The Annals of Statistics
2003, Vol. 31, No. 3, 852–879

MODERATE DEVIATIONS OF MINIMUM CONTRAST ESTIMATORS UNDER CONTAMINATION

BY TADEUSZ INGLOT[1] AND WILBERT C. M. KALLENBERG

Wrocław University of Technology and Polish Academy of Sciences, and University of Twente

Since statistical models are simplifications of reality, it is important in estimation theory to study the behavior of estimators also under distributions (slightly) different from the proposed model. In testing theory, when dealing with test statistics where nuisance parameters are estimated, knowledge of the behavior of the estimators of the nuisance parameters is needed under alternatives to evaluate the power. In this paper the moderate deviation behavior of minimum contrast estimators is investigated not only under the supposed model, but also under distributions close to the model. A particular example is the (multivariate) maximum likelihood estimator determined within the proposed model. The set-up is quite general, including also, for instance, discrete distributions.

The rate of convergence under alternatives is determined both when comparing the minimum contrast estimator with a "natural" parameter in the parameter space and when comparing it with the proposed "true" value in the parameter space. It turns out that under the model the asymptotic optimality of the maximum likelihood estimator in the local sense continues to hold in the moderate deviation area.

1. Introduction. Investigating the performance of statistical tests when nuisance parameters are involved often requires knowledge of the behavior of estimators of these nuisance parameters not only under the null hypothesis, but also under alternatives. In many cases the tests are constructed by plugging in estimators of the nuisance parameters in tests which are developed by assuming that the nuisance parameters are known. Recently, this program has been performed for data driven smooth tests for location-scale families; see Inglot and Ledwina (2001a). The latter research was the starting point for studying moderate deviations of the maximum likelihood estimator (MLE) under alternatives. The present paper gives a thorough treatment of this topic, investigating moreover not only MLEs, but more generally minimum contrast estimators (MCEs) in the framework of Jensen and Wood (1998). There is considerable interest in the econometric literature in the asymptotic behavior of MLEs or related estimators under misspecified models; see, for example, White (1982, 1994) and Sin and White (1996).

Received April 2000; revised January 2002.

[1]Supported in part by Wrocław University of Technology Grant 341701.

AMS 2000 subject classifications. 62F12, 62H12, 62E20, 60F10.

Key words and phrases. Minimum contrast estimators, multivariate maximum likelihood estimators, moderate deviations, asymptotic optimality, alternatives, misspecification, nuisance parameters, robustness, score function.

Data driven tests for the simple goodness of fit problem have been introduced by Ledwina (1994). Many standard goodness of fit tests, like the Kolmogorov–Smirnov test and the Cramér–von Mises test have only one direction with the highest possible asymptotic power and behave therefore more like a parametric test for a one-dimensional alternative and not like a well-balanced test with the omnibus property usually attributed to them.

The idea behind the data driven tests can be described as follows. In order to get high power at a broad spectrum of alternatives a sequence of exponential families with growing dimension is defined to cover more and more alternatives (in an orthogonal way, thus adding new alternatives efficiently). Within a given exponential family the goodness of fit problem reduces to a standard testing problem, for which the well-known score test can be applied. However, if the dimension of the exponential family is too large, there is a (strong) power loss due to adding too much noise. Therefore it is very important to choose the "right" dimension for the alternative at hand. The appropriate dimension is chosen by the data, using Schwarz's selection rule. The combined procedure is called a data driven test. Simulation results for data driven tests for the simple goodness of fit problem show that these tests do have a nice omnibus character, giving high and stable power over broad classes of alternatives. Inglot and Ledwina (1996) have provided theoretical support for the simulation results, showing asymptotic optimality for a large set of converging alternatives.

It is argued in Inglot and Ledwina [(1996), page 1985] that to get nontrivial results the convergence of the alternatives should be (slightly) slower than under contiguity. Corresponding to this, the involved levels are not fixed, but tend to 0 as the number n of observations tends to infinity. For more information on this so-called intermediate approach and its relation to the classical Pitman and Bahadur efficiency, we refer to Kallenberg (1999) and Inglot and Ledwina (2001b).

Basic properties such as the asymptotic null distribution and consistency for data driven smooth tests for composite goodness of fit hypotheses have been proved in Inglot, Kallenberg and Ledwina (1997). The simulations presented in Kallenberg and Ledwina (1997a, b) show that the general construction of data driven tests leads to powerful tests being competitive with best known solutions for particular testing problems, like testing normality or exponentiality.

To show asymptotic optimality of data driven smooth tests for composite goodness of fit hypotheses in the intermediate sense, moderate deviation results for the estimators of the nuisance parameters under convergent alternatives are needed. For multivariate location families, such results have been derived under a restrictive condition in Inglot and Ledwina (2001a). The present paper provides a general solution of this problem.

For moderate deviation results of MLEs under the proposed model we refer to Radavichyus (1983). We consider not only the MLE, but also the more general MCEs and emphasize their behavior under departures of the model. Moderate

deviation theorems of univariate M-estimators under the proposed model are presented by Jurečková, Kallenberg and Veraverbeke (1988). Here, we consider the multivariate case, which is not always a trivial generalization of the univariate case; for instance, monotonicity arguments cannot be used. Moreover, we do not restrict consideration to the proposed model.

Large deviation results on MCEs are given in Jensen and Wood (1998), while Almudevar, Field and Robinson (2000) present approximations for tail areas for smooth functions of M-estimators. These results do not reflect on the behavior under sequences of distributions outside the proposed model converging to it, which is the main topic of the present paper. Moreover, Jensen and Wood especially are focused on exponential small probabilities, while we concern ourselves with the moderate deviation area. Although the main subject of this paper is the moderate deviation behavior under converging distributions, we do also have some new results under the proposed model, for instance, the asymptotic optimality of the MLE in the moderate deviation sense. Finally, note that our basic regularity conditions [see (R1) and (R2′) below] are essentially the same as Conditions 3.1 and 3.2 of Pfaff (1982) and are weaker and far more easy to check than (A6) of Jensen and Wood [see the discussion on the conditions and Example 2.3 in Section 2 and the Remarks 3.6 in Pfaff (1982)].

The need for knowledge of the behavior of the MLE is not restricted to data driven tests, but is also of interest when dealing with all kinds of other tests, where nuisance parameters are estimated. Moreover, apart from being needed in evaluating size and power behavior of statistical tests, the problem itself as an estimation problem is also of interest.

Suppose we have a statistical model and have determined the MLE within the model. In general, the model is only a simplification of reality and hence it is of great importance to study the behavior of the MLE under distributions (slightly) different from the proposed model, as they can easily be the true distribution in practice. This robustness aspect is covered by the present results as the alternatives considered in a testing situation can be seen as slight modifications from the assumed model in the estimation problem. In this way one can see how well the MLE behaves under (slight) misspecifications of the model, for which modifications the MLE deteriorates and by which quantity this is determined. Because robustness is often a reason to consider another MCE than the MLE, it is important to investigate the behavior of MCEs under alternatives as well.

Another way of saying this is as follows. We are interested in estimating some parameter θ, which equals θ_0 if the proposed distribution holds. Suppose that the true distribution is slightly different from the proposed distribution. The MCE is in this case close to a parameter value, obtained by a kind of projection on the parameter set Θ. In testing theory this parameter value may be seen as the "least favorable" parameter value w.r.t. the alternative. In estimation theory,

this parameter value is the "natural" parameter value for comparison with the MCE; that is, it is the parameter value on which the MCE is concentrating under the true distribution. Such parameters are often called "pseudo true" values [cf. Machado (1993)]. This "natural" parameter value (or "projection") θ_n is obtained by equating under the true distribution the expectation of the derivative of the contrast function at θ_n to 0. If the direction of the alternative and the derivative of the contrast function at θ_0 are (asymptotically) uncorrelated, the "least favorable" or "natural" parameter can be taken equal to the original parameter value θ_0.

We return to the application of the results of the present paper to prove asymptotic optimality of data driven smooth tests for composite goodness-of-fit tests, using the MLE as estimator of the nuisance parameter. Consider an alternative, say P_n, converging to some distribution belonging to the composite null hypothesis, say P_0 with nuisance parameter θ_0. Denote the power at P_n of the data driven test by β_n. The power at P_n of the Neyman–Pearson most powerful test of the simple null distribution P_0, against (the simple alternative) P_n, is denoted by β_n^+. (Hence, the power at P_n of any test of the composite null hypothesis can never be larger than β_n^+.) In Inglot and Ledwina (2001a) it is shown that $\beta_n^+ - \beta_n$ converges to 0 as $n \to \infty$, provided that the alternative P_n is orthogonal to P_0 in the sense mentioned above.

Because the data driven test has good power properties against a broad class of alternatives, the restriction to the orthogonal direction is rather unsatisfactory. For instance, when testing normality, consider an alternative P_n with density of the form $(1 - c_n) f(x; a, \sigma^2) + c_n g(x)$ with $f(x; a, \sigma^2)$ the normal density with expectation a and variance σ^2 and $g(x)$ some other density. Restriction to an orthogonal direction means that under g the expectation should be equal to a and the variance equal to σ^2. Considering only those "pure" alternatives seems unnecessarily restrictive. On the other hand, for nonorthogonal directions $\beta_n^+ - \beta_n$ does not necessarily converge to 0.

Indeed, it seems more promising to associate with the alternative P_n not simply its limit P_0, but its projection on the composite null hypothesis, say $\tilde{P}_n$, having nuisance parameter θ_n. In the preceding example this is the normal distribution with the same expectation and variance as the alternative P_n. Denote by $\tilde{\beta}_n^+$ the power at P_n of the Neyman–Pearson most powerful test of the simple null distribution $\tilde{P}_n$ against (the simple alternative) P_n. Again, the power at P_n of any test of the composite null hypothesis can never be larger than $\tilde{\beta}_n^+$. Therefore, for proving asymptotic optimality it is certainly enough to show that $\tilde{\beta}_n^+ - \beta_n$ converges to 0 as $n \to \infty$ and hence the difference $\hat{\theta}_n - \theta_n$ is more important than $\hat{\theta}_n - \theta_0$.

So, the results of the present paper can be used for investigating the extension of the asymptotic optimality of data driven tests with the MLE to other directions than the orthogonal ones. Moreover, they are also needed for studying asymptotic optimality of data driven tests using other MCEs than the MLE. Some adaptive

tests of fit using MCEs have been recently introduced in Aerts, Claeskens and Hart (1999).

