The Annals of Statistics
1996, Vol. 24, No. 1, 307–335

EMPIRICAL PROCESS OF RESIDUALS FOR HIGH-DIMENSIONAL LINEAR MODELS[1]

BY ENNO MAMMEN

Ruprecht-Karls-Universität Heidelberg

We give a stochastic expansion for the empirical distribution function $\hat{F}_n$ of residuals in a p-dimensional linear model. This expansion holds for p increasing with n. It shows that, for high-dimensional linear models, $\hat{F}_n$ strongly depends on the chosen estimator $\hat{\theta}$ of the parameter θ of the linear model. In particular, if one uses an ML-estimator $\hat{\theta}_{\mathrm{ML}}$ which is motivated by a wrongly specified error distribution function G, then $\hat{F}_n$ is biased toward G. For $p^2/n \to \infty$, this bias effect is of larger order than the stochastic fluctuations of the empirical process. Hence, the statistical analysis may just reproduce the assumptions imposed.

1. Introduction. In many statistical applications the results of a statistical analysis may be strongly influenced by the model assumptions imposed. Sometimes this influence is only a trivial matter. However, there exist examples where this effect is hidden in a more complex structure and may not be noticed in a statistical analysis. This paper gives an example of the latter case.

For high-dimensional linear models with i.i.d. errors, we consider the problem of estimating the error distribution. We show that the empirical distribution of residuals depends strongly on the used estimator for the parameter of the linear model. In particular, if one uses an ML-estimator based on the likelihood with incorrectly specified error distribution G, the empirical distribution of residuals is shifted toward G. For high-dimensional linear models this bias effect can be of larger order than the stochastic fluctuations of the empirical distribution. In particular, then the true error distribution may be rejected with high probability by goodness-of-fit tests.

These features are not apparent in an asymptotic analysis where the dimension p of the linear model is fixed. For p fixed, the asymptotics of the empirical process of residuals is well understood. A first asymptotic treatment was given in Koul (1969). General independent errors are treated in Koul (1984). An overview can be found in Koul (1992). In this paper we use an asymptotic approach where p may increase as the sample size $n \to \infty$. This approach is appropriate for many applications in which p is not small

Received September 1992; revised April 1995.

[1]This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Humboldt-Universität zu Berlin.

AMS 1991 *subject classifications*. Primary 62G30; secondary 62J05, 62J20.

Key words and phrases. Empirical processes, residuals, linear models, asymptotics with increasing dimension.

compared with the number of observations. For high-dimensional linear models this approach can offer explanations which cannot be obtained by asymptotics for fixed dimension p. An asymptotic approach with fixed p is misleading because the high dimensionality of the model gets lost asymptotically. Asymptotics with increasing p are also proposed for linear models in Huber (1981), Shorack (1982), Bickel and Freedman (1983), Portnoy (1984, 1985, 1986), Welsh (1989) and Mammen (1989, 1993) and for log-linear models in Haberman (1977a, b), Ehm (1986), Portnoy (1988) and Sauermann (1989). Consistency of bootstrap and asymptotic normality are studied in Mo (1991, 1992) for minimum contrast estimators of parameters with increasing dimension. In Kreiss (1988, 1991) autoregressive processes of infinite order are approximated by autoregressive processes with increasing order. There too, asymptotic results are given for the empirical distribution of residuals in this model.

Section 2 contains our results on empirical processes of residuals in high-dimensional linear models. Some applications of these results are discussed in Section 3. In Section 5 it is shown that, under reasonable assumptions, M-estimators fulfil the conditions used in Section 2. This section has been included because the asymptotic description of estimators in models with increasing dimension is rather different from the case of fixed dimension. The proofs are given in Sections 4, 6 and 7.

2. Empirical distribution of residuals. In this paper we consider a linear model

$$Y_i = X_i^T\theta + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

with i.i.d. errors $\varepsilon_1, \ldots, \varepsilon_n$. The design variables $X_i \in \mathbb{R}^p$ are assumed to be nonrandom. We study how the empirical distribution function $\hat{F}_n$ of residuals $\hat{\varepsilon}_i = (Y_i - X_i^T\hat{\theta})$, $i = 1, \ldots, n$,

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(\hat{\varepsilon}_i \le t)$$

works as an estimator of the error distribution function F. In particular, we are interested in seeing how the asymptotic behaviour of $\hat{F}_n$ depends on the choice of the estimator $\hat{\theta}$ of the parameter θ.

An asymptotic description of $\hat{F}_n$ can be based on the comparison of $\hat{F}_n$ with the empirical distribution function $\tilde{F}_n$ of the error variables

$$\tilde{F}_n(t) = n^{-1}\sum_{i=1}^{n} \mathbf{I}(\varepsilon_i \le t).$$

This has been done in Koul (1969, 1970, 1984, 1992) and Loynes (1980) [see also Shorack and Wellner (1986), Section 4.6]. The comparison is based on the

following expansions for $\tau \in \mathbb{R}^p$:

$$
\begin{aligned}
& n^{1/2}\left[n^{-1} \sum_{i=1}^{n} \mathbf{I}\big(Y_i - X_i^T\tau \le t\big) - \tilde{F}_n(t) \right] \\
&\quad = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{I}\big(\varepsilon_i \le t + X_i^T(\tau - \theta)\big) - \mathbf{I}(\varepsilon_i \le t) \right] \\
&\quad = n^{-1/2} \sum_{i=1}^{n} \left[F\big(t + X_i^T(\tau - \theta)\big) - F(t) \right] + o_P(1) \\
&\quad = n^{-1/2} f(t) \sum_{i=1}^{n} \left[X_i^T(\tau - \theta) \right] + o_P(1).
\end{aligned} \tag{2.2}
$$

The main step consists in showing that (2.2) holds uniformly in τ over compact sets [under the norming (A1), see below]. The application of $\tau = \hat{\theta}$ yields

$$
\sqrt{n}\left(\hat{F}_n(t) - \tilde{F}_n(t)\right) = n^{-1/2} f(t) \sum_{i=1}^{n} \left[X_i^T(\hat{\theta} - \theta) \right] + o_P(1). \tag{2.3}
$$

The expansion (2.3) even holds if $p^3/n \to 0$ [under reasonable conditions on $\hat{\theta}$, see Ioannides (1987)]. However, in general, this expansion does not hold unless p^2/n converges to 0. This has been shown in Portnoy (1986), where the finite-dimensional distribution of $\hat{F}_n$ was determined for M-estimators $\hat{\theta}$. In particular, this implies that (2.2) does not hold uniformly for the case of $p^2/n \to \infty$. For the asymptotic treatment of $\hat{F}_n$ under this condition, a new mathematical approach is necessary.

In this paper we discuss the asymptotic behaviour of $\hat{F}_n$ for $p^2/n \to \infty$. We consider estimators $\hat{\theta}$ admitting a linear approximation $\theta + (\sum_{i=1}^n X_i X_i^T)^{-1} \sum_{i=1}^n X_i \chi(\varepsilon_i)$ for some function χ. Without loss of generality we assume

$$
\sum_{i=1}^{n} X_i X_i^T = I_p. \tag{A1}
$$

Assumption (A1) corresponds to the $\sqrt{n}$-norming in the special case of a shift model ($p = 1$, $X_i \equiv 1/\sqrt{n}$). We assume the design to be roughly balanced in the following sense (note that the $\|X_i\|^2$'s are the diagonal elements of the hat matrix, where $\|\ \|$ denotes the Euclidean norm):

$$
\sup_{1 \le i \le n} \|X_i\|^2 = O\left(\frac{p}{n}\right). \tag{A2}
$$

We allow p^2/n to increase at the following rate:

$$
\frac{p^2}{n} = o(n^{1/5}). \tag{A3}
$$

Furthermore, we assume that the linear approximation of $\hat{\theta}$ has the following accuracy:

(A4) $$\left\| \hat{\theta} - \theta - \sum_{i=1}^{n} X_i \chi(\varepsilon_i) \right\| = O_P\left(\sqrt{\frac{p^2}{n}} \right)$$
holds for a function χ with $E\chi(\varepsilon_i) = 0$.

In view of (A3), condition (A4) does not imply that the Euclidean norm between $\hat{\theta}$ and its linear approximation converges to 0. The following condition ensures that the expectations Y_i are estimated consistently.

(A5) There exists a $b = b_n \in \mathbb{R}^p$ with $\|b\| = O(\sqrt{p^2/n})$ such that

$$\sup_{1 \le i \le n} \left| X_i^T \hat{\theta} - X_i^T(\theta + b) \right| = O_P\left(\sqrt{\frac{p}{n}} (\log n)^{3/2} \right).$$

The asymptotic bias of $\hat{\theta}$ is represented by b. Due to (A2) and (A3), condition (A5) implies the consistency of the linear fit, that is, $\sup_{1 \le i \le n} |X_i^T \hat{\theta} - X_i^T \theta| = o_P(1)$. Typically, $\sup_{1 \le i \le n} |X_i^T \hat{\theta} - X_i^T(\theta + b)|$ is of the order $\sqrt{(p/n)\log n}$ (see Section 5). We shortly motivate this order: In a "first approximation" $\hat{\theta}$ behaves like a p-dimensional Gaussian random variable with covariance $E\chi^2(\varepsilon_1) I_p$. Because of (A2) one expects an "asymptotic" variance of order p/n for $X_i^T \hat{\theta}$. The additional factor $\sqrt{\log n}$ is the price to be paid for the uniformity in i.

We use the following regularity conditions on χ and the error density f.

(A6) The error density f is twice differentiable with bounded second derivative and is strictly positive.

(A7) The function χ is differentiable with bounded derivative; $E \exp(t\chi(\varepsilon_1))$ exists for $|t|$ small enough; $E\chi(\varepsilon_1) = 0$.

We need one further condition describing how the exact location of the ε_i's lying near a fixed point t interacts with the value of $\hat{\theta}$.

(A8) For a sequence $\kappa_n \to \infty$ we define $T_{n,C}$ as the finite grid $T_{n,C} = \kappa_n^{-1} n^{-1/2} \mathbb{Z} \cap [-C, C]$. It is assumed that for all $t \in \kappa_n^{-1} n^{-1/2} \mathbb{Z}$ there exists a random variable $\hat{\theta}(t)$ with the following two properties: (i) $\hat{\theta}(t)$ depends only on the random set $I(t) = \{i\colon\ |\varepsilon_i - X_i^T b - t| \ge \sqrt{(p/n)}\,(\log n)^2\}$ and the values of the ε_i's with index i in $I(t)$ [b is the asymptotic bias introduced in (A5)]; (ii) for all $C > 0$, it holds thatsup$\{|X_i^T(\hat{\theta}(t) - \hat{\theta})|\colon 1 \le i \le n,\ t \in T_{n,C}\} = o_P(n^{-1/2})$.

Condition (A8) is our main key for starting the asymptotic expansion of $\hat{F}_n$. In the asymptotic treatment of $\hat{F}_n$, property (ii) allows us to replace $\hat{\theta}$ by $\hat{\theta}(t)$ in the definition of $\hat{F}_n(t)$. In the next step we condition on $I(t)$ and on the values of ε_i for $i \in I(t)$. Then we use the fact that, conditionally, $\hat{\theta}(t)$ is fixed now [see (i)].

Now we are ready to state our main result.

THEOREM 1. *Assume* (A1)–(A8). *For the empirical distribution function of residuals* $\hat{F}_n$, *the following expansion holds*:

$$\sup_{|t|\le C}\left|\sqrt{n}\left(\hat{F}_n(t) - \tilde{F}_n(t)\right) - \Delta(t)\right| = o_P(1) \quad \text{for every } 0 < C < \infty.$$

Here

$$\Delta(t) = f(t)n^{-1/2}\sum_{i=1}^{n} X_i^T(\hat{\theta} - \theta) + pn^{-1/2}f(t)\left[\chi(t) + \frac{1}{2}\frac{f'(t)}{f(t)}E\chi^2(\varepsilon_1)\right].$$

Statistical applications of the stochastic expansion of $\hat{F}_n$ given in this theorem are discussed in the next section. Let us now briefly consider the situation where a scale estimator $\hat{\sigma}$ is used additionally. The stochastic nature of the errors may then be judged by the empirical distribution function $\hat{F}_{S,n}$ of the standardized residuals $\hat{\sigma}^{-1}(Y_i - X_i^T\hat{\theta})$, $i = 1, \ldots, n$. The asymptotic behaviour of $\hat{F}_{S,n}$ is stated in the following corollary.

