The Annals of Statistics
1989, Vol. 17, No. 1, 382–400

ASYMPTOTICS WITH INCREASING DIMENSION FOR ROBUST REGRESSION WITH APPLICATIONS TO THE BOOTSTRAP[1]

BY ENNO MAMMEN

Universität Heidelberg

A stochastic expansion for M-estimates in linear models with many parameters is derived under the weak condition $\kappa n^{1/3}(\log n)^{2/3} \to 0$, where n is the sample size and κ the maximal diagonal element of the hat matrix. The expansion is used to study the asymptotic distribution of linear contrasts and the consistency of the bootstrap. In particular, it turns out that bootstrap works in cases where the usual asymptotic approach fails.

1. Introduction. The classical approach of asymptotic statistics is to embed the model being studied in a sequence of models by letting the number n of observations grow to infinity and rescaling the parameters. But there exist examples where this approach is misleading, because features of the model which are important in the finite sample case are lost asymptotically. For instance, for linear models with many parameters Huber (1981) proposes an asymptotic approach where the dimension of the linear model grows with n to infinity. In this article we continue the study of M-estimates in linear models with increasing dimension. But the emphasis will lie on cases where dimension asymptotics leads to other results than the classical approach where the dimension is fixed.

Another example is the leakage effect in time-series analysis which occurs if there is a strong peak in the spectrum of a stationary process. Then data tapering leads to an essential improvement of parametric and nonparametric estimates of the spectral density, as can be seen in simulation studies, but only an asymptotic approach which adjusts the model for each n, making the peak ever stronger, so that its influence is felt asymptotically, can explain the advantages of data tapers [Dahlhaus (1988)]. Other examples are provided by sparse contingency tables. Standard asymptotics do not take into account the sparsity of the table and may therefore lead to inaccurate approximations. Koehler (1986) proposes another approach to the asymptotics of goodness-of-fit statistics testing the overall fit of a log-linear model. This approach is based on an increasing number of categories and leads to approximations which are more accurate for sparse tables, as can be seen in simulation studies. Ehm (1986), generalising results of Haberman (1977a, 1977b), has used a dimension-asymptotics approach to study the behaviour of ML-estimates of log-linear models in sparse and unbalanced tables. Especially he has developed quantities which indicate the accuracy of the usual Gaussian approximation. For log-linear models

Received February 1987; revised May 1988.

[1]Supported by the Deutsche Forschungsgemeinschaft. The work in this article was begun while the author was a guest of the CWI in Amsterdam.

AMS 1980 *subject classifications*. Primary 62E20; secondary 62J05, 62F35.

Key words and phrases. M-estimators, linear model, bootstrap, asymptotic normality, dimension asymptotics.

with closed form ML-estimates Sauermann (1986) shows the superiority of the parametric bootstrap over Gaussian approximations in the case of sparse tables. This is done using dimension asymptotics combined with extensive Monte Carlo simulations. For arbitrary exponential families with increasing number of parameters the asymptotic behaviour of the ML-estimate and of the likelihood ratio test has been studied by Portnoy (1988).

An alternative approach in classical asymptotics is higher-order Edgeworth expansions. But in the examples cited above this approach seems to be of a limited value. This is connected with the empirical fact that Edgeworth expansions usually produce significant improvements only when the Gaussian approximation is not too inaccurate. And clearly the assumptions necessary for Edgeworth expansions are stronger than those for Gaussian approximations. Thus it seems more promising to weaken the assumptions of the classical asymptotic approach in such a way that more features of the finite sample model become asymptotically relevant.

For each n we consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where the X_i's and β are p-dimensional vectors, the Y_i's are the observations and the ε_i's are i.i.d. errors distributed according to a distribution F. The X_i's and p may depend on n.

An M-estimator $\hat{\beta}$ or $(\hat{\beta}, \hat{\sigma})$ is defined by

$$\sum X_i \psi\left(Y_i - X_i'\hat{\beta}\right) = 0 \tag{1.2}$$

or by

$$\sum X_i \psi\left(\left(Y_i - X_i'\hat{\beta}\right)/\hat{\sigma}\right) = 0, \tag{1.3}$$

$$\sum \chi\left(\left(Y_i - X_i'\hat{\beta}\right)/\hat{\sigma}\right) = 0, \tag{1.4}$$

where $\psi, \chi: \mathbb{R} \to \mathbb{R}$ are given functions. For a discussion of robustness properties of M-estimates in linear models we refer to Hampel, Ronchetti, Rousseeuw and Stahel (1986). But especially it should be remarked that several proposed estimates with high breakdown points are M-estimates with redescending ψ-function where the solution of (1.2) or (1.3) and (1.4) is chosen properly [see Rousseeuw and Yohai (1984) and Yohai (1987)].

A overview of articles connected with the asymptotics of M-estimators in linear models with increasing dimension is contained in Portnoy (1984). The most general result is stated in Portnoy (1985). He assumes that the dimension p grows with n in such a way that

$$p^{3/2}(\log n)^{3/2}/n \to 0 \quad \text{as } n \to \infty. \tag{1.5}$$

Then, under some technical conditions on ψ and complicated assumptions about the design, it is shown that certain linear contrasts are approximately Gaussian. The conditions on the design are not fulfilled by ANOVA designs. For instance, if the X_i are i.i.d. random variables the conditions hold in probability under strong assumptions on the tails of the distribution of the X_i.

The aim of this article is not to show only that classical results for M-estimates (for instance asymptotic normality) hold under weaker conditions. Rather we will give a stochastic approximation of the M-estimate which in general is different from the (first- and) higher-order stochastic expansions for fixed dimension. Our expansion for the M-estimate holds uniformly and essentially under the condition that

$$\kappa n^{1/3}(\log n)^{2/3} \to 0, \tag{1.6}$$

where κ is the maximal diagonal element of the hat matrix

$$\begin{aligned}\kappa &= \sup_{1 \le i \le n} X_i' \left(\sum_{l=1}^{n} X_l X_l' \right)^{-1} X_i \\ &= \sup_{1 \le i \le n} \left\| \left(\sum_{l=1}^{n} X_l X_l' \right)^{-1/2} X_i \right\|^2\end{aligned} \tag{1.7}$$

(see Theorems 1 and 2). For the stochastic expansion of the M-estimate no further assumptions on the design are needed. The conditions used are weaker than in Portnoy (1985). There the conditions imply in particular that the design is balanced in a complicated sense and that $\kappa = O(p/n)$. Then (1.6) follows from (1.5).

The stochastic expansion of the M-estimate will be used to show the consistency of the bootstrap estimate of the distribution of the M-estimate and its linear contrasts (see Theorem 5). This is an improvement of a result of Shorack (1982). In particular, the bootstrap works in cases where the usual asymptotic approach fails. As can be seen by the stochastic expansion the M-estimate has bias whose euclidean norm can be of order $\sqrt{\kappa p}$ $(= p/\sqrt{n}$ in the balanced case) which may tend to ∞ under (1.6). Every linear contrast of the M-estimate is asymptotically Gaussian but biased by the projection of this bias (see Theorem 4). In Theorem 3 a criterion is given which indicates when the bias vanishes asymptoticially.