Apart from the new results holding in some neighborhood of the model, also some new results within the model are presented. As illustration, consider the probability that an estimator, say T_n, deviates more than ε_n from its target θ: $P_\theta(\|T_n - \theta\| > \varepsilon_n)$. Local comparison with ε_n of the order $n^{-1/2}$ reduces to the well-known comparison based on covariance matrices with asymptotic optimality when the Fisher information bound is attained. A similar bound can be given in the strict nonlocal case, where $\varepsilon_n = \varepsilon$ is fixed [cf. Bahadur, Zabell and Gupta (1980)] and for the intermediate range, where ε_n tends to 0, but at a lower rate than $n^{-1/2}$; see Kallenberg (1983). We speak of asymptotic optimality when the estimator attains the lower bound. We show that the well-known asymptotic optimality of the MLE in the classical local sense continues to hold in the moderate deviation region. This explains why the MLE behaves so very well in regular families. If the family is exponentially convex, as for instance in exponential families, the MLE is still asymptotically optimal in the large deviation sense, but in families which are not exponentially convex, large deviation optimality fails; see Kester and Kallenberg (1986). The present moderate deviation results fill the gap between the classical local optimality results and those concerning the large deviation optimality of the MLE, thus completing the whole picture.

The paper is organized as follows. In Section 2 assumptions and exponential bounds are presented. The set-up is quite general. For instance, discrete distributions are allowed. The exponential bounds are derived under rather weak conditions, being for example satisfied in almost any location-scale family. Under somewhat stronger conditions uniqueness of the MCE is obtained, apart from a set of exponentially small probability. The main result on moderate deviations of the MCE under sequences of distributions converging to the proposed model is presented in Section 3, giving not only the exact rate of convergence, but also the rate of the second-order terms. The section starts with a rough sketch of the approach and a discussion of the "natural" parameter θ_n, the parameter value θ_0 corresponding to the proposed distribution and the notion of the "true" value of the parameter. Some corollaries describe moderate deviation results for the Euclidean distance between the MCE and the "natural" parameter θ_n as well as between the MCE and the parameter value θ_0. The section is closed by showing the asymptotic optimality of the MLE in the moderate deviation sense within the proposed model. The proofs are presented in Section 4.

2. Assumptions and exponential bounds. Let $X_1, \ldots, X_n$ be i.i.d. r.v.s with values in a measurable space $(\mathcal{X}, \mathcal{B})$ with distribution P, when the proposed model holds. We write E or Cov when we take the expectation or covariance under P.

However, we are in particular interested in (slight) departures from the proposed model, defined by the probability measure P_n with density w.r.t. P satisfying

$$\frac{dP_n}{dP}(x) = 1 + c_n A_n(x), \qquad c_n \to 0,$$

$$\text{(A)} \quad \sup_n \sup_x |A_n(x)| < \infty, \qquad \int A_n(x)\, dP(x) = 0, \qquad \int A_n^2(x)\, dP(x) = 1.$$

Note that we only require $c_n \to 0$ and no further restrictions on c_n. The sequence $\{c_n\}$ may tend to 0 as slowly as one wants. We may also take $c_n = 0$, thus getting the model distribution P.

The expectation and covariance matrix under P_n are denoted by E_n and Cov_n. For a vector $x \in \mathbb{R}^k$ its Euclidean norm is denoted by $\|x\|$. For a matrix M with elements m_{ij} its norm is defined by $|M|_* = (\sum_{i=1}^k \sum_{j=1}^k m_{ij}^2)^{1/2}$. A constant which should be large enough is denoted by C and a constant which should be small enough is denoted by c. The constants C and c may be different in each case. When referring to a particular constant, it is mostly clear from the context which constant is meant and otherwise, which constant is used is explicitly mentioned.

We are interested in estimating a parameter θ belonging to an open parameter space $\Theta \subset \mathbb{R}^k$, $k \geq 1$. Let $h(x, \theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$, be a measurable function and let

$$\gamma(\theta) = \gamma(\theta, x_1, \ldots, x_n) = -\sum_{i=1}^n h(x_i, \theta)$$

be the contrast function. In principle, the MCE is defined by choosing θ to minimize the contrast function, or, equivalently, to maximize $\sum_{i=1}^n h(x_i, \theta)$. A more precise definition is given just below Theorem 2.1. To facilitate discussion of the MLE, we prefer the formulation in terms of maximizing $\sum_{i=1}^n h(x_i, \theta)$ rather than minimizing $\gamma(\theta, x_1, \ldots, x_n)$.

An important special case occurs when the proposed model is a parametric family with densities $f(x, \theta)$ w.r.t. some σ-finite measure μ. Taking $h(x, \theta) = \log f(x, \theta)$, the MCE equals the MLE.

We put the following assumption on the function h:

(B) There exists $\theta_0 \in \Theta$ such that $\varphi(\theta) = E\{h(X_1, \theta) - h(X_1, \theta_0)\}$ is finite and attains its unique global maximum at θ_0.

When dealing with the MLE in the proposed model, assumption (B) is, as a rule, fulfilled, which can be seen as follows. Let P belong to a parametric family with densities $f(x, \theta)$ and denote the parameter value corresponding to P by θ_0. Then $-\varphi(\theta)$ is the Kullback–Leibler information number of the distribution corresponding to $f(x, \theta)$ w.r.t. P [cf. also (3.9)]. The unique global maximum at θ_0 follows from the well-known properties of Kullback–Leibler information numbers [cf., e.g., Theorem 4.1 in Bahadur (1971)]. Furthermore, in regular families the Kullback–Leibler information number is finite.

Following Zacks [(1971), pages 233–235] and Pitman [(1979), Chapter 8], we shall introduce some useful notation. For $\theta \in \Theta$ and $V \subset \Theta$ an open set write

$$Z_{sr}(\theta, V) = \sup_{\vartheta \in V} \sum_{i=s+1}^{r} \{h(X_i, \vartheta) - h(X_i, \theta)\},$$

where $r > s \geq 0$. In particular, we write

$$Z_r(\theta, V) = Z_{r-1r}(\theta, V) = \sup_{\vartheta \in V} \{h(X_r, \vartheta) - h(X_r, \theta)\}.$$

Basic regularity assumptions are the following:

(R1) There exist $r \geq 1$, $T > 0$ and a compact set $K_0 \subset \Theta$ such that $\theta_0 \in \operatorname{int} K_0$, $E Z_{0r}(\theta_0, K_0^c) < 0$ and $E \exp\{T Z_{0r}(\theta_0, K_0^c)\} < \infty$, where $K_0^c = \Theta \setminus K_0$.

(R2′) There exist a compact set $K \subset \Theta$ with $\theta_0 \in \operatorname{int} K$ and a constant $T > 0$ such that $h(x, \theta)$ is continuous w.r.t. $\theta \in K$ for almost every (a.e.) x and for each $\theta \neq \theta_0$, $\theta \in K$, there exists a neighborhood V_θ of θ with $E \exp\{T Z_1(\theta_0, V_\theta)\} < \infty$.

The first result extends Theorem 5.3.1 of Zacks (1971) and the theorem on page 65 of Pitman (1979) to $\{P_n\}$ instead of the fixed distribution P.

THEOREM 2.1. *Assume* (A), (B), (R1) *and* (R2′) *with* $K = K_0$. *For* $\varepsilon > 0$ *denote*

$$B_n = B_n(\varepsilon) = \left\{ (x_1, \ldots, x_n) : \sup_{\|\theta - \theta_0\| > \varepsilon} \sum_{i=1}^{n} h(x_i, \theta) < \sum_{i=1}^{n} h(x_i, \theta_0) \right\}.$$

Then there exist c, C *such that for all* n,

$$P_n\big((X_1, \ldots, X_n) \notin B_n\big) \leq C e^{-cn}.$$

For a.e. $(x_1, \ldots, x_n)$ *in the set* B_n *the contrast function* $\gamma(\theta)$ *attains its global minimum at some point(s)* $\tilde{\theta}_n$ *belonging to* $\{\theta : \|\theta - \theta_0\| \leq \varepsilon\}$. *Hence, for any such point* $\tilde{\theta}_n$,

$$P_n(\|\tilde{\theta}_n - \theta_0\| > \varepsilon) \leq C e^{-cn}.$$

The proof of Theorem 2.1 is given in Section 4. If the contrast function attains its global minimum on Θ at a unique point, this point is the MCE. If there are more such points, we choose one of them to be the MCE. It does not matter which is chosen. For instance, we may take from the set of solutions the one with smallest coordinates. If there is no point in Θ where the contrast function attains its global minimum on Θ, the MCE is defined as 0. The MCE $\hat{\theta}_n(X_1, \ldots, X_n)$ defined in this way is denoted by $\hat{\theta}$.

In principle there are no problems with multiple roots. A more delicate choice can be made by taking the one closest to a preliminary estimator, if such an

estimator is available [cf. Kester and Kallenberg (1986)]. For more discussion on this point see Small, Wang and Yang (2000).

It is also shown in the proof of Theorem 2.1 that if we remove (R1) in Theorem 2.1 and consider the set

$$BK_n = BK_n(\varepsilon) = \left\{(x_1, \ldots, x_n) : \sup_{\|\theta-\theta_0\|>\varepsilon, \theta\in K} \sum_{i=1}^{n} h(x_i, \theta) < \sum_{i=1}^{n} h(x_i, \theta_0)\right\},$$

we get an analogous statement for BK_n:

$$P_n\big((X_1, \ldots, X_n) \notin BK_n\big) \le Ce^{-cn}.$$

Theorem 2.1 gives the *existence* of the MCE outside a set of exponentially small probability. If we assume more regularity conditions on $h(x, \theta)$, then also *uniqueness* of the MCE can be obtained (apart from a set of exponentially small probability). To this end we replace (R2′) by the stronger (R2) and add (R3) and (R4).

Let $K \subset \Theta$ be a compact and convex set such that $\theta_0 \in \operatorname{int} K$. As far as θ occurs in the assumptions, it is supposed that $\theta \in K$.

(R2) $\frac{\partial}{\partial\theta}h(x, \theta)$ exists for a.e. x and is continuous in θ. Moreover, there exist an (w.r.t. P) integrable function $H(x)$ and a constant $T > 0$ (independent of θ), such that

$$\exp\big[T\{h(x, \theta) - h(x, \theta_0)\}\big]\left|\frac{\partial}{\partial\theta_r}h(x, \theta)\right| \le H(x), \qquad r = 1, \ldots, k. \tag{2.1}$$

(R3) $\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(x, \theta)$ exists for a.e. x, is continuous in θ and the matrices

$$I = \operatorname{Cov}\left(\frac{\partial}{\partial\theta}h(X_1, \theta_0)\right) = E\left[\frac{\partial}{\partial\theta}h(X_1, \theta_0)\right]\left[\frac{\partial}{\partial\theta^T}h(X_1, \theta_0)\right],$$

$$J = E\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(X_1, \theta_0)$$

are finite and nonsingular. Moreover, there exist $d_K > 0$ and a measurable function $G(x)$ such that for $d \in (0, d_K)$ we have $\theta_0 + u \in K$ for all u with $\|u\| \le d_K$ and

$$\sup_{\|u\|\le d}\left|\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(x, \theta_0 + u) - \frac{\partial^2}{\partial\theta\,\partial\theta^T}h(x, \theta_0)\right|_* \le CdG(x). \tag{2.2}$$

(R4) There exists $\delta > 0$ such that

$$E\exp\left\{\delta\left\|\frac{\partial}{\partial\theta}h(X_1, \theta_0)\right\|\right\} \le C,$$

$$E\exp\left\{\delta\left|\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(X_1, \theta_0)\right|_*\right\} \le C, \qquad E\exp\{\delta G(X_1)\} \le C.$$

REMARK 2.2. Conditions (R2)–(R4) are essentially versions of the classical Cramér regularity conditions.