COROLLARY 1. *Assume* (A1)–(A8). *Furthermore, suppose that the scale estimator* $\hat{\sigma}$ *has the following accuracy*: $\hat{\sigma} - \sigma = o_P(n^{-1/4})$ *for a* $\sigma > 0$. *Then the following expansion holds for the empirical distribution function of standardized residuals* $\hat{F}_{S,n}$:

$$\sup_{|t|\le C}\left|\sqrt{n}\left(\hat{F}_{S,n}(t) - \tilde{F}_n(\sigma t)\right) - \Delta_S(\sigma t)\right| = o_P(1) \quad \text{for every } 0 < C < \infty,$$

where $\Delta_S(u) = \Delta(u) + uf(u)(\hat{\sigma} - \sigma)/\sigma$.

3. Discussion of the stochastic expansion of $\hat{F}_n$. In this section we apply the result of the previous section to the case of θ being estimated by an ML-estimator $\hat{\theta}_{\mathrm{ML}}$. For the error distribution function F we consider the following parametric families:

$$\mathcal{F}_\psi = \{G\colon\ g'/g = -c\psi \text{ for a } c > 0, \text{ where } g = G' \text{ is the density corresponding to } G\}.$$

In this model the ML-estimator is an M-estimator with M-function ψ:

$$\sum_{i=1}^{n} X_i\psi\left(Y_i - X_i^T\hat{\theta}_{\mathrm{i}} - X_i^T\hat{\theta}_{\mathrm{ML}}\right) = 0.$$

From the asymptotics with p fixed, one expects that $\hat{\theta}_{\mathrm{ML}}$ has a linear stochastic expansion $\sum_{i=1}^{n} X_i\,\chi(\varepsilon_i)$ with $\chi(t) = \psi(t)/E_F\psi'(\varepsilon_i)$. The validity of this stochastic expansion will be studied in Section 5. We consider two cases:

Case 1. The model $\mathcal{F}_\psi$ holds (i.e., $F \in \mathcal{F}_\psi$).
Case 2. The parametric model $\mathcal{F}_\psi$ is misspecified (i.e., $F \notin \mathcal{F}_\psi$).

In the following discussion we will argue that, in high-dimensional models, typically goodness-of-fit tests for models $\mathscr{F}_\psi$ based on the empirical distribution of residuals $\hat{F}_n$ break down. A modification of $\hat{F}_n$ which works will be proposed. Furthermore, we will show that under misspecification (Case 2) the estimate $\hat{F}_n$ is strongly biased toward the parametric error model $\mathscr{F}_\psi$.

In the discussion of this section we assume that $p^2/n \to \infty$. Furthermore, for simplicity let us assume that $\|n^{-1}\sum_{i=1}^n X_i\| = o(n^{-1/2})$. Then $n^{-1}\sum_{i=1}^n X_i^T(\hat{\theta} - \theta) = o_P(pn^{-1})$ [because of (A4)] and, because $n^{-1/2} = o(pn^{-1})$, the stochastic fluctuations of the empirical distribution function of residuals $\hat{F}_n$ are of smaller order than the bias (see the expansion of $\hat{F}_n$ in Theorem 1), that is,

$$\hat{F}_n(t) = F(t) + pn^{-1}f(t)\left[\chi(t) + \frac{1}{2}\left(\frac{f'(t)}{f(t)}\right)E_F\chi^2(\varepsilon_1)\right] + o_P(pn^{-1}).$$

CASE 1 ($F \in \mathscr{F}_\psi$). For regular densities f, one has $E_F[(f'/f)'(\varepsilon)] = -E_F[(f'/f)^2(\varepsilon)]$. Using this equality we obtain, from the formula of Theorem 1,

$$\hat{F}_n(t) = F(t) + pn^{-1}d(t) + o_P(pn^{-1}),$$

where $d(t) = -f'(t)/\{2E_F[(f'/f)^2(\varepsilon)]\}$. Typically, the distance between $F(t) + pn^{-1}d(t)$ and the model $\mathscr{F}_\psi$ is of order $O(pn^{-1})$ and not of order $o(pn^{-1})$. This difference will force most goodness-of-fit tests to reject the true model $\mathscr{F}_\psi$ with high probability.

However, this does not hold for the Gaussian errors (where the ML-estimator is the least squares estimator). Note that $\Phi(t/(1+\delta)) = \Phi(t) - \delta t\varphi(t) + o(\delta)$, where Φ and φ are the distribution function and density, respectively, of a standard normal law. For Gaussian F we obtain $F(t) + pn^{-1}d(t) = F(t/(1 - \frac{1}{2}pn^{-1})) + o(pn^{-1})$, that is, in this case $\hat{F}_n$ is biased toward a normal distribution with smaller variance. For normal errors we can avoid this bias effect by using scaled residuals $\hat{\varepsilon}_i(1 - pn^{-1})^{-1/2}$. This modification corresponds to the use of the unbiased variance estimator $(n-p)^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2$ instead of the empirical variance $n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2$.

For other estimators of θ we suggest the following modification of $\hat{F}_n$:

$$\hat{F}_{\mathrm{MOD},n} \text{ is the empirical distribution function of } \hat{\varepsilon}_i + \tfrac{1}{2}pn^{-1}\chi(\hat{\varepsilon}_i), \quad 1 \le i \le n.$$

Typically, the function χ is not known. For instance, in the case of M-estimators, $E_F\psi'(\varepsilon_i)$ may be unknown. We suggest estimating this quantity and substituting it, in the definition of χ, by its estimate.

It can be shown that $\sup_{|t| \le C}\sqrt{n}\,|\hat{F}_{\mathrm{MOD},n}(t) - \hat{F}_n(t - \frac{1}{2}pn^{-1}\chi(t))| = o_P(1)$ for any $C > 0$. This proves the following corollary.

COROLLARY 2. *Assume* (A1)–(A8). *For the modified empirical distribution function of residuals* $\hat{F}_{\mathrm{MOD},n}$ *the following expansion holds*:

$$\sup_{|t|\le C}\left|\sqrt{n}\left(\hat{F}_{\mathrm{MOD},n}(t)-\tilde{F}_n(t)\right)-\Delta_{\mathrm{MOD}}(t)\right|=o_P(1)\quad \textit{for every } 0<C<\infty,$$

where

$$\Delta_{\mathrm{MOD}}(t)=f(t)n^{-1/2}\sum_{i=1}^{n}X_i^T(\hat{\theta}-\theta)+pn^{-1/2}f(t)\left[\chi(t)+E\chi^2(\varepsilon_1)\frac{f'(t)}{f(t)}\right]$$

$$=\Delta(t)+\frac{1}{2}pn^{-1/2}f'(t)E\chi^2(\varepsilon_1).$$

In the case where $F\in\mathscr{F}_\psi$ and the ML-estimator is used, one obtains

$$\left[\chi(t)+E\chi^2(\varepsilon_1)\frac{f'(t)}{f(t)}\right]=0,$$

which implies $\sqrt{n}\,(\hat{F}_{\mathrm{MOD},n}(t)-\tilde{F}_n(t))=n^{-1/2}f(t)\sum_{i=1}^{n}X_i^T(\hat{\theta}-\theta)+o_P(1)$, that is, after such a modification of $\tilde{F}_n$ we are back to the fixed p asymptotics [see (2.3)]. Hence, we suggest basing goodness-of-fit tests on $\hat{F}_{MOD,n}$.

CASE 2 (Model misspecification; $F\notin\mathscr{F}_\psi$). For $i=1$ and $i=2$, let G_i denote the element of $\mathscr{F}_\psi$ whose density g_i fulfils $g_i'/g_i=-c_i^{-1}\psi$ for $c_1=(E_F\psi^2)/(2E_F\psi')$ and $c_2=(E_F\psi^2)/(E_F\psi')$. We show that the empirical distribution function of residuals $\hat{F}_n$ is strongly biased toward $G_1\in\mathscr{F}_\psi$. Furthermore, $\hat{F}_{\mathrm{MOD},n}$ is biased toward $G_2\in\mathscr{F}_\psi$. Again, we apply the formulas of Theorem 1 and Corollary 1 with $\chi(t)=\psi(t)/E_F\psi'(\varepsilon_1)$. This yields

$$\hat{F}_n(t)=F(t)+pn^{-1}\frac{E_F\chi^2(\varepsilon_1)}{2}f(t)\left[\frac{f'(t)}{f(t)}-\frac{g_1'(t)}{g_1(t)}\right]+o_P(pn^{-1}),$$

$$\hat{F}_{\mathrm{MOD},n}(t)=F(t)+pn^{-1}E_F\chi^2(\varepsilon_1)f(t)\left[\frac{f'(t)}{f(t)}-\frac{g_2'(t)}{g_2(t)}\right]+o_P(pn^{-1}).$$

Now, for δ small enough, under reasonable conditions on F and G_i, $i=1,2$,

$$\sup_t\left|F(t)+\delta f(t)\left[\frac{f'(t)}{f(t)}-\frac{g_i'(t)}{g_i(t)}\right]-G_i(t)\right|<\sup_t|F(t)-G_i(t)|.$$

This implies

$$\sup_t|\hat{F}_n(t)-G_1(t)|<\sup_t|F(t)-G_1(t)|,$$

$$\sup_t|\hat{F}_{\mathrm{MOD},n}(t)-G_2(t)|<\sup_t|F(t)-G_2(t)|,$$

for n large enough with probability tending to 1. The differences between the right- and left-hand sides of these inequalities are of order p/n. Therefore, $\hat{F}_n$ and $\hat{F}_{\mathrm{MOD},n}$ are strongly biased toward the parametric error model. Under our assumptions ($p^2/n\to\infty$) this bias effect is of a larger order than the

stochastic fluctuations of $\hat{F}_n$ and $\hat{F}_{\mathrm{MOD},n}$. In particular, the least squares residuals tend to look more like normally distributed variables than they should. This effect has also been called supernormality, and, heuristically, it may also be explained by the fact that every residual is a sum of independent variables. M-estimators with bounded ψ-function produce estimates of the error distribution with heavy tails. Qualitatively, this effect follows from the smaller influence of outlying observations Y_i on the value of the parametric estimator $\hat{\theta}_\psi$.

The following simulations show this bias effect: 100 data sets of 50 observations have been generated with a normal distribution $F_1 = N(0, \frac{13}{4})$ and with a normal mixture distribution $F_2 = \frac{1}{2}N(-\frac{3}{2}, 1) + \frac{1}{2}N(\frac{3}{2}, 1)$. Independent standard normal pseudorandom variables have been chosen as design variables X_{ij} with $p = 5$. In all 100 runs of the simulation the same design variables have been used. For both data sets, residuals have been calculated using the least squares estimator $\hat{\theta}_{\mathrm{LS}}$ and the M-estimator $\hat{\theta}_\psi$ with $\psi = (-f_2'/f_2)$. (Here f_2 is the density of F_2.) This provides four sets of 5000 residuals. Kernel density estimates using these four sets are plotted in Figures 1–4. The bandwidth is 0.75. The kernel is the biweight function. In each figure, plots of the densities of F_1 and F_2 have been added.

We have used a bimodal F_2 and we have plotted densities instead of distribution functions. This may help to interpret the figures. The plotted kernel estimates are Monte Carlo estimates of the density of the expectation of the empirical distribution of residuals (based on 50 observations).

Figures 1 and 4 show that there is no strong bias effect if ML-estimates are used, which correspond to the true model. However, under model misspecification, strong bias effects appear. This is shown in Figures 2 and 3. In both cases the kernel estimates lie between the true and the "assumed" densities, that is, there are bias effects toward the assumed model. This is in accordance with the theory presented above. Note also that there is no large difference between the kernel estimates in Figures 2 and 3. This means that the bias effect is so large here that it does not have a large effect when interchanging the true and the assumed model.

4. Proof of Theorem 1. The main idea of the proof is as follows. For considerations on the value of $\hat{F}_n$ at a fixed point t, we subdivide the error variables ε_i into two groups:

$\{\varepsilon_i : i \notin I(t)\}$ "error variables with values lying near to t";

$\{\varepsilon_i : i \in I(t)\}$ "error variables with values lying far away from t."

