The Density of Multivariate M-Estimates

Anthony Almudevar*, Chris Field* and John Robinson†

May 8, 2000

Abstract

When a unique M-estimate exists its density is obtained as a corollary to a more general theorem which asserts that under mild conditions the intensity function of the point process of solutions of the estimating equations exists and is given by the density of the estimating function standardized by multiplying it by the inverse of its derivative. We apply the results to give a result for Huber's Proposal 2 applied to regression and scale estimates. We also give a saddlepoint approximation for the density and use this to give approximations for tail areas for smooth functions of the M-estimates.

Key words: intensity, M-estimator, point process.

1 Introduction

For a random vector X defined on a probability space (Ω, Γ, P), where P is a member of some family $\mathcal{P}$, and a parameter space $\Theta \subset R^p$, we are interested in M-estimates defined as

*Department of Mathematics and Statistics, Dalhousie University, Halifax, N. S., Canada, B3H 3J5

†School of Mathematics and Statistics, University of Sydney, N.S.W. 2006, Australia.

solutions to

$$\Psi(X,\theta) = 0. \tag{1.1}$$

To enable us to consider such estimates we make the following assumption:

(A1) Let $\Psi(x,\theta)$ be a map from $R^m \times \Theta$ to R^p, which is Borel measurable with respect to x for every $\theta \in \Theta$, and, for every $x \in R^m$ and for any $\theta \in \Theta$ and vector v of length 1 such that $\theta + yv \in \Theta,\ 0 \le y \le h$, let $\Psi(x, \theta + yv)$ have a derivative with respect to y, almost everywhere in $(0,h)$.

Write $\Psi'(X,\theta)$ for the $p \times p$ matrix with (i,j)th element $\partial\Psi_i(X,\theta)/\partial\theta_j$.

Multiple solutions may exist. This makes it important to determine when the set of solutions is a point process with an intensity function absolutely continuous with respect to Lebesgue measure. Conditions under which this holds are given in the next section. This becomes the density of the estimate when there is a unique solution. In the special case of minimum contrast estimators, Skovgaard (1990), and recently Jensen and Wood (1998), have used this concept.

Let

$$\Psi^*(X,\theta) = \Psi'(X,\theta)^{-1}\Psi(X,\theta),$$

when the inverse exists and take it as infinite elsewhere. If we restrict attention to a sequence of sets in Γ for which the density of $\Psi^*(X,\theta)$ exists, then a limit of this density at $x = 0$ is shown to equal the limit of intensity functions defined on the same sequence of sets. The conditions imposed here differ from those of Skovgaard (1990) and his result is a special case of ours. Jensen and Wood (1998, Section 5) have a similar result to Skovgaard (1990) and our result covers cases not included in either of these works.

In the case when X represents a set of iid random variables $X_1, \cdots, X_n$, $\Psi(X,\theta) = \sum_{i=1}^n \psi(X_i,\theta)$ and $\bar{\Psi} = \Psi(X,\theta)/n$ and $\bar{\Psi}' = \Psi'(X,\theta)/n$, so $\Psi^*(X,\theta)$ is a smooth function of means and thus we can apply the usual tilting methods to obtain a saddlepoint approximation to the density of $\Psi^*(X, \theta)$. From this we are able to obtain accurate saddlepoint approximations of the P-values for studentized M-estimates. These saddlepoint approximations are typically very accurate in the extreme tails.

It is not necessary in this treatment to consider the parameter which may be regarded as given by solutions of $\lambda(\theta) = E\Psi(X, \theta) = 0$. This again may have multiple solutions but if for a solution θ^* we assume that $\lambda'(\theta)$ exists and is continuous in a neighbourhood of θ^* and that $\det(\lambda'(\theta^*)) \neq 0$, where $\det(M)$ is the determinant of a square matrix M, then the inverse function theorem ensures that there is an open neighbourhood of θ^* in which θ^* is the unique solution.

As noted earlier, Jensen and Wood (1998) consider a problem similar to that addressed in this paper. They look at the probabilistic behaviour of a minimum contrast estimator in the case of independent identically distributed random vectors and restrict attention to contrast functions γ which are the sum of n terms. In addition to giving a proof of Skovgaard's results, they obtain results which show that sets of the form $\{|\hat{\theta} - \theta_0| > \delta\}$ have exponentially small probability with several selection criteria for choosing the minimum contrast estimator. Also they give a tilting argument to approximate the density of $\Psi^*(X, \theta)$ at 0 in our notation. Since they are considering a situation very similar to that considered in this paper, it is important to contrast both our setting and results with theirs.

As a first point, we are considering M-estimates which are defined as solutions of the score equation. Many of the robust M-estimates in common use cannot be viewed as minimum contrast estimators since we cannot integrate the score function Ψ to get a criterion or contrast function γ. In their proof of Skovgaard's result and the resulting tilting argument, they have assumed that the joint density of $(\bar{\Psi}, \bar{\Psi}')$ (D_1 and D_2 in their notation) is continuous and bounded. For many robust estimates which use Huber's score function, the distribution of $(\bar{\Psi}, \bar{\Psi}')$ is made up of a continuous and discrete part and the continuous part may be degenerate in $\mathbb{R}^{p+p^2}$, the dimension of $(\bar{\Psi}, \bar{\Psi}')$. Huber's proposal 2 either for location-scale or regression-scale, falls into this group. In section 2, we have obtained Skovgaard's result without the necessity of having a joint continuous density but rather under the assumption of a density for Ψ^* near 0. Our result, as well as our saddlepoint argument, is able to handle the more general case where we have a discrete part of $\bar{\Psi}'$ and hence can be applied to Huber's proposal 2 and other similar robust estimators. It is worth commenting on the conditions used in both papers. Jensen and Wood have several conditions on the contrast function and its local behaviour for which we have no analogue. They also require existence of the second derivative of the score function and moments on the supremum of its local behaviour (see their theorem 5.1). We have not required the existence of the second derivative of Ψ for our results, nor do we impose any moment conditions. Finally we should point out that they have addressed the behaviour of the minimum contrast estimator under various selection criteria. This is not an issue we've addressed directly in our paper.

In section 2 we consider the general case and examine the point process of the solutions to the score function $\Psi(X, \theta) = 0$. Lemma 1 proves that there is a unique solution to the score function under some conditions. This result is used for the iid case in section 3 in discussing multiple solutions. The main result of the section, theorem 1, establishes Skovgaard's result relating the density of Ψ^* at 0 to the intensity of the point process of the solutions under less restrictive conditions than either Skovgaard (1990) or Jensen and Wood (1998). In section 3, we introduce the case of Huber's proposal 2 for regression-scale estimates and in section 4, derive a saddlepoint approximation to the density of M-estimates under conditions which are satisfied by the regression-scale model of section 3. A tail area result for a smooth function of the M-estimate is given in section 5 and we conclude with some numerical results in section 6.

2 The Density of the M-Estimate

We will use the following notation. For any $x \in \mathbb{R}^p$ let $||x|| = \max_i |x_i|$ denote the modulus norm of a vector $x = (x_1, \ldots, x_p)$. For a matrix M we set $||M|| = \sup ||Mx||/||x||$. For any $x \in \mathbb{R}^p$, $\delta > 0$, let $B_\delta(x) = \{y \in \mathbb{R}^p : ||x - y|| < \delta\}$. Let $m(\cdot)$ denote Lebesgue measure on $\mathbb{R}^p$ and let $I\{E\}$, for any $E \in \Gamma$, be the indicator function on Ω. The $p \times p$ identity matrix will be written I_p. The density of a random quantity Y, when it exists, will generally be written f_Y. We will also use the shorthand $f_Y(y; E) = f_Y(y|E)P\{E\}$. We write $X =^d Y$ if X equals Y in distribution. The distribution function and density of the unit normal will be written Φ and ϕ. We use the notation $\mu_1 \ll \mu_2$ when measure μ_1 is absolutely continuous with respect to μ_2, and a Borel measure μ on $\mathbb{R}^p$ will be called *boundedly finite* if $\mu(A) < \infty$ when A is bounded.