We expect that, under weaker conditions of the type $\kappa n^{\alpha}(\log n)^{\gamma} \to 0$, stochastic expansions of the M-estimate are possible along the lines of this article if the bias is of smaller order than $\sqrt{\kappa p}$. α could then be chosen the nearer to 0 the smaller the order of the bias is. Such a result would have similar consequences for the bootstrap and the asymptotic distribution of linear contrasts as above. But the proof would be very tedious.

2. Results. First we state our main result, the stochastic expansion for the M-estimate.

(2.1) Let $Y_i = X_i'\beta + \varepsilon_i$, $i = 1, \dots, n$, where $\beta \in \mathbb{R}^p$ and ε_i, $i = 1, \dots, n$, are i.i.d. according to a distribution F. Let $\psi: \mathbb{R} \to \mathbb{R}$ be a bounded function with three bounded derivatives fulfilling $E\psi(\varepsilon_i) = 0$ and $d := E\psi'(\varepsilon_i) > 0$. Assume $n \to \infty$ and $\kappa n^{1/3}(\log n)^{2/3} \to 0$. κ is the maximal diagonal element of the hat matrix [see (1.7)]. κ and p and X_i, $i = 1, \dots, n$, may depend on n.

THEOREM 1. *Assume* (2.1). *Then there exists a solution* $\hat{\beta}$ *of*

$$\sum X_i \psi(Y_i - X_i'\beta) = 0$$

such that for $\hat{\theta} = (\sum_{i=1}^n X_i X_i')^{1/2}(\hat{\beta} - \beta)$ *the following expansion holds*:

$$\|\hat{\theta} - \hat{\theta}_1\| = o_P(1), \tag{2.2}$$

where

$$\hat{\theta}_1 = A^{-1}\hat{\theta}_0 - A^{-1}BA^{-1}\hat{\theta}_0 + \tfrac{1}{2}A^{-1}\sum_{i=1}^n \tilde{X}_i\left(\tilde{X}_i'A^{-1}\hat{\theta}_0\right)^2\psi''\left(\varepsilon_i - \tilde{X}_i'a\right) + a, \tag{2.3}$$

$$\tilde{X}_i = \left(\sum_{l=1}^n X_l X_l'\right)^{-1/2} X_i, \qquad i = 1, \dots, n,$$

$$\hat{\theta}_0 = \sum_{i=1}^n \tilde{X}_i \psi\left(\varepsilon_i - \tilde{X}_i'a\right), \tag{2.4}$$

$$A = \sum_{i=1}^n \tilde{X}_i \tilde{X}_i' E\psi'(\varepsilon_i - X_i'a), \tag{2.5}$$

$$B = \sum_{i=1}^n \tilde{X}_i \tilde{X}_i'\left(\psi'\left(\varepsilon_i - \tilde{X}_i'a\right) - E\psi'\left(\varepsilon_i - \tilde{X}_i'a\right)\right). \tag{2.6}$$

Furthermore, *for a fixed constant C the constant a* (*the asymptotic bias*) *is a solution of*

$$\|a\| \le C\sqrt{\kappa p}, \tag{2.7}$$

$$E\hat{\theta}_1 = a. \tag{2.8}$$

If ψ *is monotone the equality* (1.2) *has a unique solution. Otherwise* $\hat{\beta}$ *is uniquely defined on* $\{b: \max_{1 \le i \le n}|X_i'b - X_i'\beta| \le \delta\}$ *with probability tending to* 1 *for* δ *small enough.*

To prove Theorem 1, we will show that $\|\sum \tilde{X}_i \psi(\varepsilon_i - \tilde{X}_i'\hat{\theta}_1)\|$ is small. Then the Newton–Kantorowitsch theorem implies that there exists a solution of (1.2) in the neighborhood of $\hat{\theta}_1$. The main idea of the proof is to use only properties of $\hat{\theta}_1$ and not of $\hat{\theta}$. The proofs of Theorem 1 and of the following theorems are given in the next section.

For p fixed, and $\kappa \to 0$, $\hat{\theta}$ is asymptotically distributed according to $N(0, E\psi^2(\varepsilon_i)/(E\psi'(\varepsilon_i))^2 I_p)$. The normalisation by $(\sum_{i=1}^n X_i X_i')^{1/2}$ corresponds to the classical $\sqrt{n}$-normalization. Only in the case $a = 0$ does the stochastic expansion (2.3) coincide with the usual higher-order expansion for fixed dimension p. But under the assumptions of the theorem it may occur that $\|a\| \to \infty$. In (2.3) the norm of the first term is of order $\sqrt{p}$ and the norm of the second and the third term is of order $\sqrt{\kappa p}$.

As in Huber (1981) and Portnoy (1985) the technical differentiability conditions assumed are not fulfilled by some commonly used ψ-functions. Such smoothness conditions are necessary for a treatment of M-estimates which is based on a stochastic expansion. As indicated by Portnoy (1985) an alternative approach may be based on a study of the joint density of the M-estimate. Then the conditions on ψ could be relaxed by assuming smoothness of the density of the ε_i's.

A theorem analogous to Theorem 1 can be formulated for the case where the scale is estimated simultaneously. Before doing this we state some regularity conditions on χ.

(2.9) Let $\chi\colon \mathbb{R} \to \mathbb{R}$ be a bounded function with three bounded derivatives satisfying $E\chi(\varepsilon_i) = 0$. Furthermore, $x^k\chi^{(3)}(x)$ and $x^k\psi^{(3)}(x)$ are bounded for $k \le 3$ and $E\psi'(\varepsilon_i)\varepsilon_i E\chi'(\varepsilon_i) < E\psi'(\varepsilon_i)E\chi'(\varepsilon_i)\varepsilon_i$.

THEOREM 2. *Assume* (2.1) *and* (2.9). *Then there exists a solution* $(\hat\beta, \hat\sigma)$ *of*

$$\sum X_i\psi((Y_i - X_i'\beta)/\sigma) = 0,$$

$$\sum \chi((Y_i - X_i'\beta)/\sigma) = 0$$

such that for $\tilde\theta = (\sum_{i=1}^n X_iX_i')^{1/2}(\hat\beta - \beta)$ *and* $\hat\gamma = n^{1/2}(\hat\sigma - 1)$ *the following expansions hold*:

$$\|\tilde\theta - \tilde\theta_1\| = o_P(1), \tag{2.10}$$

$$|\hat\gamma - \hat\gamma_1| = o_P(1), \tag{2.11}$$

where

$$\begin{aligned}\begin{pmatrix}\tilde\theta_1\\ \hat\gamma_1\end{pmatrix} &= \tilde A^{-1}\begin{pmatrix}\tilde\theta_0\\ \hat\gamma_0\end{pmatrix} - \tilde A^{-1}\tilde B\tilde A^{-1}\begin{pmatrix}\tilde\theta_0\\ 0\end{pmatrix}\\ &\quad + \frac{1}{2}\tilde A^{-1}\sum_{i=1}^n \left(\tilde X_i'\bar A\tilde\theta_0\right)^2\begin{pmatrix}\tilde X_i\psi''\\ n^{-1/2}\chi''\end{pmatrix}\left(\frac{\varepsilon_i - \tilde X_i'\tilde a}{1 + cn^{-1/2}}\right) + \begin{pmatrix}\tilde a\\ c\end{pmatrix}.\end{aligned} \tag{2.12}$$