[The set $I(t)$ is defined in assumption (A8).] First, we replace $\hat{\theta}$ by $\hat{\theta}(t)$ in the definition of $\hat{F}_n(t)$. Because of assumption (A8)(ii) this only leads to asymptotically negligible changes of $\hat{F}_n(t)$.

The random variable $\hat{\theta}(t)$ depends on $I(t)$ and the second group $\{\varepsilon_i : i \in I(t)\}$ only. In the next step we condition on $I(t)$ and $\{\varepsilon_i : i \in I(t)\}$. Conditionally, $\hat{\theta}(t)$ is fixed and nonrandom. Using exponential inequalities we show that $\hat{F}_n(t)$ is asymptotically equivalent to its conditional expectation.

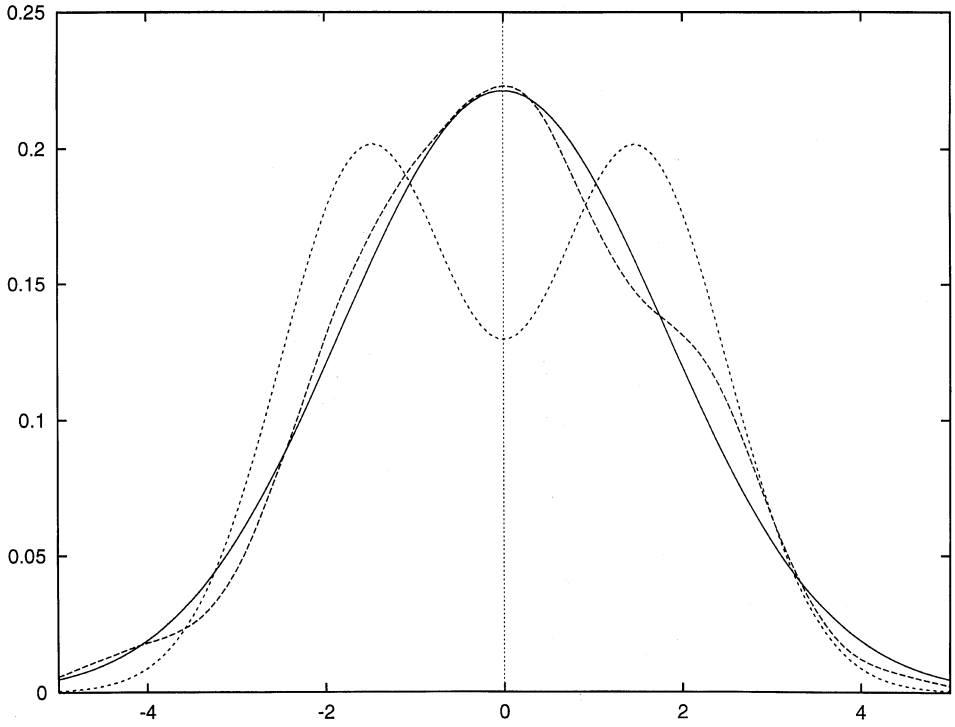

FIG. 1. *Density of the expectation of the empirical distribution of residuals* [*short dashes, error distribution* $F_1 = N(0, \frac{13}{4})$, *least squares estimator* $\hat{\theta}_{LS}$], *density of* F_1 (*solid line*) *and density of* F_2 (*long dashes*).

The conditional expectation of $\hat{F}_n(t)$ consists of a sum of smooth functions (instead of indicator functions). These functions can be treated by Taylor expansions. A more technical summary of the proof can be found after equation (4.5).

Note that (A8) remains valid with κ_n replaced by $\max\{\kappa_n/r\colon\ r \in \mathbf{N}, \kappa_n/r \le \log n\}$. (This would make the grid $T_{n,C}$ smaller.) Then

$$\kappa_n = O(\log(n)). \tag{4.1}$$

With $\gamma_n = \sqrt{(p/n)}\,(\log n)^2$ and for a sequence λ_n with $\lambda_n \to \infty$ and $\lambda_n = O(\log(n))$, define the event A_n as

$$A_n = \left\{ \sup_{1 \le i \le n} \left| X_i^T(\hat{\theta} - \theta - b) \right| \le \frac{\gamma_n}{2}, \left\| \hat{\theta} - \theta - \sum_{i=1}^{n} X_i \chi(\varepsilon_i) \right\| \le \lambda_n \sqrt{\frac{p^2}{n}}, \right.$$

$$\left. \sup_{1 \le i \le n,\, t \in T_{n,C}} \left| X_i^T\big(\hat{\theta}(t) - \hat{\theta}\big) \right| \le \lambda_n^{-1} n^{-1/2} \right\}.$$

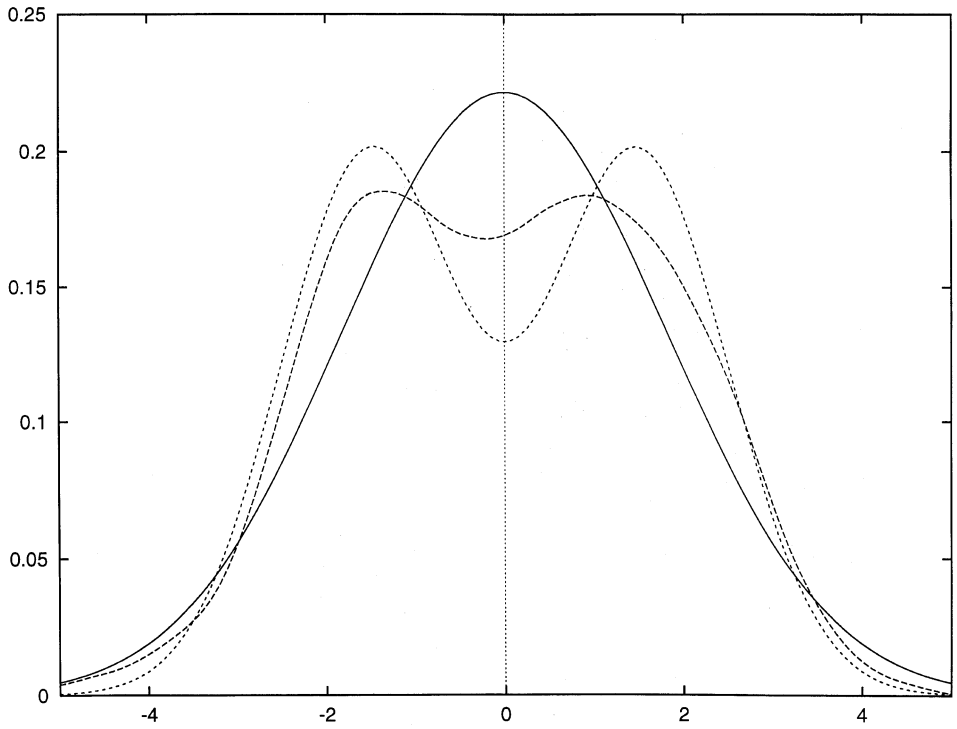

FIG. 2. *Density of the expectation of the empirical distribution of residuals* [*short dashes, error distribution* $F_1 = N(0, \frac{13}{4})$, *M-estimator* $\hat{\theta}_\psi$ *with* $\psi = (-f_2'/f_2)$], *density of* F_1 (*solid line*) *and density of* F_2 (*long dashes*).

For $\lambda_n \to \infty$ slowly enough, we obtain

$$P(A_n) \to 1. \tag{4.2}$$

We set $t_i = t_i(t) = t + X_i^T b$,

$$\hat{t}_i = \hat{t}_i(t) = \begin{cases} t + X_i^T\big(\hat{\theta}(t) - \theta\big), & \text{if } \big|X_i^T\big(\hat{\theta}(t) - \theta - b\big)\big| \le \gamma_n, \\ t_i(t), & \text{elsewhere,} \end{cases}$$

$$t_i^- = t_i^-(t) = t_i(t) - \gamma_n,$$

$$t_i^+ = t_i^+(t) = t_i(t) + \gamma_n.$$

Note that it holds on A_n (for n large enough) that $\hat{t}_i = t + X_i^T(\hat{\theta}(t) - \theta)$. By the monotonicity of $\hat{F}_n$ and $\tilde{F}_n$, it is sufficient for the proof of the theorem to show that

$$\sup_{t \in T_{n,C}} \big|\sqrt{n}\,\big(\hat{F}_n(t) - \tilde{F}_n(t)\big) - \Delta(t)\big| = o_P(1). \tag{4.3}$$

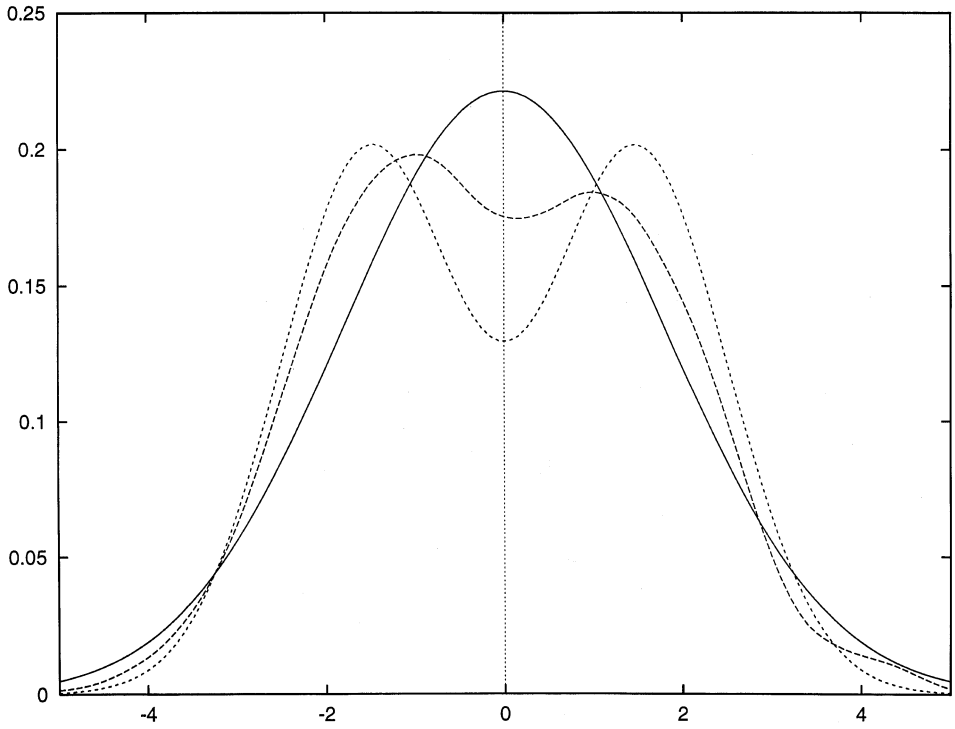

FIG. 3. *Density of the expectation of the empirical distribution of residuals* [*short dashes, error distribution* $F_2 = \frac{1}{2}N(-\frac{3}{2}, 1) + \frac{1}{2}N(\frac{3}{2}, 1)$, *least squares estimator* $\hat{\theta}_{LS}$], *density of* F_1 (*solid line*), *and density of* F_2 (*long dashes*).