Suppose we have score function $\Psi(X, \theta)$ and derivative matrix $\Psi'(X, \theta)$ where X is a random vector on a probability space $(\Omega, \mathcal{F}, P)$. Define

$$\Psi^*(X, \theta) = \begin{cases} \Psi'(X, \theta)^{-1}\Psi(X, \theta), & \det(\Psi'(X, \theta)) \neq 0, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.2}$$

Define

$$z(\theta_0, \tau) = \begin{cases} \sup_{\theta \in B_\tau(\theta_0)} ||\Psi'(X, \theta_0)^{-1}\Psi'(X, \theta) - I_p||, & \det(\Psi'(X, \theta)) \neq 0, \forall \theta \in B_\tau(\theta_0), \\ \infty, & \text{otherwise,} \end{cases} \tag{2.3}$$

and

$$H(\theta_0, \alpha, \tau) = \{z(\theta_0, \tau) < \alpha\} \subset \Omega. \tag{2.4}$$

Before proceeding with the definitions we obtain a preliminary technical result on existence of a unique solution in a neighbourhood of some θ_0.

Lemma 1 *If $X \in H(\theta_0, \alpha, \tau)$ with $\alpha < 1$, if (A1) holds and if $||\Psi^*(X, \theta_0)|| \leq (1-\alpha)\tau$, then there exists a unique solution, θ^*, of $\Psi(X, \theta) = 0$ in $B_\tau(\theta_0)$.*

Proof. Define, for each $\theta_0 \in \Theta$, and $X \in \Omega$, the mapping $T_X(\cdot; \theta_0) : \Theta \to \mathbb{R}^p$ by

$$T_X(\theta; \theta_0) = \theta - \Psi'(X, \theta_0)^{-1}\Psi(X, \theta).$$

If θ_1 and $\theta_2 = \theta_1 + hv$, with $||v|| = 1$, are in $\bar{B}_\tau(\theta_0)$, then using (A1),

$$\begin{aligned}
&||T_X(\theta_2; \theta_0) - T_X(\theta_1; \theta_0)|| \\
&= ||\Psi'(X, \theta_0)^{-1}(\Psi(X, \theta_2) - \Psi(X, \theta_1)) - (\theta_2 - \theta_1)|| \\
&= ||\Psi'(X, \theta_0)^{-1} \int_0^h \frac{d}{dy}\Psi(X, \theta + yv)dy - (\theta_2 - \theta_1)|| \\
&= ||\Psi'(X, \theta_0)^{-1} \int_0^h \sum_{i=1}^p v_i \left[\frac{\partial}{\partial t_i}\Psi(X, t)\right]_{(t=\theta_1+yv)} dy - (\theta_1 - \theta_2)|| \\
&\leq \int_0^h ||\Psi'(X, \theta_0)^{-1}\Psi'(X, \theta_1 + yv) - I||dy \\
&\leq \alpha||\theta_1 - \theta_2||.
\end{aligned}$$

Also if $\theta \in \bar{B}_\tau(\theta_0)$, then

$$\begin{aligned}
||T_X(\theta; \theta_0) - \theta_0|| &\leq ||\Psi'(X, \theta_0)^{-1}(\Psi(X, \theta) - \Psi(X, \theta_0)) - (\theta - \theta_0)|| + ||\Psi^*(X, \theta_0)|| \\
&\leq ||\Psi'(X, \theta_0)^{-1}\Psi'(X, \theta_3) - I_p|| \cdot ||\theta - \theta_0|| + (1-\alpha)\tau \\
&\leq \tau.
\end{aligned}$$

So $T_X(\cdot; \theta_0)$, restricted to $\bar{B}_\tau(\theta_0)$, is a contraction mapping. Now, by the fixed point theorem for contraction mappings (*cf* Theorem 3.1.1, Edwards (1965)) there is a unique fixed point, θ^*, say, for $T_X(\cdot; \theta_0)$. So $\theta^* = \theta^* - \Psi'(X, \theta_0)^{-1}\Psi(X, \theta^*)$. Thus $\Psi(X, \theta^*) = 0$ and this value is unique.

For $\alpha, \tau > 0$, define the point processes on Θ

$$Q_{\alpha,\tau}(A) = \{\theta \in A \subset \Theta : \Psi(X, \theta) = 0 \text{ and } X \in H(\theta, \alpha, \tau)\}. \tag{2.5}$$

Let

$$Q(A) \;\; = \;\; \cap_{\alpha>0} \cup_{\tau>0} Q_{\alpha,\tau}(A).$$

Then define the associated counting processes

$$N_{\alpha,\tau}(A) = \#Q_{\alpha,\tau}(A) \text{ and } N(A) = \#Q(A),$$

with intensity measures

$$\mu_{\alpha,\tau}(A) = E[N_{\alpha,\tau}(A)] \text{ and } \mu(A) = E[N(A)].$$

In this paper we are interested in the distributions of the locations of points in Q, which is the set of solutions to the score equation $\Psi(X, \theta) = 0$ at which the derivative matrix is invertible and locally continuous. To do this define, where the limit exists, the intensities

$$\lambda_{\alpha,\tau}(\theta) \;\; = \;\; \lim_{\delta \to 0} \frac{P(N_{\alpha,\tau}(B_\delta(\theta)) \geq 1)}{m(B_\delta(\theta))}.$$

If the intensities $\lambda_{\alpha,\tau}$ exist for small enough α and τ, then since $H(\theta, \alpha, \tau)$ is increasing as τ approaches 0 and decreasing as α approaches 0 the limit

$$\lambda(\theta) = \lim_{\alpha \to 0} \lim_{\tau \to 0} \lambda_{\alpha,\tau}(\theta) \tag{2.6}$$

will also exist. Similarly, we may define

$$f_{\Psi^*(X,\theta)}(z; H(\theta, \alpha, \tau)) = \lim_{\delta \to 0} \frac{P(\{\Psi^*(X; \theta) \in B_\delta(z)\} \cap H(\theta, \alpha, \tau))}{m(B_\delta(z))} \tag{2.7}$$

and

$$h(\theta) = \lim_{\alpha \to 0} \lim_{\tau \to 0} f_{\Psi^*(X,\theta)}(0; H(\theta, \alpha, \tau)) \tag{2.8}$$

when the limits exist. Under suitable regularity conditions, to be discussed below, we will have $\mu \ll m$ and λ will be a version of $d\mu/dm$. Then, under these conditions we will have $\lambda = h$.

We need an assumption concerning the density of $\Psi^*(X, \theta)$:

(A2) For any compact set $A \subset \Theta$ and for any $0 < \alpha < 1$, there exists $\tau > 0$ and $\delta > 0$ such that $f_{\Psi^*(X;\theta)}(z; H(\theta, \alpha, \tau))$ exists and is continuous and bounded by some fixed constant K for any $\theta \in A$ and $z \in B_\delta(0)$.

Remark 1: In situations where X represents n independent identically distributed random vectors, the constant K, and the constants α, τ, δ, in (A2) may depend on n. The scaling involved does not affect the existence of a density for each value of n.

Theorem 1 *Assume that (A1) and (A2) hold, then*

(i) $\mu_{\alpha,\tau} \ll m$ *and* $\mu \ll m$*,*

(ii) λ *is a density of* μ *on* A*,*

(iii) $\lambda(\theta) = h(\theta)$*, for* $\theta \in A$*.*

Proof. Fix $\frac{1}{3} > \alpha > 0$, choose $\tau > 0$ such that $f_{\Psi^*(X,\theta_0)}(0; H(\theta_0, \alpha, \tau))$ exists and take then $0 < \delta < \tau/2$. Suppose that

$$N_{\alpha,\tau}(B_\delta(\theta_0)) \geq 1.$$

Then there exists $\theta^* \in \bar{B}_\delta(\theta_0)$ such that $\Psi(X, \theta^*) = 0$ and $X \in H(\theta^*, \alpha, \tau)$, by the definition of $Q_{\alpha,\tau}$. So for $\theta \in \bar{B}_\delta(\theta_0)$, we must have $\theta \in \bar{B}_\tau(\theta^*)$. Thus

$$||\Psi'(X, \theta^*)^{-1}\Psi'(X, \theta_0) - I_p|| \leq \alpha \tag{2.9}$$

and

$$||\Psi'(X,\theta^*)^{-1}\Psi'(X,\theta) - I_p|| \le \alpha. \tag{2.10}$$

So from (2.9) and the Banach Lemma (*cf* Lemma 5.1, Noble and Daniel, (1977)):

$$\frac{1}{1+\alpha} \le ||[\Psi'(X,\theta^*)^{-1}\Psi'(X,\theta_0)]^{-1}|| \le \frac{1}{1-\alpha};$$

