L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Unimodality and the asymptotics of M-estimators

Marc Hallin

Université Libre de Bruxelles, Belgium

Ivan Mizera

Comenius University, Bratislava, Slovakia

Abstract: Unimodality, in its weaker and stronger forms, enters the robustness investigations somehow less often than symmetry. We point out how unimodality affects the asymptotics of M-estimators under heterogeneous ("non-i.i.d.") errors. Sufficient conditions are given for consistency, with rates, of M-estimators in unimodal heterogeneous location models. For heteroscedastic models, a particular case of heterogeneous ones, a necessary and sufficient consistency condition, with rates, is provided for the L_1 estimator - the sample median.

Key words: Sample heterogeneity, heterodasticity, unimodality, M-estimators, consistency.

AMS subject classification: Primary 62F35, 62F12; secondary 62F10, 60F05.

1 Introduction

In robustness theory, the assumption of symmetry is adopted quite regularly, although with a bit of strange taste: as pointed out by Huber (1981, page 95), "a restriction to exactly symmetric distributions ... violates the very spirit of robustness"—since it is not stable under small perturbations of the underlying probabilities. On the other hand, the symmetry assumption resolves a dilemma of estimands—the problem of finding the target of location estimation. For symmetric population distributions, the center of symmetry is widely accepted as the "natural" location parameter—see, for instance, Hoaglin, Mosteller and Tukey (1983, chapter 9). And, needless to say, symmetry considerably simplifies a number technical considerations.

Unimodality, in its weaker and stronger forms, enters the robustness investigations somehow less often. It shares the instability of the symmetry—small perturbation (in weak topology sense) of a unimodal probability can result in a non-unimodal one. However, in terms of "realism", unimodality performs better; recall only that almost all parent distributions involved in parametric models are unimodal (compared to a considerable number of asymmetric ones). And, even in the symmetric case, the center of symmetry frequently also is the mode. The impact of unimodality on the philosophy of estimation is perhaps not that unambiguous: it could be a matter of a discussion whether the mode, instead of the center of symmetry, can fulfill the need for the "natural" location in models with asymmetric (but unimodal) parent distribution. At least, the consequences of substituting unimodality assumptions for the symmetry ones deserve being closely investigated.

Skipping this problem, we shall concentrate on the technical virtues of unimodality. These are really rewarding. For instance, as shown in Mizera (1994), unimodality is most helpful in establishing the consistency of redescending M-estimators. The paper of Freedman and Diaconis (1982) pointed out that M-estimators can be inconsistent, due to non-identifiability—the lack of well-defined population value—unless the score function is monotone or the underlying distribution is symmetric and unimodal. Mizera (1994) showed that unimodality ensures the uniqueness of the population value also in asymmetric cases, for the majority of M-estimators with the non-monotone ("redescending") score functions used in practice.

In this note, we point out how unimodality affects the asymptotic of M-estimators under heterogeneous errors. The violation of the i.i.d assumption, the assumption which is central to most existing statistical models (recall i.i.d. error terms in regression or i.i.d. innovation process in time series) can arise from contamination, population heterogeneity, uncontrollable and hidden confounding factors, and variations in the measurement techniques or environmental conditions, the characteristics of which may vary through time and space. Different aspects of the asymptotics of M-estimators under heterogeneity were studied in Mizera and Wellner (1996) and Hallin and Mizera (1996). We concentrate here on the more specific consequences of unimodality. Sufficient conditions for consistency, with rates, are given for unimodal heterogeneous location models. For heteroscedastic models, a necessary and sufficient consistency condition, with rates, is established for the L1 estimator—the sample median. This condition is considerably more general than an earlier one by Sen (1968).