For $\hat{F}_n^{\mathrm{tr}}(t) = (1/n)\sum_{i=1}^n \mathbf{I}(\varepsilon_i \le \hat{t}_i)$ we obtain, on A_n,

$$\sup_{t \in T_{n,C}} \sqrt{n}\left|\hat{F}_n^{\mathrm{tr}}(t) - \hat{F}_n(t)\right| \le \sup_{t \in T_{n,C}} \sqrt{n}\left|\hat{F}_n^{\mathrm{tr}}\left(t + \lambda_n^{-1}n^{-1/2}\right) - \hat{F}_n^{\mathrm{tr}}\left(t - \lambda_n^{-1}n^{-1/2}\right)\right|.$$

Because of

$$\sup_{t \in \mathbb{R}}\left|\left(\sqrt{n}\,\tilde{F}_n + \Delta\right)\left(t + \lambda_n^{-1}n^{-1/2}\right) - \left(\sqrt{n}\,\tilde{F}_n + \Delta\right)\left(t - \lambda_n^{-1}n^{-1/2}\right)\right| = o_P(1),$$

for (4.3) it suffices to prove

$$\sup_{t \in T_{n,C}}\left|\sqrt{n}\left(\hat{F}_n^{\mathrm{tr}}(t) - \tilde{F}_n(t)\right) - \Delta(t)\right| = o_P(1). \tag{4.4}$$

We set

$$\sqrt{n}\left(\hat{F}_n^{\mathrm{tr}}(t) - \tilde{F}_n(t)\right) - \Delta(t) = S_1(t) + S_2(t) + S_3(t),$$

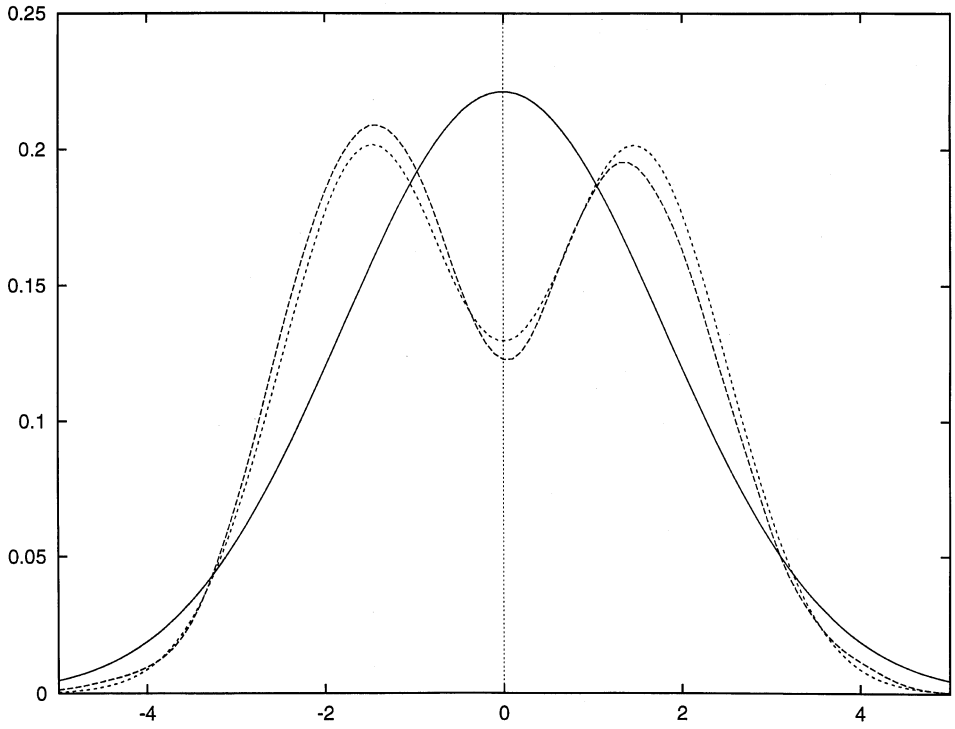

FIG. 4. *Density of the expectation of the empirical distribution of residuals* [*short dashes, error distribution* $F_2 = \frac{1}{2}N(-\frac{3}{2},1) + \frac{1}{2}N(\frac{3}{2},1)$, *M-estimator* $\hat{\theta}_\psi$ *with* $\psi = (-f_2'/f_2)$], *density of* F_1 (*solid line*) *and density of* F_2 (*long dashes*).

with

$$S_1(t) = n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{I}(\varepsilon_i \le t_i^-) - \mathbf{I}(\varepsilon_i \le t) - \left(F(t_i^-) - F(t)\right)\right],$$

$$S_2(t) = n^{-1/2} \sum_{i=1}^{n} \mathbf{I}\left(\varepsilon_i \le \hat{t}_i\right) - \mathbf{I}(\varepsilon_i \le t_i^-) - n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left(F(\hat{t}_i) - F(t_i^-)\right),$$

where $D_{ni}(t) = \mathbf{I}(t_i^- < \varepsilon_i \le t_i^+)$ and $\Gamma_{ni}(t) = F(t_i^+) - F(t_i^-) = ED_{ni}(t)$,

$$S_3(t) = n^{-1/2} \sum_{i=1}^{n} \left[\frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left(F(\hat{t}_i) - F(t_i^-)\right) + \left(F(t_i^-) - F(t)\right)\right] - \Delta(t).$$

After these preliminaries let us briefly describe the following steps of the proof. We will show

$$\sup_{t \in T_{n,C}} \left|S_j(t)\right| = o_P(1) \quad \text{for } j = 1,2,3. \tag{4.5}$$

This implies the theorem. The crucial step here is to show (4.5) for $j = 2$. For this case we condition on $I(t)$ and the values of ε_i for $i \in I(t)$. Then we use

assumption (A8), which says that, conditionally, $\hat{\theta}(t)$ (and $\hat{t}_i$) is fixed. Using exponential inequalities, we obtain for $i \notin I(t)$ that (asymptotically) $\mathbf{I}(\varepsilon_i \le \hat{t}_i) - \mathbf{I}(\varepsilon_i \le t_i^-)$ can be replaced by its conditional expectations $(F(\hat{t}_i) - F(t_i^-))/\Gamma_{ni}(t)$. The proof of (4.5) for $j = 3$ is based on Taylor expansions of $F(\hat{t}_i) - F(t_i^-)$.

In the proofs, we make use of the Bernstein inequality repeatedly [see Hoeffding (1963) or Pollard (1984)]: *For independent* $Z_1, \dots, Z_n$ *with* $|Z_i| \le M$, $EZ_i = 0$ *and* $\sum_{i=1}^n \operatorname{Var}(Z_i) \le g\,0$ it holds that

$$P(|Z_1 + \cdots + Z_n| \ge \eta) \le 2\exp\left(-\frac{\eta^2}{2V + 2M\eta/3}\right).$$

Step 1. *Proof of* (4.5) *for* $j = 1$. This can easily be seen by applying the Bernstein inequality with $Z_i = n^{-1/2}[\mathbf{I}(\varepsilon_i \le t_i^-) - \mathbf{I}(\varepsilon_i \le t) - (F(t_i^-) - F(t))]$ and the fact that the number $\#T_{n,C}$ is of polynomial order [$\#T_{n,C} = O(\log(n)n^{1/2})$, see (4.1)]. The quantity $\#T_{n,C}$ denotes the number of elements of $T_{n,C}$.

Step 2. *Proof of* (4.5) *for* $j = 2$. We obtain, for $0 < \eta \le 1$,

$$\begin{aligned} P\Big(\sup_{t\in T_{n,C}} |S_2(t)| \ge \eta\Big) &\le \sum_{t\in T_{n,C}} P(|S_2(t)| \ge \eta) \\ &= \sum_{t\in T_{n,C}} EP(|S_2(t)| \ge \eta | I(t),\ \varepsilon_i \text{ for } i \in I(t)). \end{aligned}$$

Note that $S_2(t) = \sum_{j=1}^n Z_j$ with

$$Z_j = n^{-1/2}\left[\mathbf{I}(\varepsilon_j \le \hat{t}_j) - \mathbf{I}(\varepsilon_j \le t_j^-) - \frac{D_{nj}(t)}{\Gamma_{nj}(t)}\left(F(\hat{t}_j) - F(t_j^-)\right)\right].$$

Because $|\hat{t}_j - t_j| \le \gamma_n$ for $j = 1, \dots, n$, we have $Z_j = 0$ for $j \in I(t)$. For $j \notin I(t)$ the conditional expectation of Z_j [given $I(t)$ and the values of ε_i for $i \in I(t)$] is zero and the conditional variance of Z_j is bounded by $(4n)^{-1}$. We apply the Bernstein inequality to the conditional probability with Z_j for $j \notin I(t)$. The number of Z_j's [$j \notin I(t)$] is $n - \#I(t) = \sum_{i=1}^n D_{ni}(t)$. For $C' > 2\sup_{t\in\mathbb{R}} f(t)$, this provides

$$\begin{aligned} &P\Big(\sup_{t\in T_{n,C}} |S_2(t)| \ge \eta\Big) \\ &\quad\le \sum_{t\in T_{n,C}} E\left[2\exp\left(-\frac{\eta^2}{2n^{-1}(n - \#I(t))/4 + n^{-1/2}2\eta/3}\right)\right] + o(1) \\ &\quad\le \sum_{t\in T_{n,C}} 2\exp\left(-\frac{\eta^2}{C'\gamma_n/2 + n^{-1/2}2\eta/3}\right) \\ &\qquad + 2P(n^{-1}(n - \#I(t)) \ge C'\gamma_n) + o(1) \\ &\quad= 2\sum_{t\in T_{n,C}} P\left(n^{-1}\sum_{i=1}^n D_{ni}(t) \ge C'\gamma_n\right) + o(1) = o(1). \end{aligned}$$

The latter equality follows from one further application of the Bernstein inequality (see also the proof in Step 1).

Step 3. *Proof of* (4.5) *for* $j = 3$. We set

$$S_3(t) = S_{31}(t) + S_{32}(t) + S_{33}(t),$$

where

$$S_{31}(t) = n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} \bigl(F(t_i) - F(t_i^-)\bigr),$$

$$S_{32}(t) = n^{-1/2} \sum_{i=1}^{n} \bigl[F(t_i) - F(t) - f(t)\bigl(X_i^T b\bigr)\bigr],$$

$$S_{33}(t) = n^{-1/2} \sum_{i=1}^{n} \left[\frac{D_{ni}(t)}{\Gamma_{ni}(t)}\bigl(F\bigl(\hat{t}_i\bigr) - F(t_i)\bigr)\right] - f(t) n^{-1/2} \sum_{i=1}^{n} X_i^T(\hat{\theta} - \theta - b)$$
$$- f(t) p n^{-1/2} \left[\chi(t) + \frac{1}{2}\frac{f'(t)}{f(t)} E\chi^2(\varepsilon_i)\right].$$

We have to show that

$$\sup_{t \in T_{nC}} \bigl|S_{3j}(t)\bigr| = o_P(1) \quad \text{for } j = 1, 2, 3. \tag{4.6}$$

For $j = 1$ this follows by application of the Bernstein inequality (see also the remarks at the end of Step 2). Note that $\Gamma_{ni}(t)^{-1}(F(t_i) - F(t_i^-))$ is bounded by 1. For $j = 2$, one uses a Taylor expansion of $F(t_i) - F(t)$ and $\sum_{i=1}^n (t_i - t)^2 = \sum_{i=1}^n (X_i^T b)^2 = \|b\|^2 = O(p^2 n^{-1})$. It remains to show (4.6) for $j = 3$. This will be done in the Steps 3a–3c.

Step 3a. *Proof of*

$$\sup_{t \in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)} \left[F\bigl(\hat{t}_i\bigr) - F(t_i) - f(t_i)\bigl(\hat{t}_i - t_i\bigr) - \frac{1}{2} f'(t_i)\bigl(\hat{t}_i - t_i\bigr)^2\right]\right|$$
$$= o_P(1).$$

This holds because, with t_i' between t_i and $\hat{t}_i$, this term is equal to

$$\frac{1}{6} \sup_{t \in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)} f''(t_i')\bigl(\hat{t}_i - t_i\bigr)^3 \right|$$
$$\leq \frac{1}{6} \sup_{t \in T_{n,C}} n^{-1/2} \gamma_n^3 \sup_{t \in \mathbb{R}} |f''(t)| \sup_{t \in T_{n,C}} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}$$
$$= O\bigl(n^{-1/2}\gamma_n^2\bigr)\left\{ \sup_{t \in T_{n,C}} \sum_{i=1}^{n} \bigl[D_{ni}(t) - \Gamma_{ni}(t)\bigr] + \sup_{t \in T_{n,C}} \sum_{i=1}^{n} \Gamma_{ni}(t) \right\}$$
$$= o_P\bigl(\gamma_n^2\bigr) + O\bigl(n^{1/2}\gamma_n^3\bigr).$$

For the proof of the latter inequality the Bernstein inequality can be used again. Step 3a is now complete because $n^{1/2}\gamma_n^3 \to 0$ under our conditions.