Condition (R2) ensures that we may interchange the order of integration and differentiation, leading to results like (using that φ attains its maximum at θ_0)

$$E\frac{\partial}{\partial\theta}h(X_1,\theta_0)=0.$$

It is shown in Section 4 that condition (R2) easily implies (R2′). Conditions like (R2) frequently appear in the literature [see, e.g., (A5) of Jensen and Wood (1998)]. However, their condition (A5) seems to be more restrictive than our condition (R2); see Example 2.3. Condition (R3) corresponds to (A1) and (A2) of Jensen and Wood (1998) and (R4) to (A3) of that paper.

The following example illustrates in the particular case of the MLE in location-scale families the meaning of our assumptions (R1), (R2′) and (R2).

EXAMPLE 2.3. Consider a location-scale family,

$$f(x,\theta)=\sigma^{-1}f_0\Big(\frac{x-a}{\sigma}\Big),\qquad \theta=(a,\sigma)\in\mathbb{R}\times(0,\infty),$$

with $f_0(x)$ a continuous positive density on $\mathbb{R}$. Put $h(x,\theta)=\log f(x,\theta)$, thus dealing with the MLE $\hat\theta$ of θ. Using the inequalities in the first part of the proof of Theorem III on page 71 in Pitman (1979) it can be seen that if $|x|^2 f_0(x)$ is bounded, then for any compact $K_0\subset\Theta$,

$$\begin{aligned}&E\exp\{TZ_{02}(\theta_0,K_0^c)\}\\&\quad\le C\int\!\!\int|x-y|^{-2T}\{f(x,\theta_0)f(y,\theta_0)\}^{1-T}\,dx\,dy,\end{aligned}$$

which is finite if $T\in[0,\frac12)$. From the last part of the proof of Theorem III on pages 72 and 73 in Pitman (1979) it follows that if $|x|^{2+\eta}f_0(x)$ is bounded for some $\eta>0$ then there exists a compact set $K_0\subset\Theta$ such that $\theta_0\in\operatorname{int}K_0$ and $EZ_{02}(\theta,K_0^c)<0$. So, (R1) is satisfied if $|x|^{2+\eta}f_0(x)$ is bounded on $\mathbb{R}$ for some $\eta>0$. To get (R2′) it is enough to assume that f_0 is bounded and that $\int f_0^{1-\eta}(x)\,dx<\infty$ for some $\eta>0$ while (R2) (with $T=1$) reduces to

$$\Big|\frac{\partial}{\partial\theta_r}f(x,\theta)\Big|\le H(x)\qquad\text{with }\int H(x)\,dx<\infty,\qquad r=1,2,\ \theta\in K.$$

A sufficient condition is that f_0 is continuously differentiable and that $|x|^{2+\eta}|f_0'(x)|$ and $|x|^{1+\eta}f_0(x)$ are bounded.

In particular, put $f_0(x)=\exp\{x-e^x\}$. Then (R1) and (R2) are easily satisfied and, assuming (A), Theorem 2.1 holds, but (A5) of Jensen and Wood (1998) does not hold. To see this, observe that for $h_0(x)=\log f_0(x)=x-e^x$ we get $Q_L(x)=C(|x|+1)e^{|x|}$ and $E\exp\{\delta Q_L(X_1)\}=\infty$ for every $\delta>0$. Note, however, that in this irregular example our condition (R4) is also not satisfied.

Write

$$\ell_n(\theta; x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} h(x_i, \theta)$$

and define the following sets:

$$B1_n = \left\{(x_1, \ldots, x_n) : \left| \frac{\partial}{\partial\theta^T} \ell_n(\theta_0; x_1, \ldots, x_n) - J \right|_* \le \frac{1}{4|J^{-1}|_*} \right\},$$

$$B2_n = \left\{(x_1, \ldots, x_n) : \frac{1}{n}\sum_{i=1}^{n} G(x_i) \le EG + 1 \right\}.$$

LEMMA 2.4. *Assume* (A) *and* (R2)–(R4). *Then we have*

$$P_n\big((X_1, \ldots, X_n) \notin B1_n\big) \le Ce^{-cn},$$

$$P_n\big((X_1, \ldots, X_n) \notin B2_n\big) \le Ce^{-cn}.$$

The proof of Lemma 2.4 is given in Section 4. The next lemma shows that under (R2)–(R4) essentially the contrast function is minimized at a uniquely determined point.

LEMMA 2.5. *Assume* (R2)–(R4) *and let* $0 < \delta < \min\{d_K, \frac{1}{4|J^{-1}|_* C(EG+1)}\}$ *with* C *from* (2.2). *For a.e.* $(x_1, \ldots, x_n) \in B1_n \cap B2_n \cap BK_n(\delta)$, *there exists* $\theta^* = \theta_n^*(x_1, \ldots, x_n)$ *with* $\|\theta^* - \theta_0\| \le \delta$ *and* $\ell_n(\theta^*; x_1, \ldots, x_n) = 0$. *Moreover*, θ^* *is the only solution of* $\ell_n(\theta; x_1, \ldots, x_n) = 0$ *in the set* $\{\theta : \|\theta - \theta_0\| \le \delta\}$.

The proof of Lemma 2.5 is presented in Section 4. As a corollary we get an exponential bound for the MCE.

THEOREM 2.6. *Assume* (A), (B) *and* (R1)–(R4) *with* $K = K_0$. *Let* $0 < \delta < \min\{d_K, \frac{1}{4|J^{-1}|_* C(EG+1)}\}$ *with* C *from* (2.2). *Then* (*except for a* P*-null set*) *on the set* $B1_n \cap B2_n \cap B_n(\delta)$ *the contrast function attains its global minimum in* Θ *at a uniquely determined point* $\hat{\theta}$ *which satisfies* $\|\hat{\theta} - \theta_0\| \le \delta$. *Consequently*,

$$P_n(\|\hat{\theta} - \theta_0\| > \delta) \le Ce^{-cn}.$$

3. Moderate deviation theorem. In this section we show that the assumptions, which we have posed in the previous section, do not give only exponential bounds, but are also sufficient to obtain sharp moderate deviation results for the MCE under P_n. We start with a sketch of the main ideas.

Except for a set with exponentially small probability the MCE is the unique solution of the equation $\ell_n(\theta; x_1, \ldots, x_n) = 0$ existing in each small enough neighborhood of θ_0. Essentially we deal with this solution and apply a Taylor

expansion around a point θ_n (converging to θ_0 and determined later on) of the following form:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}h(X_i,\hat{\theta}) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}h(X_i,\theta_n) - J(\hat{\theta}-\theta_n),$$

implying

$$\hat{\theta}-\theta_n \approx J^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}h(X_i,\theta_n). \tag{3.1}$$

By the law of large numbers it is seen that under P_n the MCE is close to that point $\theta_n \in \Theta$ for which

$$E_n\frac{\partial}{\partial\theta}h(X_1,\theta_n) = 0. \tag{3.2}$$

In principle, we can make a Taylor expansion around θ_0, but

$$\lim_{n\to\infty} E_n\frac{\partial}{\partial\theta}h(X_1,\theta_0) \neq 0 \qquad \text{unless } \lim_{n\to\infty} E\left(A_n(X_1)\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right) = 0.$$

Therefore, the natural point in the parameter space Θ to compare with the MCE is the point θ_n defined by (3.2). The probability measure corresponding to the point θ_n may be seen as a kind of projection of the probability measure P_n on the probability measures, parameterized by Θ. By rewriting

$$E\left(A_n(X_1)\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right) \qquad \text{as } c_n^{-1}\operatorname{Cov}\left(\frac{\partial}{\partial\theta}h(X_1,\theta_0), \frac{dP_n(X_1)}{dP(X_1)}\right),$$

it is seen that the proposed model parameter θ_0 can be taken as the projection if the score function $\frac{\partial}{\partial\theta}h(X_1,\theta_0)$ and the direction of the alternative $\frac{dP_n(X_1)}{dP(X_1)}$ are uncorrelated or "orthogonal."

As the MCE is concentrating on the projection θ_n, the more the direction of the alternative and the score function at θ_0 are correlated, the larger the distance between the MCE and θ_0.

In the statistical literature often the term "true" value of the parameter appears. For example, in Jensen and Wood [(1998), page 674] the "true" value of the parameter is defined as the limit of the estimator. Within the proposed model the parameter has a clear meaning. However, outside the model it is less clear what the parameter is. Especially in testing theory, the definition of the nuisance parameter under alternatives is not obvious.

For instance, suppose that we want to test normality. The null hypothesis contains as nuisance parameters the mean and the variance. However, instead of the mean we may also call it the median or the mode. Considering alternatives like mixtures of a normal distribution with a Laplace distribution, what are the "true" values of the nuisance parameters? More generally, when testing the

null hypothesis $\{f(x,\theta), \theta \in \Theta\}$ and investigating the power at the alternative sequence $(1-c_n)f(x,\theta_0) + c_n g(x)$, how does one define the "true" value of the nuisance parameter at these alternatives? The most appropriate candidate when using MCEs seems to be θ_n as defined in (3.2). Indeed, it turns out that the natural nuisance parameter to compare with the MCE is this θ_n; see also the discussion in the introduction on asymptotic optimality of data driven tests. Therefore, we concentrate in our theorems on the difference between the MCE and the "true" value θ_n. Nevertheless, we also present some results on the deviation between the MCE and θ_0.

The moderate deviation results are obtained by exploiting (3.1) and application of moderate deviation results for row sums of triangular arrays of rowwise i.i.d. random vectors.

After this rough sketch of the approach, we become more precise, and first we present a lemma concerning the existence of the projection θ_n and its behavior.

LEMMA 3.1. *Assume* (A), (B), (R2), (R3) *and* $EG < \infty$. *For n sufficiently large and* $\delta > 0$ *small enough there exists a uniquely determined point* $\theta_n \in \{\theta : \|\theta - \theta_0\| \le \delta\}$ *for which*

$$E_n \frac{\partial}{\partial\theta} h(X_1, \theta_n) = 0.$$

Moreover,

$$\theta_n = \theta_0 - c_n J^{-1} E\left(A_n(X_1) \frac{\partial}{\partial\theta} h(X_1, \theta_0) \right) + O(c_n^2). \tag{3.3}$$

The proof of Lemma 3.1 is given in Section 4.