$\tilde X_i$ *is defined as in Theorem* 1. *The vector* $\begin{pmatrix}\tilde\theta_0\\ \hat\gamma_0\end{pmatrix}$ *and the* $(p + 1) \times (p + 1)$ *matrices* $\tilde A$ *and* $\tilde B$ *are defined as* $H\begin{pmatrix}\tilde a\\ c\end{pmatrix}$, $EH'\begin{pmatrix}\tilde a\\ c\end{pmatrix}$ *and* $H'\begin{pmatrix}\tilde a\\ c\end{pmatrix} - EH'\begin{pmatrix}\tilde a\\ c\end{pmatrix}$, *where the random function* H *is given by*

$$H\begin{pmatrix}\theta\\ \gamma\end{pmatrix} = \sum_{i=1}^n \begin{pmatrix}\tilde X_i\psi\\ n^{-1/2}\chi\end{pmatrix}\left(\frac{\varepsilon_i - X_i'\theta}{1 + \gamma n^{-1/2}}\right). \tag{2.13}$$

The $p \times p$ *matrix* $\bar A$ *is generated from* $\tilde A^{-1}$ *by removing the last row and the last column. Furthermore, for a fixed constant* C *the constant* $(\tilde a, c)$ (*the asymptotic bias*) *is a solution of*

$$\left\|\begin{pmatrix}\tilde a\\ c\end{pmatrix}\right\| \le C\sqrt{\kappa p}, \tag{2.14}$$

$$E\begin{pmatrix}\tilde\theta_1\\ \hat\gamma_1\end{pmatrix} = \begin{pmatrix}\tilde a\\ c\end{pmatrix}. \tag{2.15}$$

$(\tilde{\beta}, \hat{\sigma})$ is uniquely defined on $\{(b, s): \max_{1 \le i \le n} |X_i'b - X_i'\beta| \le \delta,\ |s - 1| \le \delta\}$ with probability tending to 1 for δ small enough.

The solution of (1.2) [or (1.3) and (1.4)] may not be unique (for instance if ψ is redescending). Then according to Theorem 1 (or 2) there exists asymptotically only one solution which is consistent in the sense that $\max_{1 \le i \le n} |X_i'\beta|$ tends to 0 (and $\hat{\sigma}$ tends to 1). For this solution the expansion (2.2) [or (2.10) and (2.11)] holds. To apply Theorems 1 and 2 to the high breakdown point procedures of Rousseeuw and Yohai (1984) and Yohai (1987), it remains to prove consistency of these estimators.

It is possible to give an explicit asymptotic expression for the bias a. Instead of this we state only a criterion which indicates when the bias vanishes asymptotically.

Theorem 3. *Assume* (2.1) *and* $\|\sum_{i=1}^n \tilde{X}_i \|\tilde{X}_i\|^2\| = o(1)$. *Then* $\|a\| = o(1)$, *where $\tilde{X}_i$ is defined as in Theorem* 1. *Assume additionally* (2.9). *Then* $\|(\tilde{a}, c)\| = o(1)$, *if* $\|n^{-1/2}\sum_{i=1}^n \tilde{X}_i\| = o(1)$.

An example where the bias a is zero is given in the symmetric situation, where ψ is an odd function (and χ is symmetric) and the ε_i's have a symmetric distribution. But Theorem 3 shows that there exist also many asymmetric examples where the bias vanishes (at least asymptotically). In particular, Theorem 3 can be applied to the case of a random design.

Theorems 1 and 2 can be applied to prove asymptotic normality for linear contrasts.

Theorem 4. *Assume* (2.1). *Then for $\alpha_n \in \mathbb{R}^p$ with bounded* $\|\alpha_n\|$,

$$d_2\big(\mathscr{L}(\alpha_n'\hat{\theta}),\ N\big((\alpha_n' a),\ \|\alpha_n\|^2 E\psi^2(\varepsilon_i)/E^2\psi'(\varepsilon_i)\big)\big) \to 0. \tag{2.16}$$

Assume additionally (2.9). *Then*

$$d_2\left(\mathscr{L}\begin{pmatrix}\alpha_n'\tilde{\theta}\\ \hat{\gamma}\end{pmatrix},\ N\left(\begin{pmatrix}\alpha_n'\tilde{a}\\ c\end{pmatrix},\ \begin{pmatrix}\alpha_n' & 0\\ 0 & 1\end{pmatrix}\Sigma\begin{pmatrix}\alpha_n' & 0\\ 0 & 1\end{pmatrix}'\right)\right) \to 0, \tag{2.17}$$

where $\Sigma = (EH'(0))^{-1}EH(0)H(0)'(EH'(0))^{-1}$ and where for two probability measures μ, ν the quantity $d_2(\mu, \nu)$ denotes the modification of the Mallows distance,

$$d_2(\mu, \nu) = \inf\Big\{\big(E(\|U - V\|^2 \wedge 1)\big)^{1/2}: \mathscr{L}(U) = \mu,\ \mathscr{L}(V) = \nu\Big\}.$$

Another possible application of Theorems 1 and 2 is the study of bootstrap. To estimate the distribution of $\hat{\beta}$, the bootstrap proceeds as follows. Define the residuals $\hat{\varepsilon}_i := Y_i - X_i'\hat{\beta}$, $i = 1, \ldots, n$, and let F_n denote the empirical distribution function of the $\hat{\varepsilon}_i$. Now generate n i.i.d. observation $\varepsilon_1^*, \ldots, \varepsilon_n^*$ distributed according to F_n. Define $Y_i^* = X_i'\hat{\beta} + \varepsilon_i^*$, $i = 1, \ldots, n$, and $\hat{\beta}^*$ as the ("consistent") solution of

$$\sum_{i=1}^{n} X_i\left(\psi(Y_i^* - X_i'\hat{\beta}^*) - \int \psi \, dF_n\right) = 0.$$

By Monte Carlo replications of this procedure the distribution of $\hat{\beta}^* - \hat{\beta}$ (given F_n) can be determined. This is the bootstrap estimate of the distribution of $\hat{\beta} - \beta$. Similarly a bootstrap estimate of the distribution of $(\tilde{\beta} - \beta, \hat{\sigma} - \sigma)$ can be defined. It has been proved by Shorack (1982) that the bootstrap estimates the distribution of linear contrasts consistently if $\kappa p^2 \to 0$ (i.e., in the balanced case: $p^3/n \to 0$). The special case of the least-squares estimate (i.e., $\psi \equiv id$) has been considered by Bickel and Freedman (1983). They show that the bootstrap estimate is consistent in estimating the distribution of $\hat{\beta} - \beta$ if $p^2/n \to 0$ and in estimating the distribution of linear contrasts of $\hat{\beta} - \beta$ if $p/n \to 0$. We state now a generalisation of Shorack's result. Put $\hat{\theta}^* = (\sum_{i=1}^n X_i X_i')^{1/2}(\hat{\beta}^* - \hat{\beta})$ and define $\tilde{\theta}^*$ and $\hat{\gamma}^*$ analogously.

