td>
</tr>
</tbody>
</table>

n = 100    #sim = 1000
Concluding remarks

• M-estimators represent a very general and versatile approach to the problem of parametric estimation, with an important sub-class formed by proper and strictly proper scoring rules (Dawid et al., 2014a, b; Gneiting et al., 2007)

• We may lose some efficiency as compared with MLE, but obtain improved robustness, and in some cases they may offer computational advances

• Most of the theory of MLE also applies to M- or scoring rule estimation with little or no modification, and thus can be used to construct hypothesis tests and confidence intervals or distributions (Dawid et al., 2014a)
Concluding remarks

• Recent simulation studies (Dawid et al., 2014a) indicate that (adjusted) scoring and likelihood-ratio type statistics yield confidence regions whose coverage properties are satisfactory.

• The Tsallis score / BHHJ power divergence estimators offers robustness with negligible loss of efficiency for a large class of both univariate and multivariate parametric distributions.

• The Hyvärinen score offers to estimate parameters in situations where the normalising constant is unavailable, or difficult to extract, but is not robust.
References


Thanks for your attention!
Uncontaminated data
Coverage probabilities in % for location

<table>
<thead>
<tr>
<th>Cov. Prob. %</th>
<th>MLE</th>
<th>Tsallis q = 1.2</th>
<th>Tsallis q = 1.5</th>
<th>Tsallis q = 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>51.0</td>
<td>51.8</td>
<td>51.1</td>
<td>51.2</td>
</tr>
<tr>
<td>75</td>
<td>74.9</td>
<td>75.1</td>
<td>77.0</td>
<td>77.3</td>
</tr>
<tr>
<td>90</td>
<td>88.9</td>
<td>90.1</td>
<td>89.8</td>
<td>90.2</td>
</tr>
<tr>
<td>95</td>
<td>94.2</td>
<td>94.3</td>
<td>94.6</td>
<td>94.4</td>
</tr>
<tr>
<td>99</td>
<td>98.8</td>
<td>98.3</td>
<td>98.5</td>
<td>98.2</td>
</tr>
</tbody>
</table>
Uncontaminated data
Coverage probabilities in % for scale

<table>
<thead>
<tr>
<th>Prob. %</th>
<th>MLE</th>
<th>Tsallis q = 1.2</th>
<th>Tsallis q = 1.5</th>
<th>Tsallis q = 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50.1</td>
<td>48.7</td>
<td>47.7</td>
<td>50.2</td>
</tr>
<tr>
<td>75</td>
<td>75.7</td>
<td>75.3</td>
<td>73.0</td>
<td>73.8</td>
</tr>
<tr>
<td>90</td>
<td>89.9</td>
<td>90.3</td>
<td>89.9</td>
<td>89.1</td>
</tr>
<tr>
<td>95</td>
<td>94.8</td>
<td>94.9</td>
<td>94.8</td>
<td>94.5</td>
</tr>
<tr>
<td>99</td>
<td>98.7</td>
<td>98.7</td>
<td>99.0</td>
<td>99.1</td>
</tr>
</tbody>
</table>