Step 3b. *Proof of*

$$\sup_{t\in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)} f(t_i)\big(\hat{t}_i - t_i\big) - f(t) n^{-1/2} \sum_{i=1}^{n} X_i^T(\hat{\theta} - \theta - b) - pn^{-1/2} f(t)\chi(t) \right| = o_P(1). \tag{4.7}$$

Since the probability of $\hat{t}_i - t_i = X_i^T(\hat{\theta}(t) - \theta - b)$ (for $1 \le i \le n$) tends to 1, it sufficends to 1, it suffices for (4.7) to show

$$\sup_{t\in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \big(f(t_i) - f(t)\big) X_i^T(\hat{\theta} - \theta - b) \right| = o_P(1), \tag{4.8}$$

$$\sup_{t\in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)} f(t_i)\Big(X_i^T\big(\hat{\theta}(t) - \hat{\theta}\big)\Big) \right| = o_P(1), \tag{4.9}$$

$$\sup_{t\in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i)\Big(X_i^T(\hat{\theta} - \theta - b)\Big) - pn^{-1/2} f(t)\chi(t) \right| = o_P(1). \tag{4.10}$$

For the proof of (4.8) note that

$$\begin{aligned} &\left| n^{-1/2} \sum_{i=1}^{n} \big(f(t_i) - f(t)\big) X_i^T(\hat{\theta} - \theta - b) \right| \\ &\quad \le \left| n^{-1/2} \sum_{i=1}^{n} \big(f(t_i) - f(t)\big) X_i^T \sum_{j=1}^{n} X_j \chi(\varepsilon_j) \right| \\ &\qquad + \left\| n^{-1/2} \sum_{i=1}^{n} \big(f(t_i) - f(t)\big) X_i \right\| \cdot \left\| \hat{\theta} - \theta - b - \sum_{j=1}^{n} X_j \chi(\varepsilon_j) \right\| \\ &\quad = O_P\left(\left\| n^{-1/2} \sum_{i=1}^{n} \big(f(t_i) - f(t)\big) X_i \right\| \left(1 + \sqrt{\frac{p^2}{n}} \right) \right), \end{aligned}$$

where, in the latter equality, assumptions (A1), (A4) and (A7) have been applied. [Assumptions (A1) and (A7) imply $c_n^T \Sigma_{j=1}^n X_j \chi(\varepsilon_j) = O_P(\|c_n\|)$ for sequences c_n of vectors in $\mathbb{R}^p$.]

However,

$$\begin{aligned}\left\| n^{-1/2} \sum_{i=1}^{n} (f(t_i) - f(t)) X_i \right\| &\le \sup_{\|e\|=1} n^{-1/2} \sum_{i=1}^{n} \left[f(t_i) - f(t) \right] \left(X_i^T e \right) \\ &\le \sup_{\|e\|=1} \left[\sum_{i=1}^{n} \left(X_i^T e \right)^2 \right]^{1/2} \sup_{1 \le i \le n} \left| f(t_i) - f(t) \right| \\ &= \sup_{1 \le i \le n} \left| f(t_i) - f(t) \right|.\end{aligned}$$

Now claim (4.8) follows from $\sup_{t \in T_{n,C}} |t_i - t| = |X_i^T b| = O(p^{3/2} n^{-1})$ and from the assumption that f has a bounded derivative.

Claim (4.9) can be shown using

$$\sup_{t \in T_{n,C}} \left| X_i^T \left(\hat{\theta}(t) - \hat{\theta} \right) \right| = o_P(n^{-1/2}) \quad \text{and} \quad \sup_{t \in T_{n,C}} \left| n^{-1} \sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)} \right| = O_P(1).$$

To complete Step 3b, it remains to show (4.10).

Proof of (4.10). We remark first that

$$\begin{aligned}\sup_{t \in T_{n,C}} \Bigg| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i) \\ \times \left(X_i^T \left(\hat{\theta} - \theta - b - \sum_{j=1}^{n} X_j \chi(\varepsilon_j) \right) \right) \Bigg| = o_P(1).\end{aligned} \tag{4.11}$$

To prove (4.11) it suffices, in view of (A3) and (A4), to show

$$\sup_{t \in T_{n,C}} \left\| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i) X_i \right\| = o_P\left(p^{1/4} n^{-1/4} \right). \tag{4.12}$$

For (4.12) one proves

$$\begin{aligned}\sup_{1 \le j \le p,\, t \in T_{n,C}} \left[1 + \sum_{i=1}^{n} X_{ij}^2 \right]^{-1/2} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i) X_{ij} \right| \\ = o_P\left(p^{-1/4} n^{-1/4} (\log n)^{-1/2} \right).\end{aligned} \tag{4.13}$$

This can be seen by application of the Bernstein inequality.

Claim (4.10) follows now from (4.11) and

$$\sup_{t \in T_{n,C}} \left| n^{-1/2} \sum_{i \ne j} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i) \left(X_i^T X_j \right) \chi(\varepsilon_j) \right| = o_P(1) \tag{4.14}$$

and

$$\begin{aligned}\sup_{t \in T_{n,C}} \left| n^{-1/2} \sum_{i=1}^{n} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i) \|X_i\|^2 \chi(\varepsilon_i) - p n^{-1/2} f(t) \chi(t) \right| \\ = o_P(1).\end{aligned} \tag{4.15}$$

It remains to show (4.14) and (4.15).

Proof of (4.14). Set

$$Z(t) = n^{-1/2} \sum_{i \neq j} \frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} f(t_i)\left(X_i^T X_j\right)\chi(\varepsilon_j).$$

We apply now the following bound of Whittle (1960). For quadratic forms $Z = \sum_{i \neq 1} a_{ij}\xi_i\zeta_j$ with independent mean-zero random variables $\xi_1, \ldots, \xi_n, \zeta_1, \ldots, \zeta_n$, it holds for $k \geq 1$ that

$$EZ^{2k} \leq 2^{3k} C(k)\sqrt{C(2k)} \left[\sum_{i,j=1}^{n} a_{ij}^2 \left(E\left(\xi_i^{2k}\right)\right)^{1/k} \left(E\left(\zeta_i^{2k}\right)\right)^{1/k} \right]^k,$$

where $C(k) = (2^{k/2}/\sqrt{\pi})\Gamma((k+1)/2)$ (Γ is the gamma function). We apply this inequality with k equal to the integer part of $\log(n)$. For a constant b_1, Stirling's formula provides $2^{3k}C(k)\sqrt{C(2k)} \leq b_1^{\log(n)} \log(n)^{\log(n)}$. Furthermore, with some constants b_2 and b_3 we have

$$\sum_{i \neq j} \left(X_i^T X_j\right)^2 = \sum_{i=1}^{n} X_i^T \sum_{j=1}^{n} \left[X_j X_j^T - X_i X_i^T\right] X_i \leq \sum_{j=1}^{n} \|X_j\|^2 = p,$$

$$\sup_{t \in T_{n,C}} E\left(\frac{D_{ni}(t) - \Gamma_{ni}(t)}{\Gamma_{ni}(t)} \right)^{2k} \leq b_2^{\log(n)} \gamma_n^{1-2\log(n)},$$

$$E\chi(\varepsilon_j)^{2k} \leq b_3^{\log(n)} \log(n)^{2\log(n)}.$$

The latter inequality follows from assumption (A7) [and $|x| < \exp(|x|)$] because

$$E\chi(\varepsilon_j)^{2k} \leq \left(\frac{2k}{|t|}\right)^{2k} E\left[\left|\frac{t\chi(\varepsilon_j)}{2k}\right|^{2k}\right] \leq \left(\frac{2k}{|t|}\right)^{2k} E\exp\left(\left|t\chi(\varepsilon_j)\right|\right).$$

With the aid of these bounds we arrive at

$$\begin{aligned} EZ(t)^{2k} &\leq \left(\frac{p}{n} b_1 b_2 b_3\right)^{\log(n)} \log(n)^{3\log(n)} \gamma_n^{1-2\log(n)} \\ &= \left(\frac{p}{n}\right)^{1/2} (b_1 b_2 b_3)^{\log(n)} (\log n)^{2-\log(n)} \\ &= O\left(n^{-\delta \log\log(n)}\right), \end{aligned}$$

with a constant $\delta > 0$. Using this and $\#T_{n,C} = O(\log(n)n^{1/2})$ [see (4.1)] one can show (4.14) by means of the Tchebycheff inequality.

Proof of (4.15). Equation (4.15) follows from

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n} f(t_i)\|X_i\|^2\chi(\varepsilon_i)\right| = o_P(1), \tag{4.16}$$

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\|X_i\|^2 f(t_i)(\chi(\varepsilon_i)-\chi(t_i))\right| = o_P(1), \tag{4.17}$$

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)-\Gamma_{ni}(t)}{\Gamma_{ni}(t)}\|X_i\|^2 f(t_i)\chi(t_i)\right| = o_P(1), \tag{4.18}$$

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n} \|X_i\|^2\left|f(t_i)\chi(t_i)-f(t)\chi(t)\right|\right| = o(1). \tag{4.19}$$

Equations (4.16)–(4.19) can be shown by methods similar to those used above.

The proof of the theorem is now completed by the following step.

Step 3c. *Proof of*

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}f'(t_i)\left(\hat{t}_i-t_i\right)^2 - pn^{-1/2}f'(t)E\chi^2(\varepsilon_i)\right| = o_P(1). \tag{4.20}$$

Set $U=\sum_{i=1}^{n}X_i\,\chi(\varepsilon_i)$. We first show

$$\sup_{t\in T_{n,C}} n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left|f'(t_i)\right|\left|\left(\hat{t}_i-t_i\right)^2-\left(X_i^TU\right)^2\right| = o_P(1). \tag{4.21}$$

By assumptions (A5) and (A8) and $\sup_{t\in\mathbb{R}}|f'(t)| = O(1)$, for (4.21) it suffices to show

$$\sup_{t\in T_{n,C}} n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left|\left(X_i^TV\right)^2-\left(X_i^TU\right)^2\right| = o_P(1),$$

where $V=\hat{\theta}(t)-\theta-b$. Using $|(X_i^TV)^2-(X_i^TU)^2| = |X_i^T(V-U)|\,|X_i^T(V+U)|$, the Cauchy–Schwarz inequality and $p^{5/2}n^{-3/2} = o(1)$, this follows from

$$\sup_{t\in T_{n,C}} n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left(X_i^T(V-U)\right)^2 = o_P\left(\frac{p^{3/2}}{n}\right) \tag{4.22}$$

and

$$\sup_{t\in T_{n,C}} n^{-1/2}\sum_{i=1}^{n} \frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left(X_i^T(V+U)\right)^2 = O_P\left(pn^{-1/2}\right). \tag{4.23}$$

Proof of (4.22). A simple upper bound for the left-hand side is

$$\left[\inf_{t\in T_{n,C},\,1\le i\le n}\Gamma_{n,i}(t)\right]^{-1} n^{-1/2}\sum_{i=1}^{n}\left(X_i^T(V-U)\right)^2.$$

This bound is of the order $o_P(p^{3/2}/n)$ because of (A4).

Proof of (4.23). Claim (4.23) follows from (4.22) and

$$\sup_{t\in T_{n,C}} n^{-1/2}\sum_{i=1}^{n}\frac{D_{ni}(t)}{\Gamma_{ni}(t)}\left(X_i^TU\right)^2 = O_P\left(pn^{-1/2}\right). \tag{4.24}$$

Claim (4.24) can be shown using $\sum_{i=1}^{n}(X_i^TU)^2 = \|U\|^2 = O_P(p)$ and

$$\sup_{t\in T_{n,C}} |\rho(t)| = o_P(1), \tag{4.25}$$

where

$$\rho(t) = n^{-1/2}\sum_{i=1}^{n}\frac{D_{ni}(t)-\Gamma_{ni}(t)}{\Gamma_{ni}(t)}\left(X_i^TU\right)^2.$$

A lengthy calculation gives $\sup_{t\in T_{n,C}}\operatorname{Var}(\rho(t)) = O(p^{3/2}n^{-3/2})$ and $\sup_{t\in T_{n,C}}|E\rho(t)| = o(1)$. Because $\#T_{n,C} = O(\log(n)n^{1/2})$ this implies (4.25).

We now show

$$\sup_{t\in T_{n,C}}\left|n^{-1/2}\sum_{i=1}^{n}\frac{D_{ni}(t)}{\Gamma_{ni}(t)}f'(t)\left(X_i^TU\right)^2 - pn^{-1/2}f'(t)E\chi^2(\varepsilon_1)\right| = o_P(1). \tag{4.26}$$

This completes the proof of (4.20).