THEOREM 5. *Assume* (2.1). *Then for* $\alpha_n \in \mathbb{R}^p$ *with* $\|\alpha_n\|$ *bounded*

$$d_2(\alpha_n'\hat{\theta}, \alpha_n'\hat{\theta}^*) \to 0 \quad \textit{in probability}, \tag{2.18}$$

$$d_2(\hat{\theta}, \hat{\theta}^*) \to 0 \quad \textit{in probability, if additionally } p^2/n \to 0. \tag{2.19}$$

Assume additionally (2.9). *Then*

$$d_2\big((\alpha_n'\tilde{\theta}, \hat{\gamma}), (\alpha_n'\hat{\theta}^*, \hat{\gamma}^*)\big) \to 0 \quad \textit{in probability}, \tag{2.20}$$

$$d_2\big((\tilde{\theta}, \hat{\gamma}), (\tilde{\theta}^*, \hat{\gamma}^*)\big) \to 0 \quad \textit{in probability, if additionally } p^2/n \to 0. \tag{2.21}$$

Here for two random variables X, Y *the quantity* $d_2(X, Y)$ *denotes the modified Mallows distance between the distributions of* X *and* Y, *respectively* (*see Theorem* 4),

$$d_2(X, Y) = d_2(\mathscr{L}(X), \mathscr{L}(Y)).$$

In particular Theorem 5 indicates that bootstrap works in cases where the usual asymptotic approach fails. The bootstrap estimates the bias of a linear contrast consistently—also in cases where the bias tends to ∞. If design points are leverage points it can happen that $p^2/n \to 0$ but that κp does not converge to 0. Then the distribution of $\hat{\theta}$ cannot be approximated by a Gaussian distribution (measured by the Mallows distance d_2). But nevertheless the bootstrap estimate of the distribution of $\hat{\theta}$ is consistent. The proof of Theorem 5 is based on the fact that a stochastic expansion analogous to (2.3) also holds for $\hat{\theta}^*$. A similar approach has been used by Sauermann (1986) in the case of large sparse contingency tables. Also there the bootstrap estimates not only the distribution of the first-order linear term of the stochastic expansion of an estimator but also the distribution of the higher-order terms. This phenomenon has also been recognized in studies of the bootstrap which are based on higher-order Edgeworth expansions [see, for instance, Beran (1984)].

3. Proofs.

PROOF OF THEOREM 1. Without loss of generality we assume

$$\sum_{i=1}^{n} X_i X_i' = I_p. \tag{3.1}$$

Then $\tilde{X}_i = X_i$, $i = 1, \dots, n$, and $\kappa = \sup_{1 \le i \le n} \|X_i\|^2$ [see (1.7)] and $\sum_{i=1}^{n} \|X_i\|^2 = p$. First, using Brouwer's fixed point theorem [Dunford and Schwartz (1958)], we determine the order of the bias term a.

LEMMA 1. *Assume* (2.1) *and* (3.1). *Then for C large enough there exists for every n a solution a of* (2.7) *and* (2.8): $\|a\| \le C\sqrt{\kappa p}$, $E\hat{\theta}_1 = a$.

PROOF. Define $f\colon \mathbb{R}^p \to \mathbb{R}^p$ exactly as $E(A(\hat{\theta}_1 - a))$ [see (2.3)–(2.6)], but as a function of a. Then

$$\begin{aligned} f(b) = {} & \sum_{i=1}^{n} X_i E\psi(\varepsilon_i - X_i'b) \\ & - \sum_{i=1}^{n} X_i \big(X_i' A(b)^{-1} X_i\big)\{E\psi(\varepsilon_i - X_i'b)\psi'(\varepsilon_i - X_i'b) \\ & \qquad - E\psi(\varepsilon_i - X_i'b) E\psi'(\varepsilon_i - X_i'b)\} \\ & + \tfrac{1}{2} \sum_{i,k,l=1}^{n} X_i \big(X_i' A(b)^{-1} X_l\big)\big(X_i' A(b)^{-1} X_k\big) \\ & \qquad \times E\psi(\varepsilon_l - X_l'b)\psi(\varepsilon_k - X_k'b)\psi''(\varepsilon_i - X_i'b), \end{aligned} \tag{3.2}$$

where $A(b) := \sum_{i=1}^{n} X_i X_i' E\psi'(\varepsilon_i - X_i'b)$. Put $g(b) := (1/d)f(b) + b$. Then $f(a) = 0$ and $g(a) = a$.

We will show, that—if C is chosen large enough—for all n the inequality $\|b\| \le C\sqrt{\kappa p}$ implies $\|g(b)\| \le C\sqrt{\kappa p}$. Then with Brouwer's fixed-point theorem the statement of the lemma will follow.

Now assume

$$\|b\| \le C\sqrt{\kappa p}. \tag{3.3}$$

Then for θ with $\|\theta\| \le 1$ [note (3.1)]

$$\begin{aligned} & \left| \frac{1}{d} \sum_{i=1}^{n} \theta' X_i E\psi(\varepsilon_i - X_i'b) + \theta'b \right| \\ & \quad \le \text{const.} \sum_{i=1}^{n} |X_i'\theta| (X_i'b)^2 \\ & \quad \le \text{const.} \sup_{1 \le i \le n} |X_i'\theta| \sum_{i=1}^{n} (X_i'b)^2 \\ & \quad \le \text{const.} \sqrt{\kappa}\,\kappa p = o\big(\sqrt{\kappa p}\big), \end{aligned} \tag{3.4}$$

where we have used boundedness of ψ'' and $\sqrt{\kappa^2 p} \le \sqrt{\kappa^3 n} = o(1)$.

Furthermore, using (3.3) by tedious but straightforward calculations, it can be shown that the second and third terms in (3.2) are of the form const. $\sum_{i=1}^n X_i \|A(b)^{-1/2}X_i\|^2 + e_1$ and const. $\sum_{i=1}^n X_i\|A(b)^{-1}X_i\|^2 + e_2$, respectively, where $\|e_1\| = o(\sqrt{\kappa p})$ and $\|e_2\| = o(\sqrt{\kappa p})$. Now $\|g(b)\| \le$ const. $\sqrt{\kappa p}$ is implied by

$$\left\| \sum_{i=1}^n X_i \|DX_i\|^2 \right\| \le \text{const.} \sqrt{\kappa p}, \tag{3.5}$$

where $D = A(b)^{-1/2}$ or $A(b)^{-1}$. □

The proof of the expansion (2.2) is based on an application of a Newton–Kantorowitsch theorem. We cite a version which will be used here [Kantorowitsch and Akilow (1964)].

THEOREM (Newton–Kantorowitsch theorem). *Assume that a function* G: $\mathbb{R}^p \to \mathbb{R}^p$ *has two continuous derivatives for* $\|x - x_0\| \le r$. *Furthermore, assume*

$$\Gamma := (G'(x_0))^{-1} \text{ exists}, \tag{3.6}$$

$$\|\Gamma G(x_0)\| \le \eta, \tag{3.7}$$

$$\|\Gamma G''(x)\| \le \lambda \quad \text{for } \|x - x_0\| \le r, \tag{3.8}$$

$$h = \lambda\eta \le 1/2, \tag{3.9}$$

$$r_0 := \frac{1 - \sqrt{1 - 2h}}{h}\eta \le r. \tag{3.10}$$

Then the equation $G(x) = 0$ *has a solution* x^* *with*

$$\|x^* - x_0\| \le r_0.$$

This theorem will be applied as follows. We set $G(\theta) = \sum_{i=1}^n X_i \psi(\varepsilon_i - X_i'\theta)$. We will prove (3.6)–(3.8) for $\eta = O_P(\kappa n^{1/3}(\log n)^{2/3})$ and $\lambda = O_P(\sqrt{\kappa})$ and $x_0 = \hat{\theta}_1$. Then (3.9) and (3.10) follow automatically with $r_0 = O_P(\kappa n^{1/3}(\log n)^{2/3})$. This proves (2.2).