2 Consistency of M-estimators in heterogeneous location models

Associated with a nondecreasing score function ψ we define

$$\lambda_n(\psi, t) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_{ni} - t).$$

An M-estimate is defined to be a "solution" of the equation $\lambda_n(\psi, t) = 0$; since $\lambda_n(\psi, t)$ is monotone, the values of t at which $\lambda_n(\psi, t)$ crosses the zero level constitute an interval. To avoid ambiguity, we define the *M-estimate* to be the infimum of this interval:

$$T_{n,\psi} = \sup\{t : \lambda_n(\psi, t) > 0\}.$$

A *heterogeneous location model* consists of

(H1) a set of data $X_{n1}, X_{n2}, \ldots, X_{nn}$, which can be viewed as realizations of independent random variables

(H2) with distribution functions $F_{n1}, F_{n2}, \ldots, F_{nn}$, respectively,

(H3) such that $E[\psi(X_{ni} - \theta)] = 0$, for all $i = 1, 2, \ldots, n$ and for all $n = 1, 2, \ldots$.

The framework of (H1)–(H3) reduces to the standard i.i.d. one whenever $F_{n1} = F_{n2} = \ldots = F_{nn} = F_\theta$, where $F_\theta(x) = F(x - \theta)$. In such a case, the Fisher consistency condition (H3) reduces to a much simpler and traditional one involving F_θ only. However, in heterogeneous situation we need (H3) as the only thread connecting all X_{ni}'s and F_{ni}'s together, ensuring that our estimation of θ makes any sense at all, that our data are not "a bizarre melange without much statistical relevance" (Le Cam 1986, page 529).

A heterogeneous location model is called *symmetric* if all F_{ni}'s are symmetric about θ, and *unimodal* if all F_{ni}'s are unimodal with mode θ, $i = 1, 2, \ldots, n$, $n = 1, 2, \ldots$. A distribution G is called unimodal (with mode θ) if it possesses a density g which is nondecreasing on $(-\infty, \theta]$ and nonincreasing on $[\theta, \infty)$.

Various other unimodality concepts could have beeen considered. The weakest one, merely requiring the existence of a unique global maximum for the density, is too weak for most purposes. The slightly stronger definition formulated in terms of convexity and concavity of the distribution function G before and after the mode is closely related to ours—the only difference is that, unlike ours, it allows for an atom located at the mode. As for

the *strong unimodality* concepts (log-concavity, for instance), they are not needed here.

Note that, in symmetric models, (H3) automatically holds whenever ψ is odd (that is, $\psi(-x) = -\psi(x)$). This, together with a natural interest in estimating the center of symmetry in the symmetric models explains the almost exclusive choice of odd score functions ψ in the practice of M-estimation. Thus, we shall assume about our score functions ψ that

(P1) ψ is a non-decreasing and odd function.

To ensure robustness, we adopt boundedness; since a multiple of ψ yields the same M-estimates, we set

(P2) $\psi(-\infty) = -1$, $\psi(\infty) = 1$.

To avoid pathologies, we also suppose that

(P3) the set of discontinuity points of ψ is finite

and, finally, that

(P4) ψ is increasing at 0: for every $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that $\psi(\varepsilon) - \psi(-\varepsilon) = 2\psi(\varepsilon) \geq 2\delta(\varepsilon)$.

We remark that both (P3) and (P4) are satisfied by all score functions used in practice. For unimodal distributions, (P4) can ensure identifiability (uniqueness of the "population value") of the M-estimator.

Consistency holds whenever the model is *conservative*: that is, whenever the sequence of average distribution functions

$$\bar{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} F_{ni}(x)$$

is tight (weakly sequentially compact; recall that a sequence G_n is tight if and only if for any $\varepsilon > 0$ there is a $K_\varepsilon > 0$ such that $G_n(-K_\varepsilon) \leq \varepsilon/2$ and $G_n(K_\varepsilon) \geq 1 - \varepsilon/2$ for all n). An important special case of conservative model is the *mixture model*: the sequence $\bar{F}_n$ converges weakly to a (proper) distribution function $\bar{F}$. The behavior of robust estimators in mixture models was studied by Stigler (1976).