Then since $0 < \alpha < 1/3$.

$$\begin{aligned}
&||\Psi'(X,\theta_0)^{-1}\Psi'(X,\theta) - I_p|| \\
&\le \; ||\Psi'(X,\theta_0)^{-1}\Psi'(X,\theta^*)|| \times ||\Psi'(X,\theta^*)^{-1}\Psi'(X,\theta) - I_p - (\Psi'(X,\theta^*)^{-1}\Psi'(X,\theta_0) - I_p)|| \\
&\le \; \frac{2\alpha}{1-\alpha} \le 3\alpha,
\end{aligned}$$

Thus

$$X \in H(\theta_0, 3\alpha, \tau/2).$$

Also from the argument in Lemma 1 leading to (2.5) for any $\theta \in B_\delta(\theta_0)$,

$$||\Psi'(X,\theta_0)^{-1}(\Psi(X,\theta) - \Psi(X,\theta_0)) - (\theta - \theta_0))|| \le 3\alpha\delta.$$

Also since $\Psi(X,\theta^*) = 0$

$$|| - \Psi'(X,\theta_0)^{-1}\Psi(X,\theta_0) - (\theta^* - \theta_0))|| \le 3\alpha\delta.$$

Thus

$$||\Psi^*(X,\theta_0)|| \le (1+3\alpha)\delta.$$

So

$$\{N_{\alpha,\tau}(B_\delta(\theta_0)) \ge 1\} \subset \{\Psi^*(X,\theta_0) \in B_{(1+3\alpha)\delta}(0)\} \cap H(\theta_0, 3\alpha, \tau/2)$$

and so

$$P(N_{\alpha,\tau}(B_\delta(\theta_0)) \ge 1) \le P(\{\Psi^*(X,\theta_0) \in B_{(1+3\alpha)\delta}(0)\} \cap H(\theta_0, 3\alpha, \tau/2)). \tag{2.11}$$

We use this to show that $\lambda_{\alpha,\tau}$ exists and is a density of $\mu_{\alpha,\tau}$. For any $\epsilon > 0$ we may define for each $n \geq 1$ a sequence of collections of balls $\mathcal{A}^{(n)} = \{A_1^{(n)}, A_2^{(n)}, \ldots\}$ with centers in A, of maximum radius $1/n$ with $A \subset \cup_i A_i^{(n)}$ and $\sum_i m(A_i^{(n)}) \leq m(A) + \epsilon$. Suppose the balls in $\mathcal{A}^{(n)}$ have centers $\{\theta_i^{(n)}; i \geq 1\}$ and radii $\{\delta_i^{(n)} : i \geq 1\}$. Then define

$$N_{\alpha,\tau}^{(n)}(A) = \sum_i I\{N_{\alpha,\tau}(A_i^{(n)}) \geq 1\}.$$

By (2.11) we have

$$\begin{aligned} E[N_{\alpha,\tau}^{(n)}(A)] &= \sum_i P(N_{\alpha,\tau}(A_i^{(n)}) \geq 1) \\ &\leq \sum_i P(\{\Psi^*(X,\theta) \in B_{\delta_i^{(n)}}(0)\} \cap H(\theta, 3\alpha, \tau/2)). \end{aligned}$$

From (A2) we have for some fixed K

$$\sup_{\theta \in A} \sup_{z \in B_\delta(0)} f_{\Psi^*(X,\theta)}(z; H(\theta, 3\alpha, \tau/2)) \leq K$$

hence

$$\begin{aligned} E[N_{\alpha,\tau}^{(n)}(A)] &\leq \sum_i K m(A_i^{(n)}) \\ &\leq K(m(A) + \epsilon). \end{aligned}$$

The points in $Q_{\alpha,\tau}(A)$ are locally unique, that is they have no accumulation points, since by the inverse function theorem there is an open neighbourhood of any point in $Q_{\alpha,\tau}(A)$, which may depend on X, in which the solution is unique. So we have

$$N_{\alpha,\tau}(A) \leq \liminf_{n\to\infty} N_{\alpha,\tau}^{(n)}(A)$$

so by Fatou's Lemma

$$E[N_{\alpha,\tau}(A)] \leq K(m(A) + \epsilon).$$

By definition

$$N(A) = \lim_{\alpha \to 0} \lim_{\tau \to 0} N_{\alpha,\tau}(A)$$

and hence, by monotone convergence, since $H(\theta, \alpha, \tau)$ is increasing as τ approaches 0 and decreasing as α approaches 0,

$$E[N(A)] \leq K(m(A) + \epsilon). \tag{2.12}$$

Thus $\mu_{\alpha,\tau}$ and μ are boundedly finite and by letting ϵ approach 0 we may conclude that $\mu_{\alpha,\tau} \ll m$ and $\mu \ll m$. We then have the existence of the Radon-Nikodym derivative calculable by

$$\frac{d\mu_{\alpha,\tau}}{dm}(\theta) = \lim_{\delta \to 0} \frac{\mu_{\alpha,\tau}(B_\delta(\theta))}{m(B_\delta(\theta))}$$

a.e.$[m]$ on A. Then, since $\mu_{\alpha,\tau}$ is boundedly finite and $N_{\alpha,\tau}$ is simple (*cf* Definition 3.3.II, Daley and Vere-Jones (1988)) we have (*cf* Proposition 7.2.VIII, Daley and Vere-Jones (1988))

$$\lim_{\delta \to 0} \frac{P(N(B_\delta(\theta)) \geq 1)}{\mu_{\alpha,\tau}(B_\delta(\theta))} = 1$$

a.e.$[\mu_{\alpha,\tau}]$. Hence $\lambda_{\alpha,\tau}$ exists *a.e.*$[\mu_{\alpha,\tau}]$, and so *a.e.*$[m]$, and is a density for $\mu_{\alpha,\tau}$.

Now in (2.11) divide both sides by $m(B_\delta(0))$ and let δ approach 0 and we obtain

$$\lambda_{\alpha,\tau}(\theta_0) \leq (1 + 3\alpha)^p f_{\Psi^*(X,\theta_0)}(0; H(\theta_0, 3\alpha, \tau/2)), \tag{2.13}$$

a.e.$[m]$.

Conversely, suppose

$$X \in H(\theta_0, \alpha, \tau) \text{ and } \Psi^*(X, \theta_0) \in B_{(1-\alpha)\delta}(0),$$

for $\delta < \tau$. So from Lemma 1, there exists a unique solution, θ^*, of $\Psi(X, \theta) = 0$ in $\bar{B}_\delta(\theta_0)$. Thus

$$\{\Psi^*(X, \theta_0) \in B_{(1-\alpha)\delta}(0)\} \cap H(\theta_0, \alpha, \tau) \subset \{N_{\alpha,\tau}(\bar{B}_\delta(\theta_0)) \geq 1\}.$$

So

$$P(\{\Psi^*(X,\theta_0) \in B_{(1-\alpha)\delta}(0)\} \cap H(\theta_0,\alpha,\tau)) \leq P(N_{\alpha,\tau}(\bar{B}_\delta(\theta_0)) \geq 1)$$

and dividing both sides by $m(B_\delta(0))$ and letting δ approach 0 gives

$$(1-\alpha)^p f_{\Psi^*(X,\theta_0)}(0; H(\theta_0,\alpha,\tau)) \leq \lambda_{\alpha,\tau}(\theta_0), \tag{2.14}$$

a.e.[m]. The results (2.13) and (2.14) are true for small enough $\alpha > 0$ and $\tau > 0$ which suffices to give $\lambda(\theta) = h(\theta)$ for $\theta \in A$.