1. STEP: PROOF OF (3.6). We apply the following two lemmas.

LEMMA 2. *Assume* (2.1) *and* (3.1). *Then*

$$\sup_{1 \le i \le n} |X_i'(\hat{\theta}_1 - a)| = O_P(\sqrt{\kappa}\sqrt{\log n}). \tag{3.11}$$

LEMMA 3. *Assume* (2.1) *and* (3.1). *Then*

$$\left\| \sum X_i X_i'(\psi'(\varepsilon_i) - d) \right\| = o_P(1). \tag{3.12}$$

APPLICATION OF LEMMAS 2 AND 3. For (3.6) it suffices to prove

$$\|G'(\hat{\theta}_1) - dI_p\| = o_P(1).$$

But this can be seen as follows:

$$\|G'(\hat{\theta}_1) - dI_p\| = \sup_{\|\theta\| \le 1} \left| \sum_{i=1}^{n} (X_i'\theta)^2 \big(\psi'(\varepsilon_i - X_i'\hat{\theta}_1) - d\big) \right|$$

$$\le \sup_{\|\theta\| \le 1} \left| \sum_{i=1}^{n} (X_i'\theta)^2 \big(\psi'(\varepsilon_i) - d\big) \right|$$

$$+ \text{const.} \sup_{\|\theta\| \le 1} \left| \sum_{i=1}^{n} (X_i'\theta)^2 (X_i'\hat{\theta}_1) \right|$$

$$= o_P(1) + \text{const.} \sup_{1 \le i \le n} |X_i'\hat{\theta}_1|.$$

This is of order $o_P(1)$ according to Lemmas 1 and 2.

PROOF OF LEMMA 2. The proof of (3.11) is divided into three steps. We will show

$$\sup_{1 \le i \le n} \left| X_i'\big(A^{-1}(\hat{\theta}_0 - E\hat{\theta}_0)\big) \right| = O_P\big(\sqrt{\kappa}\sqrt{\log n}\,\big), \tag{3.13}$$

$$\sup_{1 \le i \le n} \left| X_i'\big(A^{-1}BA^{-1}\hat{\theta}_0 - E\big(A^{-1}BA^{-1}\hat{\theta}_0\big)\big) \right| = o_P(\sqrt{\kappa}\,), \tag{3.14}$$

$$\begin{aligned} \sup_{1 \le i \le n} \Bigg| X_i' \sum_{l=1}^{n} & X_l\big(X_l'A^{-1}\hat{\theta}_0\big)^2 \psi''(\varepsilon_l - X_l'a) \\ & - E\Big(X_l\big(X_l'A^{-1}\hat{\theta}_0\big)^2 \psi''(\varepsilon_l - X_l'a)\Big) \Bigg| = o_P(\sqrt{\kappa}\,). \end{aligned} \tag{3.15}$$

PROOF OF (3.13) [Compare Lemma 3.3 in Portnoy (1985)]. Put $\overline{X}_l := A^{-1}X_l$ and $d_l = E\psi(\varepsilon_l - X_l'a)$. Then for $1 \le i \le n$ for $C \ge 0$ and for t with $t\kappa$ small enough

$$P\big(X_i'A^{-1}(\hat{\theta}_0 - E\hat{\theta}_0) \ge C\sqrt{\kappa}\sqrt{\log n}\,\big)$$

$$\le E\exp\left(t \sum_{\substack{1 \le l \le n \\ 1 \le j \le p}} \overline{X}_{ij} X_{lj}\big(\psi(\varepsilon_l - X_l'a) - d_l\big) - tC\sqrt{\kappa}\sqrt{\log n} \right)$$

$$\le \exp\big(-tC\sqrt{\kappa}\sqrt{\log n}\,\big) \prod_{1 \le l \le n} \left(1 + \text{const.}\, t^2 \left(\sum_{j=1}^{p} \overline{X}_{ij} X_{lj} \right)^2 \right)$$

(Here it has been used that ψ is bounded and that $|t\overline{X}_i'X_l| \le |t\kappa|$const. is small enough.)

$$\le \exp\big(\text{const.}\, t^2\kappa - tC\sqrt{\kappa}\sqrt{\log n}\,\big)$$

[because of $\sum_{l=1}^{n}(\overline{X}_i'X_l)^2 = \|\overline{X}_i\|^2 \le \text{const.}\|X_i\|^2 \le \text{const.}\,\kappa$]

$$= \exp(\log n(\text{const.} - C)),$$

if t is chosen as $t = \sqrt{\log n} / \sqrt{\kappa}$. Here const. does not depend on C or i. This proves (3.13). □

PROOF OF (3.14). Instead of (3.14) we show

$$\sup_{1 \le i \le n} \left| \sum_{1 \le k, l \le n} a_{kl}^i (V_k W_l - E V_k W_l) \right| = O_P(\kappa \log n), \tag{3.16}$$

$$\sup_{1 \le i \le n} \left| \sum_{1 \le k, l \le n} a_{kl}^i V_k E \psi(\varepsilon_l - X_l' a) \right| = O_P\left(\sqrt{\kappa^3 p} \sqrt{\log n}\right), \tag{3.17}$$

where

$$a_{kl}^i := \left(X_i' A^{-1} X_k\right)\left(X_k' A^{-1} X_l\right), \tag{3.18}$$

$$V_k := \psi'(\varepsilon_k - X_k' a) - E\psi'(\varepsilon_k - X_k' a), \tag{3.19}$$

$$W_l := \psi(\varepsilon_l - X_l' a) - E\psi(\varepsilon_l - X_l' a). \tag{3.20}$$

Clearly, (3.14) is implied by (3.16) and (3.17). The terms in (3.14) are just the sum of the terms in (3.16) and (3.17).

The proof of (3.17) goes along the same lines as the proof of (3.13). Additionally one has to check that

$$\sup_{1 \le i \le n} \sum_{k=1}^{n} \left(\sum_{l=1}^{n} a_{kl}^i E\psi(\varepsilon_l - X_l' a) \right)^2 = O(\kappa^3 p). \tag{3.21}$$

The proof of (3.16) can also be based on an application of the Markov inequality [as can the proofs of (3.13) and (3.17)]. But here the difficulty arises that one has to bound the Laplace transform of a quadratic form in independent variables. To do this we expand the Laplace transform in a series and use the following bound for the moments of a quadratic form.

LEMMA 4. *Assume that $V_1, \dots, V_n, W_1, \dots, W_n$ are independent variables with mean value 0. Then for $Z = \sum_{i,j=1}^n a_{ij} V_i W_j$ for $k \ge 2$*

$$E|Z - E(Z)|^k \le 2^{3k} C(k) \sqrt{C(2k)} \left(\sum_{i,j=1}^{n} a_{ij}^2 \left(E V_i^{2k}\right)^{1/k} \left(E W_j^{2k}\right)^{1/k} \right)^{k/2}, \tag{3.22}$$

where $C(k) := (2^{k/2} / \sqrt{\pi}) \Gamma((k+1)/2)$.