Theorem 1 *Suppose that X_{ni} satisfy the assumptions (H1)–(H3) of the heterogeneous location model; suppose further that this model is unimodal and conservative. If (P1)–(P4) hold, then $T_{n,\psi} - \theta = o_P(1)$ as $n \to \infty$.*

Proof: See Section 3.

Under a slightly stronger assumption about ψ

(P4') there are non-negative integers q_1, q_2 such that $q_1 + q_2 = 1$, and functions ψ_1, ψ_2 satisfying (P1) and (P2) such that $\psi = q_1\psi_1 + q_2\psi_2$, where $\psi_1(x) = \text{sign}(x)$ and ψ_2 is absolutely continuous on some interval $[-\Delta, \Delta]$, $\Delta > 0$, with a derivative ψ' satisfying $\psi'(x) \geq K$ for all $x \in [-\Delta, \Delta]$ — for some $K > 0$

Theorem 1 can be strengthened to yield consistency rates.

Theorem 2 *Suppose that X_{ni} satisfy the assumptions (H1)-(H3) of the heterogeneous location model; suppose further that this model is unimodal and conservative. If (P1)-(P3) and (P4') hold, then $T_{n,\psi} - \theta = O_P(n^{-1/2})$ as $n \to \infty$.*

Proof: This theorem directly follows from Theorem 6 of Hallin and Mizera (1996).

Theorems 1 and 2 show that robust M-estimators behave in unimodal and conservative heterogeneous location models like in the i.i.d. case, where it can be said that they are always consistent—as soon as the corresponding population values are identifiable (see Huber 1981, page 54). The assumptions of Theorems 1 and 2 are easily checked in the particular case of *heteroscedastic location models*: heterogeneous location models with distribution functions satisfying

$$F_{ni}(x) = F\left(\frac{x-\theta}{c_{ni}}\right),$$

where $c_{n1}, c_{n2}, \ldots, c_{nn}$ are positive *scaling constants* and F is the distribution function of a fixed *parent distribution.* Note that every heteroscedastic model with symmetric and/or unimodal F is itself symmetric and/or unimodal.

For heteroscedastic models, we are able to state necessary and sufficient consistency conditions, with rates, for the special case of the L1 estimator—the sample median. Compared to the general conditions established in Mizera and Wellner (1996), our condition is specially tailored for heteroscedastic models, since, under a very mild regularity condition

(S) the parent distribution admits a density f which is bounded, and there are $\lambda > 0$, $L > 0$ such that $f(x) \geq L$ for $x \in [-\lambda, \lambda]$,

which is clearly satisfied by any unimodal parent distribution with bounded density, it involves only the "empirical distribution" of the scaling constants. The conservativeness of the model is no longer required; our results hold, in particular, when some part of "probability mass" is allowed to "escape to infinity". Let Φ_c be the function from $(0,\infty)$ to $(0,\infty)$ defined by $\Phi_c(x) = 1/c$ if $x \leq c$ and $1/x$ if $x \geq c$.

Theorem 3 *Let $T_{n,\psi}$ be the sample median (that is, $\psi(x) = \text{sign}(x)$). If a heteroscedastic model satisfies (S), then $T_{n,\psi} - \theta = o_P(r_n^{-1})$ if and only if*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi_c(r_n c_{ni}) \to \infty \qquad \text{as } n \to \infty, \tag{1}$$

for any fixed $c > 0$.

Proof: See Section 3.

