Annals of the Institute of Statistical Mathematics

Inference about the slope in linear regression: an empirical likelihood approach

Ursula U. Müller · Hanxiang Peng ·
Anton Schick

November 8, 2017

Abstract We present a new, efficient maximum empirical likelihood estimator for the slope in linear regression with independent errors and covariates. The estimator does not require estimation of the influence function, in contrast to other approaches, and is easy to obtain numerically. Our approach can also be used in the model with responses missing at random, for which we recommend a complete case analysis. This suffices thanks to results by Müller and Schick (2017), which demonstrate that efficiency is preserved. We provide confidence intervals and tests for the slope, based on the limiting chi-square distribution of the empirical likelihood, and a uniform expansion for the empirical likelihood ratio. The article concludes with a small simulation study.

Keywords efficiency · estimated constraint functions · infinitely many constraints · maximum empirical likelihood estimator · missing responses · missing at random

U.U. Müller
Texas A&M University, Department of Statistics, College Station, TX 77843-3143, USA
E-mail: uschi@stat.tamu.edu

H. Peng
Indiana University Purdue University at Indianapolis, Department of Mathematical Sciences, Indianapolis, IN 46202-3267, USA
E-mail: hpeng@math.iupui.edu

A. Schick
Binghamton University, Department of Mathematical Sciences, Binghamton, NY 13902-6000, USA
E-mail: anton@math.binghamton.edu

1 Introduction

We consider the homoscedastic regression model in which the response variable Y is linked to a covariate X by the formula

$$Y = \beta X + \varepsilon. \tag{1}$$

For reasons of clarity we focus on the case where X is one-dimensional and β an unknown real number. We will assume throughout that ε and X are independent, and that X has a finite positive variance. Our goal is to make inferences about the slope β, treating the density f of the error ε and the distribution of the covariate X as nuisance parameters. We shall do so by using an empirical likelihood approach based on independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of the base observation (X, Y).

Model (1) is the usual linear regression model with a non-zero intercept, even though it is written without an explicit intercept parameter. Since we do not assume that the error variable is centered, the mean $E[\varepsilon]$ plays the role of the intercept parameter. Working with this model and notation simplifies the explanation of the method and the presentation of the proofs. The generalization to the multivariate case is straightforward; see Remark 1 in Section 2.

The linear regression model is one of the most useful statistical models, and many simple estimators for the slope are available, such as the ordinary least squares estimator (OLSE) which takes on the form

$$\frac{\sum_{j=1}^n (X_j - \bar{X}) Y_j}{\sum_{j=1}^n (X_j - \bar{X})^2} \tag{2}$$

rather than $\sum_{j=1}^n X_j Y_j / \sum_{j=1}^n X_j^2$, because we do not assume that the errors are centered. However, these estimators are usually inefficient. The construction of *efficient* (least dispersed) estimators is in fact quite involved. The reason for this is the assumed independence between covariates and errors, which is a structural assumption that has to be taken into account by the estimator to obtain efficiency. Efficient estimators for β in model (1) were first introduced by Bickel (1982), who used sample splitting to estimate the efficient influence function. To establish efficiency we must assume that f has finite Fisher information for location. This means that f is absolutely continuous and the integral $J_f = \int \ell_f^2(y) f(y)\, dy$ is finite, where $\ell_f = -f'/f$ denotes the score function for location. It follows from Bickel (1982) that an efficient estimator $\hat{\beta}$ of β is characterized by the stochastic expansion

$$\hat{\beta} = \beta + \frac{1}{n} \sum_{j=1}^n \frac{(X_j - E[X]) \ell_f (Y_j - \beta X_j)}{J_f \operatorname{Var}(X)} + o_P(n^{-1/2}). \tag{3}$$

Further efficient estimators of the slope which require estimating the influence function were proposed by Schick (1987) and Jin (1992). Koul and Susarla

(1983) studied the case when f is also symmetric about zero. See also Schick (1993) and Forrester et al. (2003), who achieved efficiency without sample splitting and instead used a conditioning argument. Efficient estimation in the corresponding (heteroscedastic) model *without* the independence assumption (defined by $E(\varepsilon|X) = 0$) is much easier: Müller and Van Keilegom (2012), for example, proposed weighted versions of the OLSE to efficiently estimate β in the model with fully observed data and in a model with missing responses. See also Schick (2013), who proposed an efficient estimator using maximum empirical likelihood with infinitely many constraints.

Like Müller and Van Keilegom (2012), we are interested in the common case that responses are *missing at random* (MAR). This means that we observe copies of the triplet $(\delta, X, \delta Y)$, where δ is an indicator variable with $\delta = 1$ if Y is observed, and where the probability π that Y is observed depends only on the covariate,

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X),$$

with $E[\pi(X)] = E[\delta] > 0$; we refer to the monographs by Little and Rubin (2002) and Tsiatis (2006) for further reading. Note that the "MAR model" we have just described covers the "full model" (in which all data are completely observed) as a special case with $\pi(X) = 1$. To estimate β in the MAR model we propose a complete case analysis, i.e., only the $N = \sum_{j=1}^{n} \delta_j$ observations $(X_{i_1}, Y_{i_1}), \ldots, (X_{i_N}, Y_{i_N})$ with observed responses will be considered.

Complete case analysis is the simplest approach to dealing with missing data, and is frequently disregarded as naive and wasteful. In our application, however, the contrary is true: Müller and Schick (2017) showed that general functionals of the conditional distribution of Y given X can be estimated efficiently (in the sense of Hájek and Le Cam) by a complete case analysis. Since the slope β is covered as a special case, this means that an estimator of β that is efficient in the full model is also efficient in the MAR model if we simply omit the incomplete cases. This property is called "efficiency transfer". To construct efficient maximum empirical likelihood estimators for β, it therefore suffices to consider the model with completely observed data. We write $\hat{\beta}_c$ for the complete case version of $\hat{\beta}$ from (3). It follows from the *transfer principle for asymptotically linear statistics* by Koul et al. (2012) that $\hat{\beta}_c$ satisfies

$$\hat{\beta}_c = \beta + \frac{1}{N} \sum_{j=1}^{n} \frac{\delta_j (X_j - E[X|\delta = 1]) \ell_f (Y_j - \beta X_j)}{J_f \operatorname{Var}(X|\delta = 1)} + o_P(n^{-1/2}), \qquad (4)$$

and is therefore consistent for β. That $\hat{\beta}_c$ is also *efficient* follows from Müller and Schick (2017, Section 5.1). The efficiency property can alternatively be deduced from arguments in Müller (2009), who gave the efficient influence function for β in the MAR model, but with the additional assumption that the errors have mean zero; see Lemma 5.1 in that paper.

In this paper we use an empirical likelihood approach with an increasing number of estimated constraints to derive various inferential procedures about the slope. Our approach is similar to Schick (2013), but our model requires different constraints. We obtain a suitable Wilks' theorem (see Theorem 1) to derive confidence sets for β and tests about a specific value of β, and a point estimator of β via maximum empirical likelihood, i.e., by maximizing the empirical likelihood. This estimator is shown to be semiparametrically efficient.

Empirical likelihood was introduced by Owen (1988, 2001) for a *fixed* number of *known* linear constraints to construct confidence intervals in a nonparametric setting. More recently, his results have been generalized to a fixed number of estimated constraints by Hjort et al. (2009), who further studied the case of an increasing number of known constraints; see also Chen et al. (2009). Peng and Schick (2013) generalized the approach to the case of an increasing number of estimated constraints. The idea of maximum empirical likelihood goes back to Qin and Lawless (1994), who treated the case with a fixed number of known constraints. Peng and Schick (2017) generalized their result to the case with estimated constraints. Schick (2013) and Peng and Schick (2016) treated examples with an increasing number of estimated constraints and showed efficiency of the maximum empirical likelihood estimators.

The empirical likelihood is similar to the one considered for the symmetric location model in Peng and Schick (2016). We shall derive results that are analogous to those in that paper. In Section 3 we provide the asymptotic chi-square distribution of the empirical log-likelihood for both the full model and the MAR model. This facilitates the construction of confidence intervals and tests about the slope β. In Section 4 we propose a new method for estimating β efficiently, namely a guided maximum empirical likelihood estimator, as suggested by Peng and Schick (2017) for the general model with estimated constraints. Efficiency of this estimator is entailed by a uniform expansion for the local empirical likelihood (see Theorem 2), which follows from a local asymptotic normality condition. Section 5 contains a simulation study. The proofs are in Section 6.

2 Empirical likelihood approach

The construction of the empirical likelihood is crucial since we need to incorporate the independence between the covariates and the errors to obtain efficiency. Let us explain it for the full model. The corresponding approach for the missing data model is then straightforward: in that case we will proceed in the same way, now with the analysis based on the N complete cases, and with the random sample size N treated like n.

Our empirical likelihood $\mathscr{R}_n(b)$, which we want to maximize with respect to $b \in \mathbb{R}$, is of the form

$$\mathscr{R}_n(b) = \sup\Big\{ \prod_{j=1}^n n\pi_j : \pi \in \mathscr{P}_n,\ \sum_{j=1}^n \pi_j (X_j - \bar{X}) v_n(\mathbb{F}_b(Y_j - bX_j)) = 0 \Big\}.$$

Here $\mathscr{P}_n$ is the probability simplex in dimension n, defined by

$$\mathscr{P}_n = \Big\{ \pi = (\pi_1, \dots, \pi_n)^\top \in [0,1]^n : \sum_{j=1}^n \pi_j = 1 \Big\},$$