Lemma 4 is a slight modification of a result of Whittle (1960). The proof is essentially the same. Before applying (3.22) observe that $(EV_i^{2k})^{1/k}$ and $(EW_j^{2k})^{1/k}$ are bounded (say by b_0) because ψ and ψ' are assumed to be bounded. Furthermore, for a constant b_1 one gets with Stirling's formula

$$\frac{C(k)\sqrt{C(2k)}}{k!} \le b_1^k. \tag{3.23}$$

Finally, for a constant b_2,

$$\begin{aligned}\sum_{k,\,l=1}^{n}\left(a_{kl}^{i}\right)^2 &= \sum_{k,\,l=1}^{n}\left(X_i'A^{-1}X_k\right)^2\left(X_k'A^{-1}X_l\right)^2\\ &\le \text{const.}\,\kappa\sum_{k=1}^{n}\left(X_i'A^{-1}X_k\right)^2 \qquad (3.24)\\ &\le b_2\kappa^2.\end{aligned}$$

Now, applying (3.22) for $Z = Z^i = \sum_{k,\,l=1}^{n} a_{kl}^i V_k W_l$ for κt small enough, one gets

$$\begin{aligned}E\exp\left(t\left(Z^i - E(Z^i)\right)\right) &\le 1 + \sum_{k\ge 2} 2^{3k}\frac{t^k}{k!}C(k)\sqrt{C(2k)}\,b_0^k\left(b_2\kappa^2\right)^{k/2}\\ &\le 1 + \sum_{k\ge 2}\left(2^3 b_1 b_0 b_2^{1/2}\right)^k(\kappa t)^k \qquad (3.25)\\ &\le 1 + \text{const.}\,\kappa^2 t^2.\end{aligned}$$

Now proceeding similarly as in the proof of (3.13) proves (3.16). □

PROOF OF (3.15). To finish the proof of Lemma 2, we have to show (3.15). This can be done by first noting that

$$E\|D - E(D) - (\tilde{D} - E\tilde{D})\|^2 = o_P(1), \qquad (3.26)$$

where D is the third term of the expansion of $\hat{\theta}$, that is,

$$D := A^{-1}\sum_{k=1}^{n} X_k\left(X_k'A^{-1}\hat{\theta}_0\right)^2\psi''(\varepsilon_k - X_k'a)$$

and $\tilde{D}$ is the quadratic form

$$\tilde{D} := \sum_{k\ne l\ne m(\ne k)} A^{-1}X_k\left(X_k'A^{-1}X_l\right)\left(X_k'A^{-1}X_m\right)E\left(\psi''(\varepsilon_k - X_k'a)\right)W_l W_m$$

[W_l is defined in (3.20)]. (3.26) implies that for the proof of (3.15) one can treat the quadratic form $X_i'\tilde{D}$ instead of the cubic form $X_i'D$. But this can be done as in the proof of (3.14). The proof of (3.26) is straightforward but lengthy and consists essentially only of applications of the Cauchy–Schwarz inequality. □

PROOF OF LEMMA 3. We want to show that

$$\|\tilde{B}\|^2 = \lambda_{\max} = o_P(1), \qquad (3.27)$$

where $\lambda_{\max}$ is the maximal eigenvalue of $\tilde{B}^2$ and the random matrix $\tilde{B}$ is defined by

$$\tilde{B} = \sum_{i=1}^{n} X_i X_i'\left(\psi'(\varepsilon_i) - d\right).$$

If τ_1 is the (random) normed eigenvector of $\tilde{B}$ to the eigenvalue $\sqrt{\lambda_{\max}}$, then one has $\lambda_{\max} = \|\tilde{B}\tau_1\|^2$. The idea of this proof is based on an approximation procedure of τ_1. Define $e := \|\tilde{B}\tilde{B}T\|^2 = T'\tilde{B}^4T$, where T is a Gaussian random variable distributed according to $N(0, I_p/p)$ and independent of $\tilde{B}$.

Now

$$e \geq \|\tilde{B}\|^4(T'\tau_1)^2 = \lambda_{\max}^2(T'\tau_1)^2.$$

Therefore

$$\lambda_{\max}^2 \leq e/(T'\tau_1)^2 = O_P(p)e.$$

Furthermore, the proof of (3.27) is finished by

$$e = O_P\left(\kappa^2 + \sqrt{\kappa^3/p}\right). \tag{3.28}$$

The proof of (3.28) is straightforward but very lengthy (e is a form of order 6 in independently but not identically distributed variables). □

We continue with the second condition (3.7) of the Newton–Kantorowitsch theorem.

2. Step: Proof of (3.7). (3.7) follows from Lemma 5.

Lemma 5. *Assume* (2.1) *and* (3.1). *Then*

$$\|G(\hat{\theta}_1)\| = \left\| \sum_{i=1}^n X_i\psi(\varepsilon_i - X_i'\hat{\theta}_1)\right\| = O_P\left(\kappa n^{1/3}(\log n)^{2/3}\right).$$

Proof. A Taylor expansion gives

$$G(\hat{\theta}_1) = \sum_{i=1}^n X_i\psi(\varepsilon_i - X_i'\hat{\theta}_1) = T_1 + T_2 + T_3 \tag{3.29}$$

with

$$T_1 = \sum_{i=1}^n X_i\psi(\varepsilon_i - X_i'a) - \sum_{i=1}^n X_iX_i'(\hat{\theta}_1 - a)\psi'(\varepsilon_i - X_i'a),$$

$$T_2 = \tfrac{1}{2}\sum_{i=1}^n X_i\left(X_i'(\hat{\theta}_1 - a)\right)^2\psi''(\varepsilon_i - X_i'a),$$

$$T_3 = -\tfrac{1}{6}\sum_{i=1}^n X_i\left(X_i'(\hat{\theta}_1 - a)\right)^3\psi'''(\tilde{\varepsilon}_i),$$

where $\tilde{\varepsilon}_i$ lies between ε_i and $\varepsilon_i - X_i'a$. First we treat T_1.