Let

$$I_n = \mathrm{Cov}_n\left(\frac{\partial}{\partial\theta} h(X_1, \theta_n) \right) \quad \text{and} \quad J_n = E_n \frac{\partial^2}{\partial\theta\, \partial\theta^T} h(X_1, \theta_0).$$

The main result is as follows.

THEOREM 3.2. *Assume* (A), (B) *and* (R1)–(R4) *with* $K = K_0$. *Let* $\{z_n\}$ *be a sequence satisfying* $z_n \to \infty$ *and* $n^{-1/2} z_n \to 0$. *Then*

$$\begin{aligned} &P_n\big(n^{1/2}\|I^{-1/2} J(\hat\theta - \theta_n)\| \ge z_n\big) \\ &\qquad = \exp\left\{ -\frac{z_n^2}{2} + O(c_n z_n^2) + O\left(\frac{z_n^3}{\sqrt{n}} \right) + O(\log z_n) \right\}. \end{aligned} \tag{3.4}$$

In particular,

$$\lim_{n\to\infty} z_n^{-2} \log\left\{ P_n\big(n^{1/2}\|I^{-1/2} J(\hat\theta - \theta_n)\| \ge z_n\big) \right\} = -\tfrac{1}{2}. \tag{3.5}$$

Moreover, I, J *may be replaced by* I_n, J_n, *respectively*, *in* (3.4) *and* (3.5).

The proof of Theorem 3.2 is given in Section 4.

Moderate and large deviation results concerning the Euclidean distance of the MCE to θ_n can be inferred from Theorem 3.2. This leads to the following corollary, which is proved in Section 4.

COROLLARY 3.3. *Assume* (A), (B) *and* (R1)–(R4) *with* $K = K_0$. *Let* $\{z_n\}$ *be a sequence satisfying* $z_n \to \infty$ *and* $n^{-1/2}z_n \to 0$. *Then*

$$P_n(n^{1/2}\|\hat{\theta} - \theta_n\| \geq z_n) = \exp\biggl\{-\frac{\lambda_1 z_n^2}{2} + O(c_n z_n^2) + O\biggl(\frac{z_n^3}{\sqrt{n}}\biggr) + O(\log z_n)\biggr\},$$

where λ_1 *is the smallest eigenvalue of* $JI^{-1}J$. *In particular*,

$$\lim_{n\to\infty} z_n^{-2} \log\bigl\{P_n(n^{1/2}\|\hat{\theta} - \theta_n\| \geq z_n)\bigr\} = -\tfrac{1}{2}\lambda_1.$$

Although θ_n is the natural parameter to compare with the MCE, we also present moderate deviation results on the Euclidean distance between the MCE and θ_0.

The proofs of Theorem 3.4 and Corollary 3.5 use the same types of argument as those applied in the proofs of Theorem 3.2 and Corollary 3.3. Therefore, we omit them here and refer to Inglot and Kallenberg (2001) for the complete proof.

THEOREM 3.4. *Assume* (A), (B) *and* (R1)–(R4) *with* $K = K_0$. *Let* $\{z_n\}$ *be a sequence satisfying* $z_n \to \infty$ *and* $n^{-1/2}z_n \to 0$. *Denote*

$$\Delta_n = \frac{n^{1/2}c_n}{z_n} I^{-1/2} E\biggl(A_n(X_1)\frac{\partial}{\partial\theta}h(X_1, \theta_0)\biggr)$$

and suppose $\Delta_n \to \Delta$ *as* $n \to \infty$ *with* $\|\Delta\| \in [0, 1)$. *Then we have*

$$\begin{aligned} &P_n\bigl(n^{1/2}\|I^{-1/2}J(\hat{\theta} - \theta_0)\| \geq z_n\bigr) \\ &\quad = \exp\biggl\{-\frac{(1 - \|\Delta|)^2 z_n^2}{2} + O(\|\Delta_n - \Delta\| z_n^2) \\ &\qquad + O(c_n z_n^2) + O\biggl(\frac{z_n^3}{\sqrt{n}}\biggr) + O(\log z_n)\biggr\}. \end{aligned} \tag{3.6}$$

In particular,

$$\lim_{n\to\infty} z_n^{-2} \log\bigl\{P_n\bigl(n^{1/2}\|I^{-1/2}J(\hat{\theta} - \theta_0)\| \geq z_n\bigr)\bigr\} = -\tfrac{1}{2}(1 - \|\Delta\|)^2. \tag{3.7}$$

Moreover, I, J *may be replaced by* I_n, J_n, *respectively*, *in* (3.6) *and* (3.7).

COROLLARY 3.5. *Assume* (A), (B) *and* (R1)–(R4) *with* $K = K_0$. *Let* $\{z_n\}$ *be a sequence satisfying* $z_n \to \infty$ *and* $n^{-1/2}z_n \to 0$.

If for Δ_n *defined in Theorem* 3.4,

$$\lim_{n\to\infty} \Delta_n = \Delta \qquad \text{with } 0 \leq \|J^{-1}I^{1/2}\Delta\| < 1,$$

then

$$P_n\big(n^{1/2}\|\hat{\theta}-\theta_0\| \geq z_n\big) = \exp\Big\{-\frac{r^2 z_n^2}{2} + O(\|\Delta_n - \Delta\| z_n^2) + O(c_n z_n^2) + O\Big(\frac{z_n^3}{\sqrt{n}}\Big) + O(\log z_n)\Big\} \tag{3.8}$$

and

$$\lim_{n\to\infty} z_n^{-2} \log\big\{P_n(n^{1/2}\|\hat{\theta}-\theta_0\| \geq z_n)\big\} = -\frac{r^2}{2},$$

where

$$r = \inf\big\{\|I^{-1/2}Ju - \Delta\| : \|u\| \geq 1\big\}.$$

If

$$\liminf_{n\to\infty} \|J^{-1}I^{1/2}\Delta_n\| \geq 1$$

then

$$\lim_{n\to\infty} z_n^{-2} \log\big\{P_n(n^{1/2}\|\hat{\theta}-\theta_0\| \geq z_n)\big\} = 0.$$

The above moderate deviation results can be immediately applied to the case of MLEs.

Suppose $f(x,\theta)$, $x \in \mathcal{X}$, $\theta \in \Theta$, is a family of probability densities on $\mathcal{X}$ with respect to some σ-finite measure μ, where $\Theta \subset \mathbb{R}^k$, $k \geq 1$, is an open set. Assume

$$\int \big[f^{1/2}(x,\theta) - f^{1/2}(x,\vartheta)\big]^2 d\mu(x) > 0$$

for every $\theta \neq \vartheta$. Let $\theta_0 \in \Theta$. Let P denote the probability measure corresponding to $f(x,\theta_0)$ and $h(x,\theta) = \log f(x,\theta)$. Assume (A) and (R1)–(R4). Then all results of Section 2 and Section 3 hold true. We need not assume (B); see also the comments just below assumption (B). [Note that $\varphi(\theta)$ is by (R2)–(R4) finite at least in some neighborhood of θ_0.] Assume, in addition to (R2) and (R3), that $|\frac{\partial^2}{\partial\theta_r\,\partial\theta_s} f(x,\theta)|$, $r,s = 1,\ldots,k$, $\theta \in K$, are bounded by $H(x)f(x,\theta_0)$ with $H(x)$ as in (R2). Then $J = -I$ and in the theorems and corollaries we may write $I^{1/2}$ instead of $I^{-1/2}J$, $I^{-1/2}$ instead of $J^{-1}I^{1/2}$ and I instead of $JI^{-1}J$.

Let us also note that Inglot and Ledwina (2001a) have made the first attempt to get results like the ones above. More precisely, they obtained (cf. Theorem 4.7 in their paper) a weaker version of (3.8) for MLEs under $\Delta_n = 0$ and some strengthening of (R1).

Now, define the Kullback–Leibler information number by

$$K(\theta,\theta_0) = E_\theta \log\Big\{\frac{f(X_1,\theta)}{f(X_1,\theta_0)}\Big\}. \tag{3.9}$$

By (R2)–(R4) and Taylor's formula, it is easily checked that

$$K(\theta,\theta_0)=\tfrac{1}{2}(\theta-\theta_0)^T I(\theta-\theta_0)+o(\|\theta-\theta_0\|^2) \qquad \text{as } \theta\to\theta_0.$$

Noting that conditions 1, 2 and 3 of Kallenberg (1983) are easily verified, an estimator U_n may be called first-order asymptotically optimal [cf. (2.16) on page 502 of Kallenberg (1983)] if

$$\frac{-\log P(n^{1/2}\|U_n-\theta_0\|>z_n)}{\frac{1}{2}\lambda_1 z_n^2}\to 1. \tag{3.10}$$

By taking $c_n=0$, it is seen from either Corollary 3.3 ($\theta_n=\theta_0$ in this case) or Corollary 3.5 ($\Delta_n=\Delta=0$ in this case) that the MLE satisfies (3.10).

COROLLARY 3.6. *Assume* (A) *and* (R1)–(R4) *with* $K=K_0$. *The MLE in the model* $\{f(x,\theta):\theta\in\Theta\}$ *is first-order asymptotically optimal in the moderate deviation sense*.

Corollary 3.6 can be seen as an extension of the well-known asymptotic optimality of the MLE in the local sense. It states that this optimality continues to hold in the moderate deviation region.

Since for $\Delta\neq 0$,

$$\inf\{\|I^{1/2}(u-\Delta)\|^2:\|u\|\geq 1\}<\inf\{\|I^{1/2}u\|^2:\|u\|\geq 1\}=\lambda_1,$$

the optimal rate λ_1 within the model is under P_n obtained by the MLE only if $\Delta=0$. In particular, the rate of the MLE continues to hold under P_n if

$$\lim_{n\to\infty} E\left\{A_n(X_1)\frac{\partial}{\partial\theta}\log f(X_1,\theta_0)\right\}=0.$$

4. Proofs. In this section we present all proofs of the theorems and corollaries of Section 2 and 3. Before proving Theorem 2.1 we need an auxiliary lemma.

LEMMA 4.1. *For every* $r\geq 1$ *and* $n\geq r(r+1)$ *there exist* $\alpha,\beta\in\mathbb{N}\cup\{0\}$ *such that* $n=\alpha r+\beta(r+1)$ *and* $\alpha\geq\frac{n}{2r}-\frac{r+1}{2}$ *and* $\beta>\frac{n}{2r+2}-\frac{r}{2}$.