Now it remains only to identify λ with the density of μ. For any measurable $B \subset A$, using monotone convergence as in (2.12).

$$\begin{aligned} \mu(B) = E(N(B)) &= \lim_{\alpha\to 0}\lim_{\tau\to 0} E(N_{\alpha,\tau}(B)) \\ &= \lim_{\alpha\to 0}\lim_{\tau\to 0} \int_B \lambda_{\alpha,\tau}(\theta)d\theta \\ &= \int_B \lambda(\theta)d\theta \end{aligned}$$

for measurable $B \subset A$.

Corollary 1 *If (A1) and (A2) hold and with probability 1 there is a unique M-estimate, $\hat{\theta}$, then $h(\theta)$ is a density of $\hat{\theta}$.*

Remark 2: If there is either a unique M-estimate with probability p_0 or no solution such that $X \in H(\theta,\alpha,\tau)$ for some α, τ with probability $1-p_0$, and (A1) and (A2) hold, then $h(\theta)$ is an improper density with $\int_\Theta h(\theta)d\theta = p_0$. This is the case, for example, if $X_1,\cdots,X_n$ are independent identically distributed random variables with densities and $\Psi(X,\theta) = \sum_{j=1}^n h_b(X_j-\theta)$, where $h_b(x) = \max(-b, \min(x,b))$. Then there is a unique solution unless n is even and the set of $X_1,\cdots,X_n$ can be divided into two equal sized sets with minimum distance between the sets at least $2b$, in which case the point processes $Q_{\alpha,\tau}(A)$ are all null.

3 Density of the Huber Estimate of Regression and Scale

Huber (1964) proposed the following score equations for the problem of estimating the regression and scale parameters for the independent real valued random vectors (X_i, C_i) from densities $\sigma^{-1} f(\sigma^{-1}(x_i - g_i(\gamma))|c_i) f_1(c_i)$, where $g_i(\gamma) = \sum_{i'=1}^{k} c_{ii'}\gamma_{i'}$, for fixed values $c_{ii'},\ i = 1, \cdots, n,\ i' = 1, \cdots, k$, such that the $n \times k$ matrix $C = (c_{ii'})$ is of rank $k < n$ and $\theta = (\gamma, \sigma)$ are the unknown parameters:

$$\Psi(X, \gamma, \sigma) = \begin{bmatrix} \sum_{i=1}^{n} h_b(\frac{x_i - g_i(\gamma)}{\sigma}) g_i'(\gamma) \\ \frac{1}{2}\sum_{i=1}^{n}(h_b^2(\frac{x_i - g_i(\gamma)}{\sigma}) - \beta) \end{bmatrix} \tag{3.15}$$

setting

$$\beta = E[h_b^2(Z)] \tag{3.16}$$

where Z has density f and $h_b(x) = \max(-b, \min(x, b))$ is the Huber function. This corresponds to Huber's "Proposal 2" applied to multiple regression. Define $I_{i\theta} = I\left(\frac{X_i - g_i(\gamma)}{\sigma} \in [-b, b]\right)$, $I_{i\theta}^{+} = I\left(\frac{X_i - g_i(\gamma)}{\sigma} > b\right)$ and $I_{i\theta}^{-} = I\left(\frac{X_i - g_i(\gamma)}{\sigma} < -b\right)$. Put $n(\theta) = \sum_{i=1}^{n} I_{i\theta}$, $n^{+}(\theta) = \sum_{i=1}^{n} I_{i\theta}^{+}$ and $n^{-}(\theta) = \sum_{i=1}^{n} I_{i\theta}^{-}$. Then

$$\begin{aligned} \Psi(X, \gamma, \sigma) &= \begin{bmatrix} \sum_{i=1}^{n} I_{i\theta}\frac{X_i - g_i(\gamma)}{\sigma} g_i'(\gamma) + \sum_{i=1}^{n}(I_{i\theta}^{+} - I_{i\theta}^{-}) b g_i'(\gamma) \\ \frac{1}{2}\sum_{i=1}^{n} I_{i\theta}\left(\left(\frac{X_i - g_i(\gamma)}{\sigma}\right)^2 - \beta\right) + \frac{1}{2}\sum_{i=1}^{n}(I_{i\theta}^{+} + I_{i\theta}^{-}) g_i'(\gamma)(b^2 - \beta) \end{bmatrix} \\ \Psi'(X, \gamma, \sigma) &= \sigma^{-1}\begin{bmatrix} -A & -\sum_{i=1}^{n} I_{i\theta}\frac{X_i - g_i(\gamma)}{\sigma} g_i'(\gamma) \\ -\sum_{i=1}^{n} I_{i\theta}\frac{X_i - g_i(\gamma)}{\sigma} g_i'(\gamma)^T & -\sum_{i=1}^{n} I_{i\theta}\left(\frac{X_i - g_i(\gamma)}{\sigma}\right)^2 \end{bmatrix} \end{aligned}$$

where $A = (a_{jj'}) = (\sum_{i=1}^{n} c_{ij} c_{ij'} I_{i\theta})$. Then we have $\det(\Psi'(X, \gamma, \sigma))$ is equal to

$$\frac{\det A}{\sigma^2}\left(\sum_{i=1}^{n}\left(\frac{X_i - g_i(\gamma)}{\sigma}\right)^2 I_{i\theta} - \left(\sum_{i=1}^{n}\frac{X_i - g_i(\gamma)}{\sigma} I_{i\theta}\right)^T A^{-1}\left(\sum_{i=1}^{n}\frac{X_i - g_i(\gamma)}{\sigma} I_{i\theta}\right)\right)$$

which is nonnegative and we can choose b so that this determinant equals zero if and only if $n(\theta) \leq k$.

In Huber (1964) conditions are given under which the score equation for Huber's proposal 2 has a unique solution. For the regression case given here we can show by an analogous argument, largely given in Huber (1981, Section 7.7), that there is a unique solution if

$$b^2 > \beta \text{ and } n > k(1 - \beta/b^2)^{-1}. \tag{3.17}$$

In this case if $H(\theta, \alpha, \tau)$ is defined as in (2.4) then with probability 1 for any $\alpha,\ \tau,$

$$\{n(\theta) > k\} \supset H(\theta, \alpha, \tau) \supset \{n(\theta) > k\} \cap \cap_{i=1}^{n} \{|\frac{X_i - g_i(\gamma)}{\sigma} - b| > \epsilon(\alpha, \tau)\},$$

for some $\epsilon(\alpha, \tau)$, since this excludes values of X such that discontinuities occur in $\Psi'(X, \gamma, \sigma)$ in a small neighbourhood of (γ, σ) or $\det(\Psi'(X, \gamma, \sigma)) = 0$; also we may take $\epsilon(\alpha, \tau)$ tending to 0 as α and τ tend to 0. So, if (3.17) holds and if $X_1, \cdots, X_n$ have densities, then $f_{\Psi^*(X,\gamma,\sigma)}(z; H(\theta, \alpha, \tau))$ can be defined by (2.7) for any $\alpha > 0$ and $\tau > 0$, for small enough z. This is also true if we consider only the conditional densities of $X_1, \cdots, X_n$ conditionally on fixed C. Also (A1) and (A2) are both satisfied and the solution is unique, so Corollary 1 gives the density of $\theta = (\gamma, \sigma)$ as $h(\theta)$, defined by the limit in (2.8).

Remark 3: In the particular case when $b = \infty$ the density clearly exists if the densities of $X_1, \cdots, X_n$ are bounded and continuous. If we assume that the random variables are normal with mean γ_0 and variance σ_0 then elementary application of the result leads to the usual density of the estimate.

4 The Saddlepoint Approximation

Consider the case where we have iid observations $X_1, X_2, \cdots, X_n$ from a distribution F_0. We have a function $\psi(X_1, \theta)$ which assumes values in $\mathbb{R}^p$, and a score function

$$\Psi(X, \theta) = \sum_{i=1}^{n} \psi(X_i, \theta).$$

Suppose that $\int \psi(x, \theta) dF_0(x) = 0$ has a solution θ_0. Suppose that $\psi(X_1, \theta)$ has a derivative $\psi'(X_1, \theta)$ with respect to θ, with probability 1, and assume

(A3) $\det\left(\int \psi'(x, \theta_0) dF_0(x)\right) \neq 0$.