Proof of (4.26). This follows from (4.25) and

$$E\left|n^{-1/2}\sum_{i=1}^{n}\left(X_i^TU\right)^2 - pn^{-1/2}E\chi^2(\varepsilon_1)\right|^2 = o(1). \qquad \square$$

5. *M*-estimators in high-dimensional linear models. In this section we show that M-estimators fulfil the conditions of Section 2 under reasonable assumptions. We start with the special case of the least squares estimator $\hat{\theta}_{\mathrm{LS}}$. We assume (A1), (A2) and the following:

(A3′) $$p^2/n = o\left(n^{1/5}(\log n)^{-28/5}\right);$$

(A4′) ε_i has a finite Laplace transform: $E(\exp(t\varepsilon_i)) < \infty$ for $|t|$ small enough, and $E(\varepsilon_i) = 0$.

The next theorem states that, under these assumptions, least squares estimates fulfil the conditions of Theorem 1. (In this case we have no asymptotic bias; $b = 0$.)

THEOREM 2. *Assume* (A1), (A2), (A3′) *and* (A4′). *Then*

$$\sup_{1\le i\le n}|X_i^T\hat{\theta}_{\mathrm{LS}} - X_i^T\theta| = O_P\left(\sqrt{(p/n)\log n}\right). \tag{5.1}$$

Furthermore, there exists a sequence $\kappa_n \to \infty$ *such that*

$$\sup\left\{\left|X_i^T\left(\hat{\theta}_{\mathrm{LS}}(t) - \hat{\theta}_{\mathrm{LS}}\right)\right| : 1 \le i \le n,\ t \in T_{n,C}\right\} = o_P(n^{-1/2}) \tag{5.2}$$

for every $0 < C < +\infty$.

Here $\hat{\theta}_{\mathrm{LS}}(t)$ *is defined as*

$$\theta + \sum_{i \in I(t)} X_i \varepsilon_i + \sum_{i \notin I(t)} X_i E\left[\varepsilon_i \middle| |\varepsilon_i - t| < \sqrt{(p/n)}\,(\log n)^2\right];$$

$T_{n,C}$ *and* $I(t)$ *are defined as in Theorem* 1; *in particular, because* $b = 0$, *we have* $I(t) = \{i\colon |\varepsilon_i - t| \ge \sqrt{(p/n)}\,(\log n)^2\}$ *now.*

Now we come to the case of arbitrary M-estimators $\hat{\theta}_\psi$ defined to be a solution of an M-equation

$$\sum_{i=1}^{n} X_i \psi\left(Y_i - X_i^T \hat{\theta}_\psi\right) = 0. \tag{5.3}$$

We now assume (A1), (A2) and the following:

(A3″) $\qquad p^2/n = o(n^{-\delta + 1/14})$ for a $\delta > 0$;

(A4″) ψ is a bounded function with three bounded derivatives and with $E(\psi(\varepsilon_i)) = 0$.

The next theorem states that under these assumptions M-estimators fulfil the conditions of Theorem 1.

THEOREM 3. *Assume* (A1), (A2), (A3″) *and* (A4″). *Set* $\chi(x) = \psi(x)/E\psi'(\varepsilon_1)$. *Then there exist a solution* $\hat{\theta}_\psi$ *of* (5.3) *and a* $b = b_n \in \mathbb{R}^p$ *fulfilling* $\|b\| = O(\sqrt{p^2/n})$ *and*

$$\sup_{1 \le i \le n} \left|X_i^T \hat{\theta}_\psi - X_i^T(\theta + b)\right| = O_P\left(\sqrt{(p/n)\log n}\right), \tag{5.4}$$

$$\left\|\hat{\theta}_\psi - \sum_{i=1}^{n} X_i \chi(\varepsilon_i)\right\| = O_P\left(\sqrt{p^2/n}\right). \tag{5.5}$$

(5.6) *Furthermore, there exists a* $\kappa_n \to \infty$ *and, for all* $t \in \kappa_n^{-1} n^{-1/2} \mathbb{Z}$, *there exists a random variable* $\hat{\theta}_\psi(t)$ *that depends only on* $I(t) = \{i\colon |\varepsilon_i - X_i^T b - t| \ge \sqrt{(p/n)}\,(\log n)^2\}$ *and* ε_i, $i \in I(t)$, *with*

$$\sup\left\{\left|X_i^T\left(\hat{\theta}_\psi(t) - \hat{\theta}_\psi\right)\right| : 1 \le i \le n,\ t \in T_{n,C}\right\} = o_P(n^{-1/2})$$

for every $0 < C < +\infty$.

This generalizes a result of Portnoy (1986), where the finite-dimensional distribution of $\hat{F}_n$ was determined for M-estimators $\hat{\theta}_\psi$ under more restrictive assumptions on the function ψ and the design vectors $X_1, \ldots, X_n$.

6. Proof of Theorem 2. For the proof of (5.1) note first that

$$X_i^T\big(\hat\theta_{\mathrm{LS}} - \theta\big) = \sum_{j=1}^{n} X_i^T X_j \varepsilon_j.$$

Set $\rho = ((n/p)\log(n))^{1/2}$. For $C > 0$ the Markov inequality gives

$$\begin{aligned} P\Big(X_i^T\big(\hat\theta_{\mathrm{LS}} - \theta\big) > C\sqrt{(p/n)\log n}\Big) &\\ \le E\exp\Big(\rho\Big[X_i^T\big(\hat\theta_{\mathrm{LS}} - \theta\big) - C\sqrt{(p/n)\log n}\Big]\Big) &\\ \le n^{-C}\prod_{j=1}^{n} E\exp\Big(\rho\big(X_i^T X_j\big)\varepsilon_j\Big). & \end{aligned}$$

Now, by (A2), $\sup_{1\le i,\, j\le n}|\rho(X_i^T X_j)| = O(\sqrt{(p/n)\log n}) = o(1)$. Hence, with a uniform constant C', we obtain the following bound for the above expression:

$$\begin{aligned} &\le n^{-C}\prod_{j=1}^{n}\Big(1 + C'\rho^2\big(X_i^T X_j\big)^2\Big) \\ &\le n^{-C}\exp\Big(C'\rho^2\sum_{j=1}^{n}\big(X_i^T X_j\big)^2\Big) \\ &\le n^{-C+C'}. \end{aligned}$$

With C chosen large enough, this and the corresponding inequality for lower tail probabilities show (5.1) to hold.

For the proof of (5.2) note first that

$$X_i^T\big(\hat\theta_{\mathrm{LS}}(t) - \hat\theta_{\mathrm{LS}}\big) = \sum_{j=1}^{n}\big(X_i^T X_j\big)\Big[\varepsilon_j - E\big(\varepsilon_j \,\big|\, |\varepsilon_j - t| < \gamma_n\big)\Big]\mathbf{I}\big(|\varepsilon_j - t| < \gamma_n\big),$$

where, as in the latter proof, $\gamma_n = \sqrt{(p/n)}\,(\log n)^2$. The upper and lower tails of the distribution of this expression can be bounded using the Bernstein inequality (see Section 4) with $M = O(\gamma_n p/n)$, $V = O(\gamma_n^3 p/n)$ and $\eta = O(\sqrt{V\log n})$. For the proof of (5.2) note that the supremum in $\sup\{|X_i^T(\hat\theta_{\mathrm{LS}}(t) - \hat\theta_{\mathrm{LS}})|\colon 1 \le i \le n,\ t \in T_{n,C}\}$ is taken over an index set that grows polynomially in n. □

7. Proof of Theorem 3. Claims (5.4) and (5.5) are immediate consequences of results in Mammen (1989). In particular, (5.4) follows from Theorem 1 and Lemma 2 in Mammen (1989). For the proof of (5.5) one can use the stochastic expansion $\hat\theta_1$ of $\hat\theta_\psi$ given in Mammen [(1989), Theorem 1]. It remains to show (5.6). We choose $u_i(t)$ (for $t \in T_{n,C}$) with $|u_i(t) - t_i| < \gamma_n$ such that

$$E\big(\psi\big(\varepsilon_i - X_i^T b\big)\big|\,|\varepsilon_i - t_i| < \gamma_n\big) = \psi\big(u_i(t)\big).$$

As in the proof of Th$n = \sqrt{(p/n)}\,(\log n)^2$ and $t_i = t_i(t) = t + X_i^T b$.

We will show that for every $t \in T_{n,C}$ there exists a solution $\hat{\theta}_\psi(t)$ of the following equation which fulfils (5.6):

$$\sum_{i \in I(t)} X_i \psi\Big(\varepsilon_i - X_i^T\big[\hat{\theta}_\psi(t) - \theta\big]\Big) + \sum_{i \notin I(t)} X_i \psi\Big(u_i(t) - X_i^T\big[\hat{\theta}_\psi(t) - \theta - b\big]\Big) = 0.$$

To verify (5.6) we approximate $\hat{\theta}_\psi(t)$ by a variable $\hat{\tau}_\psi(t)$ given by an explicit formula. The variable $\hat{\tau}_\psi(t)$ is defined as

$$\hat{\tau}_\psi(t) = \hat{\theta}_\psi + A^{-1}\Bigg[-\sum_{i=1}^n X_i\big(\psi(\tilde{\varepsilon}_i) - \psi(u_i(t))\big)\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big) - \sum_{i,k=1}^n X_i X_i^T\big(\psi'(\tilde{\varepsilon}_i) - \psi'(u_i(t))\big)\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big) X_k\, \chi(\varepsilon_k)\Bigg],$$

where A is the matrix $A = \sum_{i=1}^n X_i X_i^T E\psi'(\tilde{\varepsilon}_i)$ and $\tilde{\varepsilon}_i = \varepsilon_i - X_i^T b$. Note that $\tilde{\varepsilon}_i - t = \varepsilon_i - t_i$.

We show, for $C > 0$,

$$\sup_{1 \le i \le n,\, t \in T_{n,C}} \Big|X_i^T\Big(\hat{\theta}_\psi - \hat{\tau}_\psi(t)\Big)\Big| = O_P\big(p^{5/4} n^{-5/4} (\log n)^{7/2}\big), \tag{7.1}$$

$$\sup_{t \in T_{n,C}} \big\|\hat{\theta}_\psi(t) - \hat{\tau}_\psi(t)\big\| = o_P\big(p^{-1/2}\big). \tag{7.2}$$

Now $p^{5/4} n^{-5/4} (\log n)^{7/2} = o(n^{-1/2})$ and $\sup_{1 \le i \le n} \|X_i\| = O(p^{1/2} n^{-1/2})$. Consequently, (7.1) and (7.2) imply (5.6).

Proof of (7.1). It suffices to show the following, with $\tilde{X}_j = A^{-1} X_j$:

$$\begin{aligned} &\sup_{1 \le j \le n,\, t \in T_{n,C}} \Big|\tilde{X}_j^T \sum_{i=1}^n X_i\big[\psi(\tilde{\varepsilon}_i) - \psi(u_i(t))\big]\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big)\Big| \\ &\qquad = O_P\big(p^{5/4} n^{-5/4} (\log n)^{7/2}\big); \end{aligned} \tag{7.3}$$

$$\begin{aligned} &\sup_{1 \le j \le n,\, t \in T_{n,C}} \Big|\tilde{X}_j^T \sum_{i,k=1}^n X_i\big(X_i^T X_k\big) \\ &\qquad \times\big[E\big(\psi'(\tilde{\varepsilon}_i)\,\big|\,|\tilde{\varepsilon}_i - t| < \gamma_n\big) - \psi'(u_i(t))\big] \\ &\qquad \times\big[\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big) - E\big[\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big)\big]\big]\psi(\varepsilon_k)\Big| \\ &\qquad = O_P\big(p^{3/2} n^{-3/2} (\log n)^{5/2} + p^{5/2} n^{-2} (\log n)^4\big); \end{aligned} \tag{7.4}$$

$$\begin{aligned} &\sup_{1 \le j \le n,\, t \in T_{n,C}} \Big|\tilde{X}_j^T \sum_{i,k=1}^n X_i\big(X_i^T X_k\big) E\big(\psi'(\tilde{\varepsilon}_i)\,\big|\,|\tilde{\varepsilon}_i - t| < \gamma_n\big) \\ &\qquad - \psi'(u_i(t)) E\big[\mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big)\big]\psi(\varepsilon_k)\Big| \\ &\qquad = O_P\big(p^{3/2} n^{-3/2} (\log n)^{9/2}\big); \end{aligned} \tag{7.5}$$

$$\sup_{1\le j\le n,\, t\in T_{n,C}} \Bigg| \tilde{X}_j^T \sum_{i,k=1}^{n} X_i\big(X_i^T X_k\big)\big[\psi'(\tilde{\varepsilon}_i) - E\big(\psi'(\tilde{\varepsilon}_i)\,\big|\,|\tilde{\varepsilon}_i - t| < \gamma_n\big)\big] \times \mathbf{I}\big(|\tilde{\varepsilon}_i - t| < \gamma_n\big)\psi(\varepsilon_k)\Bigg| \tag{7.6}$$

$$= O_P\big(p^{3/2}n^{-3/2}(\log n)^{5/2} + p^{5/2}n^{-2}(\log n)^4\big).$$

Claims (7.3) and (7.5) can be proved by application of the Bernstein inequality. For upper estimates of the quadratic forms in (7.4) and (7.6), one can proceed as in the proof of (3.16) in Mammen (1989): one applies the Markov inequality and uses the bounds for the Laplace transform of quadratic forms given in Mammen (1989).