Note that the choice of c for Φ_c is inessential, due to the following elementary inequality, holding for any $c \leq d$

$$\Phi_c(x) \geq \Phi_d(x) \geq \frac{c}{d}\Phi_c(x).$$

In other words, (1) holds for all $c > 0$ as soon as it holds for one $c > 0$. Condition (1) implies that

$$\frac{1}{r_n\sqrt{n}} \sum_{i=1}^{n} \frac{1}{c_{ni}} \to \infty \quad \text{as } n \to \infty. \tag{2}$$

Under a non-degeneracy assumption, together with conservativeness and some additional regularity requirements, Sen (1968) proved an asymptotic normality result, from which (2) follows as a necessary and sufficient consistency condition. In fact, (1) and (2) are equivalent as soon as all $c_{ni} \geq c$ for some $c > 0$. For the particular case of plain $o_P(1)$ consistency ($r_n = 1$), we obtain that (1) is equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi_c(c_{ni}) \to \infty \quad \text{as } n \to \infty, \tag{3}$$

the condition already established in Hallin and Mizera (1996). Finally, Lemma 6 of Hallin and Mizera (1996) yields the following corollary.

Theorem 4 *Under the assumptions of Theorem 3, $T_{n,\psi} - \theta = O_P(s_n^{-1})$ if and only if (1) holds (for some $c > 0$) for any sequence r_n such that $r_n = o(s_n)$.*

Proof: A direct consequence of Theorem 3 and Lemma 6 of Hallin and Mizera (1996).

3 Proofs

In the proofs, we write $\theta = 0$, without loss of generality.

Lemma 1 *Suppose that the sequence $\bar{F}_n$, $n = 1, 2, \ldots$ is tight and that the corresponding densities $\bar{f}_n$ are unimodal with the common mode θ. Then, for every $\eta > 0$, there is a K_η such that*

$$1 - \eta \leq \int_{-K_\eta}^{K_\eta} \bar{f}_n(x)\, dx \tag{4}$$

and

$$\max\{\bar{f}_n(-K_\eta), \bar{f}_n(K_\eta)\} \leq \eta \tag{5}$$

for all $n = 1, 2, \ldots$.

Proof: First note that (4) is just the tightness condition rewritten in terms of densities. Turning to (5), assume that it does not hold: then, there is an $\eta > 0$ such that for all K

$$\text{either} \qquad \bar{f}_n(-K) > \eta \qquad \text{or} \qquad \bar{f}_n(K) > \eta. \tag{6}$$

Set $K = 2/\eta$ and suppose that $\bar{f}_n(-K) > \eta$, say. By unimodality,

$$\int_{-K}^{0} \bar{f}_n(x)\, dx \leq \int_{-K}^{0} \eta\, dx = 2$$

— a contradiction; the other case in (6) is treated analogously. At this point, we could possibly have one value of K for (4) and another one for (5) — but the maximum of them two works at both.

Proof of Theorem 1: For any $\varepsilon > 0$, let

$$a_n(\psi, \varepsilon) = E\lambda_n(\psi, \theta - \varepsilon) = \frac{1}{n} \sum_{i=1}^{n} \int \psi(x - \theta + \varepsilon)\, dF_{ni}(x)$$

and

$$b_n(\psi, \varepsilon) = E\lambda_n(\psi, \theta + \varepsilon) = \frac{1}{n} \sum_{i=1}^{n} \int \psi(x - \theta - \varepsilon)\, dF_{ni}(x).$$

In view of Theorem 1 of Hallin and Mizera (1996), it is sufficient to show that for any $\varepsilon > 0$, $a_n(\psi, \varepsilon)$ and $b_n(\psi, \varepsilon)$ are bounded away from zero for sufficiently large n. We give a proof for $a_n(\psi, \varepsilon)$; the proof for $b_n(\psi, \varepsilon)$ is entirely similar. In view of (H3), it is sufficient to show that, for any $\varepsilon > 0$ (setting again $\theta = 0$),

$$\int [\psi(x + \varepsilon) - \psi(x)]\, \bar{f}_n(x)\, dx > 0 \tag{7}$$

for sufficiently large n (note that, due to monotonicity, the integrand in (7) is non-negative).