With

$$U := \sum_{i=1}^n X_i\left(X_i'A^{-1}\hat{\theta}_0\right)^2\psi''(\varepsilon_i - X_i'a),$$

we get [see (2.3)–(2.6)]

$$T_1 + \tfrac{1}{2}U = \hat{\theta}_0 - (A + B)\left(A^{-1}\hat{\theta}_0 - A^{-1}BA^{-1}\hat{\theta}_0 + \tfrac{1}{2}A^{-1}U\right) + \tfrac{1}{2}U = S_1 + S_2, \tag{3.30}$$

where $S_1 := BA^{-1}BA^{-1}\hat{\theta}_0$ and $S_2 := -\tfrac{1}{2}BA^{-1}U$. By straightforward calculations one gets

$$\|S_1\| = O_P(\kappa^{3/2}\sqrt{n}), \tag{3.31}$$

$$\|S_2\| = O_P(\kappa^{3/2}\sqrt{n}). \tag{3.32}$$

For T_2 one gets

$$T_2 - \tfrac{1}{2}U = \tfrac{1}{2}\sum_{i=1}^{n} X_i\psi''(\varepsilon_i - X_i'a)\left(\left(X_i'(\hat{\theta}_1 - a)\right)^2 - \left(X_i'A^{-1}\hat{\theta}_0\right)^2\right), \tag{3.33}$$

$$\begin{aligned}\|T_2 - \tfrac{1}{2}U\| &= \sup_{\|\theta\|\le 1} \tfrac{1}{2}\sum_{i=1}^{n} (X_i'\theta)\psi''(\varepsilon_i - X_i'a)\left(X_i'(\hat{\theta}_1 - a - A^{-1}\hat{\theta}_0)\right) \\ &\qquad \times\left(X_i'(\hat{\theta}_1 - a + A^{-1}\hat{\theta}_0)\right) \\ &\le \text{const.} \sup_{\|\theta\|\le 1}\sqrt{\sum_{i=1}^{n}(X_i'\theta)^2\sum_{i=1}^{n}\left(X_i'(\hat{\theta}_1 - a - A^{-1}\hat{\theta}_0)\right)^2} \\ &\qquad \times \sup_{1\le i\le n}\left|X_i'(\hat{\theta}_1 - a + A^{-1}\hat{\theta}_0)\right| \\ &= \text{const.}\|\hat{\theta}_1 - a - A^{-1}\hat{\theta}_0\| \sup_{1\le i\le n}\left|X_i'(\hat{\theta}_1 - a + A^{-1}\hat{\theta}_0)\right| \\ &= O_P\left(\sqrt{\kappa^2 n}\sqrt{\kappa}\sqrt{\log n}\right)\end{aligned} \tag{3.34}$$

because of Lemma 2 and (3.1) and

$$\|\hat{\theta}_1 - a - A^{-1}\hat{\theta}_0\| = O_P\left(\sqrt{\kappa^2 n}\right).$$

This can be proved by straightforward calculations.

Finally, we give a bound for $\|T_3\|$,

$$\begin{aligned}\|T_3\| &= \sup_{\|\theta\|\le 1} \tfrac{1}{6}\sum_{i=1}^{n}(X_i'\theta)\left(X_i'(\hat{\theta}_1 - a)\right)^3\psi'''(\tilde{\varepsilon}_i) \\ &\le \sup_{\|\theta\|\le 1} \text{const.}\sqrt{\sum_{i=1}^{n}(X_i'\theta)^2\sum_{i=1}^{n}\left(X_i'(\hat{\theta}_1 - a)\right)^6} \\ &\le O_P\left(\left(p\left(\sqrt{\kappa}\sqrt{\log n}\right)^4\right)^{1/2}\right) \\ &= O_P\left(\sqrt{p}\,\kappa(\log n)\right),\end{aligned} \tag{3.35}$$

where Lemma 2 and $\|\hat{\theta}_1 - a\| = O_P(\sqrt{p})$ have been used.

Now combining (3.30)–(3.35), one gets

$$\begin{aligned}\|T_1 + T_2 + T_3\| &\le \|S_1\| + \|S_2\| + \|T_2 - \tfrac{1}{2}U\| + \|T_3\| \\ &= O_P\Big(\sqrt{\kappa^3 n}\sqrt{\log n}\Big) + O_P\big(\sqrt{p}\,\kappa(\log n)\big) \\ &= O_P\Big(\kappa n^{1/3}(\log)^{2/3}\Big)\end{aligned}$$

because of $\sqrt{p} = o(n^{1/3}/(\log n)^{1/3})$, which is implied by

$$\frac{p}{n^{2/3}}(\log n)^{2/3} \le \kappa n^{1/3}(\log n)^{2/3} \to 0. \qquad \square$$

For the proof of Theorem 1 it remains to prove (3.8).

3. STEP: PROOF OF (3.8).

LEMMA 6. *Assume* (2.1) *and* (3.1). *Then* $\|G''(\theta)\| = O_P(\sqrt{\kappa})$.

PROOF.

$$\begin{aligned}\|G''(\theta)\| &= \sup_{\substack{\|\tau_1\|\le 1\\ \|\tau_2\|\le 1\\ \|\tau_3\|\le 1}} \sum_{i=1}^{n} (X_i'\tau_1)(X_i'\tau_2)(X_i'\tau_3)\psi''(\varepsilon_i - X_i'\theta) \\ &\le \text{const.} \sup_{\tau_1,\tau_2,\tau_3} \sqrt{\sum_{i=1}^{n} (X_i'\tau_1)^2(X_i'\tau_2)^2 \sum_{i=1}^{n} (X_i'\tau_3)^2} \\ &\le \text{const.} \sup_{\|\tau_1\|\le 1} \sup_{1\le i\le n} |X_i'\tau_1| = O(\sqrt{\kappa}). \qquad \square\end{aligned}$$

PROOF OF THEOREM 2. The proof goes along the lines of the proof of Theorem 1. The Newton–Kantorowitsch theorem has to be applied to the function $H(\theta, \gamma)$. □

PROOF OF THEOREM 3. Put for a constant C,

$$C_n := C\left\|\sum_{i=1}^{n} X_i\|X_i\|^2\right\|.$$

Define f and g as in Lemma 1. Then one can show that if C is chosen large enough

$$\|b\| \le C_n \quad \text{implies} \quad \|g(b)\| \le C_n.$$

Therefore the first statement of Theorem 3 follows from Brouwer's fixed-point theorem. The second statement follows similarly. □

PROOF OF THEOREM 4. We will indicate only the proof of (2.16). Assume $\|\alpha_n\| = 1$. According to Theorem 1, it suffices to show

$$d\alpha_n' A^{-1}(\hat{\theta}_0 - E\hat{\theta}_0) - \alpha_n'(\hat{\theta}_0 - E\hat{\theta}_0) = o_P(1), \tag{3.36}$$

$$\alpha_n' A^{-1} B A^{-1} \hat{\theta}_0 - E\alpha_n' A^{-1} B A^{-1} \hat{\theta}_0 = o_P(1), \tag{3.37}$$

$$\begin{aligned} &\alpha_n' A^{-1} \sum_{i=1}^n X_i \big(X_i' A^{-1} \hat{\theta}_0\big)^2 \psi''(\varepsilon_i - X_i' a) \\ &\quad - E\alpha_n' A^{-1} \sum_{i=1}^n X_i \big(X_i' A^{-1} \hat{\theta}_0\big)^2 \psi''(\varepsilon_i - X_i' a) = o_P(1). \end{aligned} \tag{3.38}$$

(3.36) follows from $\|A - dI\| = o(1)$.