Then, if for some $\tau > 0$, $\int \psi'(x, \theta) dF_0(x)$ is continuous at all $\theta \in B_\tau(\theta_0)$, the solution θ_0 is the unique solution in $B_\tau(\theta_0)$. In order to give results which hold for cases of interest such as Huber's proposal 2 for regression and scale (see section 3), we have to allow for the fact that (ψ, ψ') may have both continuous and discrete parts and that the joint density for the continuous part may be degenerate in $\mathbb{R}^{p+p^2}$, the dimension of (ψ, ψ'). In fact, in these cases

$$\Psi^*(X, \theta) = [\sum_{i=1}^{n} \psi'(X_i, \theta)]^{-1} \sum_{i=1}^{n} \psi(X_i, \theta), \text{ for } \det(\sum_{i=1}^{n} \psi'(X_i, \theta)) \neq 0,$$

may not have a density since it is defined to be infinite when $\det(\sum_{i=1}^{n} \psi'(X_i, \theta)) = 0$.

Let m_p be Lebesgue measure on $\mathbb{R}^p$. For each θ let $M_\theta \subset \mathbb{R}^p$ be a set of Lebesgue measure zero such that, by the Lebesgue decomposition theorem,

$$P(\psi(X_1, \theta) \in A) = P(\psi(X_1, \theta) \in A \cap M_\theta) + \int_A f_\theta dm_p,$$

where f_θ is a possibly improper density. Let $I_{i\theta} = 0$ if $\psi(X_i, \theta) \in M_\theta$, and 0, otherwise. Consider the following assumption:

(A4) Assume that there are iid random vectors $U_{i\theta} = (W_{i\theta}, V_{1i\theta}, V_{2i\theta})$, where $W_{i\theta}$ are jointly continuous random vectors of dimension p and $V_{1i\theta}$ and $V_{2i\theta}$ are random vectors of dimension p and p^*, respectively, such that

$$\begin{aligned} n\bar{T}_\theta = \sum_{i=1}^{n} \psi(X_i, \theta) &= \sum_{i=1}^{n} I_{i\theta} W_{i\theta} + \sum_{i=1}^{n} (1 - I_{i\theta}) V_{1i\theta} \\ vec(n\bar{S}_\theta) = vec(\sum_{i=1}^{n} \psi'(X_i, \theta)) &= A_\theta \sum_{i=1}^{n} I_{i\theta} U_{i\theta}, \end{aligned}$$

where A_θ is of dimension p^2 by $2p + p^*$. Assume further that the components of $V_{i\theta} = (V_{1i\theta}, V_{2i\theta})$ are either continuous or lattice with dimension $d = p + p^*$ and that $d = d_1 + d_0$ where d_1 is the number of continuous variables and d_0 is the number of lattice variables.

Let $U'_{j\theta} = (W'_{j\theta}, V'_{j\theta})$ have the distribution of $U_{j\theta}$ conditional on $I_{j\theta} = 1$ and $V''_{1j\theta}$ have the distribution of $V_{1j\theta}$ conditional on $I_{j\theta} = 0$. Write

$$\tilde{U}_\theta = (\tilde{W}_\theta, \tilde{V}_\theta) = \frac{1}{n} \sum_{j=1}^{K} U'_{j\theta},$$

where K has distribution equal to the conditional distribution of a binomial variable with parameters n and $\rho = P(I_{i\theta} = 1)$ conditional on it being positive, and define

$$\tilde{T}_\theta = \frac{1}{n} \sum_{j=1}^{K} W'_{j\theta} + \frac{1}{n} \sum_{j=K+1}^{n} V''_{1j\theta}$$

when $0 < K < n$. Also set

$$vec(n\tilde{S}_\theta) = A_\theta \sum_{i=1}^{K} U'_{i\theta}.$$

(A5) Assume that $\det(\tilde{S}_\theta) \neq 0$ and that the transformation $\tilde{T}_\theta$ to $\tilde{S}_\theta^{-1} \tilde{T}_\theta$ given $\tilde{V}_\theta$ is one to one with probability 1.

Remark 4: We note that (A5) is satisfied for the case of robust regression using Huber's proposal 2 as outlined in section 3. We also note that a similar treatment of the non-iid case would be possible with a corresponding complication of notation and conditions. This would permit development of saddlepoint approximations for the regression case conditionally on fixed values of C.

Remark 5: The device of considering the density of $\tilde{T}_\theta$ was used in Embrechts et al (1985).

Let $f_{\tilde{T}_\theta,\tilde{V}_\theta}(\tilde{t},\tilde{v})$ be the density of $\tilde{T}_\theta,\tilde{V}_\theta$ and let $J_\theta(\tilde{t},\tilde{v})$ be the Jacobian of the transformation $\tilde{t}$ to $\tilde{t}^* = \tilde{s}^{-1}\tilde{t}$ for fixed $\tilde{v}$. Then the density of $\tilde{T}^*_\theta$ is

$$f_{\tilde{T}^*_\theta}(\tilde{t}^*) = \int f_{\tilde{T}_\theta,\tilde{V}_\theta}(\tilde{t}(\tilde{t}^*),\tilde{v}) J_\theta(\tilde{t},\tilde{v}) dm^*, \tag{4.18}$$

where m^* is the product measure of Lebesgue and counting measure appropriate to $\tilde{v}$ and where the inverse function $\tilde{t}(\tilde{t}^*)$ may depend on $\tilde{v}$.

Now $f_{\Psi^*(X,\theta)}(z; H(\theta,\alpha,\tau))$ exists for some $\alpha > 0, \tau > 0$, where $H(\theta,\alpha,\tau))$ is defined in (2.4), by (A2) and whenever $\Psi^*(X,\theta)$ has a density, $\tilde{T}_\theta$ has the same density. Hence, we can write

$$f_{\Psi^*(X,\theta)}(z; H(\theta,\alpha,\tau)) = f_{\tilde{T}^*_\theta}(z; H(\theta,\alpha,\tau))$$

and from Theorem 1

$$h(\theta) \quad = \quad \lim_{\alpha\to 0}\lim_{\tau\to 0} f_{\tilde{T}^*_\theta}(0; H(\theta,\alpha,\tau)).$$

In order to obtain saddlepoint approximations for the density of $(\tilde{T}_\theta,\tilde{V}_\theta)$ and so for $\tilde{T}^*_\theta$ we need to assume

(A6) Assume that $E\exp(\beta^T U_\theta) < \infty$ for $||\beta|| < a$ for some $a > 0$ and for all θ.

We now find a saddlepoint approximation to $f_{\tilde{T}^*_\theta}(z)$, and complete the proof by using this approximation for $f_{\tilde{T}^*_\theta}(0; H(\theta, \alpha, \tau))$.

Our approach is to develop a saddlepoint approximation for

$$Y = (\sum_{i=1}^{n} I_{i\theta} W_{i\theta} + \sum_{i=1}^{n}(1 - I_{i\theta})V_{1i\theta}, \sum_{i=1}^{n} I_{i\theta} V_{i\theta}).$$

However before doing that we need to relate the distribution of Y to that of $\tilde{Y} = (\tilde{T}_\theta, \tilde{V}_\theta)$. We let B be the product of B_1 a ball of radius $rad(B_1)$ centered at 0 for dimensions d_1 corresponding to the continuous variables and the point 0 for the d_0 lattice variables in Y. Note that

$$\begin{aligned}
& P(\tilde{Y} \in y' + B/n, 0 < K < n) \\
= \;& P((\sum_{i=1}^{K} W'_{i\theta} + \sum_{i=K+1}^{n} V''_{1i\theta}, \sum_{i=1}^{K} V'_{i\theta}) \in ny' + B, 0 < K < n) \\
= \;& P((\sum_{i=1}^{n} I_{i\theta} W_{i\theta} + \sum_{i=1}^{n}(1 - I_{i\theta})V_{1i\theta}, \sum_{i=1}^{n} I_{i\theta} V_{i\theta}) \in ny' + B) + O(e^{-cn}) \\
= \;& P(Y \in ny' + B) + O(e^{-cn})
\end{aligned}$$

for some $c > 0$. Since $\tilde{Y}$ has a continuous density in d_1 dimensions we can write

$$P(\tilde{Y} \in y' + B/n) = f_{\tilde{Y}}(y') vol(B_1/n)(1 + O(rad(B_1)/n))$$

Now consider the saddlepoint approximation to $P(Y \in ny' + B)$. Denote the cumulant generating function of $\psi(X_1, \theta)$ by

$$\kappa(\tau, \theta) = \log \int \exp(\tau^T \psi(x, \theta)) dF_0(x)$$

and note that $\psi(X_j, \theta) = I_{j\theta} W_{j\theta} + (1 - I_{j\theta})V_{1j\theta}$. Define $\tau(\theta)$ as the solution to

$$\frac{\partial \kappa(\tau, \theta)}{\partial \tau} = 0.$$

Let ν be the probability measure of Y and let

$$\nu_\tau(B) = \int_B e^{n\kappa(\tau,\theta)-\tau^T v} d\nu,$$

where (t, v) corresponds to the partition $(\tilde{t}, \tilde{v})$. Write Y_τ as the random variable corresponding to Y under ν_τ and let $n\mu_\tau$ be the mean of Y_τ and $n\Sigma_\tau$ be the covariance matrix of Y_τ.