Fix $\varepsilon > 0$. By (P4), there is a $\delta(\varepsilon/2)$ such that $\psi(\varepsilon/2) \geq \delta(\varepsilon/2)$; hence we have, for all $x \in [-\varepsilon/2, 0]$,

$$\psi(x+\varepsilon) - \psi(x) \geq \psi(x+\varepsilon) \geq \psi(\tfrac{1}{2}\varepsilon) \geq \delta(\tfrac{1}{2}\varepsilon). \tag{8}$$

Now, choose η and $C > 0$ such that for K_η given by Lemma 1 we have

$$\eta + CK_\eta \leq \min\{\tfrac{1}{2}\delta(\tfrac{1}{2}\varepsilon), \tfrac{1}{4}\}.$$

If $\bar{f}_n(\varepsilon/2) \geq C$, we have by (8)

$$\int [\psi(x+\varepsilon) - \psi(x)]\, \bar{f}_n(x)\, dx \geq \int_{-\varepsilon/2}^{0} \delta(\tfrac{1}{2}\varepsilon)\bar{f}_n(x)\, dx \geq \tfrac{1}{2}\varepsilon C\delta(\tfrac{1}{2}\varepsilon),$$

due to unimodality. If $\bar{f}_n(\varepsilon/2) \leq C$, Lemma 1 gives

$$\begin{aligned} \int_{-\infty}^{-\varepsilon/2} \bar{f}_n(x)\, dx &= \int_{-\infty}^{-K_\eta} \bar{f}_n(x)\, dx + \int_{-K_\eta}^{-\varepsilon/2} \bar{f}_n(x)\, dx \\ &\leq \eta + (K_\eta - \tfrac{1}{2}\varepsilon)C \leq \eta + CK_\eta \end{aligned}$$

due to unimodality again; hence,

$$\int_{-\varepsilon/2}^{\infty} \bar{f}_n(x)\, dx \geq 1 - \eta - CK_\eta$$

and, consequently,

$$\begin{aligned} &\int [\psi(x+\varepsilon) - \psi(x)]\, \bar{f}_n(x) dx = \int \psi(x+\varepsilon)\overline{f}_n(x) dx \\ &= \int_{-\infty}^{-\varepsilon/2} \psi(x+\varepsilon)\bar{f}_n(x)\, dx + \int_{-\varepsilon/2}^{\infty} \psi(x+\varepsilon)\bar{f}_n(x)\, dx \\ &\geq \int_{-\infty}^{-\varepsilon/2} -1\bar{f}_n(x)\, dx + \int_{-\varepsilon/2}^{\infty} \delta(\tfrac{1}{2}\varepsilon)\bar{f}_n(x)\, dx \\ &\geq -\eta - CK_\eta + \delta(\tfrac{1}{2}\varepsilon)(1 - \eta - CK_\eta) \\ &\geq -\tfrac{1}{2}\delta(\tfrac{1}{2}\varepsilon) + \tfrac{3}{4}\delta(\tfrac{1}{2}\varepsilon) = \tfrac{1}{4}\delta(\tfrac{1}{2}\varepsilon). \end{aligned}$$

In both cases, we have that (7) is bounded from below by

$$\min\{\tfrac{1}{2}\varepsilon C\delta(\tfrac{1}{2}\varepsilon), \tfrac{1}{4}\delta(\tfrac{1}{2}\varepsilon)\} > 0,$$

which proves the statement.

Proof of Theorem 3: Let $\psi(x) = \text{sign}(x)$, let f be a density of the parent distribution of a heteroscedastic model satisfying (S). Then, f is bounded by K; without loss of generality we may suppose that $K \geq 1$.