The proof of (3.37) and (3.38) is straightforward, but lengthy. One has to calculate the variance of the left-hand side terms. For the treatment of (3.38) (3.26) can be used. □

PROOF OF THEOREM 5. We will give only the proof of (2.18) and (2.19). Without loss of generality assume (3.1). Note that with probability tending to 1 given F_n (the distribution of the residuals) the following holds, as can be proved along the lines of the proof of Theorem 1:

$$\|\hat{\theta}^* - \hat{\theta}_1^*\| = o_{F_n}(1), \tag{3.39}$$

where

$$\begin{aligned} \hat{\theta}_1^* &= \hat{A}^{-1}\hat{\theta}_0^* - \hat{A}^{-1}\hat{B}\hat{A}^{-1}\hat{\theta}_0^* \\ &\quad + \tfrac{1}{2}\hat{A}^{-1} \sum_{i=1}^n X_i \big(X_i \hat{A}^{-1} \hat{\theta}_0^*\big)^2 \psi''(\varepsilon_i^* - X_i'\hat{a}) + \hat{a}. \end{aligned} \tag{3.40}$$

Furthermore,

$$\hat{\theta}_0^* = \sum_{i=1}^n X_i \left(\psi(\varepsilon_i^* - X_i'\hat{a}) - \int \psi \, dF_n \right), \tag{3.41}$$

$$\hat{A} = \sum_{i=1}^n X_i X_i' E_{F_n} \psi'(\varepsilon_i^* - X_i'\hat{a}), \tag{3.42}$$

$$\hat{B} = \sum_{i=1}^n X_i X_i' \big(\psi'(\varepsilon_i^* - X_i'\hat{a}) - E_{F_n}\psi'(\varepsilon_i^* - X_i'\hat{a})\big) \tag{3.43}$$

and $\hat{a}$ is chosen such that

$$\|\hat{a}\| \le C\sqrt{\kappa p} \quad \text{and} \quad E_{F_n}\hat{\theta}_i^* = \hat{a}.$$

PROOF OF (2.18). With the arguments given in the proof of Theorem 4 one sees that for (2.18) it suffices to show

$$\|\hat{a} - a\| = O_P\left(\sqrt{\kappa} \sqrt{\frac{p^2}{n}} \right) = o_P(1). \tag{3.44}$$

To prove (3.44), one shows

$$\| f(\hat{a}) - f(a) \| = O_P\left(\sqrt{\kappa}\sqrt{\frac{p^2}{n}}\right), \tag{3.45}$$

$$\| f'(b) - dI \| = o(1) \tag{3.46}$$

uniformly in $\|b\| \le \text{const.}\sqrt{\kappa p}$ (f is defined in Lemma 1). □

PROOF OF (2.19). To prove (2.19), we will show that the Mallows distance between the distributions of the corresponding terms in (3.40) and (2.3), respectively, converges in probability to 0.

Assume $R_1, \dots, R_n$ are i.i.d. according to F and $R_1^*, \dots, R_n^*$ are i.i.d. according to F_n with

$$\delta := E(R_i - R_i^*)^2 = \tilde{d}_2^2(F, F_n) \ge d_2^2(F, F_n),$$

where $\tilde{d}_2$ denotes the Mallows distance,

$$\tilde{d}_2(\mu, \nu) = \inf\left\{ \left(E\|U - V\|^2\right)^{1/2} : \mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu \right\}.$$

Then given F_n [see Bickel and Freedman (1981)]

$$\begin{aligned} & d_2^2\left(\hat{\theta}_0 - E\hat{\theta}_0, \hat{\theta}_0^* - E_{F_n}\hat{\theta}_0^*\right) \\ & \le \sum_{i=1}^n \|X_i\|^2 E\big(\psi(R_i - X_i'a) - E\psi(R_i - X_i'a) - \psi(R_i^* - X_i'\hat{a}) \\ & \qquad + E\psi(R_i^* - X_i'\hat{a})\big)^2 \\ & \le \text{const.} \sum_{i=1}^n \|X_i\|^2\left(\delta + \left(X_i'(a - \hat{a})\right)^2\right) \\ & \le \text{const.}\left(p\delta + \kappa\|a - \hat{a}\|^2\right). \end{aligned} \tag{3.47}$$

With similar arguments for the other terms one gets finally

$$\begin{aligned} & d_2^2\left(\hat{\theta}_1 - E\hat{\theta}_1, \hat{\theta}_1^* - E_{F_n}\hat{\theta}_1^*\right) \\ & \le \text{const.}\left(p\delta + \kappa\|a - \hat{a}\|^2 + \|A^{-1} - \hat{A}^{-1}\|^2 p + \|a - \hat{a}\|^2\right). \end{aligned} \tag{3.48}$$

Now because of (3.44) the second and the fourth terms converge to 0 in probability. The convergence of the first term follows by the assumption $p^2/n \to 0$ and

$$\delta = O_P(p/n). \tag{3.49}$$

To prove (3.49), note that

$$\delta = \tilde{d}_2^2(F, F_n) \le 2\tilde{d}_2^2\left(F, \hat{F}_n\right) + 2\tilde{d}_2^2\left(\hat{F}_n, F_n\right), \tag{3.50}$$

where $\hat{F}_n$ denotes the empirical distribution of $\varepsilon_1, \dots, \varepsilon_n$, and

$$\tilde{d}_2^2(\hat{F}_n, F_n) \le \frac{1}{n} \sum_{i=1}^{n} (X_i'\hat{\theta})^2 = \frac{1}{n} \|\hat{\theta}\|^2 = O_P(p/n). \tag{3.51}$$

It remains to bound the third term in (3.48). This can be done by

$$\begin{aligned} \|A^{-1} - \hat{A}^{-1}\|^2 &\le \|A^{-1}\|^2 \|\hat{A}^{-1}\|^2 \|A - \hat{A}\|^2 \\ &\le O_P(1) \left(\sup_{\|\theta\| \le 1} \sum_{i=1}^{n} (X_i'\theta)^2 \left(E|R_i - R_i^*| + |X_i'(a - \hat{a})| \right)^2 \right) \\ &\le O_P(1)\left(\delta + \kappa\|a - \hat{a}\|^2\right) \\ &= O_P\left(\frac{p}{n}\right) + O_P\left(\kappa^2 \frac{p^2}{n}\right), \end{aligned}$$

where in the last step (3.44) and (3.49) have been used. □

REFERENCES

BERAN, R. (1984). Jackknife approximations to bootstrap estimates. *Ann. Statist.* **12** 101–118.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. J. and FREEDMAN, D. A. (1983). Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 28–48. Wadsworth, Belmont, Calif.

DAHLHAUS, R. (1988). Small sample effects in time series analysis: A new asymptotic theory and a new estimate. *Ann. Statist.* **16** 808–841.

DUNFORD, N. and SCHWARTZ, J. (1958). *Linear Operators* **1**. Wiley, New York.

EHM, W. (1986). On maximum likelihood estimation in high-dimensional log-linear type models. I. The independent case. Preprint SFB 123, Univ. Heidelberg.

HABERMAN, S. J. (1977a). Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5** 1148–1169.

HABERMAN, S. J. (1977b). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* Wiley, New York.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KANTOROWITSCH, L. W. and AKILOW, G. P. (1964). *Funktionalanalysis in Normierten Räumen.* Akademie, Berlin.

KOEHLER, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* **81** 483–493.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309.

PORTNOY, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.

ROUSSEEUW, P. J. and YOHAI, V. J. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 256–272. Springer, New York.

SAUERMANN, W. (1986). Bootstrap-Verfahren in log-linearen Modellen. Dissertation, Univ. Heidelberg.

SHORACK, G. (1982). Bootstrapping robust regression. *Comm. Statist. A—Theory Methods* **11** 961–972.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
6900 HEIDELBERG
WEST GERMANY