PROOF. Let $n=ir+j$ with $0\leq j\leq r-1$. Then, by the assumption $n\geq r(r+1)$, it follows that $i\geq r+1$. So, $n=(i-j)r+j(r+1)$ and a representation $n=\alpha r+\beta(r+1)$ exists. Let α_0 be the smallest $\alpha\in\mathbb{N}\cup\{0\}$ for which $n=\alpha_0 r+\beta_0(r+1)$ and let L_0 be the smallest integer for which $\alpha_1=\alpha_0+L_0(r+1)\geq\frac{n}{2r}-\frac{r+1}{2}$. Then a simple calculation shows that $n=\alpha_1 r+\beta_1(r+1)$ and $\beta_1=\beta_0-L_0 r>\frac{n}{2r+2}-\frac{r}{2}$. This proves the lemma. □

PROOF OF THEOREM 2.1. Clearly one has

$$\sup_{\vartheta\in K_0^c}\sum_{i=1}^{r+1}\{h(x_i,\vartheta)-h(x_i,\theta_0)\}\le\frac{1}{r}\sum_{i=1}^{r+1}\sup_{\vartheta\in K_0^c}\sum_{j=1,j\neq i}^{r+1}\{h(x_j,\vartheta)-h(x_j,\theta_0)\} \tag{4.1}$$

and hence

$$EZ_{0\,r+1}(\theta_0,K_0^c)\le\frac{r+1}{r}EZ_{0r}(\theta_0,K_0^c),$$

implying by (R1) that $EZ_{0\,r+1}(\theta_0,K_0^c)<0$.

We shall prove that (R1) also implies $E\exp\{tZ_{0r+1}(\theta_0,K_0^c)\}<\infty$ for $0\le t\le Tr/(r+1)$. Indeed, by (4.1) and the inequality between the geometric and arithmetic means, we have

$$\begin{aligned}&\exp\left\{tZ_{0\,r+1}(\theta_0,K_0^c)\right\}\\&\quad\le\left(\prod_{i=1}^{r+1}\exp\left\{\frac{t(r+1)}{r}\sup_{\vartheta\in K_0^c}\sum_{j=1,j\neq i}^{r+1}\{h(x_j,\vartheta)-h(x_j,\theta_0)\}\right\}\right)^{1/(r+1)}\\&\quad\le\frac{1}{r+1}\sum_{i=1}^{r+1}\exp\left\{\frac{t(r+1)}{r}\sup_{\vartheta\in K_0^c}\sum_{j=1,j\neq i}^{r+1}\{h(x_j,\vartheta)-h(x_j,\theta_0)\}\right\}\end{aligned}$$

and therefore, as $0\le t(r+1)/r\le T$,

$$E\exp\{tZ_{0\,r+1}(\theta_0,K_0^c)\}\le E\exp\left\{\frac{t(r+1)}{r}Z_{0r}(\theta_0,K_0^c)\right\}<\infty.$$

Let $\alpha_n=\alpha$ and $\beta_n=\beta$ be as in Lemma 4.1. Then

$$\begin{aligned}\sup_{\vartheta\in K_0^c}\sum_{i=1}^{n}h(x_i,\vartheta)\le{}&\sum_{s=1}^{\alpha_n}\sup_{\vartheta\in K_0^c}\sum_{i=1+(s-1)r}^{sr}h(x_i,\vartheta)\\&+\sum_{t=1}^{\beta_n}\sup_{\vartheta\in K_0^c}\sum_{i=1+\alpha_n r+(t-1)(r+1)}^{\alpha_n r+t(r+1)}h(x_i,\vartheta).\end{aligned}$$

Consequently,

$$\begin{aligned}&\left\{\sup_{\vartheta\in K_0^c}\sum_{i=1}^{n}h(X_i,\vartheta)\ge\sum_{i=1}^{n}h(X_i,\theta_0)\right\}\\&\quad\subset\left\{\sum_{s=1}^{\alpha_n}Z_{(s-1)r\,sr}(\theta_0,K_0^c)\ge0\right\}\\&\qquad\cup\left\{\sum_{t=1}^{\beta_n}Z_{\alpha_n r+(t-1)(r+1)\,\alpha_n r+t(r+1)}(\theta_0,K_0^c)\ge0\right\}.\end{aligned}$$

Write $M_r(t) = E\exp\{tZ_{0r}(\theta_0, K_0^c)\}$, $t \in [0, T]$. Since $EZ_{0r}(\theta_0, K_0^c) < 0$ by (R1), we infer that $m_r = \inf_{0\le t\le T} M_r(t) < 1$. Now for $t \in [0, T]$,

$$P_n\left(\sum_{s=1}^{\alpha_n} Z_{(s-1)rsr}(\theta_0, K_0^c) \ge 0\right) \le \left[E_n \exp\{tZ_{0r}(\theta_0, K_0^c)\}\right]^{\alpha_n} \le \{(1 + Cc_n)M_r(t)\}^{\alpha_n}. \tag{4.2}$$

Take n so large that $(1 + Cc_n)m_r \le \frac{1}{2}(1 + m_r) < 1$. Then applying (4.2) for the point t at which M_r attains the value m_r, we get

$$\begin{aligned} &P_n\left(\sum_{s=1}^{\alpha_n} Z_{(s-1)rsr}(\theta_0, K_0^c) \ge 0\right) \\ &\quad\le \exp\left\{-\alpha_n \log\left(\frac{2}{1+m_r}\right)\right\} \\ &\quad\le \exp\left\{\frac{r+1}{2}\log\left(\frac{2}{1+m_r}\right)\right\}\exp\left\{-\frac{n}{2r}\log\left(\frac{2}{1+m_r}\right)\right\} = Ce^{-cn}. \end{aligned}$$

Repeating the same argument for $M_{r+1}(t), t \in [0, Tr/(r+1)]$, and analogously defined m_{r+1} and combining both estimates we obtain

$$P_n\left(\sup_{\vartheta\in K_0^c}\sum_{i=1}^n h(X_i, \vartheta) \ge \sum_{i=1}^n h(X_i, \theta_0)\right) \le Ce^{-cn}.$$

Recall that we have denoted

$$BK_{0n}(\varepsilon) = \left\{(x_1, \dots, x_n) : \sup_{\|\theta-\theta_0\|>\varepsilon, \theta\in K_0}\sum_{i=1}^n h(x_i, \theta) < \sum_{i=1}^n h(x_i, \theta_0)\right\}.$$

Using (R2′) and the Heine–Borel theorem, arguing as in the proof of Theorem 5.3.1 of Zacks (1971) and noting that by (B) we have $EZ_1(\theta_0, V_\theta) < 0$ for a sufficiently small neighborhood V_θ, it is seen that for some c, C,

$$P_n\big((X_1, \dots, X_n) \notin BK_{0n}\big) \le Ce^{-cn}.$$

The rest of the proof follows from the relation

$$\begin{aligned} &\{(X_1, \dots, X_n) \notin B_n(\varepsilon)\} \\ &\quad\subset \big\{(X_1, \dots, X_n) \notin BK_{0n}(\varepsilon)\big\} \cup \left\{\sup_{\vartheta\in K_0^c}\sum_{i=1}^n h(X_i, \vartheta) \ge \sum_{i=1}^n h(X_i, \theta_0)\right\}. \end{aligned}$$

This completes the proof of Theorem 2.1. □

PROOF THAT (R2) IMPLIES (R2′). Write for $\vartheta, \theta \in K$, some ξ between θ_0 and ϑ and T given in (R2),

$$\exp\bigl[T\{h(x,\vartheta)-h(x,\theta_0)\}\bigr] = 1 + T\exp\bigl[T\{h(x,\xi)-h(x,\theta_0)\}\bigr]\frac{\partial}{\partial\theta^T}h(x,\xi)(\vartheta-\theta_0).$$

Hence, using (2.1,) we get

$$\sup_{\vartheta\in V_\theta}\exp\bigl[T\{h(x,\vartheta)-h(x,\theta_0)\}\bigr] \le 1 + Tk^{1/2}H(x)\sup_{\vartheta\in V_\theta}\|\vartheta-\theta_0\|,$$

which immediately proves that $E\exp\{TZ_1(\theta_0, V_\theta)\} < \infty$ for any $V_\theta \subset K$. □

PROOF OF LEMMA 2.4. We present a proof of the second statement. The first statement can be proved in the same way. (Note that for a matrix M with elements m_{ij} the statement $|M|_* > \delta$ implies $|m_{ij}| > \frac{\delta}{k}$ for some i, j and hence the proof can be given for each component separately.) In view of (R4), Taylor expansion yields

$$E\bigl[\exp\{\eta(G(X_1)-EG-1)\}\bigr] = 1-\eta+O(\eta^2) \qquad \text{as } \eta\to 0.$$

Hence, there exists $\eta > 0$ such that

$$E\bigl[\exp\{\eta(G(X_1)-EG-1)\}\bigr] < 1-\tfrac{1}{2}\eta.$$

The dominated convergence theorem ensures that

$$\lim_{n\to\infty} E_n\bigl[\exp\{\eta(G(X_1)-EG-1)\}\bigr] = E\bigl[\exp\{\eta(G(X_1)-EG-1)\}\bigr].$$

Therefore, for all $n \ge n_1$ we have

$$E_n\bigl[\exp\{\eta(G(X_1)-EG-1)\}\bigr] < 1-\tfrac{1}{4}\eta.$$

By the Markov inequality we get

$$\begin{aligned} &P_n\left(\frac{1}{n}\sum_{i=1}^n G(X_i) > EG+1\right)\\ &= P_n\left(\sum_{i=1}^n G(X_i) > n(EG+1)\right)\\ &= P_n\left(\exp\left[\eta\sum_{i=1}^n\{G(X_i)-EG-1\}\right] > 1\right)\\ &\le \bigl(E_n\exp[\eta\{G(X_1)-EG-1\}]\bigr)^n \le \left(1-\frac{1}{4}\eta\right)^n \end{aligned}$$

for all $n \ge n_1$ and the result easily follows. □

PROOF OF LEMMA 2.5. By the definition of δ and assumption (R3) we have $\{\theta : \|\theta - \theta_0\| \le \delta\} \subset \mathrm{int}(K)$. For a.e. $(x_1, \dots, x_n) \in B1_n \cap B2_n \cap BK_n(\delta)$ the contrast function has a (local) minimum, which is attained at a point in the set $\{\theta : \|\theta - \theta_0\| \le \delta\}$. Therefore, for a.e. $(x_1, \dots, x_n) \in BK_n(\delta)$ the existence of a solution of $\ell_n(\theta; x_1, \dots, x_n) = 0$ in the set $\{\theta : \|\theta - \theta_0\| \le \delta\}$ follows.