Proof of (7.2). Set

$$G_t(\tau) = \sum_{i\in I(t)} X_i\psi\big(\varepsilon_i - X_i^T[\tau - \theta]\big) + \sum_{i\notin I(t)} X_i\psi\big(u_i(t) - X_i^T[\tau - \theta - b]\big).$$

Then $\hat{\theta}_\psi(t)$ is a solution of $G_t(\tau) = 0$. We give (7.7)–(7.9) (see below). Equation (7.7) implies that $\|G_t(\hat{\tau}_\psi(t))\| = o_P(p^{-1/2})$. Equation (7.8) shows that the matrix $G_t'(\hat{\tau}_\psi(t))$ is nondegenerate. Furthermore, in (7.9) we give bounds for the norm of the trilinear form $G_t''(\theta)$ in a neighborhood of $\hat{\tau}_\psi(t)$. The Newton–Kantorowitsch theorem (see below) implies that under these conditions there exists a solution $\hat{\theta}_\psi(t)$ of $G_t(\tau) = 0$ fulfilling (7.2);

$$\sup_{t\in T_{n,C}} \Bigg\| \sum_{i\in I(t)} X_i\psi\big(\varepsilon_i - X_i^T\big[\hat{\tau}_\psi(t) - \theta\big]\big) + \sum_{i\notin I(t)} X_i\psi\big(u_i(t) - X_i^T\big[\hat{\tau}_\psi(t) - \theta - b\big]\big)\Bigg\| \tag{7.7}$$

$$= o_P\big(p^{-1/2}\big),$$

$$\sup_{t\in T_{n,C}} \Bigg\| \sum_{i\in I(t)} X_i X_i^T\psi'\big(\varepsilon_i - X_i^T\big[\hat{\tau}_\psi(t) - \theta\big]\big) + \sum_{i\notin I(t)} X_i X_i^T\psi'\big(u_i(t) - X_i^T\big[\hat{\tau}_\psi(t) - \theta - b\big]\big) - I_p E\psi'(\varepsilon_1)\Bigg\| \tag{7.8}$$

$$= o_P(1)$$

and

$$\sup_{t,\tilde{\tau},e,f,g} \Bigg| \sum_{i\in I(t)} \big(X_i^T e\big)\big(X_i^T f\big)\big(X_i^T g\big)\psi''\big(\varepsilon_i - X_i^T[\tilde{\tau} - \theta]\big) + \sum_{i\notin I(t)} \big(X_i^T e\big)\big(X_i^T f\big)\big(X_i^T g\big)\psi''\big(u_i(t) - X_i^T[\tilde{\tau} - \theta]\big)\Bigg| \tag{7.9}$$

$$= o_P(1) \quad \forall\, d > 0.$$

The supremum in (7.9) runs over all t in $T_{n,C}$, over all $\tilde{\tau}$ with $\|\tilde{\tau} - \hat{\tau}_\psi(t)\| \le d$ and over all vectors e, f and g with unit norm $\|e\| = 1$, $\|f\| = 1$ and $\|g\| = 1$.

We now cite a version of the Newton–Kantorowitsch theorem that will be used here [Kantorowitsch and Akilow (1964)].

NEWTON–KANTOROWITSCH THEOREM. *For a point $x_0 \in \mathbb{R}^p$ and a constant $r > 0$, assume that a function $G\colon \mathbb{R}^p \to \mathbb{R}^p$ has two continuous derivatives for x with $\|x - x_0\| \le r$. Furthermore, assume that $\Gamma = (G'(x_0))^{-1}$ exists and that for some constants $\lambda, \eta > 0$ the following inequalities hold: $\|\Gamma G(x_0)\| \le \eta$, $\|\Gamma G''(x)\| \le \lambda$ for $\|x - x_0\| \le r$, $h = \lambda \cdot \eta \le \frac{1}{2}$, $r_0 = [(1 - \sqrt{1 - 2h})/h]\eta \le r$. Then the equation $G(x) = 0$ has a solution x^* with $\|x^* - x_0\| \le r_0$.*

This theorem will be applied with $G = G_t$, for $t \in T_{n,C}$. Equations (7.7)–(7.9) show that the assumptions are fulfilled with r, λ and η independent of $t \in T_{n,C}$. This proves (7.3).

Now we come to the proof of (7.7)–(7.9). Claim (7.9) can be shown using the Cauchy–Schwarz inequality and the assumption that ψ'' is bounded. For the proof of (7.7) and (7.8) we use the following two lemmas.

LEMMA 1. *For every triangular area of independent random variables $Z_{n,1}, \ldots, Z_{n,n}$ with $\sup_{1 \le i \le n,\, n \ge 1} E|Z_{n,i}|^{10} < +\infty$ and $EZ_{n,i} = 0$, one has*

$$\lambda_{\mathrm{amax}}\left(\sum_{i=1}^{n} X_i X_i^T Z_{n,i} \right) = O_P\left(p^{3/5} n^{-1/2} \right),$$

where $\lambda_{\mathrm{amax}}(B)$ denotes the maximal absolute eigenvalue of a matrix B.

LEMMA 2. *For $t \in T_{n,C}$ we consider triangular areas of random variables $Z_{n,1}(t), \ldots, Z_{n,n}(t)$ that are uniformly bounded, that is,*

$$\sup_{1 \le i \le n,\, n \ge 1,\, t \in T_{n,C}} \left| Z_{n,i}(t) \right| < +\infty \quad (a.s.).$$

For $t \in T_{n,C}$ we define the matrix $B = B_n(t) = \sum_{i=1}^{n} X_i X_i^T \mathbf{I}(|\tilde{\varepsilon}_i - t| < \gamma_n) Z_{n,i}(t)$. Then, for every $\varepsilon > 0$, it holds that

$$\sup_{t \in T_{n,C}} \lambda_{\mathrm{amax}}(B) = O_P\left(p^{5/8} n^{-1/2} n^{\varepsilon} \right).$$

PROOF OF LEMMA 1. It suffices to show

$$E\,\mathrm{trace}(D^{10}) = O\left(p^6 n^{-5} \right),$$

where $D = \sum_{i=1}^{n} X_i X_i^T Z_{n,i}$. This follows from the evaluation of

$$E\,\mathrm{trace}(D^{10}) = E \sum_{i_1, \ldots, i_{10} = 1}^{n} \left(X_{i_1}^T X_{i_2} \right)\left(X_{i_2}^T X_{i_3} \right) \cdots \left(X_{i_{10}}^T X_{i_1} \right) Z_{n,i_1} \cdots Z_{n,i_{10}}. \quad \square$$

PROOF OF LEMMA 2. It holds that

$$\sup_{e,t} |e^T B e| \le \sup_{1 \le i \le n,\, n \ge 1,\, t \in T_{n,C}} \left| Z_{n,i}(t) \right| \sup_{e,t} \left(e^T \sum_{i=1}^{n} X_i X_i^T \mathbf{I}\left(|\tilde{\varepsilon}_i - t| < \gamma_n \right) e \right),$$

where the supremum $\sup_{e,t}$ runs over all vectors e in $\mathbb{R}^p$ with unit norm $\|e\| = 1$ and all t in $T_{n,C}$. Therefore, without loss of generality, we can assume that $Z_{n,i} \equiv 1$, for $1 \le i \le n$, $n \ge 1$.

For every even integer J and for every $t \in T_{n,C}$ we show

$$E\,\mathrm{trace}(B^J) = O\Big(\Big[(\log n)^2 p^{5/8} n^{-1/2}\Big]^J p\Big), \tag{7.10}$$

where $B = \sum_{i=1}^n X_i X_i^T I_i$, $I_i = I(|\tilde{\varepsilon}_i - t| < \gamma_n)$. With J large enough this implies the statement of Lemma 2 because the number of elements of $T_{n,C}$ has polynomial growth. □

Proof of (7.10). It holds that

$$\mathrm{trace}(B^J) = \sum_{i_1,\dots,i_J} \big(X_{i_1}^T X_{i_2}\big)\big(X_{i_2}^T X_{i_3}\big)\cdots\big(X_{i_J}^T X_{i_1}\big) I_{i_1}\cdots I_{i_J}.$$

Now we divide the summation region $\{1,\dots,n\}^J$ into pairwise disjoint regions I such that every region I is defined by specifying for each pair $j \ne j'$ if $i_j = i_{j'}$ holds for all $i = (i_1,\dots,i_J)$ in I or if $i_j \ne i_{j'}$ holds for all i in I. Consider now such a region I of $i = (i_1,\dots,i_J)$. Denote the number of different indices in I by R, that is, $R = \#\{i_1,\dots,i_J\}$ for $i \in I$. We denote the different indices by $k_1,\dots,k_R$, that is, $\{k_1,\dots,k_R\} = \{i_1,\dots,i_J\}$. To every region I one can associate a pseudograph $\mathscr{G}_I$ with R nodes and J edges. (In contrast to a graph, a pseudograph may contain loops or pairs of vertices connected by more than one eresponds to an index k_r. There exists an edge from node k_r to node k_s if $(k_r,k_s) = (i_j, i_{j+1})$ or $(k_s,k_r) = (i_j, i_{j+1})$ for a j with $0 \le j \le J-1$ (where we set $i_0 = i_J$). [Indices (i_j, i_{j+1}) with $i_j = i_{j+1}$ would correspond to a loop.] The pseudographs are connected. First we treat only sets I with pseudographs $\mathscr{G}_I$ fulfilling the following:

(7.11) For each node there exist at least four edges arriving from another node.