We will use Theorem 1 of Robinson et al (1990) to write the following approximation

$$\begin{aligned} P(Y \in ny' + B) &= \frac{\exp(n\kappa(\tau(\theta), \theta))}{(2\pi/n)^{(p+d_1)/2}(2\pi n)^{d_0/2} \det \Sigma_{\tau(\theta)}^{1/2}} \\ &\quad \{\int_{y'+B/n} \exp(-ny^{*T} y^*/2)(1 + Q(y^*\sqrt{n})dy + R\} \end{aligned} \tag{4.19}$$

where $y^* = \Sigma_{\tau(\theta)}^{-1/2}(y - \mu_{\tau(\theta)})$, $R = vol(B_1/n)O(1/n)$ and dy denotes integration with respect to Lebesgue measure on R^{p+d_1}. The first and third error terms of that theorem can be easily reduced to this form if ϵ of the theorem equals $rad(B_1)/n$ and the second term can be bounded in that form by the following argument. If $\hat{\nu}_\tau$ denotes the characteristic function of the measure ν_τ, the probability measure of Y_τ then

$$\begin{aligned} |\hat{\nu}_\tau(\xi)| &= |E_\tau exp[(\tau + i\xi_1)(\sum_{j=1}^n I_{j\theta\tau} W_{j\theta\tau} + \sum_{j=1}^n (1 - I_{j\theta\tau})V_{1j\theta\tau}) + i\xi_2 \sum_{j=1}^n I_{j\theta\tau} V_{j\theta\tau})]| \\ &= \frac{|\rho E exp[(\tau + i\xi_1)W'_{j\theta\tau} + i\xi_2 V'_{j\theta\tau}] + (1-\rho)\exp[(\tau + i\xi_1)V''_{1j\theta\tau}]|^n}{|\rho E exp[\tau W'_{j\theta\tau}] + (1-\rho)\exp[\tau V''_{1j\theta\tau}]|^n}. \end{aligned}$$

Then, if L is the set of subscripts of ξ corresponding to lattice random components,

$$\begin{aligned} q_n(n) &= \sup\{|\hat{\nu}_\tau(\xi)| : |\Sigma_\tau^{\frac{1}{2}}\xi| > c,\ |\xi_k| < \pi,\ k \in L,\} \\ &< e^{-c_1 n}, \end{aligned}$$

for some $c_1 > 0$. So the second term in the errors of the theorem can also be bounded by $vol(B_1/n)O(1/n)$.

The integral in (4.19) can be approximated to give

$$\begin{aligned} P(Y \in ny' + B) &= \frac{\exp(n\kappa(\tau(\theta),\theta))}{(2\pi/n)^{(p+d_1)/2}(2\pi n)^{d_0/2}\det\Sigma_{\tau(\theta)}^{1/2}} \\ &\quad \left\{\exp(-ny'^{*T}y'^{*}/2)(1+Q(y'^{*}\sqrt{n})vol(B_1/n)(1+O(rad(B_1/n))+R\right\}. \end{aligned}$$

By choosing B_1 to have order O(1), the density of $(\tilde{T}_\theta, \tilde{V}_\theta)$ is

$$f_{\tilde{T}_\theta,\tilde{V}_\theta}(y) = \frac{\exp(n\kappa(\tau(\theta),\theta))}{(2\pi/n)^{(p+d_1)/2}(2\pi n)^{d_0/2}\det\Sigma_{\tau(\theta)}^{1/2}}\left\{\exp(-ny^{*T}y^{*}/2)(1+Q(y^{*}\sqrt{n})+O(1/n)\right\} \tag{4.20}$$

Substituting the approximation in (4.18) gives

$$\begin{aligned} f_{\tilde{T}_\theta^*}(0) &= \frac{\exp(n\kappa(\tau(\theta),\theta))}{(2\pi/n)^{p/2}\det\Sigma_{\tau(\theta)}^{1/2}}\int \det J(0,\tilde{v})\frac{\exp(-ny_2^{*T}y_2^{*}/2)}{(2\pi/n)^{d_1}(2\pi n)^{d_0}} \\ &\quad (1+Q(0,y_2^{*}\sqrt{n})+O(n^{-1}))dm^{*} \end{aligned}$$

where, if $\Sigma_{\tau(\theta)}^{22}$ is the submatrix of $\Sigma_{\tau(\theta)}^{-1}$ omitting the first p rows and columns, $y_2^* = (\Sigma_{\tau(\theta)}^{22})^{1/2}(y_2-\mu_{2\tau(\theta)})$ for $y=(y_1,y_2)$, $\mu_{\tau(\theta)}=(\mu_{1\tau(\theta)},\mu_{2\tau(\theta)})$. We note that $y_1=\mu_{1\tau(\theta)}=0$ when t, the value of $\tilde{T}_\theta$ equals 0. We can replace the sum arising from the lattice part of m^* by an integral with errors of order n^{-1}. Then using a Laplace approximation in the integral, we have

$$f_{\tilde{T}_\theta^*}(0) = \exp(n\kappa(\tau(\theta),\theta))\frac{b_\theta}{(2\pi/n)^{p/2}\det(\Sigma_{11,\tau(\theta)})^{1/2}}(1+O(n^{-1})) \tag{4.21}$$

where $\Sigma_{11,\tau(\theta)}$ is the submatrix of $\Sigma_{\tau(\theta)}$ corresponding to $T_{\tau(\theta)}$ and equals $\mathrm{cov}_{\tau(\theta)}\psi(X_1,\theta)$, and $b_\theta = \det(J(0,\mu_{2\tau(\theta)})) = \det E_{\tau(\theta)}\bar{\Psi}'(X,\theta)$. This last equality follows from the definition of the transformation of $\tilde{T}_\theta$ to $\tilde{T}_\theta^*$.

We now need to show that $f_{\tilde{T}_\theta^*}(0)$ approximates $f_{\tilde{T}_\theta^*}(0;H(\theta,\alpha,\tau))$ with an exponentially small error. If for any $0<\alpha<1$ there exist $\tau>0$ and $\delta>0$ such that

$$P(H(\theta,\alpha,\tau)) > 1-e^{-cn}, \tag{4.22}$$

for some $c > 0$, then the density of $\Psi^*(X, \theta)$ at 0, restricted to $H(\theta, \alpha, \tau)$, is just the density of $\tilde{T}^*_\theta$ at 0 up to exponentially small errors. Thus, by Theorem 1, we can approximate the intensity $h(\theta)$ by the right hand side of (4.21).

If there is a unique M-estimate, the right hand side of (4.21) approximates the density of $\hat{\theta}$. In the case where we do not have uniqueness, if we can verify that the event $\{N(B_\tau(\theta_0)) = 1\}$ converges to one exponentially fast in increasing sample size, for suitably chosen τ, then except on sets with exponentially small probability, there is a unique M-estimate, $\hat{\theta}$ in $B_\tau(\theta_0)$, and its density, $f_{\hat{\theta}}(\theta)$, exists and is equal to $f_{\tilde{T}^*_\theta}(0)$ up to exponentially small error and so is approximated by the right hand side of (4.21).