Necessity. By Theorem 3 of Hallin and Mizera (1996), $o_P(r_n^{-1})$ consistency implies that $\sqrt{n}a_n(\psi, r_n^{-1}) \to \infty$ as $n \to \infty$. Proceeding similarly as in the proof of Theorem 8 of Hallin and Mizera (1996), we obtain

$$\begin{aligned}
\sqrt{n}a_n(\psi, r_n^{-1}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_{-(r_n c_{ni})^{-1}}^{0} f(y)dy \\
&= \frac{1}{\sqrt{n}} \left[\sum_{r_n c_{ni} \leq 1} \int_{-(r_n c_{ni})^{-1}}^{0} f(y)\, dy + \sum_{r_n c_{ni} > 1} \int_{-(r_n c_{ni})^{-1}}^{0} f(y)\, dy \right] \\
&\leq K \frac{1}{\sqrt{n}} \left[\sum_{r_n c_{ni} \leq 1} 1 + \sum_{r_n c_{ni} > 1} \frac{1}{r_n c_{ni}} \right] \\
&= K \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi_1(r_n c_{ni}),
\end{aligned}$$

and (1) follows.

Sufficiency. Let (S) hold with λ and L; choose Γ and c such that $\Gamma = \lambda c$. If $0 \leq \varepsilon \leq \Gamma$, then, as in the proof of Theorem 7 of Hallin and Mizera (1996),

$$\begin{aligned}
a_n(\psi, \varepsilon r_n^{-1}) &= \frac{1}{n} \sum_{i=1}^{n} \int_{-\varepsilon(r_n c_{ni})^{-1}}^{0} f(y)dy \\
&= \frac{1}{n} \left[\sum_{r_n c_{ni} \leq c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^{0} f(y)\, dy + \sum_{r_n c_{ni} > c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^{0} f(y)\, dy \right] \\
&\geq \frac{1}{n} \left[\sum_{r_n c_{ni} \leq c} \int_{-\varepsilon c^{-1}}^{0} f(y)\, dy + \sum_{r_n c_{ni} > c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^{0} f(y)\, dy \right] \\
&\geq L\varepsilon \frac{1}{n} \left[\sum_{r_n c_{ni} \leq c} \frac{1}{c} + \sum_{r_n c_{ni} > c} \frac{1}{r_n c_{ni}} \right] \\
&= L\varepsilon \frac{1}{n} \sum_{i=1}^{n} \Phi_c(r_n c_{ni}).
\end{aligned}$$

An entirely similar argument for $b_n(\psi, \varepsilon r_n^{-1})$ and the subsequent application—assuming (1)—of Theorems 2 and 3 of Hallin and Mizera (1996) conclude the proof.

Acknowledgements

For the first author the research was supported by the Fonds d'Encouragement à la Recherche de l'Université Libre de Bruxelles and the European Human Capital contract ERB CT CHRX 940963 and for the second author, the research was supported by the Fonds National de la Recherche Scientifique, the Banque Nationale de Belgique, and Slovak GAS grant 1/1489/94.

References

[1] Freedman, D. A. and Diaconis, P. (1982). On inconsistent M-estimators. *Ann. Statist.* **10**, 454–461.

[2] Hallin, M. and Mizera, I. (1995). Sample heterogeneity and the asymptotics of M-estimators. Preprint IS-P 1996-15 (No. 49), Institut de Statistique de l'Université Libre de Bruxelles, Brussels.

[3] Hoaglin D. C., Mosteller F. M., and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

[4] Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

[5] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* New York: Wiley.

[6] Mizera, I. (1994) On consistent M-estimators: tuning constants, unimodality and breakdown. *Kybernetika* **30**, 289–300.

[7] Mizera, I. and Wellner, J. A. (1996). Necessary and sufficient conditions for the consistency of the sample median of independent but not identically distributed random variables. Preprint IS-P 1996-6 (No. 40), Institut de Statistique de l'Université Libre de Bruxelles, Brussels.

[8] Sen, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Statist.* **39**, 1724–1730.

[9] Stigler, S. M. (1976). The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. *J. Amer. Statist. Assoc.* **71**, 956–960.