$\bar{X}$ is the sample mean of the covariates $X_1, \dots, X_n$, $\mathbb{F}_b$ is the empirical distribution function constructed from 'residuals' $Y_1 - bX_1, \dots, Y_n - bX_n$, i.e.,

$$\mathbb{F}_b(t) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}[Y_j - bX_j \le t], \quad t \in \mathbb{R},$$

which serves as a surrogate for the unknown error distribution F. The function v_n maps from $[0,1]$ into $\mathbb{R}^{r_n}$ and will be described in (6) below. The constraint $\sum_{j=1}^n \pi_j (X_j - \bar{X}) v_n(\mathbb{F}_b(Y_j - bX_j)) = 0$ in the definition of $\mathscr{R}_n(b)$ is therefore a vector of r_n one-dimensional constraints, where the integer r_n tends to infinity slowly as the sample size n increases. These constraints emerge from the independence assumption as follows. Independence of X and ε is equivalent to $E[c(X)a(\varepsilon)] = 0$ for all square-integrable centered functions c and a under the distributions of X and ε, respectively. This leads to the empirical likelihood in Peng and Schick (2013). We do not work with these constraints. Instead we use constraints in the subspace

$$\{(X - E[X])a(\varepsilon) : a \in L_{2,0}(F)\} \tag{5}$$

with $L_{2,0}(F) = \{a \in L_2(F) : \int a\, dF = 0\}$, which suffices since it contains the efficient influence function; see (3). By our assumptions, F is continuous and $F(\varepsilon)$ is uniformly distributed on the interval $[0,1]$, i.e., $F(\varepsilon) \sim \mathscr{U}$. An orthonormal basis of $L_{2,0}(F)$ is $\varphi_1 \circ F, \varphi_2 \circ F, \dots$, where φ_k denotes an orthonormal basis of $L_{2,0}(\mathscr{U})$. This suggests the constraints

$$\sum_{j=1}^n \pi_j \{X_j - E(X)\} \varphi_k\{F(Y_j - bX_j)\} = 0, \quad k = 1, \dots, r_n,$$

which, however, cannot be used since neither F nor the the mean of X are known. So we replace them by empirical estimators. In this article we will work with the trigonometric basis

$$\varphi_k(x) = \sqrt{2}\cos(k\pi x), \quad 0 \le x \le 1, k = 1, 2, \dots,$$

and take

$$v_n = (\varphi_1, \dots, \varphi_{r_n})^\top. \tag{6}$$

This yields our empirical likelihood $\mathscr{R}_n(b)$ from above.

Let us briefly discuss the complete case approach that we propose for the MAR model. In the following a subscript "c" will, as before when we introduced $\hat\beta_c$, indicate that a complete case statistic is used. For example, $\mathbb{F}_{b,c}$ is the complete case version of $\mathbb{F}_b$, i.e.,

$$\mathbb{F}_{b,c}(t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}[Y_j - bX_j \le t] = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}[Y_{i_j} - bX_{i_j} \le t], \quad t \in \mathbb{R}.$$

The complete case empirical likelihood is

$$\mathscr{R}_{n,c}(b) = \sup\Big\{ \prod_{j=1}^{N} N\pi_j : \pi \in \mathscr{P}_N, \sum_{j=1}^{N} \pi_j (X_{i_j} - \bar X_c) v_N(\mathbb{F}_{b,c}(Y_{i_j} - bX_{i_j})) = 0 \Big\},$$

with $\mathscr{P}_N$ and v_n defined above. Note that we perform a complete case analysis, so the above formula must involve $\bar X_c = N^{-1}\sum_{j=1}^n \delta_j X_j$, which is a consistent estimator of the conditional expectation $E[X|\delta = 1]$, as given in (4); see also Section 3 in Müller and Schick (2017) for the general case. Moments of the covariate distribution are replaced by moments of the conditional covariate distribution given $\delta = 1$, when switching from the full model to the complete case analysis.

Remark 1 If the covariate X is a p-dimensional vector we have

$$Y_j = \beta^\top X_j + \varepsilon_j, \quad j = 1, \dots, n,$$

and construct $\mathbb{F}_b$ using the 'residuals' $Y_j - b^\top X_j$. Now we need to interpret (5) with X being p-dimensional. The empirical likelihood $\mathscr{R}_n(b)$ is then

$$\sup\Big\{ \prod_{j=1}^{n} n\pi_j : \pi \in \mathscr{P}_n, \sum_{j=1}^{n} \pi_j (X_j - \bar X) \otimes v_n(\mathbb{F}_b(Y_j - b^\top X_j)) = 0 \Big\},$$

where $\otimes$ denotes the Kronecker product. Since the Kronecker product of two vectors with dimensions p and q is a vector of dimension pq, there are pr_n random constraints in the above empirical likelihood. Working with this likelihood is notationally more cumbersome, but the proofs are essentially the same. The complete case empirical likelihood $\mathscr{R}_{n,c}(b)$ changes analogously. It equals

$$\sup\Big\{ \prod_{j=1}^{N} N\pi_j : \pi \in \mathscr{P}_N, \sum_{j=1}^{N} \pi_j (X_{i_j} - \bar X_c) \otimes v_N(\mathbb{F}_{b,c}(Y_{i_j} - bX_{i_j})) = 0 \Big\}.$$

3 A Wilks' theorem

Wilks' original theorem states that the classical log-likelihood ratio test statistic is asymptotically chi-square distributed. Our first result is a version of that theorem for the empirical log-likelihood. It is given in Theorem 1 below and proved in the first subsection of Section 6. As in the previous section we write $\mathscr{R}_n(b)$ for the empirical likelihood and $\mathscr{R}_{n,c}(b)$ for the complete case empirical likelihood. Further let $\chi_\gamma(d)$ denote the γ-quantile of the chi-square distribution with d degrees of freedom.

Theorem 1 *Consider the full model and suppose that X also has a finite fourth moment and that the number of basis functions r_n satisfies $r_n \to \infty$ and $r_n^4 = o(n)$ as $n \to \infty$. Then we have*

$$P(-2\log \mathscr{R}_n(\beta) \le \chi_u(r_n)) \to u, \quad 0 < u < 1.$$

The conclusion of this theorem is equivalent to $(-2\log\mathscr{R}_n(\beta) - r_n)/\sqrt{r_n}$ being asymptotically standard normal. This implies that the complete case version $(-2\log\mathscr{R}_{n,c}(\beta) - r_N)/\sqrt{r_N}$ is also asymptotically standard normal. This is a consequence of the *transfer principle* for complete case statistics; see Remark 2.4 in the article by Koul et al. (2012). More precisely, these authors showed that if the limiting distribution of a statistic is $\mathcal{L}(Q)$, then the limiting distribution of its complete case version is $\mathcal{L}(\tilde{Q})$, where Q is the joint distribution of (X,Y), belonging to some model, and $\tilde{Q}$ is the distribution of (X,Y) given $\delta = 1$. One only needs to assume that $\tilde{Q}$ belongs to the same model as Q, i.e., it satisfies the same assumptions. Here we assume that the responses are missing at random, i.e., δ and Y are conditionally independent given X. Therefore we only need to require that the conditional covariate distribution given $\delta = 1$ and the unconditional covariate distribution belong to the same model. Here the limiting distribution is not affected as it does not depend on Q.

Although the result for the MAR model is more general than the result for the full model (which is covered as a special case), we can now, thanks to the transfer principle, formulate it as a corollary, i.e., we only need to take the modified assumptions for the conditional covariate distribution into account, and prove Theorem 1 for the full model.

Corollary 1 *Consider the MAR model and suppose that the distribution of X given $\delta = 1$ has a finite fourth moment and a positive variance. Let the number of basis functions r_N satisfy $1/r_N = o_P(1)$ and $r_N^4 = o_P(N)$ as $n \to \infty$. Then we have*

$$P(-2\log \mathscr{R}_{n,c}(\beta) \le \chi_u(r_N)) \to u, \quad 0 < u < 1.$$

Note that the conditions on the number of basis functions r_n and r_N in the full model and the MAR model are equivalent since n and N increase proportionally,

$$\frac{N}{n} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \;\to\; E[\delta] \quad \text{almost surely,}$$

with $E[\delta] > 0$ by assumption.

The distribution of X given $\delta = 1$ has density $\pi/E[\delta]$ with respect to the distribution of X. Thus the variance of the former distribution is positive unless X is constant almost surely on the event $\{\pi(X) > 0\}$.

Remark 2 The above result shows that

$$\{b \in \mathbb{R} : -2 \log \mathscr{R}_{n,c}(b) < \chi_{1-\alpha}(r_N)\}$$

is a $1 - \alpha$ confidence region for β and that

$$\mathbf{1}[-2 \log \mathscr{R}_{n,c}(\beta_0) \geq \chi_{1-\alpha}(r_N)]$$

is a test of asymptotic size α for testing the null hypothesis $H_0 : \beta = \beta_0$. Note that both the confidence region and the test about the slope also apply to the special case of a full model with $N = n$ and $\mathscr{R}_n$ in place of $\mathscr{R}_{n,c}$. The asymptotic confidence interval for the slope, for example, is

$$\{b \in \mathbb{R} : -2 \log \mathscr{R}_n(b) < \chi_{1-\alpha}(r_n)\}.$$

4 Efficient estimation

Our next result gives a strengthened version of the uniform local asymptotic normality (ULAN) condition for the local empirical likelihood ratio

$$\mathscr{L}_n(t) = \log\left(\frac{\mathscr{R}_n(\beta + n^{-1/2}t)}{\mathscr{R}_n(\beta)}\right), \quad t \in \mathbb{R}$$

in the full model. The usual ULAN condition is established for fixed compact intervals for the local parameter t. Here we allow the intervals to grow with the sample size.

Theorem 2 *Suppose X has a finite fourth moment, f has finite Fisher information for location, and r_n satisfies $(\log n)/r_n = O(1)$ and $r_n^5 \log n = o(n)$. Then for every sequence C_n satisfying $C_n \geq 1$ and $C_n^2 = O(\log n)$, the uniform expansion*

$$\sup_{|t| \leq C_n} \frac{|\mathscr{L}_n(t) - t\varGamma_n + J_f \operatorname{Var}(X) t^2/2|}{(1 + |t|)^2} = o_P(1) \tag{7}$$

holds with

$$\varGamma_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - E[X_j]) \ell_f(X_j - \beta X_j),$$

which is asymptotically normal with mean zero and variance $J_f \operatorname{Var}(X)$.

The proof of Theorem 2 is quite elaborate and carried out in Section 6. Expansion (7) is critical to obtain the asymptotic distribution of the maximum empirical likelihood estimator. We shall follow Peng and Schick (2017) and work with a guided maximum empirical likelihood estimator (GMELE). This requires a preliminary $n^{1/2}$-consistent estimator $\tilde{\beta}_n$ of β. One possibility is the OLSE, see (2), which requires the additional assumption that the error has a finite second moment. Another possibility which avoids this assumption is the solution $\tilde{\beta}_n$ to the equation

$$\frac{1}{n}\sum_{j=1}^{n}(X_j-\bar{X})\psi(Y_j-bX_j)=0,$$

where ψ is a bounded function with a positive and bounded first derivative ψ' and a bounded second derivative as, for example, the arctangent. Then

$$\tilde{\beta}_n=\beta-\frac{1}{n}\sum_{j=1}^{n}\frac{(X_j-\mu)(\psi(\varepsilon_j)-E[\psi(\varepsilon)])}{\operatorname{Var}(X)E[\psi'(\varepsilon)]}+o_P(n^{-1/2})$$

and $n^{1/2}(\tilde{\beta}_n-\beta)$ is asymptotically normal with mean zero and variance

$$\frac{\operatorname{Var}(\psi(\varepsilon))}{(E[\psi'(\varepsilon)])^2\operatorname{Var}(X)}.$$

The GMELE associated with a $n^{1/2}$-consistent preliminary estimator $\tilde{\beta}_n$ is defined by

$$\hat{\beta}_n=\underset{n^{1/2}|b-\tilde{\beta}_n|\le C_n}{\arg\max}\ \mathscr{R}_n(b), \tag{8}$$

where C_n is proportional to $(\log n)^{1/2}$. By the results in Peng and Schick (2017) the expansion (7) implies

$$n^{1/2}(\hat{\beta}_n-\beta)=\mathit{\Gamma}_n/(J_f\operatorname{Var}(X))+o_P(n^{-1/2}).$$

Thus, under the assumptions of Theorem 2, the GMELE $\hat{\beta}_n$ satisfies (3) and is therefore efficient. The complete case estimator

$$\hat{\beta}_{n,c}=\underset{N^{1/2}|b-\tilde{\beta}_{n,c}|\le C_N}{\arg\max}\ \mathscr{R}_{n,c}(b)$$

is then efficient in the MAR model, provided the conditional distribution of X given $\delta=1$ has a finite fourth moment and a positive variance. Let us summarize our finding in the following theorem.

Theorem 3 *Suppose that the error density f has finite Fisher information for location and that r_n satisfies $(\log n)/r_n=O(1)$ and $r_n^5\log n=o(n)$.*

(a) Assume that the covariate X has a finite fourth moment and a positive variance. Then the GMELE $\hat{\beta}_n$ satisfies expansion (3) and is therefore efficient in the full model.

(b) Consider the MAR model and assume that given $\delta = 1$ the covariate X has a finite conditional fourth moment and a positive conditional variance. Then the complete case version $\hat{\beta}_{n,c}$ of the GMELE satisfies expansion (4) and is efficient in the MAR model.

The choice of r_n (and r_N) is addressed in Remark 4 in Section 5.

Remark 3 A referee suggested the following. "An alternative (but asymptotically equivalent) procedure to compute the maximum empirical likelihood estimator can be based on the set of the generalized set of estimating equations $g_j(b) = (X_j - \bar{X})v_n(\mathbb{F}_b(Y_j - bX_j))$ (with $r_n > 1$) and the following program, i.e.,

$$\hat{\beta}_n^{EE} = \operatorname*{arg\,min}_{n^{1/2}|b-\tilde{\beta}_n|\le C_n} \frac{1}{n}\sum_{j=1}^n g_j(b)^\top \Big(\frac{1}{n}\sum_{j=1}^n g_j(\overline{\beta}_n)g_j(\overline{\beta}_n)^\top\Big)^{-1}\frac{1}{n}\sum_{j=1}^n g_j(b),$$

where $\overline{\beta}_n$ is a preliminary estimator defined as

$$\overline{\beta}_n = \operatorname*{arg\,min}_{n^{1/2}|b-\tilde{\beta}_n|\le C_n} \frac{1}{n}\sum_{j=1}^n g_j(b)^\top \widehat{W}\frac{1}{n}\sum_{j=1}^n g_j(b)$$

for any positive semi-definite matrix $\widehat{W}$ (and similarly for the complete case analysis). This estimator is computationally simpler than the maximum empirical likelihood estimator, especially if the dimension of β is larger than one."

An even simpler estimator which avoids the preliminary step is the estimator $\overline{\beta}_n$ with $\widehat{W} = (\hat{\tau}_n^2 I_{r_n})^{-1}$, where I_{r_n} is the $r_n \times r_n$ identity matrix and $\hat{\tau}_n^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \bar{X})^2$. This estimator reduces to

$$\hat{\beta}_n^S = \operatorname*{arg\,min}_{n^{1/2}|b-\tilde{\beta}_n|\le C_n} \Big\|\frac{1}{\sqrt{n}}\sum_{j=1}^n g_j(b)\Big\|^2/\hat{\tau}_n^2 = \operatorname*{arg\,min}_{n^{1/2}|b-\tilde{\beta}_n|\le C_n} \Big\|\frac{1}{\sqrt{n}}\sum_{j=1}^n g_j(b)\Big\|^2.$$

Using arguments from the proof of Theorem 2, both estimators, $\hat{\beta}_n^{EE}$ and $\hat{\beta}_n^S$, can be shown to be efficient. In simulations the GMELE outperformed the alternative estimators $\hat{\beta}_n^{EE}$ and $\hat{\beta}_n^S$; see Table 1 in Section 5.

5 Simulations

Here we report the results of a small simulation study carried out to investigate the finite sample behavior of the GMELE (8) and the test from Remark 2. The simulations were carried out with the help of the R package. The R function *optimize* was used to locate the maximizers.

5.1 Comparing GMELE with the competing estimators from Remark 3

For this study we used the full model with $\beta = 1$ and sample size $n = 100$. We worked with two error distributions and two covariate distributions. As error distributions we picked the mixture normal distribution $0.25\mathscr{N}(-10, 1) + 0.5\mathscr{N}(0, 1) + 0.25\mathscr{N}(10, 1)$ and the skew normal distribution with location parameter zero, scale parameter 1 and skewness parameter 4. As covariate distributions we chose the standard normal distribution and the uniform distribution on $(-1, 3)$. Table 1 reports simulated mean squared errors of the estimators, $\hat{\beta}_n^S$, $\hat{\beta}_n^{EE}$ and the GMELE, based on 2000 repetitions, and for the choices $r_n = 1, \dots, 10$. We used the OLSE as preliminary estimator for the GMELE and $\hat{\beta}_n^S$, to specify the location of the search interval. As preliminary estimator for $\hat{\beta}_n^{EE}$ we used $\hat{\beta}_n^S$. We chose $2c_n\sqrt{\log(n)/n}$ as the length of the interval, with $c_n = 1$ for skew normal errors and $c_n = 10$ for the mixture normal errors. As can be seen from Table 1, the GMELE clearly outperforms the two competing approaches.

Table 1 Comparing the GMELE $\hat{\beta}_n$ (M) with $\hat{\beta}_n^S$ (S) and $\hat{\beta}_n^{EE}$ (EE) from Remark 3

r_n	1	2	3	4	5	6	7	8	9	10
mixture normal error, normal covariate										
S	.625	3.79	1.53	.494	.345	.333	.314	.315	.295	.314
EE	.625	4.09	2.22	.855	.629	.712	.742	.749	.801	.820
M	.123	0.16	0.33	.373	.148	.144	.132	.131	.163	.158
mixture normal error, uniform covariate										
S	.454	4.57	1.26	.339	.221	.212	.199	.197	.208	.217
EE	.454	4.83	1.90	.629	.393	.380	.395	.466	.535	.621
M	.094	0.11	0.19	.212	.067	.071	.089	.086	.077	.076
skew normal error, normal covariate										
S	.028	.020	.015	.013	.012	.012	.012	.012	.012	.012
EE	.028	.020	.015	.013	.012	.012	.012	.012	.012	.013
M	.009	.009	.007	.008	.008	.008	.008	.008	.009	.009
skew normal error, uniform covariate										
S	.027	.019	.013	.011	.009	.009	.009	.008	.009	.009
EE	.027	.019	.014	.011	.009	.009	.009	.009	.009	.009
M	.008	.006	.005	.005	.005	.005	.005	.005	.005	.006

The table entries are the simulated MSE's for the three estimators in the full model for sample size $n = 100$ based on 2000 repetitions.

5.2 Performance with missing data

Here we report on the performance of the GMELE and the OLSE with missing data. We again used the model $Y = \beta X + \varepsilon$ with $\beta = 1$ and chose

$$\pi(X) = P(\delta = 1|X) = 1/(1 + d\ \exp(X))$$

with $d = 0$, 0.1 and 0.5 to produce different missingness rates. Note that $d = 0$ corresponds to the full model.

Table 2 Simulated MSE's for OLSE and GMELE with missing data

n	MR	OLSE	1	2	3	4	5	6	7	8	9	10
mixture normal error, normal covariate												
70	0%	.745	.246	.416	.631	.815	.402	.395	.341	.382	.345	.396
	12%	.969	.355	.583	.877	1.07	.588	.599	.555	.614	.544	.547
	36%	1.43	.712	1.15	1.58	1.82	1.34	1.31	1.23	1.31	1.24	1.27
140	0%	.368	.073	.086	.136	.149	.047	.048	.044	.040	.038	.032
	12%	.461	.105	.142	.240	.260	.098	.077	.087	.094	.090	.081
	36%	.722	.188	.251	.447	.554	.245	.266	.299	.305	.274	.297
mixture normal error, uniform covariate												
70	0%	.563	.176	.257	.390	.447	.209	.194	.232	.238	.214	.216
	27%	.876	.386	.601	.846	.961	.588	.587	.578	.623	.582	.565
	56%	1.80	1.25	1.89	2.21	2.63	1.94	2.01	1.89	2.01	2.01	1.94
140	0%	.267	.051	.056	.091	.086	.020	.021	.019	.024	.018	.020
	27%	.435	.100	.116	.210	.204	.058	.042	.067	.076	.083	.075
	56%	.853	.329	.447	.696	.800	.439	.420	.425	.436	.441	.468
skew normal error, normal covariate												
70	0%	.146	.141	.141	.119	.127	.129	.138	.143	.153	.149	.158
	12%	.185	.181	.178	.159	.168	.169	.179	.182	.198	.202	.211
	36%	.281	.281	.286	.269	.280	.285	.301	.313	.328	.330	.332
140	0%	.070	.070	.061	.050	.050	.049	.050	.053	.055	.057	.056
	12%	.088	.087	.078	.062	.062	.062	.063	.066	.069	.071	.074
	36%	.142	.138	.127	.112	.117	.118	.119	.123	.127	.125	.130
skew normal error, uniform covariate												
70	0%	.114	.112	.101	.084	.086	.086	.087	.096	.101	.107	.110
	27%	.172	.167	.160	.139	.150	.152	.159	.159	.178	.187	.202
	56%	.361	.354	.395	.381	.413	.404	.430	.449	.448	.469	.485
140	0%	.053	.052	.042	.034	.033	.030	.032	.033	.033	.033	.034
	27%	.082	.081	.070	.059	.056	.054	.056	.057	.058	.060	.062
	56%	.166	.162	.155	.142	.138	.142	.154	.154	.159	.159	.176

The table entries are simulated mean squared errors for mixture normal errors, and 10 times the simulated mean squared errors for skew normal errors.

We used the same error and covariate distributions as before and worked with the search interval $\tilde{\beta}_{N,c} \pm c_N\sqrt{\log(N)/N}$ based on the complete case version of the OLSE. We chose $c_N = 1$ for the skew normal errors and $c_N = 10$ for the mixture normal errors. The reported results are based on samples of size $n = 70$ and 140, $r_n = 1, \dots, 10$ basis functions and 2000 repetitions.

Table 2 reports simulated mean squared errors of the OLSE and GMELE for $r_n = 1, \dots, 10$. The mean squared errors are multiplied by 10 for skew normal errors. We also list the average missingness rates (MR).

The GMELE performs in most cases much better (smaller MSE's) than the OLSE, except in some of the small samples. The results for the scenario with uniform covariates are better than the corresponding figures for standard normal covariates. The mean squared errors for the skew normal errors are even better than those for mixture normal errors.

5.3 Behavior for errors without finite Fisher information

A different scenario is considered in Table 3, namely when the errors are from an exponential distribution. Since the exponential distribution has no finite Fisher information for location it does not fit into our theory, but it still demonstrates superior performance of the GMELE over the OLSE.

Table 3 Simulated MSE's for exponential error

					normal covariate		
n	MR	OLSE	1	2	3	4	5
70	0%	.0157	.0085	.0092	.0073	.0075	.0086
	12%	.0198	.0118	.0125	.0115	.0120	.0123
	36%	.0306	.0242	.0235	.0232	.0216	.0209
140	0%	.0075	.0021	.0020	.0017	.0018	.0020
	12%	.0090	.0030	.0026	.0029	.0026	.0028
	36%	.0140	.0058	.0056	.0064	.0054	.0063
					uniform covariate		
70	0%	.0109	.0041	.0041	.0045	.0041	.0044
	27%	.0169	.0098	.0102	.0103	.0100	.0110
	56%	.0359	.0304	.0333	.0351	.0339	.0351
140	0%	.0054	.0009	.0010	.0009	.0011	.0011
	27%	.0086	.0023	.0026	.0021	.0023	.0025
	56%	.0179	.0100	.0096	.0088	.0088	.0097

The table entries are the MSE's for $r_n = 1, \dots, 5$ constraints when the errors are from an exponential distribution (no finite Fisher information).

Remark 4 The choice of the number of basis vectors r_n (and r_N) does affect the performance of the GMELE. This suggests using a data-driven choice. One possibility is the approach of Peng and Schick (2005, Section 5.1), who used bootstrap to select r_n in a related setting, with convincing results. The idea is to compute the bootstrap mean squared errors of the estimator (the GMELE in our case) for different values of r_n, say for $r_n = 1, \dots, 10$. Then select the r_n with the minimum bootstrap mean squared error.

5.4 Comparison of two tests

We performed a small study comparing the empirical likelihood test about the slope from Remark 2 and the corresponding bootstrap test, which uses resampling instead of the χ^2 approximation to obtain critical values. The null hypothesis is $\beta = \beta_0 = 1$ and the nominal level is .05. As in Table 1 we consider only the full model and the sample size $n = 100$. Table 4 reports the simulated significance level and power of the two tests, using $r_n = 1, 2, \dots, 5$ basis functions. The covariates X and the errors ε were generated from the same distributions as before. The bootstrap resample size was taken to be the same as the sample size (i.e. $n = 100$), while we used more repetitions than before: in order to stabilize the results obtained by the bootstrap method we worked with $10,000$ repetitions. Our simulations indicate that the results based on the χ^2 approximation (denoted by χ^2) are much more reliable than the results of the bootstrap approach (denoted by $\mathscr{B}$). For $r_n \geq 3$ the bootstrapped significance levels are far away from the nominal level 5%: they are between 11% and 60%, i.e. the test is far too liberal, which is in contrast to the χ^2 approach. The significance levels for $r_n = 1, 2$ are reasonable for both tests. In terms of power the bootstrap test is better than the χ^2 test in the upper table with normal covariates; for uniform covariates it is the other way round.

Table 4 Simulated significance level and power of the empirical likelihood test about the slope using χ^2 and bootstrap quantiles

		1	2	3	4	5	1	2	3	4	5
		normal covariate									
		mixture normal error					skew normal error				
$\beta = 1.0$	χ^2	.05	.06	.07	.09	.10	.05	.06	.07	.09	.10
	$\mathscr{B}$	.03	.07	.16	.28	.42	.01	.04	.11	.20	.33
$\beta = 1.2$	χ^2	.11	.11	.11	.13	.21	.52	.55	.64	.67	.72
	$\mathscr{B}$	.19	.33	.51	.69	.86	.54	.74	.89	.95	.98
		uniform covariate									
		mixture normal error					skew normal error				
$\beta = 1.0$	χ^2	.05	.06	.07	.09	.10	.06	.07	.08	.09	.11
	$\mathscr{B}$	.02	.06	.14	.25	.39	.08	.15	.28	.44	.60
$\beta = 1.2$	χ^2	.11	.10	.11	.13	.21	.54	.56	.65	.67	.71
	$\mathscr{B}$	.05	.10	.19	.31	.54	.33	.26	.48	.62	.75

The table shows simulated significance level and power figures of the empirical likelihood test with null hypothesis $\beta = 1$ at the nominal level $\alpha = 0.05$. We consider the full model; the sample size is $n = 100$. The test uses approximative χ^2 quantiles (χ^2) and bootstrap quantiles ($\mathscr{B}$).

6 Proofs

This section contains the proofs of Theorem 1 (given in the first subsection) and of Theorem 2. The proof of the uniform expansion that is provided in Theorem 2 is split into three parts. In Subsection 6.2 we give six conditions and show that they are sufficient for the expansion. That the conditions are indeed satisfied is shown separately in Subsections 6.3 and 6.4. Subsection 6.5

contains an auxiliary result. As explained in the introduction, we only need to prove the results for the full model, i.e., the case when $\pi(X)$ equals one.

6.1 Proof of Theorem 1

Let μ denote the mean and τ denote the standard deviation of X. We should point out that $\mathscr{R}_n(b)$ does not change if we replace $(X_j - \bar{X})$ by $(X_j - \bar{X})/\tau = V_j - \bar{V}$, where

$$V_j = \frac{X_j - \mu}{\tau} \quad \text{and} \quad \bar{V} = \frac{1}{n}\sum_{j=1}^{n} V_j.$$

Thus, for the purpose of our proofs, we may assume that $\mathscr{R}_n(b)$ is given by

$$\mathscr{R}_n(b) = \sup\Big\{\prod_{j=1}^{n} n\pi_j : \pi \in \mathscr{P}_n,\ \sum_{j=1}^{n} \pi_j (V_j - \bar{V}) v_n(\mathbb{F}_b(Y_j - bX_j)) = 0\Big\}.$$

In what follows we shall repeatedly use the bounds

$$|v_n(y)|^2 \le 2r_n, \quad |v_n'(y)|^2 \le 2\pi^2 r_n^3, \quad \text{and} \quad |v_n''(y)|^2 \le 2\pi^4 r_n^5$$

for all real y.

Let us set $Z_j = V_j v_n(F(\varepsilon_j))$ and $\hat{Z}_j = (V_j - \bar{V}) v_n(\mathbb{F}_\beta(\varepsilon_j))$, $j = 1, \dots, n$. With $Z = Z_1$, we find the identities $E[Z] = 0$ and $E[ZZ^\top] = I_{r_n}$, where I_{r_n} is the $r_n \times r_n$ identity matrix, and the bound $E[|Z|^4] \le (2r_n)^2 E[V^4] = O(r_n^2)$. As shown in Peng and Schick (2013), these results yield

$$\tilde{Z}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_j = O_P(r_n^{1/2}) \tag{9}$$

and

$$\sup_{|u|=1} \Big| \frac{1}{n} \sum_{j=1}^{n} (u^\top Z_j)^2 - 1 \Big| \le \Big| \frac{1}{n} \sum_{j=1}^{n} Z_j Z_j^\top - I_{r_n} \Big| = O_P(r_n n^{-1/2}). \tag{10}$$

From Corollary 7.6 in Peng and Schick (2013) and $r_n^4 = o(n)$, the desired result follows if we verify

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\hat{Z}_j - Z_j) = o_P(1) \quad \text{and} \quad \frac{1}{n} \sum_{j=1}^{n} |\hat{Z}_j - Z_j|^2 = o_P(r_n^3/n).$$

Let

$$\varDelta_j = v_n(\mathbb{F}_\beta(\varepsilon_j)) - v_n(F(\varepsilon_j)), \quad j = 1, \dots, n.$$

In view of the identity $\hat{Z}_j - Z_j = V_j \Delta_j - \bar{V}\Delta_j - \bar{V}v_n(F(\varepsilon_j))$, the bound $|v_n|^2 \le 2r_n$, and the fact $n^{1/2}\bar{V} = O_P(1)$, it is easy to see the desired results follow from the following rates:

$$S_1 = \frac{1}{\sqrt{n}}\sum_{j=1}^n V_j \Delta_j = O_P(r_n^{3/2}n^{-1/2}),$$

$$S_2 = \frac{1}{n}\sum_{j=1}^n \Delta_j = o_P(r_n^{3/2}n^{-1/2}),$$

$$S_3 = \frac{1}{n}\sum_{j=1}^n v_n(F(\varepsilon_j)) = O_P(r_n^{1/2}n^{-1/2}),$$

$$S_4 = \frac{1}{n}\sum_{j=1}^n V_j^2|\Delta_j|^2 = O_P(r_n^3 n^{-1}).$$

Note that $\Delta_1, \dots, \Delta_n$ are functions of the errors $\varepsilon_1, \dots, \varepsilon_n$ only and satisfy

$$M_n = \max_{1\le j\le n} |\Delta_j|^2 \le 2\pi^2 r_n^3 \sup_{t\in\mathbb{R}} |\mathbb{F}_\beta(t) - F(t)|^2 = O_P(r_n^3/n).$$

Conditioning on the errors thus yields

$$E[|S_1|^2|\varepsilon_1, \dots, \varepsilon_n] = E[S_4|\varepsilon_1, \dots, \varepsilon_n] \le M_n.$$

This establishes the rates for S_1 and S_4. The other rates follow from $|S_2|^2 \le M_n$ and $nE[|S_3|^2] = E[|v_n(F(\varepsilon))|^2] = r_n$.

6.2 Proof of Theorem 2

For $t \in \mathbb{R}$, we let $\hat{F}_{nt} = \mathbb{F}_{\beta+n^{-1/2}t}$ and note that $\hat{F}_{nt}$ is the empirical distribution function of the random variables

$$\varepsilon_{jt} = \varepsilon_j - n^{-1/2}tX_j, \quad j = 1, \dots, n.$$

These random variables are independent with common distribution function F_{nt} given by

$$F_{nt}(y) = E[\hat{F}_{nt}(y)] = E[F(y + n^{-1/2}tX)], \quad y \in \mathbb{R}.$$

To simplify notation we introduce

$$\hat{R}_{jt} = \hat{F}_{nt}(\varepsilon_{jt}), \quad R_{jt} = F_{nt}(\varepsilon_{jt}), \quad R_j = F(\varepsilon_j),$$

and

$$\hat{Z}_{jt} = (V_j - \bar{V})v_n(\hat{R}_{jt}), \quad Z_{jt} = V_j v_n(R_{jt}), \quad Z_j = V_j v_n(R_j).$$

Since we are working with the form of the empirical likelihood given in the previous section, we have

$$\mathscr{R}_n(\beta + n^{-1/2}t) = \sup\Big\{ \prod_{j=1}^n n\pi_j : \pi \in \mathscr{P}_n,\ \sum_{j=1}^n \pi_j \hat{Z}_{jt} = 0 \Big\}, \quad t \in \mathbb{R}.$$

Fix a sequence C_n such that $C_n \geq 1$ and $C_n = O((\log n)^{1/2})$. The desired result follows if we verify the uniform expansion

$$\sup_{|t| \leq C_n} \frac{|-2\log \mathscr{R}_n(\beta + n^{-1/2}t) - |\tilde{Z}_n|^2 + 2t\varGamma_n - t^2\tau^2 J_f|}{(1+|t|)^2} = o_P(1) \tag{11}$$

with $\tilde{Z}_n$ as in (9). To verify (11) we introduce

$$\nu_n = E[X\ell_f(\varepsilon) V v_n(F(\varepsilon))].$$

We shall establish (11) by verifying the following six conditions.

$$\sup_{|t| \leq C_n} \sup_{|u|=1} \Big| \frac{1}{n} \sum_{j=1}^n (u^\top \hat{Z}_{jt})^2 - 1 \Big| = o_P(1/r_n), \tag{12}$$

$$\sup_{|t| \leq C_n} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n (\hat{Z}_{jt} - Z_{jt}) \Big| = o_P(r_n^{-1/2}), \tag{13}$$

$$\sup_{|t| \leq C_n} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n (Z_{jt} - Z_j - E[Z_{jt} - Z_j]) \Big| = o_P(r_n^{-1/2}), \tag{14}$$

$$\sup_{|t| \leq C_n} |n^{1/2} E[Z_{1t} - Z_1] + t\nu_n| = o(r_n^{-1/2}), \tag{15}$$

$$|\nu_n|^2 \to \tau^2 J_f, \tag{16}$$

$$\nu_n^\top \tilde{Z}_n - \varGamma_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n [\nu_n^\top Z_j - (X_j - \mu)\ell_f(\varepsilon_j)] = o_P(1). \tag{17}$$

These six conditions are proved in the next two subsections. We first establish their sufficiency.

Lemma 1 *The conditions (12)–(17) imply (11).*

To prove this lemma, we use the following result which is a special case of Lemma 5.2 in Peng and Schick (2013). This version was used in Schick (2013).

Lemma 2 *Let $x_1, \dots, x_n$ be m-dimensional vectors. Set*

$$\bar{x} = \frac{1}{n}\sum_{j=1}^n x_j, \quad x^* = \max_{1\le j\le n} |x_j|, \quad \nu_4 = \frac{1}{n}\sum_{j=1}^n |x_j|^4, \quad S = \frac{1}{n}\sum_{j=1}^n x_j x_j^\top,$$

and let λ and Λ denote the smallest and largest eigenvalue of the matrix S. Then the inequality $\lambda > 5|\bar{x}|x^*$ implies

$$\Big| -2\log(\mathscr{R}) - n\bar{x}^\top S^{-1}\bar{x}\Big| \le \frac{n|\bar{x}|^3(\Lambda\nu_4)^{1/2}}{(\lambda - |\bar{x}|x^*)^3} + \frac{4n\Lambda^2|\bar{x}|^4\nu_4}{\lambda^2(\lambda - |\bar{x}|x^*)^4}$$

with

$$\mathscr{R} = \sup\Big\{\prod_{j=1}^n n\pi_j : \pi \in \mathscr{P}_n, \ \sum_{j=1}^n \pi_j x_j = 0\Big\}.$$

Proof of Lemma 1 We introduce

$$\mathbb{T}(t) = \frac{1}{n}\sum_{j=1}^n \hat{Z}_{jt} \quad \text{and} \quad \mathbb{S}(t) = \frac{1}{n}\sum_{j=1}^n \hat{Z}_{jt}\hat{Z}_{jt}^\top,$$

and let $\lambda_n(t)$ and $\Lambda_n(t)$ denote the smallest and largest eigenvalues of $\mathbb{S}(t)$, i.e.,

$$\lambda_n(t) = \inf_{|u|=1} u^\top \mathbb{S}(t) u = \inf_{|u|=1} \frac{1}{n}\sum_{j=1}^n (u^\top \hat{Z}_{jt})^2$$

and

$$\Lambda_n(t) = \sup_{|u|=1} u^\top \mathbb{S}(t) u = \sup_{|u|=1} \frac{1}{n}\sum_{j=1}^n (u^\top \hat{Z}_{jt})^2.$$

By (12), we have

$$\sup_{|t|\le C_n} |\lambda_n(t) - 1| = o_P(1) \quad \text{and} \quad \sup_{|t|\le C_n} |\Lambda_n(t) - 1| = o_P(1).$$

The conditions (13)–(15) imply

$$\sup_{|t|\le C_n} |n^{1/2}\mathbb{T}(t) - \tilde{Z}_n + t\nu_n| = o_P(r_n^{-1/2}). \tag{18}$$

This, together with (9) and (16) yields

$$\sup_{|t|\le C_n} n|\mathbb{T}(t)|^2 = O_P(r_n). \tag{19}$$

Next, we find

$$\sup_{|t|\le C_n} \max_{1\le j\le n} |\hat{Z}_{jt}| \le (2r_n)^{1/2} \max_{1\le j\le n} |V_j - \bar{V}| = o_P(r_n^{1/2} n^{1/4})$$

and

$$\sup_{|t|\le C_n} \frac{1}{n}\sum_{j=1}^n |\hat{Z}_{jt}|^4 \le (2r_n)^2 \frac{1}{n}\sum_{j=1}^n |V_j - \bar{V}|^4 = O_P(r_n^2).$$

Thus we derive

$$\sup_{|t|\le C_n}\Big|-2\log\mathscr{R}_n(\beta+n^{-1/2}t)-n\mathbb{T}(t)^\top(\mathbb{S}(t))^{-1}\mathbb{T}(t)\Big|=o_P(1), \tag{20}$$

since by Lemma 2 the left-hand side is of order $O_P(r_n^{5/2}n^{-1/2}+r_n^4/n)$. For a positive definite matrix A and a compatible vector x, we have

$$|x^\top A^{-1}x-x^\top x|\le x^\top A^{-1}x\sup_{|u|=1}|1-u^\top Au|\le\frac{|x|^2}{\lambda}\sup_{|u|=1}|1-u^\top Au|$$

with λ the smallest eigenvalue of A. This, together with (12) and (19) yields

$$\sup_{|t|\le C_n} n|\mathbb{T}(t)^\top(\mathbb{S}(t))^{-1}\mathbb{T}(t)-\mathbb{T}(t)^\top\mathbb{T}(t)|=o_P(1). \tag{21}$$

With the help of (9), (16) and (18) we verify

$$\sup_{|t|\le C_n}\Big|n|\mathbb{T}(t)|^2-|\tilde Z_n|^2+2t\nu_n^\top\tilde Z_n-t^2|\nu_n|^2\Big|=o_P(1). \tag{22}$$

The results (20)–(22) yield the expansion

$$\sup_{|t|\le C_n}\Big|-2\log\mathscr{R}_n(\beta+n^{-1/2}t)-|\tilde Z_n|^2+2t\nu_n^\top\tilde Z_n-t^2|\nu_n|^2\Big|=o_P(1).$$

From (16) and (17) we derive the expansion

$$\sup_{|t|\le C_n}\frac{|2t(\nu_n^\top\tilde Z_n-\Gamma_n)-t^2(|\nu_n|^2-\tau^2J_f)|}{(1+|t|)^2}=o_P(1).$$

The desired result (11) follows from the last two expansions. □

6.3 Proofs of (14)-(17)

We begin by mentioning properties of f and F that are crucial to the proofs. Since f has finite Fisher information for location, we have

$$\int|f(y+t)-f(y+s)|\,dy\le B_1|t-s|, \tag{23}$$

$$|F(t)-F(s)|\le B_1|t-s|, \tag{24}$$

$$|F(t+s)-F(t)-sf(t)|\le B_2|s|^{3/2}, \tag{25}$$

$$\int|F(y+s)-F(y)-sf(y)|\,dy\le B_1s^2 \tag{26}$$

for all real s and t, and some constants B_1 and B_2, see, e.g., Peng and Schick (2016).

Next, we look at the process

$$H_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (h_{nt}(X_j, Y_j) - E[h_{nt}(X, Y)]), \quad t \in \mathbb{R},$$

where h_{nt} are measurable functions from $\mathbb{R}^2$ to $\mathbb{R}^{m_n}$ such that $h_{n0} = 0$. We are interested in the cases $m_n = 1$ and $m_n = r_n$. A version of the following lemma was used in Peng and Schick (2016).

Lemma 3 *Suppose that the map $t \mapsto h_{nt}(x, y)$ is continuous for all $x, y \in \mathbb{R}$ and*

$$E[|h_{nt}(X, Y) - h_{ns}(X, Y)|^2] \le K_n|t - s|^2, \quad s, t \in \mathbb{R} \tag{27}$$

for some positive constants K_n. Then we have the rate

$$\sup_{|t| \le C_n} |H_n(t)| = O_P(C_n K_n^{1/2}).$$

Proof of (14) The desired result follows from Lemma 3 applied with

$$h_{nt}(X, Y) = V[v_n(F_{nt}(\varepsilon - n^{-1/2}tX)) - v_n(F(\varepsilon))], \quad t \in \mathbb{R},$$

and $K_n = 2\pi^2 r_n^3 B_1^2 E[V^2(X_1 - X)^2]/n$. Indeed, we have $h_{n0} = 0$ and (27) in view of (24). Note also that $r_n C_n^2 K_n \to 0$. □

Proof of (15) Since V and ε are independent and V has mean zero, we obtain the identity

$$n^{1/2}E[Z_{1t} - Z_1] + t\nu_n = n^{1/2}E[V_1 v_n(F_{nt}(\varepsilon_{1t}))] + t\nu_n = n^{1/2}(\Delta_1(t) + \Delta_2(t))$$

with

$$\Delta_1(t) = E\Big[V \int [v_n(F_{nt}(y)) - v_n(F(y))][f(y + n^{-1/2}tX) - f(y)]\,dy\Big]$$

and

$$\Delta_2(t) = E\Big[V \int v_n(F(y))[f(y + n^{-1/2}tX) - f(y) - n^{-1/2}tXf'(y)]\,dy\Big].$$

It follows from (23) and (24) that

$$|\Delta_1(t)| \le (2\pi^2 r_n^3)^{1/2} B_1 E[|X|] B_1 E[|VX|] t^2/n.$$

Integration by parts shows that

$$\Delta_2(t) = -E\Big[V \int (v_n'(F(y))f(y)[F(y + n^{-1/2}tX) - F(y) - n^{-1/2}tXf(y)]\,dy\Big].$$

It follows from (24) that f is bounded by B_1. This, together with (26), yields the bound

$$|\Delta_2(t)| \le (2\pi^2 r_n^3)^{1/2} B_1^2 E[|VX^2|] t^2/n.$$

From these bounds we conclude

$$\sup_{|t|\le C_n}\left|n^{1/2}E[Z_{1t}-Z_1]+t\nu_n\right| = O(r_n^{3/2}(\log n)n^{-1/2}) = o(r_n^{-1/2}),$$

which is the desired (15). □

Proof of (16) and (17) Note that ν_n can be written as

$$\nu_n = E[X\ell_f(\varepsilon)Vv_n(F(\varepsilon))] = \tau E[V\ell_f(\varepsilon)Vv_n(F(\varepsilon))].$$

The functions $V\varphi_1(F(\varepsilon)), V\varphi_2(F(\varepsilon)), \dots$ form an orthonormal basis of the space $\mathscr{V} = \{Va(\varepsilon) : a \in L_{2,0}(F)\}$. Thus ν_n is the vector consisting of the first r_n Fourier coefficients of $(X-\mu)\ell_f(\varepsilon) = \tau V\ell_f(\varepsilon)$ with respect to this basis. Because $(X-\mu)\ell_f(\varepsilon)$ is a member of $\mathscr{V}$, Parseval's theorem yields

$$|\nu_n|^2 \to E[((X-\mu)\ell_f(\varepsilon))^2] = \tau^2 J_f$$

and

$$E[(\nu_n^\top Vv_n(F(\varepsilon)) - (X-\mu)\ell_f(\varepsilon))^2] \to 0.$$

The former is (16) and the latter implies (17). □

6.4 Proofs of (12) and (13)

We begin by deriving properties of $\hat{R}_{jt}$ and R_{jt} which we need in the proofs of (12) and (13). For this we introduce the leave-one-out version $\tilde{R}_{jt}$ of $\hat{R}_{jt}$ defined by

$$\tilde{R}_{jt} = \frac{1}{n-1}\sum_{i:i\ne j}\mathbf{1}[\varepsilon_{it}\le\varepsilon_{jt}] = \frac{n}{n-1}\hat{R}_{jt} - \frac{1}{n-1}\mathbf{1}[\varepsilon_{jt}\le\varepsilon_{jt}],$$

which satisfies

$$|\hat{R}_{jt}-\tilde{R}_{jt}| \le \frac{2}{n-1}. \tag{28}$$

We abbreviate $\tilde{R}_{j0}$ by $\tilde{R}_j$. In the ensuing arguments we rely on the following properties of these quantities, where B_1 and B_2 are the constants appearing in (24) and (25):

$$\max_{1\le j\le n}\sup_{|t|\le C_n}|\tilde{R}_{jt}-R_{jt}-\tilde{R}_j+R_j| = O_P(n^{-5/8}(C_n\log n)^{1/2}), \tag{29}$$

$$\max_{1\le j\le n}|\tilde{R}_j-R_j| = O_P(n^{-1/2}), \tag{30}$$

$$\sup_{|t|\le C_n}|R_{jt}-R_j| \le B_1C_nn^{-1/2}(|X_j|+E[|X|]), \tag{31}$$

$$\begin{aligned}\sup_{|t|\le C_n}&|R_{jt}-R_j+n^{-1/2}t(X_j-\mu)f(\varepsilon_j)|\\ &\le B_2C_n^{3/2}n^{-3/4}\sqrt{2}(|X_j|^{3/2}+E[|X|^{3/2}]).\end{aligned} \tag{32}$$

The second statement follows from properties of the empirical distribution function and the last two statements from (24) and (25), respectively. To prove (29) we use Lemma 4 from Subsection 6.5. Let $\zeta_j(t) = \tilde{R}_{jt} - R_{jt} - \tilde{R}_j + R_j$ and $m = n-1$. These random variables are identically distributed, and $(n-1)\zeta_n(t)$ equals $\tilde{N}(n^{-1/2}t, X_n, \varepsilon_n)$ from the beginning of Subsection 6.5, with the role of Y_i played by ε_i. Lemma 4 gives

$$
\begin{aligned}
&P(\max_{1\le j\le n}\sup_{|t|\le C_n}|\zeta_j(t)| > 4KC_n^{1/2}(n-1)^{-5/8}(\log(n-1))^{1/2})\\
&\le nP(\sup_{|t|\le C_n}|\zeta_n(t)| > 4KC_n^{1/2}m^{-5/8}(\log m)^{1/2})\\
&\le nP(|X_n| > m^{1/4}) + nE[\mathbf{1}[|X_n| \le m^{1/4}]p_m(\varepsilon_n, C_n, K)]\\
&\le 2E[|X|^4\mathbf{1}[|X| > m^{1/4}] + Cn^2\exp(-K\log(m))
\end{aligned}
$$

for $m > 2$ and $K > 6B_1(1 + E[|X|])$ and some constant C. The desired (29) is now immediate.

Note that statements (28) – (31) yield the bounds

$$\sup_{|t|\le C_n}|\hat{R}_{jt} - R_j| \le B_1C_n n^{-1/2}(|X_j| + E[|X|]) + n^{-1/2}\xi_n, \quad j = 1,\dots,n, \tag{33}$$

which we need for the next proof. Here ξ_n is a positive random variable which satisfies $\xi_n = O_P(1)$.

Proof of (12) Given (10) and the properties of r_n, it suffices to verify

$$\sup_{|u|=1}\sup_{|t|\le C_n}\left|\frac{1}{n}\sum_{j=1}^n(u^\top\hat{Z}_{jt})^2 - \frac{1}{n}\sum_{j=1}^n(u^\top Z_j)^2\right| = o_P(1/r_n). \tag{34}$$

Using the Cauchy-Schwarz inequality we bound the left-hand side of (34) by $2(D_n\Lambda_n)^{1/2} + D_n$ with

$$\Lambda_n = \sup_{|u|=1}\frac{1}{n}\sum_{j=1}^n(u^\top Z_j)^2 \quad \text{and} \quad D_n = \sup_{|t|\le C_n}\frac{1}{n}\sum_{j=1}^n|\hat{Z}_{jt} - Z_j|^2.$$

Given (10), it therefore suffices to prove $D_n = o_P(1/r_n^2)$. This follows from (33), the inequality

$$
\begin{aligned}
D_n &\le \sup_{|t|\le C_n}\frac{1}{n}\sum_{j=1}^n(2\bar{V}^2|v_n(\hat{R}_{jt})|^2 + 2V_j^2|v_n(\hat{R}_{jt}) - v_n(R_j)|^2)\\
&\le 4r_n\bar{V}^2 + 4\pi^2r_n^3\frac{1}{n}\sum_{j=1}^n V_j^2\sup_{|t|\le C_n}|\hat{R}_{jt} - R_j|^2 = O_P(r_n^3C_n^2/n),
\end{aligned}
$$

and the rate $r_n^5\log n = o(n)$. □

Proof of (13) In view of the rate $\bar{V} = O_P(n^{-1/2})$ and the identity

$$\hat{Z}_{jt} - Z_{jt} = V_j(v_n(\hat{R}_{jt}) - v_n(R_{jt})) - \bar{V}(v_n(\hat{R}_{jt}) - v_n(R_j)) - \bar{V}v_n(R_j),$$

the desired (13) is implied by the following three statements:

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} v_n(R_j) = O_P(r_n^{1/2}), \tag{35}$$

$$\sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(v_n(\hat{R}_{jt}) - v_n(R_j))\Big| = O_P(C_n r_n^{3/2}), \tag{36}$$

$$\sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j[v_n(\hat{R}_{jt}) - v_n(R_{jt})]\Big| = o_P(r_n^{-1/2}). \tag{37}$$

We obtain (35) from $E[v_n(F(\varepsilon))] = 0$ and $E[|v_n(F(\varepsilon)|^2] = r_n$. Also, (36) follows from (33) and the fact that its left-hand side is bounded by

$$(2\pi^2 r_n^3)^{1/2}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\sup_{|t|\le C_n}|\hat{R}_{jt} - R_j|.$$

Using (28) we find

$$\sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j[v_n(\hat{R}_{jt}) - v_n(\tilde{R}_{jt})]\Big| = O_P(r_n^{3/2}n^{-1/2}).$$

Taylor expansions, the bound $|v_n'''|^2 \le 2\pi^6 r_n^7$ and equations (28), (31) and (33) show that

$$\sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j[v_n(\tilde{R}_{jt}) - v_n(R_j) - v_n'(R_j)(\tilde{R}_{jt} - R_j) - \frac{1}{2}v_n''(R_j)(\tilde{R}_{jt} - R_j)^2]\Big|$$

and

$$\sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}V_j[v_n(R_{jt}) - v_n(R_j) - v_n'(R_j)(R_{jt} - R_j) - \frac{1}{2}v_n''(R_j)(R_{jt} - R_j)^2]\Big|$$

are of order $r_n^{7/2}C_n^3 n^{-1}$. Using the identity

$$(a+b+c)^2 - a^2 - b^2 + 2db = c^2 + 2(a+d)b + 2(a+b)c$$

with $a = R_{jt} - R_j$, $b = \tilde{R}_j - R_j$, $c = \tilde{R}_{jt} - R_{jt} - \tilde{R}_j + R_j$ and $d = n^{-1/2}t(X_j - \mu)f(\varepsilon_j) = n^{-1/2}t\tau V_j f(\varepsilon_j)$, together with the properties (29)–(32), we derive the bounds

$$\sup_{|t|\le C_n}|(\tilde{R}_{jt} - R_j)^2 - (R_{jt} - R_j)^2 - (\tilde{R}_j - R_j)^2 + 2n^{-1/2}t\tau V_j f(\varepsilon_j)(\tilde{R}_j - R_j)|$$
$$\le \zeta_n(1 + |X_j|)^{3/2}, \quad j = 1, \ldots, n,$$

with $\zeta_n = O_P(n^{-9/8}C_n^{3/2}(\log n)^{1/2})$. If follows that the left-hand side of (37) is bounded by $|T_1|/2 + C_n\tau|T_2| + T_3 + T_4$, where

$$T_1 = \frac{1}{\sqrt{n}}\sum_{j=1}^n V_j v_n''(R_j)(\tilde{R}_j - R_j)^2,$$

$$T_2 = \frac{1}{n}\sum_{j=1}^n V_j^2 v_n''(R_j) f(\varepsilon_j)(\tilde{R}_j - R_j),$$

$$T_3 = \sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^n V_j v_n'(R_j)(\tilde{R}_{jt} - R_{jt})\Big|,$$

and

$$T_4 = O_P(r_n^{3/2}n^{-1/2} + r_n^{7/2}C_n^3 n^{-1} + r_n^{5/2}n^{-5/8}C_n^{3/2}(\log n)^{1/2}) = o_P(r_n^{-1/2}).$$

We calculate

$$E[|T_1|^2|\varepsilon_1,\dots,\varepsilon_n] = \frac{1}{n}\sum_{j=1}^n |v_n''(R_j)|^2(\tilde{R}_j - R_j)^4 = O_P(r_n^5 n^{-2}).$$

Thus $|T_1| = o_P(r_n^{-1/2})$. Next, we write T_2 as the vector U-statistic

$$T_2 = \frac{1}{n(n-1)}\sum_{i\ne j} V_j^2 v_n''(F(\varepsilon_j)) f(\varepsilon_j)(\mathbf{1}[\varepsilon_i \le \varepsilon_j] - F(\varepsilon_j))$$

and obtain

$$E[|T_2|^2] \le \frac{E[|k(\varepsilon)|^2]}{n} + \frac{2E[V_2^4|v_n''(F(\varepsilon_2)|^2 f^2(\varepsilon_2)(\mathbf{1}[\varepsilon_1 \le \varepsilon_2] - F(\varepsilon_2))^2]}{n(n-1)}$$

with $k(x) = E[v_n''(F(\varepsilon))f(\varepsilon)(\mathbf{1}[x \le \varepsilon] - F(\varepsilon))]$. Using the representation $f(y) = \int_y^\infty \ell_f(z)f(z)\,dz$ and Fubini's theorem, we calculate

$$\begin{aligned}k(x) &= \int_{-\infty}^\infty v_n''(F(y))f(y)(1[x \le y] - F(y))f(y)\,dy\\ &= \int_x^\infty (v_n'(F(z)) - v_n'(F(x))\ell_f(z)f(z)\,dz\\ &\quad - \int_{-\infty}^\infty [v_n'(F(z))F(z) - v_n(F(z))])\ell_f(z)f(z)\,dz.\end{aligned}$$

Thus $|k|$ is bounded by a constant times $r_n^{3/2}$ and we see that $E[|T_2|^2] = O(r_n^3/n + r_n^5/n^2)$. This proves $C_n|T_2| = O_P(r_n^{-1/2})$.

We bound T_3 by the sum $T_{31} + T_{32} + T_{33}$, where

$$T_{31} = \sup_{|t|\le C_n}\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^n W_j v_n'(R_j)(\tilde{R}_{jt} - R_{jt})\Big|,$$

$$T_{32} = \sup_{|t|\le C_n} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n V_j \mathbf{1}[|V_j| > n^{1/4}] v_n'(R_j)(\tilde{R}_{jt} - R_{jt}) \Big|,$$

$$T_{33} = \sup_{|t|\le C_n} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n E[V\mathbf{1}[|V| > n^{1/4}]] v_n'(R_j)(\tilde{R}_{jt} - R_{jt}) \Big|,$$

and $W_j = V_j \mathbf{1}[|V_j| \le n^{1/4}] - E[V\mathbf{1}[|V| \le n^{1/4}]$. Since V has a finite fourth moment, we obtain the rates $\max_{1\le j\le n} |V_j| = o_P(n^{1/4})$ and $E[|V|\mathbf{1}[|V| > n^{1/4}]] \le n^{-3/4} E[V^4 \mathbf{1}[|V| > n^{1/4}]] = o(n^{-3/4})$. Thus we find $P(T_{32} > 0) \le P(\max_{1\le j\le n} |V_j| > n^{1/4}) \to 0$ and $T_{33} = o_P(n^{-3/4} r_n^{3/2})$, using (29) and (30). This shows $T_{32} + T_{33} = o_P(r_n^{-1/2})$.

To deal with T_{31} we express it as

$$T_{31} = \sup_{|t|\le C_n} n^{1/2} \Big| \frac{1}{n(n-1)} \sum_{i\ne j} W_j v_n'(F(\varepsilon_j)) \Big(\mathbf{1}[\varepsilon_{it} \le \varepsilon_{jt}] - F_{nt}(\varepsilon_{jt}) \Big) \Big|.$$

Let us set

$$k_{nt}(z) = E[W v_n(F(z + n^{-1/2} t X))], \quad z \in \mathbb{R}.$$

Using (24) we obtain the bound

$$E[|k_{nt}(\varepsilon_{jt}) - k_{ns}(\varepsilon_{js})|^2] \le 2\pi^2 r_n^3 B_1^2 E[W^2 (X_j - X)^2] |t-s|^2/n$$

and derive with the help of Lemma 3

$$\sup_{|t|\le C_n} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n (k_{nt}(\varepsilon_{jt}) - E[k_{nt}(\varepsilon_{jt})]) \Big| = O_P(r_n^{3/2} C_n n^{-1/2}).$$

We therefore obtain the rate $T_{31} = o_P(r_n^{-1/2})$, if we verify

$$\sup_{|t|\le C_n} |U(t)| = O_P(r_n^{3/2} n^{-1} \log n), \tag{38}$$

where $U(t)$ is the vector U-statistic equaling

$$\frac{1}{n(n-1)} \sum_{i\ne j} \Big[W_j v_n'(F(\varepsilon_j)) \Big(\mathbf{1}[\varepsilon_{it} \le \varepsilon_{jt}] - F_{nt}(\varepsilon_{jt}) \Big) + k_{nt}(\varepsilon_{it}) - E[k_{nt}(\varepsilon_{it})] \Big].$$

It is easy to verify that $U(t)$ is degenerate. Let $t_k = -C_n + 2kC_n/n$, $k = 0, \dots, n$. Then we have

$$\sup_{|t|\le C_n} |U(t)| \le \max_{1\le k\le n} \Big(|U(t_k)| + \sup_{t_{k-1}\le t\le t_k} |U(t) - U(t_k)| \Big). \tag{39}$$

For $t \in [t_{k-1}, t_k]$, we find

$$|U(t) - U(t_k)| \le (2\pi^2 r_n^3)^{1/2} \Big(2n^{1/4} (N_k^+ + N_k^-) + 2B_1 C_n n^{-3/2} S \Big) \tag{40}$$

with

$$S = \frac{1}{n}\sum_{j=1}^{n}\big(|W_j|(|X_j| + E[|X|]) + E[|W|]|X_j| + 2E[|WX|] + E[|W|]E[|X|]\big),$$

$$N_k^+ = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{1}[t_{k-1}D_{ij} < \varepsilon_i - \varepsilon_j \le t_k D_{ij}]\mathbf{1}[D_{ij} \ge 0],$$

$$N_k^- = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{1}[t_k D_{ij} < \varepsilon_i - \varepsilon_j \le t_{k-1} D_{ij}]\mathbf{1}[D_{ij} < 0],$$

and $D_{ij} = n^{-1/2}(X_i - X_j)$. We write $U_l(t)$ for the l-th component of the vector $U(t)$. Then we have

$$P(\max_{1\le k\le n}|U(t_k)| > \eta) \le \sum_{k=1}^{n}\sum_{l=1}^{r_n} P(|U_l(t_k)| > \eta r_n^{-1/2}), \quad \eta > 0.$$

Since $U_l(t)$ is a degenerate U-statistic whose kernel is bounded by

$$b_l = 2n^{1/4}(\sqrt{2}\pi l + \sqrt{2}) \le 27 n^{1/4} l$$

and has second moment bounded by $2(\pi l)^2$, we derive from part (c) of Proposition 2.3 of Arcones and Giné (1993) that

$$\sup_{|t|\le C} P((n-1)|U_l(t)| > \eta) \le c_1 \exp\Big(-\frac{c_2\eta}{\sqrt{2}\pi l + b_l^{2/3}\eta^{1/3}n^{-1/3}}\Big)$$

for universal constants c_1 and c_2. Using the above we obtain

$$\begin{aligned} P(\max_{1\le k\le n}|U(t_k)| > \frac{K^3 r_n^{3/2}\log n}{n-1}) &\le \sum_{k=1}^{n}\sum_{l=1}^{r_n} P((n-1)|U_l(t_k)| > K^3 r_n \log n) \\ &\le n r_n c_1 \exp\Big(\frac{-c_2 K^3 \log(n)}{\sqrt{2}\pi + 9K(\log n)^{1/3} n^{-1/6}}\Big), \qquad K > 0. \end{aligned}$$

This shows that

$$\max_{1\le k\le n}|U(t_k)| = O_P(r_n^{3/2} n^{-1}\log n). \tag{41}$$

To deal with N_k^+ we introduce the degenerate U-statistic

$$\tilde{N}_k^+ = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{1}[D_{ij}\ge 0]\xi_k(i,j)$$

with

$$\begin{aligned} \xi_k(i,j) =& \mathbf{1}[t_{k-1}D_{ij} < \varepsilon_i - \varepsilon_j \le t_k D_{ij}] - F(\varepsilon_j + t_k D_{ij}) + F(\varepsilon_j + t_{k-1}D_{ij}) \\ & - F(\varepsilon_i - t_{k-1}D_{ij}) + F(\varepsilon_i - t_k D_{ij}) + F_2(t_k D_{ij}) - F_2(t_{k-1}D_{ij}) \end{aligned}$$

and F_2 the distribution function of $\varepsilon_1 - \varepsilon_2$. It is easy to see that

$$|N_k^+ - \tilde{N}_k^+| \leq 6B_1 C_n n^{-3/2} \frac{1}{n(n-1)} \sum_{i \neq j} |X_i - X_j|.$$

The kernel of the U-statistic $\tilde{N}_k^+$ is bounded by 8 and has second moment bounded by $D_n n^{-3/2}$ with $D_n = 2B_1 C_n E[|X_1 - X_2|]$. Thus, by part (c) of Proposition 2.3 in Arcones and Giné (1993), we see that the corresponding degenerate U-statistic $\tilde{N}_k^+$ satisfies

$$\sum_{k=1}^n P(|\tilde{N}_k^+| > \frac{K^3(\log n)^{3/2} n^{-1/2}}{n-1}) \leq nc_1 \exp\Big(-\frac{c_2 K^3 (\log n)^{3/2}}{D_n^{1/2} n^{-1/4} + 4K(\log n)^{1/2}}\Big).$$

The above shows that

$$\max_{1 \leq k \leq n} N_k^+ = O_P(n^{-3/2}(\log n)^{3/2}). \tag{42}$$

Similarly one obtains

$$\max_{1 \leq k \leq n} N_k^- = O_P(n^{-3/2}(\log n)^{3/2}). \tag{43}$$

The desired (38) follows from (39)-(43) and $S = O_P(1)$. This concludes the proof of (13). □

6.5 Auxiliary Results

Let X and Y be independent random variables. Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be independent copies of (X, Y). For reals t, x and y, set

$$N(t, x, y) = \sum_{i=1}^m (\mathbf{1}[Y_i - tX_i \leq y - tx] - \mathbf{1}[Y_i \leq y])$$

and

$$\tilde{N}(t, x, y) = N(t, x, y) - E[N(t, x, y)].$$

Lemma 4 *Suppose X has finite expectation and the distribution function F of Y is Lipschitz: $|F(y) - F(x)| \leq \Lambda|y - x|$ for all x, y and some finite constant Λ. Then the inequality*

$$P\Big(\sup_{|t| \leq \delta} |\tilde{N}(t, x, y)| > 4\eta\Big) \leq (8M + 4) \exp\Big(\frac{-\eta^2}{2m\Lambda\delta E[|X - x|] + 2\eta/3}\Big)$$

holds for $\eta > 0$, $\delta > 0$, real x and y and every integer $M \geq m\Lambda\delta E[|X - x|]/\eta$. In particular, for $C \geq 1$ and $K \geq 6\Lambda(1 + E[|X|])$, we have

$$\begin{aligned} p_m(y, C, K) &= \sup_{|x| \leq m^{1/4}} P\Big(\sup_{|t| \leq C/m^{1/2}} |\tilde{N}(t, x, y)| > 4KC^{1/2} m^{3/8} (\log m)^{1/2}\Big) \\ &\leq \Big(12 + \frac{8m^{3/8} C^{1/2})}{6(\log m)^{1/2}}\Big) \exp(-K \log(m)), \quad y \in \mathbb{R}. \end{aligned}$$

Proof Fix x and y and set $\nu = E[|X - x|]$. Abbreviate $N(t, x, y)$ by $N(t)$ and $\tilde{N}(t, x, y)$ by $\tilde{N}(t)$, set

$$N_+(t) = \sum_{i=1}^{m} (\mathbf{1}[Y_j - t(X_j - x) \le y] - \mathbf{1}[Y_j \le y])\mathbf{1}[X_j - x \ge 0],$$

$$N_-(t) = \sum_{i=1}^{m} (\mathbf{1}[Y_j - t(X_j - x) \le y] - \mathbf{1}[Y_j \le y])\mathbf{1}[X_j - x < 0]$$

and let $\tilde{N}_+(t) = N_+(t) - E[N_+(t)]$ and $\tilde{N}_-(t) = N_-(t) - E[N_-(t)]$. Since F is Lipschitz, we obtain

$$|E[N_+(t_1)] - E[N_+(t_2)]| \le m\Lambda |t_1 - t_2| \nu.$$

For $s \le t \le u$, we have

$$N_+(s) - E[N_+(u)] \le N_+(t) - E[N_+(t)] \le N_+(u) - E[N_+(s)]$$

and thus

$$\tilde{N}_+(s) - m\Lambda|u - s|\nu \le \tilde{N}_+(t) \le \tilde{N}_+(u) + m\Lambda|u - s|\nu.$$

It is now easy to see that

$$\sup_{|t| \le \delta} |\tilde{N}_+(t)| \le \max_{k=-M,\dots,M} |N_+(k\delta/M)| + m\Lambda\delta\nu/M$$

for every integer M. From this we obtain the bound

$$P(\sup_{|t|\le\delta} |\tilde{N}_+(t)| \ge 2\eta) \le \sum_{k=-M}^{M} P(|\tilde{N}_+(k\delta/M) > \eta) + P(m\Lambda\delta\nu/M > \eta).$$

The Bernstein inequality and the fact that the variance of

$$(\mathbf{1}[Y - t(X - x) \le y] - \mathbf{1}[Y \le y])\mathbf{1}[X \ge x]$$

is bounded by $\Lambda|t|\nu$ yield

$$P(|\tilde{N}_+(k\delta/M)| > \eta) \le 2 \exp\Big(-\frac{\eta^2}{2m\Lambda\delta\nu + 2\eta/3}\Big).$$

Thus we have

$$P(\sup_{|t|\le\delta} |\tilde{N}_+(t)| > 2\eta) \le 2(2M+1) \exp\Big(-\frac{\eta^2}{2m\Lambda\delta\nu + 2\eta/3}\Big)$$

for $M \ge m\Lambda\delta\nu/\eta$. Similarly, one verifies for such M,

$$P(\sup_{|t|\le\delta} |\tilde{N}_-(t)| > 2\eta) \le 2(2M+1) \exp\Big(-\frac{\eta^2}{2m\Lambda\delta\nu + 2\eta/3}\Big).$$

Since $\tilde{N}(t) = \tilde{N}_+(t) + \tilde{N}_-(t)$, we obtain the first result. The second result follows from the first one by taking $\delta = Cm^{-1/2}$, $\eta = KC^{1/2}m^{3/8}(\log m)^{1/2}$ and observing the inequality $(\log m)^{1/2}m^{-3/8} \le 1$. □

Acknowledgements We thank two reviewers and the Associate Editor for their knowledgeable comments and suggestions, which helped us improve the paper.

References

1. Arcones, M. and Giné E. (1993). Limit theorems for U-processes. *Ann. Probab.*, **21**, 1494-1542.
2. Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.*, **10**, 647-671.
3. Chen, S.X., Peng, L. and Qin, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, **26**, 711-722.
4. Forrester, J., Hooper, W., Peng H. and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions*, **21**, 109-138.
5. Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079-1111.
6. Jin, K. (1992) Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.* **20**, 1844-1874.
7. Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.* **60**, 932-945.
8. Koul, H.L. and Susarla, V. (1983). Adaptive estimation in linear regression. *Statist. Decisions*, **1**, 379-400.
9. Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Second edition. Wiley Series in Probability and Statistics, Wiley, Hoboken.
10. Müller, U.U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, **37**, 2245-2277.
11. Müller, U.U. and Schick, A. (2017). Efficiency transfer for regression models with responses missing at random. *Bernoulli*, **23**, 2693-2719.
12. Müller, U.U. and Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electron. J. Stat.*, **6**, 1200-1219.
13. Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.
14. Owen, A.B. (2001). *Empirical Likelihood.* Monographs on Statistics and Applied Probability, **92**, Chapman & Hall.
15. Peng, H. and Schick, A. (2005). Efficient estimation of linear functionals of a bivariate distribution with equal, but unknown marginals: the least-squares approach. *J. Multivariate Anal.*, **95**, 385-409.
16. Peng, H. and Schick, A. (2013). An empirical likelihood approach to goodness of fit testing. *Bernoulli*, **19**, 954-981.
17. Peng, H. and Schick, A. (2016). Inference in the symmetric location model: an empirical likelihood approach. Preprint.
18. Peng, H. and Schick, A. (2017). Maximum empirical likelihood estimation and related topics. Preprint.
19. Qin, J. and Lawless J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300-325.
20. Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, **16**, 89-105.
21. Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.*, **21**, 1486-521. Correction and addendum: **23** (1995), 1862-1863.
22. Schick, A. (2013). Weighted least squares estimation with missing responses: An empirical likelihood approach. *Electron. J. Statist.*, **7**, 932-945.
23. Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer, New York.