It now remains to prove that the set $H(\theta, \alpha, \tau)$ and the set $\{N(B_\tau(\theta_0)) = 1\}$ converge to one exponentially fast. To prove these results, we will use a single proof based on both Cramér's and Sanov's theorems on large deviations. Let $\mathcal{M}$ be the class of all probability measures on the sample space of X_1, endowed with the topology of weak convergence. We denote the empirical measure of the random sample X by F_n. Let

$$\Lambda(F; \theta', \theta) = \begin{cases} \|(\int \psi'(x, \theta)\, dF(x))^{-1} \int \psi'(x, \theta')\, dF(x)) - I_p\|, & \det \int (\psi'(x, \theta))\, dF(x) \neq 0, \\ \infty, & \text{otherwise,} \end{cases}$$

and

$$\Lambda^*(F; \theta, \tau) = \sup_{\theta' \in B_\tau(\theta)} \Lambda(F; \theta', \theta)$$

We make the following assumptions which we use to show that the event $\{N(B_\tau(\theta_0)) = 1\}$ converges to one exponentially and then that (4.22) holds.

(A7) Given $0 < \alpha < 1$ there is a τ such that $\sup_{\theta \in \bar{B}_\tau(\theta_0)} \Lambda^*(F_0; \theta, \tau) < \alpha$.

(A8) For fixed $\theta \in B_\tau(\theta_0)$, $\Lambda(\cdot; \cdot, \theta)$ is continuous at (F_0, θ) in the product topology.

Remark. The conditions (A7) and (A8) will be satisfied if the derivative of the score function is bounded and is continuous as a function of θ. The conditions are also satisfied for the case of Huber's robust regression outlined in section 3.

Lemma 2 *If (A1)-(A8) hold, then the probability that there is exactly one solution to the score equation in $B_\tau(\theta_0)$ approaches one exponentially quickly in n.*

Proof. Given $0 < \alpha < 1$ select τ such that $\sup_{\theta \in \bar{B}_\tau(\theta_0)} \Lambda^*(F_0; \theta, \tau) = \alpha < 1$. Then select $\alpha' \in (\alpha, 1)$ and define

$$\begin{aligned}
\Gamma_1(\theta) &= \{F \in \mathcal{M} : \Lambda^*(F; \theta, \tau) > \alpha'\}, \\
\Gamma_2 &= \{F \in \mathcal{M} : ||[\int \psi'(x, \theta_0)\, dF(x)]^{-1} \int \psi(x, \theta_0)\, dF(x)|| > \tau_0(1 - \alpha)\} \\
\Gamma_3 &= \{F \in \mathcal{M} : \exists \text{ unique } \theta \in B_\tau(\theta_0) \text{ such that } \int \psi(x, \theta)\, dF(x) = 0\}.
\end{aligned}$$

Now by the results of lemma 1, which give conditions for F_n to have a unique solution to (1.1), we have that $\Gamma_3^c \subset \Gamma_1(\theta_0) \cup \Gamma_2$,

By Sanov's Theorem, which gives the large deviation principle for empirical measures (see Dembo and Zeitouni, 1993), we may state

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{F_n \in \Gamma_1(\theta)\} \leq - \inf_{F \in \bar{\Gamma}_1(\theta)} H(F|F_0)$$

where

$$H(F|G) = \begin{cases} \int \frac{dF}{dG} log \frac{dF}{dG} dF, & \text{if } \frac{dF}{dG} \text{ exists,} \\ \infty, & \text{otherwise,} \end{cases}$$

which is a good, convex rate function. We will verify that for all $\theta \in \bar{B}_\tau(\theta_0)$, $F_0 \notin \bar{\Gamma}_1(\theta)$ in which case $\inf_{F \in \bar{\Gamma}_1(\theta)} H(F|F_0)$ is positive since $H(F|G) = 0$ if and only if $F = G$. Suppose there is a sequence of measures in $\mathcal{M}$, say $\{F^m\}$ in $\Gamma_1(\theta)$ converging to F_0. Select $\epsilon > 0$ so that $\alpha < \alpha' - \epsilon$. Then we can find $\theta_m \in B_\tau(\theta_0)$ such that $\Lambda(F^m; \theta_m, \theta) \geq \alpha' - \epsilon$. There is a convergent subsequence $\{\theta_{m_i}\}$ of the θ_m with limit $\theta_1 \in B_\tau(\theta_0)$. Now condition (A8) asserts that $\Lambda(F_0; \theta_1, \theta) \geq \alpha' - \epsilon$ which contradicts assumption (A7). Hence $F_0 \notin \bar{\Gamma}_1(\theta)$ and $\inf_{F \in \bar{\Gamma}_1(\theta)} H(F|F_0) > 0$. In particular, this holds for $\Gamma_1(\theta_0)$. Using assumptions (A4) and (A6), we can apply Cramér's theorem and conclude that $P(F_n \in \Gamma_2)$ approaches zero exponentially as n approaches 0. Hence

$$\begin{aligned} P(N(B_\tau(\theta_0)) \neq 1) &= P(F_n \in \Gamma_3^c) \\ &\leq P(F_n \in \Gamma_2) + P(F_n \in \Gamma_1(\theta_0)) \end{aligned}$$

Since both terms on the right hand side are exponentially small, this completes the proof.

It remains only to prove (4.22). Now we can write $H(\theta, \alpha, \tau)^c = \{F_n \in \Gamma_1(\theta)\}$, for $\theta \in \bar{B}_\tau(\theta_0)$ and we showed in the proof of Lemma 2 that the probability of this set is exponentially small. Thus we have from (4.21):

Theorem 2 *Under conditions (A1)-(A8), there is, with probability* $1 - e^{-cn}$ *for some* $c > 0$, *a uniquely defined M-estimate* $\hat{\theta}$ *on* $B_\tau(\theta_0)$ *which has a density, restricted to* $B_\tau(\theta_0)$,

$$f_{\hat{\theta}}(\theta) = \exp(n\kappa(\tau(\theta), \theta)) \frac{\det E_{\tau(\theta)} \psi'(X_1, \theta)}{(2\pi/n)^{p/2} \det(\mathrm{cov}_{\tau(\theta)} \psi(X_1, \theta))^{1/2}} (1 + O(n^{-1})). \tag{4.23}$$

It should be noted that this approximation, (4.23), is the same as that derived in Field (1982). His result was obtained under more restrictive conditions and technically does not cover the estimation of location and scale with Huber's proposal 2.

5 Smooth Functions of M-Estimates

In this section we derive tail probability approximations for smooth real valued functions of multivariate M-estimates, proceeding as in Jing and Robinson (1994). Let $g_1(\theta)$ be an infinitely differentiable function from $\Re^p$ to $\Re^1$ with $|\nabla g_1(\theta)| > 0$ for $\theta = \theta_0$, where $E[\Psi(X, \theta_0)] = 0$. We wish to find an approximation for the tail area $P_{\theta_0}(g_1(\theta) \geq \eta_1)$. We proceed by finding a differentiable function $g_2(\theta)$ from $\Re^p$ to $\Re^{p-1}$ such that $g(\theta) = (g_1(\theta), g_2(\theta))$ has Jacobian matrix

$$J(\theta) = [\nabla g(\theta)]^{-1}.$$

(We may assume without loss of generality that $\partial g_1(\theta)/\partial \theta_1 > 0$ and choose $g_2(\theta) = (\theta_2, \ldots, \theta_p)$. Then we can find an ϵ- neighborhood of θ_0, $B_\epsilon(\theta_0)$, in which $J(\theta) > 0$. Let $B_\epsilon^*(\theta_0)$ be the image of $B_\epsilon(\theta_0)$ under g. Then g is a one-to-one transformation from $B_\epsilon(\theta_0)$ onto $B_\epsilon^*(\theta_0)$. Let $\theta(\eta) = g^{-1}(\eta)$ for any $\eta \in B_\epsilon^*(\theta_0)$, and

$$L(\eta) = -\kappa(\tau(\theta(\eta)), \theta(\eta)).$$

Then, as in Jing and Robinson (1994), we can choose ϵ small enough so that $L(\eta)$ is convex in $B_\epsilon^*(\theta_0)$. Assuming that (A1)-(A8) hold then Theorem 2 asserts that there is, with probability greater than $1 - e^{-cn}$ for some $c > 0$, a unique solution $\hat{\theta}$ of (1.1) with density given in (4.23). If $\hat{\eta} = g(\hat{\theta})$, then the density of $\hat{\eta}$ is

$$f_{\hat{\eta}}(\eta) = \frac{\exp(-nL(\eta))}{(2\pi/n)^{p/2}} A(\eta)(1 + O(1/n)),$$

fot η close enough to $g(\theta_0)$, where

$$A(\eta) = \frac{\det E_{\tau(\theta(\eta))} \psi'(X_1, \theta(\eta)) \det(J(\theta(\eta)))}{\det(\mathrm{cov}_{\tau(\theta(\eta))} \psi(X_1, \theta(\eta)))^{1/2}}.$$

Now let $\eta = (\eta_1, \eta_2)$ where $\eta_1 \in \Re^1$, and set $\hat{\eta}_1 = g_1(\hat{\theta})$. Let

$$H(\eta_1) = \inf_{\eta_2} L(\eta) = L(\tilde{\eta}).$$

Following Jing and Robinson (1994) we have, using the Laplace approximation,

$$f_{\hat{\eta}_1}(\eta_1) = \frac{\exp(-nH(\eta_1))}{(2\pi/n)^{1/2}} \det(\text{Ł}_{22}(\tilde{\eta}))^{-1/2} A(\tilde{\eta})(1 + O(1/n)) \tag{5.24}$$

where

$$L_{22}(\eta) = \frac{\partial^2 L(\eta)}{\partial \eta_2^2}.$$

This density can be integrated as in Jing and Robinson (1994) to give:

Theorem 3 *If (A1)-(A8) hold and g is defined as above then for some $\delta > 0$ and $0 < \eta_1 - g_1(\theta_0) < \delta$*

$$P(g_1(\hat{\theta}) \geq \eta_1) = [1 - \Phi(n^{1/2}s^*)][1 + O(1/n)] \tag{5.25}$$

where $s^* = (2H(\eta_1))^{1/2} - \log[(2H(\eta_1))^{1/2}\det(Ł_{22}(\tilde{\eta}))^{-1/2}A(\tilde{\eta})/H'(\eta_1))]/(n(2H(\eta_1))^{1/2})$.

Remark. The equation (5.25) gives lower tail probabilities when H' is replaced with $-H'$.

6 Numerical Example

We consider a numerical example in which we apply the approximation to the case of Huber's proposal 2 as given in section 3 for the special case of location and scale. We will assume that the condition holds which guarantees a unique solution (see section 3). Then, since the derivatives are piecewise continuous and globally bounded, conditions (A1)-(A8) of Section 3 hold. So to apply Theorems 2 and 3 we note that the calculations are based on the moment generating function

$$\begin{aligned}
\kappa(\theta,\tau) &= E[\exp(\tau_1 h_b((Z-\mu)\sigma^{-1}) + \tau_2 h_b^2((Z-\mu)\sigma^{-1}) - \tau_2\beta)] \\
&= \exp(-\tau_2\beta)\{\exp(\tau_1 b + \tau_2 b^2)(1 - \Phi(\mu + b\sigma)) + \exp(-\tau_1 b + \tau_2 b^2)\Phi(\mu + b\sigma) \\
&\quad + \frac{\sigma}{\sqrt{2}c_1(\theta,\tau)} \exp\left(-\frac{\mu^2}{2} + \left(\frac{c_2(\theta,\tau)}{2c_1(\theta,\tau)}\right)^2\right) \times \\
&\quad + \left(\Phi\left(\sqrt{2}c_1(\theta,\tau)b + \frac{c_2(\theta,\tau)}{\sqrt{2}c_1(\theta,\tau)}\right) - \Phi\left(-\sqrt{2}c_1(\theta,\tau)b + \frac{c_2(\theta,\tau)}{\sqrt{2}c_1(\theta,\tau)}\right)\right)\}
\end{aligned}$$

where Z is a standard normal random variable and

$$\tau = (\tau_1, \tau_2),\ \theta = (\mu, \sigma),\ c_1(\theta, \tau) = (\frac{\sigma^2}{2} - \tau_2)^{1/2},\ c_2(\theta, \tau) = (\mu\sigma - \tau_1) \tag{6.26}$$

for $\sigma^2/2 \geq \tau_2$. For Huber's Proposal 2 we set

$$\beta = E[h_b^2(Z)] = b^2 * \Phi(-b) - 2b\phi(b) + \Phi(b) - \Phi(-b). \tag{6.27}$$

For $b = 1.345$ we have $\beta = 0.71$, so that (3.17) is satisfied for any $n \geq 2$.

Letting $(\hat{\mu}, \hat{\sigma})$ be the M-estimates defined above, we now examine the distributions of $S_1 = \hat{\mu}/\hat{\sigma}$ and $S_2 = \hat{\sigma}$. The method of Section 5 was applied to approximate tail probabilities of S_1 and S_2. In addition, tail probabilities were estimated by simulations using 200,000 replications each for $n = 5, 10, 20$. Samples are simulated from the standard normal distribution. In Figures 1 and 2, the approximate marginal density of the studentized mean $(\hat{\mu}/\hat{\sigma})$ and the scale parameter estimate $(\hat{\sigma})$ obtained from equation (5.24) is compared to an empirical estimate of the densities obtained from the simulations. In Tables 1 and 2 various tail probabilties for $\hat{\mu}/\hat{\sigma}$ and $\hat{\sigma}$ are estimated using the Monte Carlo simulations, by integrating the approximate marginal density of (5.24) and by using the tail probability formula (5.25).

Before evaluating the results, we note that the Monte Carlo tail areas will have standard errors of 0.00016 with a tail area of .005 and a standard error of 0.00003 with a tail area of .0001. The graphs of of the empirical and the approximate marginal densities show very good agreement for $n = 10$ and $n = 20$. There is a systematic error at $n = 5$ but even here the approximation is quite reasonable. The empirical density for the standard deviation $\hat{\sigma}$ shows some irregularities at 0.5 and 1.5. It is not clear exactly what causes this behavior but it may represent some underlying discontinuity with small samples.

The tail area approximations obtained by integrating the marginal density numerically and those from the tail area approximation itself are quite comparable and give similar levels of accuracy. The tail area approximations for both $\hat{\mu}/\hat{\sigma}$ and $\hat{\sigma}$ are very good for $n = 20$ and $n = 10$ except perhaps in the extreme tail. For $n = 5$ the accuracy deteriorates as we might expect due to the errors in the marginal density approximation.

7 References

DALEY, D.J. and VERE-JONES, D. (1988). An Introduction to the Theory of Point Processes. Springer-Verlag, New York.

DEMBO, A. and ZEITOUNI, O. (1993). Large Deviations Techniques and Applications. Jones and Bartlett Publishers, Boston.

EDWARDS, R.E. (1965). Functional Analysis; Theory and Applications. Holt, Rinehart and Winston, New York.

EMBRECHTS, P., JENSEN, J.L., MAEJIMA, M. and TEUGELS, J.L. (1985). Approximations for compound Poisson and Pólya processes. *Adv. Appl. Prob.* **17**, 623-637.

FIELD, C.A. (1982). Small sample asymptotic expansions for multivariate M-estimates. *Ann. Statist.*, **10**, 672-689.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73-101.

HUBER, P.J. (1981). Robust Statistics, John Wiley and Sons, New York.

JENSEN, J.L. and WOOD, A.T.A. (1998). Large deviation results for minimum contrast estimators. *Ann. Inst. Statist. Math.*, **50**, 673-695.

JING, B.Y. and ROBINSON, J. (1994). Saddlepoint approximations for marginal and conditional probabilities of transformed variables. *Ann. Statist.* **22**, 1115-1132.

NOBLE, B. and DANIEL, J.W. (1977). Applied Linear Algebra. Prentice-Hall, Englewood Cliffs, N.J.

ROBINSON, J., HÖGLUND, T., HOLST, L. and QUINE, M.P. (1990) On approximating probabilities for large and small deviations in R^d. *Ann. Prob.* **18**, 727-753.

SKOVGAARD, IB.M. (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18**, 779-789.