That for a.e. $(x_1, \dots, x_n) \in B1_n \cap B2_n \cap BK_n(\delta)$ there is only one solution of $\ell_n(\theta; x_1, \dots, x_n) = 0$ in the set $\{\theta : \|\theta - \theta_0\| \le \delta\}$ follows from the fact that, for a.e. $(x_1, \dots, x_n)$, $\ell_n(\theta; x_1, \dots, x_n)$ is one-to-one on the set $\{\theta : \|\theta - \theta_0\| \le \delta\}$. The latter is seen from the following inequalities. Let $\|\theta - \theta_0\| \le \delta$ and $\|\theta + \vartheta - \theta_0\| \le \delta$. Then, if $\vartheta \ne 0$, by (R3) for a.e. $(x_1, \dots, x_n) \in B1_n \cap B2_n \cap BK_n(\delta)$,

$$\begin{aligned}
&\|\ell_n(\theta + \vartheta; x_1, \dots, x_n) - \ell_n(\theta; x_1, \dots, x_n)\| \\
&\quad \ge \|J\vartheta\| - \Big\| \ell_n(\theta + \vartheta; x_1, \dots, x_n) - \ell_n(\theta; x_1, \dots, x_n) \\
&\qquad\qquad - \Big(\frac{\partial}{\partial \theta^T} \ell_n(\theta_0; x_1, \dots, x_n)\Big)\vartheta \Big\| \\
&\qquad - \Big\| \Big(\frac{\partial}{\partial \theta^T} \ell_n(\theta_0; x_1, \dots, x_n) - J\Big)\vartheta \Big\| \\
&\quad \ge \frac{\|\vartheta\|}{|J^{-1}|_*} - C(EG + 1)\delta\|\vartheta\| - \frac{\|\vartheta\|}{4|J^{-1}|_*} \ge \frac{\|\vartheta\|}{2|J^{-1}|_*} > 0. \qquad \square
\end{aligned}$$

PROOF OF LEMMA 3.1. Denote

$$\varphi_n(\theta) = E_n[h(X_1, \theta) - h(X_1, \theta_0)].$$

By (B) and the uniform boundedness of $A_n(x)$ it follows that $\varphi_n(\theta)$ is well defined on Θ. Using (R2), (R3) and a Taylor expansion we have for $\theta, \vartheta \in D = \{\theta : \|\theta - \theta_0\| < d_K\} \cap \Theta$ and some ξ between θ and ϑ,

$$\begin{aligned}
h(x, \vartheta) - h(x, \theta) &= \frac{\partial}{\partial \theta^T} h(x, \theta)(\vartheta - \theta) + \frac{1}{2}(\vartheta - \theta)^T \frac{\partial^2}{\partial \theta\, \partial \theta^T} h(x, \theta)(\vartheta - \theta) \\
&\quad + \frac{1}{2}(\vartheta - \theta)^T \Big[\frac{\partial^2}{\partial \theta\, \partial \theta^T} h(x, \xi) - \frac{\partial^2}{\partial \theta\, \partial \theta^T} h(x, \theta)\Big](\vartheta - \theta).
\end{aligned}$$

By (2.2) and the integrability of G (under P as well as P_n), it follows that all terms in the above expansion are integrable with respect to P_n. Hence

$$\begin{aligned}
|\varphi_n(\vartheta) - \varphi_n(\theta)| &\le (1 + Cc_n)\Big\| E \frac{\partial}{\partial \theta} h(X_1, \theta)\Big\| \, \|\vartheta - \theta\| \\
&\quad + (1 + Cc_n)\Big| E \frac{\partial}{\partial \theta\, \partial \theta^T} h(X_1, \theta)\Big|_* \, \|\vartheta - \theta\|^2 \\
&\quad + CEG(X_1)\|\vartheta - \theta\|^3,
\end{aligned}$$

which proves the continuity of φ_n (and φ) in D.

Similarly, taking $\vartheta = \theta + \varepsilon e_i$, $\varepsilon \neq 0$, $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^k$ the ith unit vector, we get

$$\left| \frac{\varphi_n(\vartheta) - \varphi_n(\theta)}{\varepsilon} - E_n \frac{\partial}{\partial \theta_i} h(X_1, \theta) \right| \leq \frac{1}{2} |\varepsilon| (1 + C c_n) E \frac{\partial^2}{\partial \theta_i^2} h(X_1, \theta) + C |\varepsilon|^2 E G(X_1),$$

which, in turn, proves the differentiability of φ_n on D with the derivative $\frac{\partial \varphi_n}{\partial \theta}(\theta) = E_n \frac{\partial}{\partial \theta} h(X_1, \theta) = g_n(\theta)$, say.

Let $0 < \delta < d_K$ and $V = \{\theta : \|\theta - \theta_0\| < \delta, \varphi(\theta) > -\varepsilon\}$, where $\varepsilon > 0$ is small enough that $\operatorname{cl} V \subset \{\theta : \|\theta - \theta_0\| < \delta\}$. Then on $\{\theta : \|\theta - \theta_0\| < \delta\} \setminus V$,

$$\varphi_n(\theta) = \varphi(\theta) + c_n E A_n(X_1) \{h(X_1, \theta) - h(X_1, \theta_0)\} < -\frac{\varepsilon}{2}$$

for n sufficiently large. As $\varphi_n(\theta_0) = 0$ it follows that for n sufficiently large φ_n attains its global maximum in $\operatorname{cl} V$ at some point $\theta_n \in V$ [in which $\varphi_n(\theta_n) \geq 0$] and consequently $\frac{\partial \varphi_n}{\partial \theta}(\theta_n) = E_n \frac{\partial}{\partial \theta} h(X_1, \theta_n) = 0$. This proves the existence.

Now, let $0 < \delta < \min\{d_K, \frac{1}{4|J^{-1}|_* C \sup_n E_n G}\}$ with C from (2.2) and recall that $g_n(\theta) = E_n \frac{\partial}{\partial \theta} h(X_1, \theta)$.

That there is only one solution of $g_n(\theta) = 0$ in the set $\{\theta : \|\theta - \theta_0\| \leq \delta\}$ follows from the fact that $g_n(\theta)$ is one-to-one on the set $\{\theta : \|\theta - \theta_0\| \leq \delta\}$. The latter is seen from the following inequalities. Let $\|\theta - \theta_0\| \leq \delta$ and $\|\theta + \vartheta - \theta_0\| \leq \delta$. Then, if $\vartheta \neq 0$,

$$\begin{aligned} &\|g_n(\theta + \vartheta) - g_n(\theta)\| \\ &\quad \geq \|J\vartheta\| - \left\| g_n(\theta + \vartheta) - g_n(\theta) - \left(\frac{\partial}{\partial \theta^T} g_n(\theta_0) \right) \vartheta \right\| \\ &\qquad - \left\| \left(\frac{\partial}{\partial \theta^T} g_n(\theta_0) - J \right) \vartheta \right\| \\ &\quad \geq \frac{\|\vartheta\|}{|J^{-1}|_*} - C(E_n G)\delta \|\vartheta\| - \frac{\|\vartheta\|}{4|J^{-1}|_*} \geq \frac{\|\vartheta\|}{2|J^{-1}|_*} > 0. \end{aligned}$$

Since, for $\vartheta \to 0$,

$$\begin{aligned} &E_n \frac{\partial}{\partial \theta} h(X_1, \theta_0 + \vartheta) \\ &\quad = \int \left\{ \frac{\partial}{\partial \theta} h(x, \theta_0 + \vartheta) \right\} \{1 + c_n A_n(x)\} \, dP(x) \\ &\quad = \int \left\{ \frac{\partial}{\partial \theta} h(x, \theta_0) \right\} \{1 + c_n A_n(x)\} \, dP(x) \end{aligned}$$

$$+\int\left(\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(x,\theta_0)\right)\vartheta\{1+c_nA_n(x)\}\,dP(x)+O(|\vartheta\|^2)$$

$$=c_nE\left\{A_n(X_1)\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right\}+J\vartheta+O(c_n\|\vartheta\|+\|\vartheta\|^2),$$

it follows that

$$\theta_n=\theta_0-c_nJ^{-1}E\left\{A_n(X_1)\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right\}+O(c_n^2).$$

This gives (3.3) and the proof is complete. □

Next we state and prove two auxiliary lemmas which will be used in the proof of Theorem 3.2, our main theorem. Define

$$Y_{ni}=I_n^{-1/2}\frac{\partial}{\partial\theta}h(X_i,\theta_n).$$

LEMMA 4.2. *Assume* (A), (B) *and* (R2)–(R4). *Let* $\{z_n\}$ *be a sequence satisfying* $z_n\to\infty$ *and* $n^{-1/2}z_n\to 0$. *Then*

$$P_n\left(\left\|n^{-1/2}\sum_{i=1}^n Y_{ni}\right\|\ge z_n\right)=\exp\left\{-\frac{z_n^2}{2}+O\left(\frac{z_n^3}{\sqrt{n}}\right)+O(\log z_n)\right\}.$$

PROOF. By definition of θ_n we have that $E_nY_{ni}=0$ and that $\mathrm{Cov}_n(Y_{ni})$ equals the identity matrix. Moreover, we have [cf. (4.7), below]

$$\|Y_{n1}\|\le|I_n^{-1/2}|_*\left\|\frac{\partial}{\partial\theta}h(X_1,\theta_n)\right\|,$$

$$\lim_{n\to\infty}|I_n^{-1/2}|_*=|I^{-1/2}|_*,\qquad f(x,\theta_0)\{1+c_nA_n(x)\}\le Cf(x,\theta_0)$$

and by (R3),

$$\begin{aligned}\left\|\frac{\partial}{\partial\theta}h(X_1,\theta_n)\right\|&\le\left\|\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right\|+\left\|\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(X_1,\theta_0)(\theta_n-\theta_0)\right\|\\&\quad+C\|\theta_n-\theta_0\|^2G(X_1)\\&\le\left\|\frac{\partial}{\partial\theta}h(X_1,\theta_0)\right\|+\left|\frac{\partial^2}{\partial\theta\,\partial\theta^T}h(X_1,\theta_0)\right|_*\|\theta_n-\theta_0\|\\&\quad+C\,\|\theta_n-\theta_0\|^2\,G(X_1).\end{aligned}$$

Hence by (R4), there exists $\eta>0$ and a constant C such that for all n,

$$E_n\exp(\eta\|Y_{n1}\|)\le C.$$

Therefore, for all $y \in \mathbb{R}^k$, we get

$$|E_n(y^T Y_{n1})^j| \le \|y\|^j E_n \|Y_{n1}\|^j = \|y\|^j \eta^{-j} j! E_n \left(\frac{\|\eta Y_{n1}\|^j}{j!} \right)$$
$$\le \|y\|^j \eta^{-j} j! E_n \exp(\eta \|Y_{n1}\|).$$

Applying Theorem 4.9 of Inglot and Ledwina (2001a) [cf. also Prokhorov (1973) and Theorem 3.1 of Yurinskii (1976)], we get

$$P_n\left(\left\| n^{-1/2} \sum_{i=1}^n Y_{ni} \right\| \ge z_n \right) \le \exp\left\{ -\frac{z_n^2}{2} + O\left(\frac{z_n^3}{\sqrt{n}} \right) + O(\log z_n) \right\}.$$