Clearly, (7.11) implies $R \le J/2$. We choose now a cycle (a closed path) $\wp$ which pays exactly one visit to every node in $\mathscr{G}_I$. The Cauchy–Schwarz inequality provides

$$\sum_{i\in I} \big(X_{i_1} X_{i_2}\big)\big(X_{i_2}^T X_{i_3}\big)\cdots\big(X_{i_J}^T X_{i_1}\big) \le S_1^{1/2} S_2^{1/2},$$

where

$$S_1 = \sum_{i\in I} \prod_{(i_{j-1},i_j)\in\wp} \big(X_{i_{j-1}}^T X_{i_j}\big)^2 \quad\text{and}\quad S_2 = \sum_{i\in I} \prod_{(i_{j-1},i_j)\notin\wp} \big(X_{i_{j-1}}^T X_{i_j}\big)^2,$$

and where the notation $i_0 = i_J$ was used again. Using $\sum_{i=1}^n X_i X_i^T = I$, $\sum_{i=1}^n \|X_i\|^2 = p$ and $|X_i^T X_{i'}| = O(p/n)$, for $1 \le i,\ i' \le n$, iteratively one can show

$$S_1 = O\big(p(p/n)^R\big).$$

For an estimate of S_2 note first that, after removing the edges of $\mathcal{P}$, the pseudograph may divide into L components with $R_1, \ldots, R_L$ nodes, where $R_1 + \cdots + R_L = R$. Because of $R_m \geq 2$, for $1 \leq m \leq L$, it holds that $L \leq R/2$. By similar calculations as for S_1, we obtain

$$S_2 = O\left((p/n)^{2(J-2R)} \prod_{m=1}^{L} \left(p(p/n)^{R_m}\right)\right) = O\left((p/n)^{2J-3R} p^{R/2}\right).$$

These bounds give

$$\begin{aligned} \sum_{i \in I} \left(X_{i_1}^T X_{i_2}\right)\left(X_{i_2}^T X_{i_3}\right) \cdots \left(X_{i_J}^T X_{i_1}\right) &= O\left((p/n)^{J-R} p^{1+R/4}\right) \\ &= O\left((p/n)^{J-R} p^{1+J/8}\right). \end{aligned} \tag{7.12}$$

A simple calculation shows that equation (7.12) holds also for the case of $R = 1$. Now we consider sets not fulfilling (7.11). Then there exists a node that is connected with other nodes by only two edges. This means that there exist integers $j(1)$ and $j(2)$ with $j(1) \leq j(2)$ such that we obtain for all i in I that $i_j = i_{j(1)}$ if and only if $j(1) \leq j \leq j(2)$. For an i in I, consider

$$\begin{aligned} &\sum_{r \notin R(i)} X_r \left(X_r^T X_r\right)^{j(2)-j(1)+1} X_r^T \\ &\quad = (p/n)^{j(2)-j(1)} \Gamma_n - \sum_{r \in R(i)} X_r \left(X_r^T X_r\right)^{j(2)-j(1)} X_r^T, \end{aligned} \tag{7.13}$$

where $R(i)$ is the set $R(i) = \{i_1, \ldots, i_{j(1)-1}, i_{j(2)+1}, \ldots, i_J\}$; Γ_n is a matrix with uniformly absolutely bounded eigenvalues. Without loss of generality, one can suppose $\Gamma_n = I_p$. Formula (7.13) can be plugged into $S = \sum_{i \in I} (X_{i_1}^T X_{i_2})(X_{i_2}^T X_{i_3}) \cdots (X_{i_J}^T X_{i_1})$ repeatedly. Each application replaces S by a sum of $[1 + \#R(i)]$ terms S_{NEW} that are of the same type as S. The first term is multiplied by a factor $F = (p/n)^{j(2)-j(1)}$. For the first term the old quantities J and R are replaced by $J_{\mathrm{NEW}} = J_{\mathrm{OLD}} - (j(2) - j(1) + 1)$ and $R_{\mathrm{NEW}} = R_{\mathrm{OLD}} - 1$. Note that $F(p/n)^{J_{\mathrm{NEW}} - R_{\mathrm{NEW}}} = (p/n)^{J_{\mathrm{OLD}} - R_{\mathrm{OLD}}}$. For the other terms we obtain $J_{\mathrm{NEW}} = J_{\mathrm{OLD}}$ and $R_{\mathrm{NEW}} = R_{\mathrm{OLD}} - 1$, that is, $(p/n)^{J_{\mathrm{NEW}} - R_{\mathrm{NEW}}} < (p/n)^{J_{\mathrm{OLD}} - R_{\mathrm{OLD}}}$. After repeated use of (7.13) we arrive at terms S_{NEW} fulfilling (7.11) or $R_{\mathrm{NEW}} = 1$. This shows for $S = S_{\mathrm{OLD}}$, $J = J_{\mathrm{OLD}}$ and $R = R_{\mathrm{OLD}}$ that (because of $R_{\mathrm{NEW}} \leq J/2$)

$$\begin{aligned} \sum_{i \in I} \left(X_{i_1}^T X_{i_2}\right)\left(X_{i_2}^T X_{i_3}\right) \cdots \left(X_{i_J}^T X_{i_1}\right) &= O\left((p/n)^{J-R} p^{1+R_{\mathrm{NEW}}/4}\right) \\ &= O\left((p/n)^{J-R} p^{1+J/8}\right). \end{aligned}$$

Now the application of $(p/n)^{J-R} \gamma_n^R = O((\log n)^{2J} (p/n)^{J/2})$, for $1 \leq R \leq J$, yields (7.10).

Next we prove (7.7) and (7.8).

Proof of (7.8). Because $\sup_{1 \leq i \leq n,\, t \in T_{n,C}} (|u_i(t) - \varepsilon_i|, |X_i^T[\hat{\tau}_\psi(t) - \theta]|) = o_P(1)$ and (A1), it suffices to show $\|\sum_{i=1}^n X_i X_i^T (\psi'(\varepsilon_i) - E\psi'(\varepsilon_i))\| = o_P(1)$. This follows from Lemmis follows from Lemma 1.

Proof of (7.7). The left-hand side of (7.7) is the norm of the following expression:

$$\begin{aligned}
&\sum_{i\in I(t)} X_i\psi\big(\varepsilon_i - X_i^T\big[\hat\tau_\psi(t) - \theta\big]\big) + \sum_{i\notin I(t)} X_i\psi\big(u_i(t) - X_i^T\big[\hat\tau_\psi(t) - \theta - b\big]\big)\\
&\quad= \sum_{i=1}^n X_i\Big[\psi\big(\varepsilon_i - X_i^T\big[\hat\tau_\psi(t) - \theta\big]\big) - \psi\big(\varepsilon_i - X_i^T\big[\hat\theta_\psi - \theta\big]\big)\Big]\\
&\qquad+ \sum_{i\notin I(t)} X_i\Big[\psi\big(u_i(t) - X_i^T\big[\hat\tau_\psi(t) - \theta - b\big]\big)\\
&\qquad\qquad - \psi\big(\varepsilon_i - X_i^T\big[\hat\tau_\psi(t) - \theta\big]\big)\Big]\\
&\quad= V_1(t) + \cdots + V_9(t),
\end{aligned}$$

where

$$\begin{aligned}
V_1(t) &= \sum_{i=1}^n X_iX_i^TE\psi'(\tilde\varepsilon_i)\Big[\hat\theta_\psi - \hat\tau_\psi(t)\Big],\\
V_2(t) &= \sum_{i=1}^n X_iX_i^T\big[\psi'(\tilde\varepsilon_i) - E\psi'(\tilde\varepsilon_i)\big]\Big[\hat\theta_\psi - \hat\tau_\psi(t)\Big],\\
V_3(t) &= \sum_{i=1}^n X_iX_i^T\Big[\psi'\big(\varepsilon_i - X_i^T\big[\hat\theta_\psi - \theta\big]\big) - \psi'(\tilde\varepsilon_i)\Big]\Big[\hat\theta_\psi - \hat\tau_\psi(t)\Big],\\
V_4(t) &= \sum_{i=1}^n X_i\psi''(\varepsilon_i^*)\Big[X_i^T\big(\hat\theta_\psi - \hat\tau_\psi(t)\big)\Big]^2,
\end{aligned}$$

for an ε_i^* lying between $\varepsilon_i - X_i^T\hat\tau_\psi(t)$ and $\varepsilon_i - X_i^T\hat\theta_\psi$,

$$\begin{aligned}
V_5(t) &= \sum_{i\notin I(t)} X_i\big[\psi(u_i(t)) - \psi(\tilde\varepsilon_i)\big],\\
V_6(t) &= \sum_{i\notin I(t)} X_i\big[\{\psi'(u_i(t) + \delta_i) - \psi'(\tilde\varepsilon_i + \delta_i)\}\\
&\qquad\qquad - \{\psi'(u_i(t)) - \psi'(\tilde\varepsilon_i)\}\big]X_i^T\big(\hat\tau_\psi(t) - \theta - b\big),
\end{aligned}$$

for a δ_i lying between 0 and $X_i^T(\hat\tau_\psi(t) - \theta - b)$,

$$\begin{aligned}
V_7(t) &= \sum_{i\notin I(t)} X_iX_i^T\big[\psi'(u_i(t)) - \psi'(\tilde\varepsilon_i)\big]\Big[\hat\tau_\psi(t) - \hat\theta_\psi\Big],\\
V_8(t) &= \sum_{i\notin I(t)} X_iX_i^T\big[\psi'(u_i(t)) - \psi'(\tilde\varepsilon_i)\big]\Big[\hat\theta_\psi - \theta - b - \sum_{j=1}^n X_j\chi(\varepsilon_j)\Big],\\
V_9(t) &= \sum_{i\notin I(t)} X_iX_i^T\big[\psi'(u_i(t)) - \psi'(\tilde\varepsilon_i)\big]\sum_{j=1}^n X_j\chi(\varepsilon_j).
\end{aligned}$$

By definition of $\hat\tau_\psi(t)$ it holds that

$$V_1(t) + V_5(t) + V_9(t) = 0.$$

It remains to show

$$\sup_{t \in T_{n,C}} \|V_j(t)\| = o_P(p^{-1/2}) \qquad \text{for } j \neq 1, 5, 9. \tag{7.14}$$

This can be shown by calculations that we indicate briefly. For $j = 2$, (7.14) follows from Lemma 1 and

$$\|\hat{\theta}_\psi - \hat{\tau}_\psi(t)\| = O_P(p^{5/4} n^{-3/4} (\log n)^{7/2}). \tag{7.15}$$

Equation (7.15) is an immediate consequence of (7.1). For $j = 3$ one applies (7.15) and (5.4). For $j = 4$, (7.14) follows from (7.15), (5.4) and (A1). For $j = 6$, one applies Lemma 2 with

$$Z_{n,i} = \left[\left(\psi'(u_i(t) + \delta_i) - \psi'(\tilde{\varepsilon}_i + \delta_i)\right) - \left(\psi'(u_i(t)) - \psi'(\tilde{\varepsilon}_i)\right)\right][\delta_i \gamma_n]^{-1}.$$

Because of (7.15) and (5.4), we have here

$$\sup_{1 \le i \le n} |\delta_i| = O_P(p^{1/2} n^{-1/2} (\log n)^{1/2}).$$

For $j = 7$ and 8, one uses (7.15) (for $j = 7$) or (5.5) (for $j = 8$), Lemma 2 and

$$\sup_{1 \le i \le n,\, t \in T_{n,C}} |\psi'(u_i(t)) - \psi'(\tilde{\varepsilon}_i)| = O_P(p^{1/2} n^{-1/2} (\log n)). \qquad \square$$

Acknowledgments. The comments of the referees resulted in a significant improvement over a previous version of the paper. The author would also like to thank one of the referees for pointing out many inaccuracies.

REFERENCES

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

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

HABERMAN, S. J. (1977a). Log-linear 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.

HOEFFDING (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

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

IOANNIDES, E. (1987). Asymptotik des empirischen Prozesses der Residuen und Schätzen der Form der Fehlerverteilung im linearen Regressionsmodell. Diplomarbeit, Fakultät für Mathematik, Univ. Heidelberg.

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

KOUL, H. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40** 1950–1979.

KOUL, H. (1970). Some convergence theorems for ranks and weighted empirical cumulatives. *Ann. Math. Statist.* **41** 1768–1773.

KOUL, H. (1984). Tests of goodness-of-fit in linear regression. *Colloq. Math. Soc. János Bolyai* **45** 279–315.

KOUL, H. (1992). *Weighted Empiricals and Linear Models*. IMS, Hayward, CA.
KREISS, J.-P. (1988). Asymptotic statistical inference for a class of stochastic processes. Habilitationsschrift, Fachbereich Mathematik, Univ. Hamburg.
KREISS, J.-P. (1991). Estimation of the distribution function of noise in stationary processes. *Metrika* **38** 285–297.
LOYNES, R. M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.* **8** 285–298.
MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400.
MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285.
MO, M. (1991). Asymptotic normality of minimum contrast estimators. Unpublished manuscript, Dept. Statistics, Univ. Toronto.
MO, M. (1992). Bootstrapping with increasing dimension. Unpublished manuscript, Dept. Statistics, Univ. Toronto.
POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
PORTNOY, S. (1984). Asymptotic behaviour of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309.
PORTNOY, S. (1985). Asymptotic behaviour of M-estimator of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417.
PORTNOY, S. (1986). Asymptotic behaviour of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Statist.* **14** 1152–1170.
PORTNOY, S. (1988). Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.
SAUERMANN, W. (1989). Bootstrapping the maximum likelihood estimator in high-dimensional log-linear models. *Ann. Statist.* **17** 1198–1216.
SHORACK, G. R. (1982). Bootstrapping robust regression. *Comm. Statist. Theory Methods* **11** 961–972.
SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
WELSH, A. H. (1989). On M-processes and M-estimation *Ann. Statist.* **17** 337–361.
WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

INSTITUT FÜR ANGEWANDTE MATHEMATIK
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY