Ann. Inst. Statist. Math.
Vol. 50, No. 4, 673–695 (1998)

LARGE DEVIATION AND OTHER RESULTS FOR MINIMUM CONTRAST ESTIMATORS

JENS LEDET JENSEN[1] AND ANDREW T. A. WOOD[2]

[1]*Department of Theoretical Statistics, Institute of Mathematics, Aarhus University, DK-8000 Aarhus C, Denmark*
[2]*School of Mathematical Sciences, University of Bath, Bath BA2 7AY, U.K.*

(Received January 17, 1997; revised October 21, 1997)

Abstract. A number of authors have been concerned with constructing large deviation approximations to densities and probabilities associated with minimum contrast estimators (equivalently, M-estimators) using a tilting approach due to Field. These developments are an interesting and important extension of saddlepoint-type methodology. However, in the case of a multivariate parameter, the theoretical picture has remained incomplete in certain respects, as explained below. In this paper we present results which provide rigorous justification of the tilting argument, using conditions which it is feasible to check. These results include a new formulation and proof of Skovgaard's theorem for the intensity of minimum contrast estimators, but under conditions which are typically straigtforward to check in practice. Our most detailed application is to multivariate location-scatter models.

Key words and phrases: Elliptical distribution, exponentially small, intensity, location-scatter model, M-estimator.

1. Introduction

Let β be a p-dimensional parameter and let Y denote a vector of observations. Given an objective function $\gamma(Y,\beta)$, a minimum contrast estimator of β is a $\hat{\beta}$ which minimises $\gamma(Y,\beta)$ over β for given Y. In most situations, γ will be such that $\hat{\beta}$ satisfies

$$\frac{\partial\gamma}{\partial\beta}(Y,\hat{\beta}) = 0, \tag{1.1}$$

i.e. the minimum of $\gamma(Y,\beta)$ will be a stationary minimum.

In this paper, we shall focus on the case in which the data, Y, consists of a sample of IID random vectors $X_1,\ldots,X_n$, and restrict attention to contrast functions of the form

$$\gamma(Y,\beta) = -\sum_{i=1}^{n} h(X_i,\beta), \tag{1.2}$$

where h is a given function. [Note: the minus sign is introduced in (1.2) because later on we shall maximise $-\gamma$ rather than minimise γ.]

Our purpose is to present some results concerning the large deviation properties of minimum contrast estimators. Two particular problems will concern us:

(a) A proof, under suitable conditions, that the probability $P(|\hat{\beta} - \beta_0| > \gamma)$ is exponentially small for any $\gamma > 0$ (in a sense which is made precise in Definition 2.1 below). In the above, β_0 is the "true" value of β, defined as the limiting value of $\hat{\beta}$, this limit being assumed to exist in probability; and $|.|$ denotes the usual Euclidean norm on $\boldsymbol{R}^p$.

(b) A proof of Skovgaard's (1990) existence and representation theorem for the intensity of local minima of the contrast function, but under conditions which are far easier to check in practice than Skovgaard's.

Problems (a) and (b) are both of intrinsic interest. However, we have additional motivation for considering them: they are the key results which are required for establishing the large deviation properties of Field's (1982) tilting argument for approximating the distribution of $\hat{\beta}$. See also Daniels (1983) for related developments. For further details and references concerning the tilting argument, see Field and Ronchetti (1990) and Jensen (1995). We note that rigorous proofs of the large deviation properties of the tilting approach are lacking when β is multi-dimensional and (1.1) has multiple solutions; see Jensen ((1995), p. 114) for further discussion of this point.

There is, of course, a substantial literature on problem (a) in the case of maximum likelihood estimators in exponential families, starting with Cramér's classical large deviation theorem (see e.g. Varadhan (1984) for an account of the latter). We also mention Bahadur's (1961) large deviation result for the maximum likelihood estimator of a scalar parameter which lies in a finite interval. In this paper we prove corresponding results for multivariate minimum contrast estimators ranging in an unbounded parameter space.

As an alternative to deriving the large deviation results given here, it may be possible to use the general theorem of Sieders and Dzhaparidze (1987), which is based on earlier work of Ibragimov and Has'minskii (1981). However, some of the conditions assumed by Sieders and Dzhaparidze are not very explicit and appear to be rather difficult to check.

In constructing the tilted approximation to the distribution of $\hat{\beta}$, we make use of the intensity of local minima, rather than the density of a particular solution of (1.1). The first rigorous result on the existence of the intensity was proved by Skovgaard (1990). Unfortunately, one of his conditions (which he refers to as (C2)) is virtually impossible to check in practice. In this paper, we prove Skovgaard's (1990) theorem under conditions which are much easier to check.

On a first impression, it may appear strange that the intensity (as opposed to the density) plays such a prominant role. Two points are worth noting: (i) it is the intensity (rather than the density) which appears naturally in Field's (1982) tilting approach; (ii) due to (a), the intensity provides a good approximation to the distribution of a suitable solution of (1.1) in a sense which we will make more precise later. We also note that it appears to be extremely difficult to prove that the density of $\hat{\beta}$ exists; the example given by Jensen (1995, p. 114) indicates the

subtlety of the existence problem. The principal source of the difficulty is that (1.1) may have multiple solutions, which rules out a direct application of the implicit function theorem. Even in the case of maximum likelihood estimators in curved exponential families, the problem is non-trivial; see Pazman (1986). Also, even if we could prove the existence of the density of $\hat{\beta}$, we would expect it to be too complicated for practical use.

As already indicated, if (1.1) has multiple solutions, as it will in many applications we have in mind, then the technical difficulties are greatly increased. One point is worth emphasising: we should be clear about which solution to (1.1) we wish to use as our estimator. In Clarke's (1991) terminology, we need to decide which selection functional to use. In this paper, we present theorems for two types of selection functional: (i) the solution to (1.1) which globally minimises $\gamma(Y,.)$ (Theorem 2.1); and (ii) the solution to (1.1) which is closest to a given ("quick-and-dirty") preliminary estimator (Theorem 2.3). In approach (ii), which is based on an idea due to Kester and Kallenberg (1986), the important point is that the appropriate solution of (1.1) will inherit the large deviation properties of the preliminary estimator.

The layout of the paper is as follows. In Section 2, we present some large deviation theorems relating to (a), with proofs given in Section 3. Section 3 contains some auxillary results which may be of independent interest. In Section 4, we focus on the case in which h in (1.2) is the log-likelihood ratio for an elliptical model and prove two intermediate results which enable us to apply the general theorems in Section 2. In Section 5 we prove Skovgaard's theorem under an alternative (and simpler) set of conditions, and in Section 6 we give a brief summary of the tilting argument.

2. Main results

Before stating our main results in Subsection 2.3, we introduce the basic framework and notation in Subsection 2.1, and state some assumptions in Subsection 2.2.

2.1 *Framework and notation*

In this paper we shall focus exclusively on the case in which the observations consist of a set of independent and identically distributed (IID) random vectors $X_1, \ldots, X_n$, where the X_i lie in $\mathcal{X}$, a subset of finite-dimensional Euclidean space. It is assumed that we are interested in a p-dimensional parameter $\beta = (\beta_1, \ldots, \beta_p)^T \in \boldsymbol{R}^p$. Consider a function $h(x, \beta)$. We shall consider minimum contrast estimators of β defined by choosing β to minimise $-\bar{h}(\beta) = -n^{-1} \sum_{i=1}^n h(X_i, \beta)$ or, equivalently, to maximise $\bar{h}(\beta)$. In this paper it will be convenient to work with the latter formulation to facilitate discussion of the maximum likelihood estimators.

For $u \in \boldsymbol{R}^q$, viewed as a column vector, we define the norm $|u| = (u^T u)^{1/2}$, where u^T denotes the transpose to u. We shall denote by $B_\xi(a)$ the open ball $\{u : |u - a| < \xi\} \subset \boldsymbol{R}^q$. The closure of a set $A \subseteq \boldsymbol{R}^q$ will be denoted $cl(A)$.

A bar will always indicate a sample mean, as in $\bar{h}(\beta)$ defined above, or e.g. $\bar{Q}_L = n^{-1} \sum_{i=1}^n Q_L(X_i)$, where the function $Q_L(x)$ is referred to in Assumption (A.5) below.

Probabilities and expectations will always be calculated with respect to the true underlying distribution of the X_i's. It will be assumed that $E\{h(X_1, \beta)\}$ has a unique global maximum, β_0, which will be referred to as the true β. We define

$$\bar{H}(\beta) = n^{-1} \sum_{i=1}^{n} h(X_i, \beta) - h(X_i, \beta_0),$$

with expected value

$$e_0(\beta) = E\{\bar{H}(\beta)\};$$

the $p \times 1$ column vector

$$d_1(x, \beta) = \frac{\partial h(x, \beta)}{\partial \beta},$$

and the $p \times p$ matrix of second derivatives

$$d_2(x, \beta) = \frac{\partial^2 h(x, \beta)}{\partial \beta \partial \beta^T}.$$

The derivatives $d_1(x, \beta)$ and $d_2(x, \beta)$ are assumed to exist for all x and β, except possibly on a null x-set which does not depend on β. Note that $\bar{H}(\beta_0) = e_0(\beta_0) = 0$. We also define, for a set $\Omega \subseteq \boldsymbol{R}^p$ to be specified later,

$$Y_{\beta,\delta} \equiv Y_{\beta,\delta}(X) = \sup_{\beta_1 \in B_\delta(\beta) \cap \Omega} h(X, \beta_1) - h(X, \beta_0), \tag{2.1}$$

$$e_\delta(\beta) = E(Y_{\beta,\delta}), \tag{2.2}$$

$$M_{\beta,\delta}(\theta) = E[\exp\{\theta Y_{\beta,\delta}\}], \quad \text{and} \quad K_{\beta,\delta}(\theta) = \log M_{\beta,\delta}(\theta).$$

In (2.1), the distribution of X is the same as the common distribution of the X_i. Detailed assumptions concerning these quantities will be given in the next subsection.

The set $\Omega \subseteq \boldsymbol{R}^p$ will be open to choice. It will not necessarily consist of all possible values of β. Further clarification concerning appropriate choice of Ω in a particular setting will be given in Section 4.

2.2 *Assumptions*

(A1) At $\beta_0 \in \Omega$, $E\{d_1(X, \beta_0)\} = 0$ and $i_0 = -E\{d_2(X, \beta_0)\}$ is positive definite.

(A2) There exists an $\xi > 0$ and a function $R(x)$ such that

$$|[d_2(x, \beta) - d_2(x, \beta_0)]_{jk}| \leq R(x)|\beta - \beta_0|, \qquad 1 \leq j, k \leq p,$$

for all $x \in \mathcal{X}$, and all $\beta \in B_\xi(\beta_0)$.

(A3) The moment generating functions of $d_1(X, \beta_0)$, $d_2(X, \beta_0)$ and $R(x)$ all exist in a neighbourhood of the origin.

(A4) The unique global supremum of $e_0(\beta)$ over $\beta \in \Omega$ is $e_0(\beta_0) = 0$. In other words, for any $\epsilon > 0$, $\sup_{\beta:|\beta-\beta_0|>\epsilon} e_0(\beta) < 0$.

(A5) For any compact set $L \subset cl(\Omega)$, there exists a function $Q_L(x)$ such that

$$|h(x, \beta_1) - h(x, \beta_2)| \leq Q_L(x)|\beta_1 - \beta_2|$$

for all $x \in \mathcal{X}$, and all $\beta_1, \beta_2 \in L$. Moreover, there exists a $\theta_L > 0$ such that $E\exp\{\theta_L Q_L(X)\} < \infty$, and for each $\beta \in \Omega$, there exists a $\theta_\beta > 0$ such that $E\exp[\theta_\beta\{h(X,\beta) - h(X,\beta_0)\}] < \infty$.

(A6) Let $\delta_0 > 0$. For any $0 < \delta < \delta_0$, there exist constants $\theta_0 > 0$, $C > 0$ and $\alpha > 0$ such that

$$M_{\beta,\delta}(\theta_0) \leq \frac{C}{(1 + |\beta - \beta_0|)^\alpha}$$

for all $\beta \in \Omega$; and also, for any $0 < \delta < \delta_0$,

$$\limsup_{\beta\in\Omega:|\beta|\to\infty} e_\delta(\beta) \to -\infty.$$

2.3 *Main results*

Our main results are summarised in three theorems. Proofs are given in subsection 3.5. First, we give a definition.

DEFINITION 2.1. Let $\{A_n\}_{n\geq 1}$ be a sequence of events in a probability space with probability measure $P[.]$. We shall say that the sequence $\{A_n\}$ *has exponentially small probability* or, more briefly, *the event* A_n has ESP, if there exist constants $\alpha > 0$ and $\rho > 0$ such that

$$P[A_n] \leq \alpha\exp(-\rho n) \quad \text{for all} \quad n \geq 1.$$

THEOREM 2.1. *Suppose that Assumptions* (A1)–(A6) *are satisfied, and write* $\hat{\beta}_\Omega$ *for the global maximiser of* $\bar{H}(\beta)$ *over* $\beta \in \Omega$. *Then, for any* $\gamma > 0$,

$$\textit{the event} \quad \{|\hat{\beta}_\Omega - \beta_0| > \gamma \ \textit{ or } \ \hat{\beta}_\Omega \ \textit{ is not well-defined}\} \quad \textit{has ESP}$$

in the sense of Definition 2.1.

Remark 2.1. In saying that $\hat{\beta}_\Omega$ is well-defined, we mean that it exists and is unique, in the sense that for some $\hat{\beta}_\Omega \in \Omega$, $\bar{H}(\hat{\beta}_\Omega) > \bar{H}(\beta)$ for all $\beta \in \Omega$ such that $\beta \neq \hat{\beta}_\Omega$. Otherwise, we say that $\hat{\beta}_\Omega$ is not well-defined.

Remark 2.2. Note that $\hat{\beta}_\Omega$ is only assumed to be well-defined with high probability. In continuous models, one would expect that, typically, $\hat{\beta}_\Omega$ would be well-defined with probability one when n is sufficiently large. However, no general results of this kind seem to be known, but see Pazman (1986) for such a result in curved exponential models.

With certain choices of the contrast function h, the equation $\bar{d}_1(\beta) = 0$ may have a unique solution $\hat{\beta}_U \in \boldsymbol{R}^p$ with probability one. In such cases, we have the following simplified form of Theorem 2.1.

THEOREM 2.2. *Suppose that Assumptions* (A1)–(A3) *hold. If, with probability one, the equation* $\bar{d}_1(\beta)$ *has a unique solution* $\hat{\beta}_U \in \boldsymbol{R}^p$ *then, for any given* $\gamma > 0$,

$$\{|\hat{\beta}_U - \beta_0| > \gamma\} \quad \textit{has ESP}$$

in the sense of Definition 2.1.

Finally, we consider a situation in which we have a preliminary estimator $\hat{\beta}_P$ which, for any fixed $\gamma > 0$, is such that

$$\{|\hat{\beta}_P - \beta_0| > \gamma \text{ or } \hat{\beta}_P \text{ is not well-defined}\} \quad \text{has ESP} \tag{2.3}$$

in the sense of Definition 2.1.

THEOREM 2.3. *Suppose that Assumptions* (A1)–(A3) *hold, and that we have a preliminary estimator* $\hat{\beta}_P$ *which satisfies* (2.3). *Define* $\hat{\beta}_S$ *to be the* $\hat{\beta}$ *satisfying* $\bar{d}_1(\hat{\beta}) = 0$ *for which* $|\hat{\beta} - \hat{\beta}_P|$ *is minimised. Then, for any* $\gamma > 0$,

$$\{|\hat{\beta}_S - \beta_0| > \gamma \text{ or } \hat{\beta}_S \text{ is not well-defined}\} \quad \text{has ESP.}$$

3. Proofs of theorems

We begin by recalling an elementary fact which will be used repeatedly in what follows. Suppose that $Z, Z_1, \ldots, Z_n$ are independent and identically distributed random variables such that $E(e^{\theta Z}) < \infty$ for some $\theta > 0$. Write $\bar{Z} = n^{-1} \sum_{i=1}^{n} Z_i$. Then for any $z > E(Z)$, there exist $\alpha > 0$ and $\rho > 0$ such that

$$P(\bar{Z} > z) \leq \alpha \exp(-\rho n) \tag{3.1}$$

for all n.

3.1 *Some intermediate results*

The proofs of Theorems 2.1 2.3 depend on three results which may be of independent interest.

PROPOSITION 3.1. *Suppose that Assumptions* (A1)–(A3) *hold. Then there exists a* $\gamma_1 > 0$ *such that for all* $0 < \epsilon < \gamma_1$, *there exist positive constants* α *and* ρ *with*

$$P[\text{there exists a unique } \hat{\beta} \in B_\epsilon(\beta_0) \text{ satisfying } \bar{d}_1(\hat{\beta}) = 0] \geq 1 - \alpha \exp(-\rho n).$$

Moreover, if $\hat{\beta}_C$ *minimises* $|\hat{\beta} - \beta_0|$ *over solutions,* $\hat{\beta}$, *of* $\bar{D}_1(\hat{\beta}) = 0$, *then for each fixed* $\gamma > 0$,

$$\{|\hat{\beta}_C - \beta_0| > \gamma \text{ or } \hat{\beta}_C \text{ is not well-defined}\} \quad \text{has ESP.}$$

PROPOSITION 3.2. *Suppose that Assumptions* (A1)–(A5) *hold. Let* $L \subset \Omega$ *be any compact set which contains a ball* $B_\xi(\beta_0)$ *for some* $\xi > 0$, *and write* $\hat{\beta}_L$ *for the* β *which maximises* $\bar{H}(\beta)$ *over* $\beta \in L$ *if this maximum is well-defined in the sense of Definition* 2.1. *Then, for any* $\gamma > 0$,

$$\{|\hat{\beta}_L - \beta_0| > \gamma \text{ or } \hat{\beta}_L \text{ is not well-defined}\} \quad \text{has ESP.}$$

PROPOSITION 3.3. *Suppose that Assumptions* (A1)–(A6) *hold. Let* L *be any compact set satisfying the conditions of Proposition* 3.2 *and write* $K = \Omega - L$. *Then there exists a* $y < 0$ *such that*

$$\left\{ \sup_{\beta \in K} \bar{H}(\beta) > y \right\} \quad \text{has ESP.}$$

3.2 *Proof of Proposition* 3.1.

Put $i_0 = -E\{d_2(X, \beta_0)\}$. From (A1) the eigenvalues of i_0 lie in the interval (k_1, k_2), for some $k_1 > 0$. Choose δ such that $0 < \delta < k_1/(4p)$.

From (A3), we have the existence of positive constants k_3, α_1 and ρ_1 such that

$$P[\bar{R} > k_3 \text{ or } |[\bar{d}_2(\beta_0) + i_0]_{rs}| > \delta \text{ for some } r, s] \leq \alpha_1 \exp(-\rho_1 n). \tag{3.2}$$

On the complement of the set

$$\{\bar{R}_n > k_3 \text{ or } |[\bar{d}_2(\beta_0) + i_0]_{rs}| > \delta\},$$

we have from (A2)

$$\begin{aligned} |[\bar{d}_2(\beta) + i_0]_{rs}| &\leq |[\bar{d}_2(\beta) - \bar{d}_2(\beta_0)]_{rs}| + |[\bar{d}_2(\beta_0) + i_0]_{rs}| \\ &\leq k_3|\beta - \beta_0| + \delta \\ &\leq 2\delta \end{aligned}$$

when $|\beta - \beta_0| \leq \delta/k_3$ and $|\beta - \beta_0| < \xi$ where ξ is given in (A2). This bound implies that there can be at most one solution to $\bar{d}_1(\beta) = 0$ with $|\beta - \beta_0| < \delta/k_3$. If β_1 is such a solution then, for $\beta \neq \beta_1$ with $|\beta - \beta_0| < \delta/k_3$, we have

$$\begin{aligned} \bar{d}_1(\beta)(\beta - \beta_1) &= \{\bar{d}_1(\beta) - \bar{d}_1(\beta_1)\}(\beta - \beta_1) \\ &= -\int_0^1 (\beta - \beta_1)^T [\bar{d}_2(\beta_1 + \psi(\beta - \beta_1))](\beta - \beta_1) d\psi \\ &\leq -\{(\beta - \beta_0)^T i_0 (\beta - \beta_0) - 2\delta p|\beta - \beta_0|\} \\ &\leq -(k_1 - 2\delta p)|\beta - \beta_1| \\ &< 0, \end{aligned}$$

for $\beta \neq \beta_1$. Consequently, when $\beta \neq \beta_1$ and $|\beta - \beta_0| < \delta/k_3$, $\bar{d}_1(\beta) \neq 0$.

Let ϵ satisfying $0 < \epsilon < \delta/k_3$ be given. If

$$|\bar{d}_1(\beta_0)| \leq \frac{1}{2}\epsilon(k_1 - 2\delta p)$$

and we are in the complement of the set given in (3.2), then for any unit vector $v \in \boldsymbol{R}^p$,

$$\begin{aligned} v^T \bar{d}_1(\beta_0 + \epsilon v) &= v^T[\bar{d}_1(\beta_0 + \epsilon v) - \bar{d}_1(\beta_0)] + v^T \bar{d}_1(\beta_0) \\ &\leq \int_0^\epsilon v^T \bar{d}_2(\beta_0 + uv) v du + \frac{1}{2}\epsilon(k_1 - 2\delta p) \\ &\leq -\frac{1}{2}\epsilon(k_1 - 2\delta p) \\ &< 0. \end{aligned} \tag{3.3}$$

Brouwer's fixed point theorem and (3.3) imply that there exists a β_1 such that $|\beta_1 - \beta_0| < \epsilon$ and $\bar{d}_1(\beta_1) = 0$. From (A3), there exist $\alpha_2 > 0$ and $\rho_2 > 0$ such that

$$P\left[|\bar{d}_1(\beta_0)| > \frac{1}{2}\epsilon(k_1 - 2\delta p)\right] \leq \alpha_2 \exp(-\rho_2 n). \tag{3.4}$$

The first result in Proposition 3.1 now follows from (3.2) and (3.4) with $\gamma_1 = \delta/k_3$; the second result follows from (3.2) and (3.4) with $\epsilon = \gamma$. □

3.3 *Proof of Proposition* 3.2

Choose $\epsilon > 0$ so that Proposition 3.1 holds, and let $\hat{\beta}_1$ be the unique maximum of $\bar{H}(\beta)$ for $|\beta - \beta_0| < \epsilon$ on the set (3.1). Note that $\bar{H}(\hat{\beta}_1) \geq \bar{H}(\beta_0) = 0$. Write $L_\epsilon = L/B_\epsilon(\beta_0)$. We will show that, under the given assumptions,

$$P\left[\sup_{\beta \in L_\epsilon} \bar{H}(\beta) > y\right] \leq \alpha_1 \exp(-\rho_1 n) \tag{3.5}$$

for some $y < 0$ and $\alpha_1, \rho_1 > 0$. Consider the function Q_L in (A5). Assumption (A5) implies that there exist constants $z \in \boldsymbol{R}$ and $\alpha_2, \rho_2 > 0$ such that

$$P(\bar{Q}_L > z) \leq \alpha_2 \exp(-\rho_2 n).$$

Let $g = \sup_{\beta \in L_\epsilon} e_0(\beta) < 0$, and choose $y \in (g, 0)$. Then choose δ so that $2z\delta < y - g$, and let $\{B_\delta(\beta) : \beta \in L_\epsilon\}$ be a covering of L_ϵ. Since L_ϵ is compact, we may choose a finite subcover

$$B_\delta(\beta_1), \ldots, B_\delta(\beta_J)$$

for some J and β_j, $j = 1, \ldots, J$. On $L_\epsilon \cap B_\delta(\beta_j)$ we have

$$\sup_{\beta \in L_\epsilon \cap B_\delta(\beta_j)} [h(x, \beta) - h(x, \beta_0)] \leq h(x, \beta_j) - h(x, \beta_0) + Q_L(x)\delta$$

for $j = 1, \ldots, J$ and all x, and therefore

$$\sup_{\beta \in L_\epsilon \cap B_\delta(\beta_j)} \bar{H}(\beta) \leq \bar{H}(\beta_j) + \bar{Q}_L \delta.$$

Consequently,

$$\begin{aligned}
P\left[\sup_{\beta \in L_\epsilon} \bar{H}(\beta) > y\right] &\leq \sum_{j=1}^{J} P\left[\sup_{\beta \in L_\epsilon \cap B_\delta(\beta_j)} \bar{H}(\beta) > y\right] \\
&\leq \sum_{j=1}^{J} P[\bar{H}(\beta_j) + \bar{Q}_L \delta > y] \\
&\leq \sum_{j=1}^{J} \{P[\bar{Q}_L > z] + P(\bar{H}(\beta_j) + z\delta > y)\} \\
&\leq \sum_{j=1}^{J} \{P(\bar{Q}_L > z) \\
&\qquad + P[\bar{H}(\beta_j) - E(\bar{H}(\beta_j) > y - z\delta - g]\} \\
&\leq J \times P[\bar{Q}_L > z] \\
&\qquad + \sum_{j=1}^{J} P\left[\bar{H}(\beta_j) - E(\bar{H}(\beta_j)) > \frac{y - g}{2}\right].
\end{aligned} \tag{3.6}$$

But $y > g$ and J stays fixed, so all $J + 1$ terms on the right hand side of (3.6) are exponentially small by Assumptions (A4) and (A5). □

3.4 *Proof of Proposition* 3.3

By the second part of Assumption (A6), there exists a $\delta_1 > 0$ and a $t > 0$ such that

$$y_0 = \sup_{\beta \in \Omega / B_t(\beta_0)} e_\delta(\beta) < 0$$

for all $0 < \delta < \delta_1$. Define

$$K_0 = K \cap B_t(\beta_0) \quad \text{and} \quad K_1 = K / K_0$$

where t is chosen as before. We have

$$P\left[\sup_{\beta \in K} \bar{H}(\beta) > y\right] \leq P\left[\sup_{\beta \in K_0} \bar{H}(\beta) > y\right] + P\left[\sup_{\beta \in K_1} \bar{H}(\beta) > y\right], \tag{3.7}$$

and since $K_0 \subseteq B_t(\beta_0)/B_\xi(\beta_0) \subseteq \Omega$ for some $\xi > 0$, the first term in (3.7) is exponentially small according to (3.5) for some $y < 0$.

Let now $G(j), j \in \boldsymbol{Z}^p$, be a partition of $\boldsymbol{R}^p$. Then

$$\begin{aligned} P\left[\sup_{\beta \in K_1} \bar{H}(\beta) > y\right] &\leq \sum_{j \in \boldsymbol{Z}^p} P\left[\sup_{\beta \in G(j) \cap K_1} \bar{H}(\beta) > y\right] \\ &= \sum_{|j| > r} + \sum_{|j| \leq r} \end{aligned} \tag{3.8}$$

for any $r > 0$. Let us take

$$G(j) = \left\{u = (u_1, \ldots, u_p)^T \in \boldsymbol{R}^p : \left(j_q - \frac{1}{2}\right)\delta_0 \leq u_q < \left(j_q + \frac{1}{2}\right)\delta_0, q = 1, \ldots, p\right\}.$$

Then

$$\sup_{\beta \in G(j) \cap K_1} \bar{H}(\beta) \leq \bar{Y}_{\beta_j, \delta} \quad \text{for any} \quad \beta_j \in G(j) \cap K_1,$$

and

$$P\left[\sup_{\beta \in G(j) \cap K_1} \bar{H}(\beta) > y\right] \leq P\left[\bar{Y}_{\beta_j, \delta} > y\right]. \tag{3.9}$$

Then for $y > y_0$ the term (3.9) is, according to (A6), exponentially small since $\beta_j \notin B_t(\beta_0)$. The second term in the sum (3.8) is therefore exponentially small. If $r > t + \delta$, we get for the first term in (3.8) the bound

$$\sum_{|j| \geq r} P[\bar{Y}_{\beta_j, \delta} > y] \tag{3.10}$$

where $\beta_j = \beta_0 + j\delta/(2p^{1/2})$.

Finally, from (A6), we have for all β

$$P[\bar{Y}_{\beta, \delta_0} > y] \leq [\exp\{K_{\beta, \delta_0}(\theta_0) - \theta_0 y\}]^n$$

$$\leq C_0^n M_{\beta,\delta_0}(\theta_0)^n$$
$$\leq \frac{C_1^n}{(1+|\beta-\beta_0|)^{n\alpha}}$$

for some constant $C_1 > 0$. Taking r so large that for some n_0 we have

$$\sum_{|j|\geq r} \frac{C_1^{n_0}}{\{1+\delta|j|/(2p^{1/2})\}^{n_0\alpha}} < 1,$$

we find that (3.10) is exponentially small. Consequently both terms in (3.7) are exponentially small. □

3.5 *Proofs of Theorems* 2.1–2.3

The proofs of Theorems 2.1–2.3 are easy consequences of Propositions 3.1–3.3.

PROOF OF THEOREM 2.1. Let $K \subset \boldsymbol{R}^p$ be any compact set which satisfies $B_\xi(\beta_0) \subset K \subset \Omega$ for some $\xi > 0$. Let $\hat{\beta}_\Omega$ and $\hat{\beta}_K$ be the maximisers of $\bar{H}(\beta)$ over $\beta \in \Omega$ and $\beta \in K$, respectively. A direct consequence of Proposition 3.3 is that $\hat{\beta}_K = \hat{\beta}_\Omega$ on a set whose complement has exponentially small probability. Now apply Proposition 3.2 to obtain Theorem 2.1. □

PROOF OF THEOREM 2.2. This follows from the uniqueness of $\hat{\beta}_U$, the solution of the equation $\bar{D}_1(\beta) = 0$, combined with Proposition 3.1. □

PROOF OF THEOREM 2.3. Let $\hat{\beta}_0$ be the solution of $\bar{D}_1(\beta) = 0$ whose existence on a set whose complement has exponentially small probability is guaranteed by Proposition 3.1. We have, for any $\gamma > 0$,

$$\begin{aligned} &P(|\hat{\beta}_S - \beta_0| > \gamma) \\ &\quad = P\left(|\hat{\beta}_S - \beta_0| > \gamma, |\hat{\beta}_P - \beta_0| \leq \frac{\gamma}{4}\right) + P\left(|\hat{\beta}_S - \beta_0| > \gamma, |\hat{\beta}_P - \beta_0| > \frac{\gamma}{4}\right) \\ &\quad \leq P\left(|\hat{\beta}_0 - \beta_0| > \frac{\gamma}{4}\right) + P\left(|\hat{\beta}_P - \beta_0| > \frac{\gamma}{4}\right), \end{aligned}$$

and so Theorem 2.3 follows from (2.3) and Proposition 3.1. □

4. Application to elliptical contrast functions

In this section, $X_1, \ldots, X_n$ will denote IID d-dimensional random vectors, and we shall use X to denote a generic X_i. We consider "elliptical" constrast functions of the form

$$h(x,\beta) = -\frac{1}{2}\log\det(V) + \log f\{(x-\mu)^T V^{-1}(x-\mu)\} \tag{4.1}$$

where $f : [0,\infty) \to [0,\infty)$ is a given function. In (4.1), $\beta = (\mu, V)$ where μ is a d-dimensional location vector, and V is a symmetric positive-definite $d \times d$ scatter matrix. The effective dimension of β is given by $p = d + \frac{1}{2}d(d+1) = \frac{1}{2}d(d+3)$.

Remark 4.1. If f is normalised so that $\int_0^\infty r^{d-1} f(r^2) dr = 1$, then

$$g(x; \mu, V) = \{\det(V)\}^{-1/2} f\{(x-\mu)^T V^{-1}(x-\mu)\} \tag{4.2}$$

is a probability density function on $\boldsymbol{R}^d$ for all $\mu \in \boldsymbol{R}^d$ and all $V > 0$ (where $V > 0$ means V is positive definite). Families of distributions generated via (4.2) are known as elliptical models, because the contours of constant density are ellipses. When $d = 1$, the elliptical models are precisely the symmetric location-scale models. See, for example, Chmielewski (1981), Khatri (1988) and Mitchell (1988, 1989) for further details of elliptic distributions.

The remainder of this section is divided into two parts. In Subsection 4.1, we present two general results which facilitate the application of Theorem 2.1 to elliptical contrast functions, while in Subsection 4.2 we briefly specialise the discussion to maximum likelihood estimators (i.e. it is assumed that the contrast function is the true log-likelihood).

4.1 *Two general results*

The natural parameter space for the contrast function (4.1) is

$$\{\beta = (\mu, V) : \mu \in \boldsymbol{R}^d; V(d \times d), V > 0\}. \tag{4.3}$$

The set (4.3) has a boundary

$$\{\beta = (\mu, V) : \mu \in \boldsymbol{R}^d; V \geq 0, \det(V) = 0\}, \tag{4.4}$$

where $V > 0$ ($V \geq 0$) means that V is positive (non-negative) definite. It turns out that some of the assumptions stated in Subsection 2.2, in particular (A5) and (A6), are very difficult to check at the boundary (4.4), and may even fail to hold there.

For this reason, it is not a good idea to take the set Ω in Theorem 2.1 to be the whole parameter space (4.3). Instead, we adopt the following choice for Ω:

$$\Omega = \{\beta = (\mu, V) : \mu \in \boldsymbol{R}^p; V > 0, \lambda_1(V) \geq \sigma^\dagger\}$$

where $\lambda_1(V)$ is the smallest eigenvalue of V, and $\sigma^\dagger > 0$ is "sufficiently small". We deal with the complementary set

$$\Omega^\dagger = \{\beta = (\mu, V) : \mu \in \boldsymbol{R}^p; V > 0, \lambda_1(V) \in (0, \sigma^\dagger)\} \tag{4.5}$$

separately. More specifically, we use a direct argument (Proposition 4.1 below) to show that, barring exponentially small probability, no maximum of the contrast function occurs in the region (4.5), provided $\sigma^\dagger$ is sufficiently small.

Define the ellipse $E_\eta = E_\eta(\mu, A) \subset \boldsymbol{R}^d$ by

$$E_\eta(\mu, A) = \{x \in \boldsymbol{R}^p : (x-\mu)^T A^{-1}(x-\mu) \leq \eta\}$$

for any $\mu \in \boldsymbol{R}^d$ and any symmetric positive-definite $d \times d$ matrix A. We shall also write $\beta = (\mu, \sigma, A)$ where $V = \sigma A$, $\sigma = \lambda_1(A)$, and $\lambda_1(A) = 1$, and $\mathcal{A}$ will denote

the set of symmetric $d \times d$ matrices with $\lambda_1(A) = 1$. We shall identify $\bar{H}(\mu, \sigma, A)$ with $\bar{H}(\beta) = n^{-1} \sum_{i=1}^{n} h(X_i, \beta)$, where h is the contrast function defined in (4.1).

Consider the following conditions on f and the distribution of X.

(C1) For each unit vector $\xi \in \boldsymbol{R}^p$, the distribution of $U = \xi^T X$ is absolutely continuous with respect to Lebesgue measure on $\boldsymbol{R}$, with density $g_\xi(u)$ say. In addition,

$$\sup_{\xi : \xi^T \xi = 1} \sup_{u \in \boldsymbol{R}} g_\xi(u) < \infty.$$

(C2) The function $f(y)$ in (4.1) is strictly positive, differentiable for all $y > 0$, and satisfies (i) and (ii) below:

(i) $$\sup_{y \in (0,\infty)} \frac{|y f'(y)|}{f(y)} < M$$

for some constant $M < \infty$; and

(ii) $$\limsup_{y \to \infty} \frac{y f'(y)}{f(y)} < -p/2.$$

Remark 4.2. At first glance, Condition (C2)(ii) may seem unfamiliar. However, it is easy to show that it is satisfied by any elliptical distribution whose support is the whole of $\boldsymbol{R}^p$, provided the function f in (4.2) is ultimately monotonic decreasing.

PROPOSITION 4.1. *Under Conditions* (C1) *and* (C2), *there exists a* $\sigma^\dagger > 0$ *such that*

$$\left\{ \sup_{0 < \sigma < \sigma^\dagger} \bar{H}(\mu, \sigma, A) > \bar{H}(\mu, \sigma^\dagger, A) \quad \text{for some} \quad (\mu, A) \in \mathbf{R}^p \times \mathcal{A} \right\} \quad \text{has ESP.}$$

PROOF. Using Condition (C2), choose $x_0 > 0$ and $\gamma > 0$ such that

$$x > x_0 \quad \text{implies} \quad \frac{x f'(x)}{f(x)} < -p/2 - \gamma;$$

and choose $\pi > 0$ so that

$$p/2 + \pi M - (1 - \pi)(p/2 + \gamma) < 0.$$

Take a fixed $\eta > 0$ and choose $\sigma^\dagger$ so that $\eta / \sigma^\dagger > x_0$; and define

$$u_j = (X_j - \mu) A^{-1} (X_j - \mu), \qquad j = 1, \ldots, n.$$

Let Γ_n and Γ denote the empirical and true probability measures, respectively. So if $C \subseteq \boldsymbol{R}^d$, then

$$\Gamma_n(C) = \frac{\#\{i : X_i \in C\}}{n}$$

and, for measurable C, $\Gamma(C) = P[X \in C]$. On the set

(4.6) $$\sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}} \Gamma_n\{E_\eta(\mu, A)\} \leq \pi,$$

the number of u_j's with $u_j < \eta$ is less than $n\pi$. When $u_j > \eta$ and $0 < \sigma < \sigma^\dagger$ then, still on the set (4.7), we have $u_j/\sigma > \eta/\sigma^\dagger > x_0$, and consequently

$$\frac{\partial \bar{H}}{\partial \sigma}(\mu, \sigma, A) = -\frac{n}{\sigma}\left[p/2 + n^{-1}\sum_{i=1}^{n}\left(\frac{u_j}{\sigma}\right)\frac{f'\left(\frac{u_j}{\sigma}\right)}{f\left(\frac{u_j}{\sigma}\right)}\right] > 0$$

for all σ satisfying $0 < \sigma < \sigma^\dagger$, and all $(\mu, A) \in \boldsymbol{R}^p \times \mathcal{A}$. Therefore on (4.6) we have

$$\bar{H}(\mu, \sigma^\dagger, A) \geq \sup_{\sigma \in (0, \sigma^\dagger)} \bar{H}(\mu, \sigma, A)$$

for each fixed $(\mu, A) \in \boldsymbol{R}^p \times \mathcal{A}$. We must therefore show that the complement of (4.6) has exponentially small probability.

In Pollard's (1984) terminolgy, the ellipses in $\boldsymbol{R}^d$ constitute a class of sets with polynomial discrimination. Consequently, an exponential inequality of the following type holds (see Pollard (1984), Chapter 2, for more general results): for any $\epsilon > 0$, there exist positive constants α and ρ such that

$$(4.7) \qquad P\left[\sup_{\eta > 0}\sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}}\left|\Gamma_n\{E_\eta(\mu, A)\} - \Gamma\{E_\eta(\mu, A)\}\right| > \epsilon\right] \leq \alpha \exp(-\rho n).$$

However, since for any $\eta > 0$ we have

$$\sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}}[\Gamma_n\{E_\eta(\mu, A)\} - \Gamma\{E_\eta(\mu, A)\}] \geq \sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}} \Gamma_n\{E_\eta(\mu, A)\} - \sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}} \Gamma\{E_\eta(\mu, A)\},$$

it follows that

$$P\left[\sup_{(\mu, A) \in \boldsymbol{R}^d} \Gamma_n\{E_\eta(\mu, A)\} > \sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}} \Gamma\{E_\eta(\mu, A)\} + \epsilon\right] \leq \alpha \exp(-\rho n),$$

where ϵ, α and ρ are the same as in (4.7).

Let c_1 be the supremum in Condition (C1). Since

$$E_\eta(\mu, A) \subseteq \{x : |\xi^T(x - \mu)| \leq \eta\}$$

for a suitable unit vector $\xi = \xi(\mu, A)$, we have the bound $\Gamma\{E_\eta(\mu, A)\} \leq 2c_1\eta$. Therefore

$$\lim_{\eta \to 0} \sup_{(\mu, A) \in \boldsymbol{R}^d \times \mathcal{A}} \Gamma\{E_\eta(\mu, A)\} = 0.$$

Consequently, given any $\pi > 0$, there exist $\eta > 0$, $\alpha > 0$ and $\rho > 0$ such that

$$P\left[\sup_{(\mu, A) \in \mathbf{R}^d \times \mathcal{A}} \Gamma_n\{E_\eta(\mu, A)\} > \pi\right] \leq P[\sup \Gamma_n > \sup \Gamma + \pi/2] \leq \alpha \exp(-\rho n),$$

as required. □

Recall that, for given $\sigma^\dagger > 0$, we define

$$\Omega = \boldsymbol{R}^d \times \{V(d \times d) : V \text{ symmetric}, \ \lambda_1(V) \geq \sigma^\dagger\}.$$

In Proposition 4.2 below, it will be assumed, without loss of generality, that $\beta_0 = (0, I_d)$.

PROPOSITION 4.2. *Suppose that the true density g and the function f in* (4.1) *satisfy the inequalities*

$$0 \leq f(x^T x) \leq C_1(1 + |x|^2)^{-\alpha_1}, \qquad g(x) \leq C_2(1 + |x|^2)^{\alpha_2} f(x^T x), \tag{4.8}$$

where $C_1 > 0$, $C_2 > 0$ and $\alpha_1 > 1$, $\alpha_2 \in \boldsymbol{R}$ are constants, and $\alpha_1 - \alpha_2 > d/2$. Then for any choice of $\sigma^\dagger > 0$, both parts of Assumption (A6) *are satisfied.*

PROOF. Fix $\sigma^\dagger > 0$. Then by definition of Ω, $\lambda_1(V) \geq \sigma^\dagger$ for all $\beta = (\mu, V) \in \Omega$. Let $s^2 = s^2(V)$ denote the largest eigenvalue of V and write $s_1^2 = s^2(V_1)$. Then since $V_1^{-1} \geq s_1^{-2} I_d$, where I_d is the $d \times d$ identity matrix, it follows that

$$\begin{aligned} 1 + (x - \mu)^T V_1^{-1}(x - \mu) &\geq 1 + |x - \mu|^2 / s_1^2 \\ &\geq 1 + (|x| - |\mu|)^2 / s_1^2. \end{aligned} \tag{4.9}$$

We recall the following fact (see e.g. Mirsky (1955), p. 211): the eigenvalue with largest absolute value of a $d \times d$ matrix $W = (w_{rs})$ is bounded above by $d \sup_{r,s} |w_{rs}|$. Using this fact, an elementary argument shows that there exists a $\delta_0 > 0$ so small that, for any $\beta \in \Omega$, β_1 such that $|\beta_1 - \beta| < \delta$, and $0 < \delta < \delta_0$, we have $\lambda_1(V_1) \geq \sigma^\dagger / 2$. A similar elementary argument shows that there exists a $\delta_1 > 0$ and constants $C_i = C_i(\delta_1, \sigma^\dagger) > 0$, $i = 1, 2$, such that

$$\det(V_1^{-1}) \leq C_1 \det(V^{-1}) \quad \text{and} \quad s_1^{-2} \geq C_2 s^{-2} \tag{4.10}$$

for all $\beta \in \Omega$ and all β_1 such that $|\beta_1 - \beta| < \delta_1$.

Using (4.9) and (4.10), with $\delta > 0$ sufficiently small, we obtain

$$\begin{aligned} \inf_{\beta_1 : |\beta_1 - \beta| < \delta} [1 + (x - \mu_1)^T V_1^{-1}(x - \mu_1)] &\geq \inf_{|\beta_1 - \beta| < \delta} [1 + (|x| - |\mu_1|)^2 / s_1^2] \\ &\geq C^* [1 + (|x| - |\mu|)^2 / s^2] \end{aligned} \tag{4.11}$$

for some $C^* = C^*(\delta, \sigma^\dagger) > 0$ independent of β.

We now show that for $0 < \delta < \delta_1$,

$$\limsup_{\beta \in \Omega, |\beta| \to -\infty} e_\delta(\beta) \to -\infty. \tag{4.12}$$

By definition,

$$e_\delta(\beta) = \int_{\boldsymbol{R}^d} g(x) \sup_{|\beta_1 - \beta| < \delta} [h(x, \beta_1) - h(x, \beta_0)] dx.$$

Using (4.8)–(4.11), we find that

$$\sup_{\beta_1:|\beta_1-\beta|<\delta} h(\beta_1, x) \le C_3 - \frac{1}{2}\log\det(V) - \alpha_1 \log[(1 + (|x| - |\mu|)^2/s^2],$$

and therefore

$$e_\delta(\beta) \le C_4 - \frac{1}{2}\log\det(V) - \alpha_1 \int_{\boldsymbol{R}^d} g(x)\log[1 + (|x| - |\mu|)^2/s^2]dx. \tag{4.13}$$

If $|\beta| \to \infty$, then there are two possibilities: either (i) $|V - I_d| \to \infty$, or (ii) $|V - I_d|$ stays bounded above and $|\mu| \to \infty$. For $\beta = (\mu, V) \in \Omega$, $|V - I_d| \to \infty$ if and only if $s^2 \to \infty$ and $\det(V) \to \infty$, because of the lower bound on $\lambda_1(V)$. But the right hand side of (4.13) goes to minus infinity as $\det(V) \to \infty$, since the integral in (4.13) is always non-negative, and so (4.12) holds in case (i). In case (ii), we may assume that s^2 is bounded above. It is then clear that the integral in (4.13) goes to plus infinity as $|\mu| \to \infty$, and therefore (4.12) also holds in case (ii).

The remainder of the proof is concerned with obtaining the desired bound for

$$M_{\beta,\delta}(\theta) = \int_{\boldsymbol{R}^d} g(x)\left\{\sup_{|\beta_1 \ \ \beta|<\delta} \{\det(V_1)\}^{-1/2} \frac{f\{(x-\mu_1)^T V_1^{-1}(x-\mu_1)\}}{f(x^T x)}\right\}^\theta dx. \tag{4.14}$$

The integrand in (4.14) may be written in the form

$$\frac{g(x)}{f(x^T x)} f(x^T x)^{1-\theta}\left\{\sup_{|\beta_1-\beta|<\delta} \{\det(V_1)\}^{-1/2} f\{(x-\mu_1)^T V_1^{-1}(x-\mu_1)\}\right\}^\theta,$$

and using (4.8), (4.10) and (4.11) we obtain the bound

$$M_{\beta,\delta}(\theta) \le C^*\{\det(V)\}^{-\theta/2} \int_{\boldsymbol{R}^d} \left(\frac{1}{1+|x|^2}\right)^{(1-\theta)\alpha_1-\alpha_2} \left(\frac{1}{1+(|x|-|\mu|)^2/s^2}\right)^{\theta\alpha_1} dx$$

where here and below C^* is a generic positive constant. Transforming to polar coordinates and integrating out the directional component we obtain

$$\begin{aligned} M_{\beta,\delta}(\theta) &\le C^*\{\det(V)\}^{-\theta/2} \int_0^\infty r^{d-1}\left(\frac{1}{1+r^2}\right)^{(1-\theta)\alpha_1-\alpha_2} \\ &\quad \cdot \left(\frac{1}{1+(r-|\mu|)^2/s^2}\right)^{\theta\alpha_1} dr \\ &\le C^*\{\det(V)\}^{-\theta/2} \int_0^\infty \left(\frac{1}{1+r^2}\right)^{\alpha_3} \left(\frac{1}{1+(r-|\mu|)^2/s^2}\right)^{\theta\alpha_1} dr \end{aligned} \tag{4.15}$$

where

$$\alpha_3 = (1-\theta)\alpha_1 - \alpha_2 - \frac{(d-1)}{2}.$$

Since, by hypothesis, $\alpha_1 - \alpha_2 > d/2$, we may choose $\theta_0 > 0$ so that $\alpha_3 > \frac{1}{2}$.

Fix $a > 0$. In the remainder of the proof we obtain two separate bounds for $M_{\beta,\delta}(\theta_0)$, one in the region $|\mu| \leq a$ and the other in the region $|\mu| > a$. Since $\lambda_1(V)$ is bounded away from zero on Ω, it follows that for some $\gamma_1 > 0$ we have

$$
\begin{aligned}
M_{\beta,\delta}(\theta_0) &\leq C^* \{\det(V)\}^{-\theta/2} \\
&\leq \frac{C^*}{(1 + |V - I_d|)^{\gamma_1}} \\
&\leq \frac{C_1^*}{(1 + |\beta - \beta_0|)^{\gamma_1}}
\end{aligned} \tag{4.16}
$$

on $|\mu| \leq a$, where C_1^* is a positive constant which depends on $a, \sigma^\dagger$ and δ.

To obtain the bound on $|\mu| > a$ we bound the integral in (4.15) as follows:

$$
\begin{aligned}
&\left(\int_0^{|\mu|/2} + \int_{|\mu|/2}^{\infty} \right) \left(\frac{1}{1 + r^2} \right)^{\alpha_3} \left(\frac{1}{1 + (r - |\mu|)^2/s^2} \right)^{\theta\alpha_1} dr \\
&\quad \leq \int_0^{|\mu|/2} \left(\frac{1}{1 + r^2} \right)^{\alpha_3} \left(\frac{1}{1 + |\mu|^2/(4s^2)} \right)^{\theta\alpha_1} dr \\
&\qquad + \int_{|\mu|/2}^{\infty} \left(\frac{1}{1 + r^2} \right)^{\alpha_3} dr \\
&\quad \leq C^* \left\{ \left(\frac{1}{1 + |\mu|^2/(4s^2)} \right)^{\theta\alpha_1} + \left(\frac{1}{|\mu|} \right)^{2\alpha_3 - 1} \right\}.
\end{aligned} \tag{4.17}
$$

Putting (4.15)–(4.17) together we obtain the bound of the form

$$
M_{\beta,\delta}(\theta_0) \leq \frac{C_2^*}{(1 + |\beta - \beta_0|)^{\gamma_2}}
$$

on $|\mu| > a$, for some $\gamma_2 > 0$ and some positive constant C_2^* depending on a, δ and $\sigma^\dagger$. Finally, if we put C and α in the statement of (A6) equal to $\max\{C_1^*, C_2^*\}$ and $\min\{\gamma_1, \gamma_2\}$, respectively, then we have a bound for $M_{\beta,\delta}(\theta_0)$ of the desired form over the whole of Ω. □

Remark 4.3. A noteworthy feature of Condition (4.8) is that the tail of the true density g should not be "too much heavier" than the tail of the contrast density f. This condition seems quite natural and, presumably, is not very far from being a necessary condition for the conclusion of Proposition 4.2 to hold.

Before moving on to maximum likelihood estimators, we spell out the strategy for checking that a minimum constrast estimator, $\hat{\beta}$, obtained using an elliptic contrast function f, is such that

$$
\{|\hat{\beta} - \beta| > \gamma\} \quad \text{has ESP.} \tag{4.18}
$$

Given g, the true density of the observations, check that the Conditions (C1) and (C2) in Proposition 4.1, and the conditions of Proposition 4.2, hold. Then check Assumptions (A1)–(A5) in Subsection 2.2.

4.2 *Maximum likelihood estimators*

Conditions (A1)–(A5), (C1) and (C2), and the conditions of Proposition 4.2 are often especially easy to check in the case of maximum likelihood estimators (i.e. when $f = g$). Consider, for example, the logistic density, and the t density with known degrees of freedom. Rigorous large deviation results of the form (4.18) have not previously been established for location-scatter models based on these densities. However, it is a straighforward matter to check all the conditions are satisfied by these two densities, so that (4.18) does hold in both cases.

For the t distribution with ν degrees of freedom a generalization of the proof in Copas (1975) shows that there is only one solution to the likelihood equations for the location and scale parameters when $\nu \geq 1/2$. [Note that $\nu = 1/2$ corresponds to the Cauchy distribution.] Thus when $\nu \geq 1/2$ is known we may appeal directly to Theorem 2.2.

Assume now that $\nu \geq 1/2$ is also unknown, but an estimate $\tilde{\nu}$ is available. If $\tilde{\nu}$ has exponentially small large deviation probabilities, we can use $\tilde{\nu}$ in the likelihood equations for the location and scale parameters μ and σ to obtain estimates of μ and σ which also have exponentially small large deviation probabilities. In practice, an estimate of ν can be based on the configuration

$$\frac{X_1 - m}{r}, \ldots, \frac{X_n - m}{r}$$

where m is the sample median and r is the inter-quartile range of the sample. A particularly simple example of the use of this kind of statistic can be found in Resek (1976).

When $\nu < 1/2$ the situation is more complicated and we do not consider this possibility here.

5. Skovgaard's theorem

In this section we derive a representation of the intensity of a local minimum of the contrast function. The result is very close to the result given in Skovgaard (1990), but our conditions are more easily checked. Instead of requiring a bound on a conditional density we require instead a bound on a marginal density. In Section 6, the relevance of Theorem 5.1 to the tilting argument should become clear.

We first introduce some notation. For a $p \times p$ symmetric matrix C we let $S(C)$ and $\|C\|$ denote the smallest eigenvalue, respectively the lagest eigenvalue. The first and second derivatives of the contrast function $\gamma(Y, \beta)$ are denoted by $D_1(Y, \beta)$ and $D_2(Y, \beta)$ respectively. For a fixed value $\beta = b$ we write $D_1 = D_1(Y, b)$ and $D_2 = D_2(Y, b)$. Define the sets

$$L(\epsilon, \delta) = \{y : D_1(y, \beta) = 0, S(D_2(y, \beta)) > \delta \text{ for some } |\beta - b| < \epsilon\},$$
$$\tilde{L}(\epsilon, \delta) = \{y : |D_1 D_2^{-1}| < \epsilon, S(D_2) > \delta\},$$
$$L(\epsilon, 0) = L(\epsilon), \tilde{L}(\epsilon, 0) = \tilde{L}(\epsilon), \quad \text{and}$$
$$M(y, \epsilon) = \sup \left\{ \left| \frac{\partial^3 \gamma(y, \beta + h\nu)}{\partial h^3} \right|_{h=0} | : |\nu| = 1, |\beta - b| < \epsilon \right\}.$$

THEOREM 5.1. *Suppose that* (i) *the joint density* f *of* (D_1, D_2) *is continuous and satisfies*

$$f(d_1, d_2) \leq \frac{c}{(1 + ||d_2||)^{\xi}}$$

for some constant c *and some* $\xi > p + p(p-1)/2$ *and* (ii) *for some positive* α *and* τ *such that* $\alpha(p+\tau) > p(p-1)(p+\alpha)$,

$$E\{M(Y, \epsilon_0)\}^{p+\alpha} < \infty, \qquad E(||D_2||^{p+\tau}).$$

Then the intensity $g(b)$ *of local minimas of* γ *satisfies*

$$g(b) = \lim_{\epsilon \to 0} \epsilon^{-p} P\{L(\epsilon)\} = \int_{S(d_2)>0} f(0, d_2)|d_2| d(d_2).$$

To prove this theorem we use the following lemmas.

LEMMA 5.2. *Assume that*

$$EM(Y, \epsilon_0)^{p+\alpha} < \infty, \qquad E||D_2||^{p+\tau} < \infty,$$

for some positive ϵ_0, α, τ *which satisfy*

$$\alpha(p+\tau) > p(p-1)(p+\alpha),$$

and assume that the density of D_1 *is bounded. Then for all* λ *with* $p(p-1)/(p+\tau) < \lambda < \alpha/(p+\alpha)$ *we have*

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^p} P\{L(\epsilon) \setminus L(\epsilon, \epsilon^{\lambda})\} = 0.$$

PROOF. Consider $\epsilon \in (0, \epsilon_0)$, and $y \in L(\epsilon)$, and let $\hat{\beta} = \hat{\beta}(y)$ denote a β for which $|\hat{\beta} - b| < \epsilon$ and $D_1(y, \hat{\beta}) = 0$. Then Taylor's theorem yields

$$0 = D_1(y, \hat{\beta}) = D_1(y, b) + (\hat{\beta} - b) D_2(y, b) + \frac{1}{2} M(y, \epsilon_0)|\hat{\beta} - b|^2 w_1, \tag{5.1}$$

where w_1 is a p-vector with $|w_1| \leq 1$, and

$$D_2(y, \hat{\beta}) = D_2(y, b) + M(y, \epsilon_0)|\hat{\beta} - b| Q, \tag{5.2}$$

where $Q = (Q_i j)$ is a symmetric $p \times p$ matrix with $|Q_{ij}| \leq 1$ for $i, j = 1, \ldots, p$.

Consider $p/(p+\alpha) < \zeta < 1$, $\eta > p/(p+\tau)$ and $0 < \lambda < 1 - \zeta$. Then, from the moment assumptions combined with Markov's inequality, we have

$$\lim_{\epsilon \to 0} \epsilon^{-p} P\{M(Y, \epsilon_0) > \epsilon^{-\zeta}\} = \lim_{\epsilon \to 0} \epsilon^{-p} P\{||D_2|| > \epsilon^{-\eta}\} = 0,$$

and we only need to consider the set

$$\{L(\epsilon) \setminus L(\epsilon, \epsilon^{\lambda})\} \bigcap \{y : M(y, \epsilon_0) \leq \epsilon^{-\zeta}\} \bigcap \{y : ||D_2|| \leq \epsilon^{-\eta}\}. \tag{5.3}$$

From (5.2) we have on this set

$$S(D_2) < \epsilon^\lambda + p\epsilon^{1-\zeta} < 2\epsilon^\lambda \tag{5.4}$$

for ϵ small. And from (5.1) we find

$$D_1 \in \{\epsilon v D_2 + \epsilon^{2-\zeta} w : |v| \le 1, |w| \le 1\} = A(\epsilon),$$

say. Using (5.4) and $||D_2|| \le \epsilon^{-\eta}$ we find that the volume of $A(\epsilon)$ is bounded by

$$\begin{aligned} C_p(2\epsilon^{1+\lambda} + \epsilon^{2-\zeta})(\epsilon^{1-\eta} + \epsilon^{2-\zeta})^{p-1} &= \epsilon^p C_p(2 + \epsilon^{1-\zeta-\lambda})(1 + \epsilon^{1-\zeta+\eta})^{p-1}\epsilon^{\lambda-\eta(p-1)} \\ &= \epsilon^p O(\epsilon^{\lambda-\eta(p-a)}). \end{aligned}$$

Using that $\alpha(p+\tau) > p(p-1)(p+\alpha)$ we can take η and ζ sufficiently small and λ sufficiently large so that $\lambda - \eta(p-a) > 0$. Since by hypothesis the density of D_1 is bounded we have shown that the probability of (5.3) is $o(\epsilon^p)$ and the result follows. □

LEMMA 5.3. *Suppose that for fixed $\epsilon_0 > 0$ and $\alpha > 0$, we have $EM(Y, \epsilon_0)^{p+\alpha} < \infty$. Then for all $0 < \lambda < \alpha/(p+\alpha)$ and all $\delta > 0$,*

$$\lim_{\epsilon\to 0} \epsilon^{-p} P\{\tilde{L}(\epsilon, \delta)\} \le \lim_{\epsilon\to 0} e^{-p} P\{L(\epsilon, \epsilon^\lambda)\} \le \lim_{\epsilon\to 0} \epsilon^{-p} P\{\tilde{L}(\epsilon)\}.$$

PROOF. Let $y \in L(\epsilon, \epsilon^\lambda)$. From (5.2) we have

$$S(D_2) \ge \epsilon^\lambda - \epsilon p M(y, \epsilon_0) = \epsilon^\lambda\{1 - \epsilon^{1-\lambda} p M(y, \epsilon_0)\}.$$

Given $0 < \omega < 1/(2p)$ we have therefore $S(D_2) \ge \frac{1}{2}\epsilon^\lambda$ if $M(y, \epsilon_0) \le \epsilon^{\lambda-1}\omega$. Then from (5.1) we obtain

$$|D_1 D_2^{-1}| \le \epsilon + \frac{1}{2} M(y, \epsilon_0)\epsilon^2 |w_1 D_2^{-1}| \le \epsilon + \frac{1}{2}\epsilon^{\lambda-1}\omega\epsilon^2 \frac{2}{\epsilon^\lambda} = \epsilon(1+\omega).$$

We have then shown that

$$L(\epsilon, \epsilon^\lambda) \subseteq \tilde{L}\{\epsilon(1+\omega)\} \bigcup \{y : M(y, \epsilon_0) > \epsilon^{\lambda-1}\omega\}.$$

Therefore, using Markov's inequality, we have,

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{1}{\epsilon^p} P\{L(\epsilon, \epsilon^\lambda)\} &\le (1+\omega)^p \lim_{\epsilon\to 0} \frac{1}{\epsilon^p} P[\tilde{L}(\epsilon)] \\ &\quad + \lim_{\epsilon\to 0} \frac{1}{\epsilon^p} \frac{\epsilon^{(1-\lambda)(p+\alpha)} EM(Y, \epsilon_0)^{p+\alpha}}{\omega^{p+\alpha}} \\ &= (1+\omega)^p \lim_{\epsilon\to 0} \frac{1}{\epsilon^p} P\{\tilde{L}(\epsilon)\}, \end{aligned}$$

and letting $\omega \to 0$ we obtain

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon^p} P\{L(\epsilon, \epsilon^\lambda)\} \le \lim_{\epsilon\to 0} \frac{1}{\epsilon^p} P[\tilde{L}(\epsilon)].$$

For the reverse inequality, we start with $y \in \tilde{L}\{\epsilon, \delta\}$ and $\omega > 0$. It follows from (5.2) that, if

$$M(y, \epsilon_0) < \delta\omega/[2\epsilon p(1+\omega)^2] \quad \text{and} \quad |\beta - b| < \epsilon(1+\omega),$$

then

$$\begin{aligned} S\{D_2(y,\beta)\} &\geq \delta - \frac{\delta\omega}{2\epsilon p(1+\omega)^2} p\epsilon(1+\omega) \\ &\geq \frac{1}{2}\delta. \end{aligned}$$

Defining

$$R(v) = \frac{1}{\epsilon(1+\omega)} D_1(y, b - \epsilon(1+\omega)v) D_2^{-1} + v,$$

we find from the expansion in (5.1) that for $|v| \leq 1$,

$$\begin{aligned} |R(v)| &\leq \left| \frac{D_1 D_2^{-1}}{\epsilon(1+\omega)} + \frac{1}{2} M(y, \epsilon_0)\epsilon(1+\omega) w D_2^{-1} \right| \\ &\leq \frac{1}{1+\omega} + \frac{1}{2} \frac{\delta\omega}{2\epsilon p(1+\omega)^2} \epsilon(1+\omega) \frac{1}{\delta} \leq \frac{1+\omega/4}{1+\omega} \\ &< 1. \end{aligned}$$

Using the fixed point theorem we find that there exists a v with $|v| < 1$ such that $R(v) = v$ or, equivalently,

$$D_1(y, \beta) = 0 \quad \text{for some} \quad |b - \beta| < \epsilon(1+\omega).$$

Thus for $\{\epsilon(1+\omega)\}^\lambda < \frac{1}{2}\delta$

$$\tilde{L}\{\epsilon, \epsilon^\lambda\} \subseteq L\{\epsilon(1+\omega), [\epsilon(1+\omega)]^\lambda\} \bigcup \left\{ y : M(y, \epsilon_0) > \frac{\delta\omega}{2\epsilon p(1+\omega)^2} \right\}.$$

Therefore

$$\begin{aligned} \lim_{\epsilon \to 0} \epsilon^{-p} P\{\tilde{L}(\epsilon, \delta)\} &\leq (1+\omega)^p \lim_{\epsilon \to 0} \epsilon^{-p} P\{L(\epsilon, \epsilon^\lambda)\} \\ &\quad + \lim_{\epsilon \to 0} \epsilon^{-p} \left[\frac{2\epsilon p(1+\omega)}{\delta\omega} \right]^{p+\alpha} E M(Y, \epsilon_0)^{p+\alpha} \\ &= (1+\omega)^p \lim_{\epsilon \to 0} \epsilon^{-p} P\{L(\epsilon, \epsilon^\lambda)\}. \end{aligned}$$

Finally, letting $\omega \to 0$, we obtain

$$\lim_{\epsilon \to 0} \epsilon^{-p} P\{\tilde{L}(\epsilon, \delta)\} \leq \lim_{\epsilon \to 0} P\{L(\epsilon, \epsilon^\lambda)\}. \qquad \square$$

PROOF OF THEOREM 5.1. Let C_p be the volume of the unit ball in $\boldsymbol{R}^p$. Note that the bound on f implies that the marginal density of D_1 is bounded, so that

we can use Lemma 5.2. From Lemma 5.2 and Lemma 5.3 we have for any $\delta > 0$

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{L(\epsilon)\} &= \lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{L(\epsilon,\epsilon^\lambda)\} \geq \lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{\tilde{L}(\epsilon,\delta)\}\\ &= \lim_{\epsilon\to 0}\epsilon^{-p}C_p^{-1}\int_{|d_1 d_2^{-1}|<\epsilon, S(d_2)>\delta} f(d_1,d_2)d(d_1)d(d_2)\\ &= \lim_{\epsilon\to 0} C_p^{-1}\int_{|u|<1, S(d_2)>\delta} f(\epsilon u d_2, d_2)|d_2|du d(d_2).\end{aligned}$$

With the bound on f we can use the dominated convergence theorem and get

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{L(\epsilon)\} \geq \int_{S(d_2)>\delta} f(0,d_2)|d_2|d(d_2).$$

Similarly, from Lemma 5.3 we get

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{L(\epsilon)\} &= \lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{L(\epsilon,\epsilon^\lambda)\} \leq \lim_{\epsilon\to 0}\frac{1}{\epsilon^p C_p}P\{\tilde{L}(\epsilon)\}\\ &= \lim_{\epsilon\to 0} C_p^{-1}\int_{|u|<1, S(d_2)>0} f(\epsilon u d_2, d_2)|d_2|du d(d_2)\\ &= \int_{S(d_2)>\delta} f(0,d_2)|d_2|d(d_2),\end{aligned}$$

and the result of the theorem has been proved. □

6. The tilting argument

In this section we provide a brief account of the tilting argument, indicating the relevance of Theorem 5.1 and the theorems given in Section 2. Consider an estimate $\hat{\beta}$ for which the complement of the event

$$A_\epsilon = \{\hat{\beta} \text{ is the unique solution of (1.1) in } B_\epsilon(\beta_0)\}$$

has exponentially small probability for sufficiently small $\epsilon > 0$; relevant results are given in Sections 2 and 3. Let us argue heuristically for the moment and assume that the density $f(b)$ of $\hat{\beta}$ exists. Then we can write $f(b) = g(b) - h(b)$, where $h(b)$ relates to events where either $|\hat{\beta} - \beta_0|$ is large and there is a second solution to (1.1) at b or there are multiple solutions to (1.1) in a small neighbourhood of β_0. From Proposition 3.1 we have that the integral of h over a neighbourhood of β_0 is exponentially small. There exists therefore a neighbourhood of β_0 for which we can do probability calculations as though $g(b)$ is the density of $\hat{\beta}$, thereby only making an exponentially small error.

However, as noted in the Introduction, *no useful results on the existence of $f(b)$* seem to be available. Consequently, we should interpret the exponential closeness of the distribution of $\hat{\beta}$ and the (approximate) distribution determined by $g(b)$ in terms of *probabilities rather than densities.*

In order to approximate the integral in Theorem 5.1 we use the tilting idea. Define a new measure $P_{\zeta,b}$ by letting each of the n observations (assumed independent) have density $\exp\{\zeta \cdot d_1(x,b)\}/\psi(\zeta,b)$ with respect to the original measure $P_{0,b}$, where

$$\psi(\zeta,b) = \int \exp(\zeta \cdot d_1) dP_{0,b}$$

is a norming constant. Then

$$\begin{aligned} (6.1) \qquad \int_{S(d_2)>0} f(0,d_2)|d_2|d(d_2) &= \psi(\zeta,b)^n \int_{S(d_2)>0} \frac{\exp(\zeta \cdot 0)}{\psi(\zeta,b)^n} f(0,d_2)|d_2|d(d_2) \\ &= \psi(\zeta,b)^n \int_{S(d_2)>0} f_{\zeta,b}(0,d_2)|d_2|d(d_2), \end{aligned}$$

where $f_{\zeta,b}$ is the density of (D_1, D_2) under the tilted measure $P_{\zeta,b}$. We then take $\zeta = \zeta(b)$ such that the mean of D_1 under $P_{\zeta,b}$ is zero. What is important here is that for b in a neighbourhood of β_0 the mean of D_2 under $P_{\zeta,b}$ is positive definite, and we can transform an Edgeworth expansion for $f_{\zeta,b}$ to an expansion of (6.1). The relevant Edgeworth expansion results may be found in Bhattacharya and Rao (1976); the only requirements are moment conditions and mild smoothness conditions on the joint distribution of (D_1, D_2). The main term in the approximation for (6.1) then becomes

$$\psi(\zeta,b)^n \left(\frac{n}{2\pi}\right)^{p/2} |\Sigma_b|^{-1/2},$$

where

$$(6.2) \qquad |\Sigma_b| = \left| \frac{\partial^2 \log \psi(\zeta(b),b)}{\partial\zeta\partial\zeta^*} \right| |E_{P_{\zeta(b),b}}(\bar{D}_2)|^{-2}.$$

where $\bar{D}_2 = n^{-1}D_2$. The variance matrix formula in (6.2) is based on the two results

$$n^{-1/2}D_1 \to N_p\left(0, \frac{\partial^2 \log \psi(\zeta(b),b)}{\partial\zeta\partial\zeta^*}\right) \quad \text{in distribution,}$$

and

$$n^{-1}D_2 \to E_{P_{\zeta(b),b}}(\bar{D}_2) \quad \text{in probability.}$$

Note that the tilting formula (6.1) is derived here without the assumptions made in Field (1982). As discussed above, it is difficult to establish a rigorous expansion for the density of a minimum contrast estimator. Moreover, the conditions given in Field (1982) do not seem to be sufficient to guarantee the expansion given there. To get around this problem, we have chosen to interpret Field's tilting idea in terms of probabilities rather than densities.

REFERENCES

Bahadur, R. R. (1961). On the asymptotic efficiency of tests and estimates, *Sankhyā*, **22**, 229–252.

Bhattacharya, R. N. and Rao, R. R. (1976). *Normal Approximation and Asymptotic Expansion*, Wiley, New York.

Chmielewski, M. A. (1981). Elliptically symmetric distributions: a review and bibliography, *Internat. Statist. Rev.*, **49**, 67–74.

Clarke, B. R. (1991). The selection functional, *Probab. Math. Statist.*, **11**, 149–156.

Copas, J. B. (1975). On the unimodality of the likelihood function for the Cauchy distribution, *Biometrika*, **62**, 701–704.

Daniels, H. E. (1983). Saddlepoint approximations for estimating equations, *Biometrika*, **70**, 89–96.

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

Field, C. A. and Ronchetti, E. (1990). *Small sample asymptotics, IMS Lecture Notes-Monograph Ser.*,**13**.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer, New York.

Jensen, J. L. (1995). *Saddlepoint Approximations*, Oxford University Press, Oxford.

Kester, A. D. M. and Kallenburg, W. C. M. (1986). Large deviations of estimators, *Ann. Statist.*, **14**, 648–664.

Khatri, C. G. (1988). Some inferential problems connected with elliptic distributions, *J. Multivariate Anal.*, **27**, 319–333.

Mirsky, L. (1955). *An Introduction to Linear Algebra*, Clarendon Press, Oxford.

Mitchell, A. F. S. (1988). Statistical manifolds of univariate elliptic distributions, *Internat. Statist. Rev.*, **56**, 1–16.

Mitchell, A. F. S. (1989). The information matrix, skewness tensor and α-connections for the general multivariate elliptic distribution, *Ann. Inst. Statist. Math.*, **41**, 289–304.

Pazman, A. (1986). On the uniqueness of the M.L. estimates in curved exponential families, *Kybernetika*, **22**, 124–132.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer, New York.

Resek, R. W. (1976). Estimation of the parameters of a general student's t distribution, *Comm. Statist. Theory Methods.*, **A5**(7), 635–645.

Sieders, A. and Dzhaparidze, K. (1987). A large deviation result for parameter estimators and its application to nonlinear regression analysis, *Ann. Statist.*, **15**, 1031–1049.

Skovgaard, I. M. (1990). On the density of minimum contrast estimators, *Ann. Statist*, **18**, 779–789.

Varadhan, S. R. S. (1984). *Large Deviations and Applications*, SIAM, Philadelphia.