Let W_{ni} be the first component of Y_{ni}. Then we have $E_n W_{ni} = 0$, $E_n W_{ni}^2 = 1$ and

$$\left| n^{-1/2} \sum_{i=1}^n W_{ni} \right| \le \left\| n^{-1/2} \sum_{i=1}^n Y_{ni} \right\|$$

and hence,

$$P_n\left(\left\| n^{-1/2} \sum_{i=1}^n Y_{ni} \right\| \ge z_n \right) \ge P_n\left(\left| n^{-1/2} \sum_{i=1}^n W_{ni} \right| \ge z_n \right). \tag{4.3}$$

Application of Corollary 2.22 in Book (1976) [cf. also Lemma 4.1(ii) in Jurečková, Kallenberg and Veraverbeke (1988)] yields

$$P_n\left(\left| n^{-1/2} \sum_{i=1}^n W_{ni} \right| \ge z_n \right) = \exp\left\{ -\frac{z_n^2}{2} + O\left(\frac{z_n^3}{\sqrt{n}} \right) + O(\log z_n) \right\}. \tag{4.4}$$

The lemma follows from a combination of (4.3) and (4.4). □

LEMMA 4.3. *Assume* (A), (B) *and* (R2)–(R4). *Then for* $y_n > 0$,

$$P_n\left(\left| \frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial\theta\, \partial\theta^T} h(X_i, \theta_0) - J_n \right|_* \ge y_n \right) \le 2 \exp\left\{ -\frac{n y_n^2 \delta_0^2}{8 C_0} (1 + c y_n)^{-1} \right\},$$

where as previously,

$$J_n = E_n \frac{\partial^2}{\partial\theta\, \partial\theta^T} h(X_1, \theta_0),$$

and C_0, δ_0 *are the constants appearing in* (R4),

$$E \exp\left\{ \delta_0 \left\| \frac{\partial^2}{\partial\theta\, \partial\theta^T} h(X_1, \theta_0) \right\| \right\} \le C_0. \tag{4.5}$$

The proof of Lemma 4.3 is simply an application of Theorem 3.1 of Yurinskii (1976) [cf. Lemma 5.5 in Inglot and Ledwina (2001a)], so we omit the details. Note that by (A), (R2) and (R3),

$$|J_n - J|_* \le Cc_n.$$

PROOF OF THEOREM 3.2. Let δ_0 and C_0 be the constants in (R4); see (4.5). W.l.o.g. assume $\delta_0 < \min\{d_K, \frac{1}{4|J^{-1}|_* C(EG+1)}\}$ with C from (2.2). Take $y_n = (8C_0)^{1/2} n^{-1/2} \delta_0^{-1} z_n$ and define

$$\widetilde{B1}_n = \left\{(x_1, \dots, x_n) : \left| \frac{\partial}{\partial \theta^T} \ell_n(\theta_0; x_1, \dots, x_n) - J_n \right|_* \le y_n \right\}.$$

Then for n sufficiently large $\widetilde{B1}_n \subset B1_n$ and by Lemma 4.3,

$$\begin{aligned} &P_n\big((X_1, \dots, X_n) \notin \widetilde{B1}_n\big) \\ &\quad \le 2\exp\left\{-z_n^2\left(1 + c\frac{z_n}{\sqrt{n}}\right)^{-1}\right\} = 2\exp\{-z_n^2(1 + o(1))\}. \end{aligned} \tag{4.6}$$

By (R3) and Taylor expansion we have with some ξ between θ_0 and θ_n,

$$\begin{aligned} I_n = {} & E_n\left(\frac{\partial}{\partial \theta} h(X_1, \theta_0) + \frac{\partial^2}{\partial \theta\, \partial \theta^T} h(X_1, \xi)(\theta_n - \theta_0)\right) \\ & \times \left(\frac{\partial}{\partial \theta} h(X_1, \theta_0) + \frac{\partial^2}{\partial \theta\, \partial \theta^T} h(X_1, \xi)(\theta_n - \theta_0)\right)^T. \end{aligned}$$

Hence, using (A), (2.2) and (R4), we get

$$|I_n - I|_* \le Cc_n \tag{4.7}$$

and consequently,

$$\big|I_n^{1/2} - I^{1/2}\big|_* \le Cc_n, \qquad \big|I_n^{-1/2} J_n - I^{-1/2} J\big|_* \le Cc_n. \tag{4.8}$$

Now restrict attention to $\{(X_1, \dots, X_n) \in B_n(\delta_0) \cap \widetilde{B1}_n \cap B2_n\}$. By (R2)–(R4) and Taylor expansion of ℓ_n around θ_n we obtain with some ξ between $\hat{\theta}$ and θ_n,

$$\begin{aligned} 0 &= I_n^{-1/2} \ell_n(\hat{\theta}; X_1, \dots, X_n) \\ &= \frac{1}{n} \sum_{i=1}^n Y_{ni} + I_n^{-1/2} J_n(\hat{\theta} - \theta_n) \\ &\quad + I_n^{-1/2}\left(\frac{\partial}{\partial \theta^T} \ell_n(\xi; X_1, \dots, X_n) - \frac{\partial}{\partial \theta^T} \ell_n(\theta_0; X_1, \dots, X_n)\right)(\hat{\theta} - \theta_n) \\ &\quad + I_n^{-1/2}\left(\frac{\partial}{\partial \theta^T} \ell_n(\theta_0; X_1, \dots, X_n) - J_n\right)(\hat{\theta} - \theta_n). \end{aligned} \tag{4.9}$$

Since $\|\xi-\theta_0\| \le \|\theta_n-\theta_0\| + \|\hat\theta-\theta_n\|$, we have, using (R3) and (4.7),

$$\begin{aligned}(4.10)\quad &\left\| I_n^{-1/2}\left(\frac{\partial}{\partial\theta^T}\ell_n(\xi; X_1,\dots,X_n) - \frac{\partial}{\partial\theta^T}\ell_n(\theta_0; X_1,\dots,X_n)\right)(\hat\theta-\theta_n)\right\| \\ &\qquad \le C\|\theta_n-\theta_0\|\,\|I_n^{1/2}(\hat\theta-\theta_n)\| + C\|I_n^{1/2}(\hat\theta-\theta_n)\|^2.\end{aligned}$$

By the definition of $\widetilde{B1}_n$ and (4.7), we get

$$\begin{aligned}(4.11)\quad &\left\| I_n^{-1/2}\left(\frac{\partial}{\partial\theta^T}\ell_n(\theta_0; X_1,\dots,X_n) - J_n\right)(\hat\theta-\theta_n)\right\| \\ &\qquad \le C y_n\|I_n^{1/2}(\hat\theta-\theta_n)\|.\end{aligned}$$

Combining (4.9)–(4.11), we arrive at

$$\begin{aligned}(4.12)\quad &\left\| I_n^{-1/2}J_n(\hat\theta-\theta_n) + \frac{1}{n}\sum_{i=1}^n Y_{ni}\right\| \\ &\qquad \le C_1(y_n+c_n)\|I_n^{1/2}(\hat\theta-\theta_n)\| + C_2\|I_n^{1/2}(\hat\theta-\theta_n)\|^2 \\ &\qquad \le C_1(y_n+c_n)\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\| + C_2\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\|^2\end{aligned}$$

for some constants C_1 and C_2. This implies two inequalities,

$$\begin{aligned}(4.13)\quad &\left\|\frac{1}{n}\sum_{i=1}^n Y_{ni}\right\| \ge (1-C_1y_n-C_1c_n)\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\| \\ &\qquad\qquad - C_2\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\|^2, \\ &\left\|\frac{1}{n}\sum_{i=1}^n Y_{ni}\right\| \le (1+C_1y_n+C_1c_n)\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\| \\ &\qquad\qquad + C_2\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\|^2,\end{aligned}$$

which, in turn give the following inclusions, holding for sufficiently large n,

$$\begin{aligned}(4.14)\quad &\{z_n n^{-1/2} \le \|I_n^{-1/2}J_n(\hat\theta-\theta_n)\| \le (4C_2)^{-1}\} \\ &\qquad \subset \left\{\left\|\frac{1}{n}\sum_{i=1}^n Y_{ni}\right\| \ge (1-C_1y_n-C_1c_n)z_n n^{-1/2} - C_2 z_n^2 n^{-1}\right\}, \\ &\{\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\| \ge z_n n^{-1/2}\} \\ &\qquad \supset \left\{\left\|\frac{1}{n}\sum_{i=1}^n Y_{ni}\right\| \ge (1+C_1y_n+C_1c_n)z_n n^{-1/2} + C_2 z_n^2 n^{-1}\right\}.\end{aligned}$$

Applying Lemma 4.2, Theorem 2.6 and the definition of y_n we obtain

$$
\begin{aligned}
&P_n\big((X_1,\dots,X_n)\in B_n(\delta_0)\cap\widetilde{B1}_n\cap B2_n, n^{1/2}\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\|\ge z_n\big)\\
&\le P_n\left(\left\|n^{-1/2}\sum_{i=1}^n Y_{ni}\right\|\ge(1-Cc_n-Cz_nn^{-1/2})z_n\right)+Ce^{-cn} \qquad (4.15)\\
&\le\exp\left\{-\frac{z_n^2}{2}+O(c_nz_n^2)+O\left(\frac{z_n^3}{\sqrt{n}}\right)+O(\log z_n)\right\}
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
&P_n\big((X_1,\dots,X_n)\in B_n(\delta_0)\cap\widetilde{B1}_n\cap B2_n, n^{1/2}\|I_n^{-1/2}J_n(\hat\theta-\theta_n)\|\ge z_n\big)\\
&\ge P_n\left(\left\|n^{-1/2}\sum_{i=1}^n Y_{ni}\right\|\ge(1+Cc_n+Cz_nn^{-1/2})z_n\right)\\
&\quad -P_n\big((X_1,\dots,X_n)\notin B_n(\delta_0)\cap\widetilde{B1}_n\cap B2_n\big) \qquad (4.16)\\
&\ge\exp\left\{-\frac{z_n^2}{2}+O(c_nz_n^2)+O\left(\frac{z_n^3}{\sqrt{n}}\right)+O(\log z_n)\right\}\\
&\quad -P_n\big((X_1,\dots,X_n)\notin B_n(\delta_0)\cap\widetilde{B1}_n\cap B2_n\big).
\end{aligned}
$$

In view of Theorem 2.1, Lemma 2.4 and (4.6), we have

$$
\begin{aligned}
&P_n\big((X_1,\dots,X_n)\notin B_n(\delta_0)\cap\widetilde{B1}_n\cap B2_n\big)\\
&\qquad\le Ce^{-cn}+2\exp\big\{-z_n^2\big(1+o(1)\big)\big\}
\end{aligned}
$$

and (3.4) follows with $I_n^{-1/2}J_n$ instead of $I^{-1/2}J$. In view of (4.8), we have for all $x\in\mathbb{R}^k$,

$$
\begin{aligned}
\|I_n^{-1/2}J_nx\|&\le\|I^{-1/2}Jx\|+\|(I_n^{-1/2}J_n-I^{-1/2}J)x\|\\
&\le\|I^{-1/2}Jx\|+|I_n^{-1/2}J_n-I^{-1/2}J|_*\|x\|\\
&\le\|I^{-1/2}Jx\|+Cc_n\|x\|\\
&\le\|I^{-1/2}Jx\|+Cc_n\|I^{-1/2}Jx\|\\
&=(1+Cc_n)\|I^{-1/2}Jx\|
\end{aligned}
$$

and similarly,

$$
\|I_n^{-1/2}J_nx\|\ge(1-Cc_n)\|I^{-1/2}Jx\|.
$$

The replacement of $I_n^{-1/2}J_n$ by $I^{-1/2}J$ in (3.4) now immediately follows.

Noting that (3.5) is an immediate consequence of (3.4) completes the proof of Theorem 3.2. $\square$

PROOF OF COROLLARY 3.3. Since

$$\|I^{-1/2}J(\hat{\theta}-\theta_n)\| \geq \lambda_1^{1/2}\|\hat{\theta}-\theta_n\|,$$

where λ_1 is the smallest eigenvalue of $JI^{-1}J$, Theorem 3.2 implies

$$\begin{aligned} &P_n\big(n^{1/2}\|\hat{\theta}-\theta_n\| \geq z_n\big) \\ &\quad \leq P_n\big(n^{1/2}\|I^{-1/2}J(\hat{\theta}-\theta_n)\| \geq \lambda_1^{1/2}z_n\big) \\ &\quad = \exp\left\{-\frac{\lambda_1 z_n^2}{2} + O(c_n z_n^2) + O\left(\frac{z_n^3}{\sqrt{n}}\right) + O(\log z_n)\right\}. \end{aligned} \tag{4.17}$$

Similarly as in the proof of Theorem 3.2 [cf. (4.9)–(4.16)] it is seen that

$$\begin{aligned} &P_n\big((X_1,\ldots,X_n) \in B_n(\delta_0)\cap \widetilde{B}1_n \cap B2_n, n^{1/2}\|\hat{\theta}-\theta_n\| \geq z_n\big) \\ &\quad \geq P_n\left(\left\|n^{-1/2}\sum_{i=1}^n J_n^{-1}I_n^{1/2}Y_{ni}\right\| \geq (1+Cc_n+Cz_nn^{-1/2})z_n\right) \\ &\qquad - P_n\big((X_1,\ldots,X_n)\notin B_n(\delta_0)\cap \widetilde{B}1_n\cap B2_n\big). \end{aligned} \tag{4.18}$$

Let Λ_n be the eigenvector of $I_n^{1/2}J_n^{-2}I_n^{1/2}$ with $\|\Lambda_n\|=1$ corresponding to λ_{n1}^*, where λ_{n1}^* is the largest eigenvalue of $I_n^{1/2}J_n^{-2}I_n^{1/2}$. Then, noting that $\lambda_{n1}^* = \lambda_{n1}^{-1}$ with λ_{n1} the smallest eigenvalue of $J_nI_n^{-1}J_n$, we have

$$n^{-1/2}\left\|\sum_{i=1}^n J_n^{-1}I_n^{1/2}Y_{ni}\right\| \geq \lambda_{n1}^{-1/2}\left|n^{-1/2}\sum_{i=1}^n \Lambda_n^T Y_{ni}\right|. \tag{4.19}$$

Since in view of (4.8), for any $x\neq 0$

$$\left|\frac{x^TJ_nI_n^{-1}J_nx}{\|x\|^2} - \frac{x^TJI^{-1}Jx}{\|x\|^2}\right| \leq Cc_n$$

and

$$\lambda_{n1} = \inf_x \frac{x^TJ_nI_n^{-1}J_nx}{\|x\|^2}, \qquad \lambda_1 = \inf_x \frac{x^TJI^{-1}Jx}{\|x\|^2},$$

it follows that

$$|\lambda_{n1}-\lambda_1| \leq Cc_n.$$

Application of Corollary 2.22 in Book (1976) [cf. also Lemma 4.1(ii) in Jurečková, Kallenberg and Veraverbeke (1988)] yields in combination with (4.18) and (4.19),

$$\begin{aligned} &P_n\big((X_1,\ldots,X_n)\in B_n(\delta_0)\cap\widetilde{B}1_n\cap B2_n, n^{1/2}\|\hat{\theta}-\theta_n\|\geq z_n\big) \\ &\quad \geq P_n\left(\left|n^{-1/2}\sum_{i=1}^n \Lambda_n^TY_{ni}\right| \geq \lambda_{n1}^{1/2}(1+Cc_n+Cz_nn^{-1/2})z_n\right) \end{aligned}$$

$$
\begin{aligned}
&\quad - P_n\big((X_1,\dots,X_n) \notin B_n(\delta_0) \cap \widetilde{B1}_n \cap B2_n\big) \\
&\geq \exp\left\{-\frac{\lambda_{n1} z_n^2}{2} + O(c_n z_n^2) + O\left(\frac{z_n^3}{\sqrt{n}}\right) + O(\log z_n)\right\} \\
&\quad - P_n\big((X_1,\dots,X_n) \notin B_n(\delta_0) \cap \widetilde{B1}_n \cap B2_n\big) \\
&= \exp\left\{-\frac{\lambda_1 z_n^2}{2} + O(c_n z_n^2) + O\left(\frac{z_n^3}{\sqrt{n}}\right) + O(\log z_n)\right\} \\
&\quad - P_n\big((X_1,\dots,X_n) \notin B_n(\delta_0) \cap \widetilde{B1}_n \cap B2_n\big).
\end{aligned}
$$

Noting that

$$
P_n\big((X_1,\dots,X_n) \notin B_n(\delta_0) \cap \widetilde{B1}_n \cap B2_n\big) \leq Ce^{-cn} + 2\exp\{-z_n^2(1+o(1))\},
$$

we obtain

$$
P_n\big(n^{1/2}\|\hat{\theta} - \theta_n\| \geq z_n\big) \geq \exp\left\{-\frac{\lambda_1 z_n^2}{2} + O(c_n z_n^2) + O\left(\frac{z_n^3}{\sqrt{n}}\right) + O(\log z_n)\right\},
$$

which in combination with (4.17) proves the corollary. □

Acknowledgments. We thank the referees for the useful comments, the reference to Jensen and Wood (1998) and the careful reading of the manuscript.

REFERENCES

AERTS, M., CLAESKENS, G. and HART, J. D. (1999). Testing the fit of a parametric function. *J. Amer. Statist. Assoc.* **94** 869–879.

ALMUDEVAR, A., FIELD, C. and ROBINSON, J. (2000). The density of multivariate M-estimates. *Ann. Statist.* **28** 275–297.

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

BAHADUR, R. R., ZABELL, S. L. and GUPTA, J. C. (1980). Large deviations, tests and estimates. In *Asymptotic Theory of Statistical Tests and Estimation. In Honor of Wassily Hoeffding* (I. M. Chakravarti, ed.) 33–64. Academic Press, New York.

BOOK, S. A. (1976). The Cramér–Feller–Petrov large deviation theorem for triangular arrays. Technical report, Dept. Mathematics, California State College, Dominguez Hills.

INGLOT, T. and KALLENBERG, W. C. M. (2001). Moderate deviations of minimum contrast estimators under contamination. Memorandum 1598, Univ. Twente, Faculty of Mathematical Sciences, Enschede, The Netherlands.

INGLOT, T., KALLENBERG, W. C. M. and LEDWINA, T. (1997). Data driven smooth tests for composite hypotheses. *Ann. Statist.* **25** 1222–1250.

INGLOT, T. and LEDWINA, T. (1996). Asymptotic optimality of data driven Neyman's tests for uniformity. *Ann. Statist.* **24** 1982–2019.

INGLOT, T. and LEDWINA, T. (2001a). Asymptotic optimality of data driven smooth tests for location-scale family. *Sankhyā Ser. A* **63** 41–71.

INGLOT, T. and LEDWINA, T. (2001b). Intermediate approach to comparison of some goodness of fit tests. *Ann. Inst. Statist. Math.* **53** 810–834.

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

JUREČKOVÁ, J., KALLENBERG, W. C. M. and VERAVERBEKE, N. (1988). Moderate and Cramér-type large deviation theorems for M-estimators. *Statist. Probab. Lett.* **6** 191–199.

KALLENBERG, W. C. M. (1983). On moderate deviation theory in estimation. *Ann. Statist.* **11** 498–504.

KALLENBERG, W. C. M. (1999). Efficiency, intermediate or Kallenberg. In *Encyclopedia of Statistical Sciences*, Update **3** (S. Kotz, C. B. Read and D. L. Banks, eds.) 192–197. Wiley, New York.

KALLENBERG, W. C. M. and LEDWINA, T. (1997a). Data driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* **92** 1094–1104.

KALLENBERG, W. C. M. and LEDWINA, T. (1997b). Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simulation* **59** 101–121.

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

LEDWINA, T. (1994). Data driven version of Neyman's smooth tests of fit. *J. Amer. Statist. Assoc.* **89** 1000–1005.

MACHADO, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory* **9** 478–493.

PFAFF, T. (1982). Quick consistency of quasi maximum likelihood estimators. *Ann. Statist.* **10** 990–1005.

PITMAN, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. Wiley, New York.

PROKHOROV, A. V. (1973). On sums of random vectors. *Theory Probab. Appl.* **18** 186–188.

RADAVICHYUS, M. È. (1983). On the probability of large deviations of maximum likelihood estimators. *Soviet Math. Dokl.* **27** 127–131.

SIN, C. Y. and WHITE, H. (1996). Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71** 207–225.

SMALL, C. G., WANG, J. and YANG, Z. (2000). Eliminating multiple root problems in estimation (with discussion). *Statist. Sci.* **15** 313–341.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25.

WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Univ. Press.

WOODALL, W. H. and MONTGOMERY, D. C. (1999). Research issues and ideas in statistical process control. *J. Quality Technology* **31** 376–387.

YURINSKII, V. V. (1976). Exponential inequalities for sums of random vectors. *J. Multivariate Anal.* **6** 473–499.

ZACKS, S. (1971). *The Theory of Statistical Inference.* Wiley, New York.

INSTITUTE OF MATHEMATICS
WROCŁAW UNIVERSITY
OF TECHNOLOGY
WYBRZEŻE WYSPIAŃSKIEGO 27
50-370 WROCŁAW
POLAND
AND
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCE
UL. KOPERNIKA 18
51-617 WROCŁAW
POLAND

FACULTY OF ELECTRICAL ENGINEERING
MATHEMATICS AND COMPUTER SCIENCE
UNIVERSITY OF TWENTE
P.O. BOX 217
7500 AE ENSCHEDE
THE NETHERLANDS
E-MAIL: W.C.M.Kallenberg@math.utwente.nl