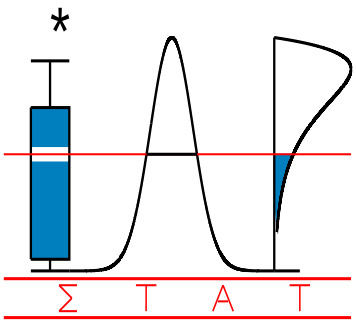
T E C H N I C A L

R E P O R T

0415

EXTENDING THE SCOPE OF EMPIRICAL LIKELIHOOD

N.L. HJORT, I.W. MCKEAGUE and I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Extending the Scope of Empirical Likelihood

Nils Lid Hjort, Ian W. McKeague, and Ingrid Van Keilegom

University of Oslo, Florida State University,
and Université catholique de Louvain

– April 2004 –

ABSTRACT. This paper extends the scope of empirical likelihood methodology in three directions: to allow for plug-in estimates of nuisance parameters in estimating equations, slower than $\sqrt{n}$-rates of convergence, and settings in which there are a relatively large number of estimating equations compared to the sample size. Calibrating empirical likelihood confidence regions with plug-in is sometimes intractable due to the complexity of the asymptotics, so we introduce a bootstrap approximation that can be used in such situations. We provide a range of examples from survival analysis and nonparametric statistics to illustrate the main results.

KEY WORDS: *bootstrap calibration, current status data, empirical processes, estimating equations, growing number of parameters, nonparametric regression, nuisance parameters, orthogonal series, plug-in*

1. Introduction

Empirical likelihood (Owen, 1990, 2001) has traditionally been used for providing confidence regions for multivariate means and, more generally, for parameters in estimating equations, under various standard assumptions: the number of estimating equations is fixed, they do not involve nuisance parameters, and the parameters of interest are estimable at $\sqrt{n}$-rate, where n is the sample size. Under such assumptions and with i.i.d. observations (or even dependent observations, see e.g. Ch. 8 of Owen, 2001), empirical likelihood (EL) based confidence regions can be calibrated using a nonparametric version of Wilks's theorem involving a chi-squared limiting distribution.

The aim of the present paper is to develop adaptations when the traditional assumptions are violated. More specifically, under certain weak asymptotic stability conditions, we establish generalisations of the basic theorem of EL to allow for plug-in estimates of nuisance parameters in the estimating equations, for slower than $\sqrt{n}$-rates of convergence, and for i.i.d. settings in which there are a relatively large number of estimating equations compared to the sample size. Several of our examples share the characteristic that they would be harder to analyse with other methods. In particular, the method of profile EL (see e.g. Owen, 2001, page 42) for dealing with nuisance parameters in estimating equations is often not applicable for infinite dimensional nuisance parameters, and even when it is applicable, implementation can be computationally difficult. The triangular array EL theorem of Owen (2001, page 85) applies under slower than $\sqrt{n}$-rates, and has been useful

in the context of nonparametric density estimation, for instance, but is not flexible enough to handle estimating functions with plug-in.

The use of plug-in for nuisance parameters in EL confidence regions is not new. It has recently been applied in various survival analysis contexts, see Qin and Jing (2001a, 2001b), Wang and Jing (2001), Li and Wang (2003) and Qin and Tsao (2003). The technique has also been used in survey sampling with imputation for missing response, see Wang and Rao (2002). Our aim here, however, is to provide a more widely applicable version of this approach, that can accommodate a wide array of examples, allowing both plug-in and slower than $\sqrt{n}$-rates of convergence. We take the point of view that it is preferable to derive a general result using generic assumptions, that can be checked fairly easily in specific applications, rather than reinventing the basic theory on each occasion. Calibrating EL confidence regions with plug-in is sometimes intractable due to the complexity of the asymptotics, so we introduce a bootstrap approximation that can be used in such situations.

To illustrate our general results we consider a range of examples from survival analysis and nonparametric statistics in settings where the inference is based on estimating functions. In particular, we look at functionals of survival distributions with right censored data (treated via EL in Wang and Jing, 2001), the error distribution in nonparametric regression (Akritas and Van Keilegom, 2001), density estimation (treated by EL in Hall and Owen, 1993, and Chen, 1996), and survival function estimation from current status data (van der Laan and van der Vaart, 2000).

Standard maximum likelihood theory for parametric models, as well as EL theory, keeps the dimension of the parameter (or the number of estimating equations) fixed, say at p, as sample size n grows. This is what leads to asymptotic normality, Wilks type theorems for likelihood ratio statistics and Owen type theorems for EL. Portnoy (1986, 1988) and others have investigated the extent to which maximum likelihood theory based results still hold, when p is allowed to increase with n. The canonical growth restriction for normal approximations to hold is that $p^2/n \to 0$, while $p^{3/2}/n \to 0$ typically suffices for certain quadratic approximations associated with Wilks theorems to hold.

In this article we investigate the similar problem of finding conditions under which the EL methods continue to work adequately when p grows. The canonical growth condition will be seen to be $p^3/n \to 0$. Under this condition, in addition to other requirements that have to do with stability of eigenvalues of covariance matrices, minus twice the log-EL can be approximated well enough with a certain quadratic form that in itself is close to a χ^2_p.

We should add that in situations with a high number of parameters the typical aim is not to provide a simultaneous confidence region for the full parameter vector, say $(\mu_1, \ldots, \mu_p)$. It could rather be to test whether a subset of the parameters have zero values, or to compare one distribution with another, or to make inference for a focus parameter, say $\phi = f(\mu_1, \ldots, \mu_p)$. Each of these tasks can be done with EL methods, inside

our framework for growing p. Setting a confidence for such a focus parameter ϕ may e.g. be done via profile EL methods, as in Owen (2001, Section 3.4).

The paper is organised as follows. Section 2 develops the EL theory with plug-in under $\sqrt{n}$-rate of convergence, including the bootstrap approximation of the limiting distribution of the EL statistic. Four examples are discussed in Section 3. The theory is extended to slower than $\sqrt{n}$-rate in Section 4. In Section 5 we examine the limiting behavior of the EL statistic in situations where the number of estimating functions is allowed to increase with growing sample size. Some examples are presented in Section 6, including setups with 'growing polynomial regression' and 'growing exponential families'. Proofs can be found in the Appendix.

2. Plug-in empirical likelihood

We first describe the general framework. The basic idea of empirical likelihood (EL) is to regard the observations $Z_1, \ldots, Z_n$ as if they are i.i.d. from a fixed and unknown d-dimensional distribution P, and to model P by a multinomial distribution P_n concentrated on the observations. Inference for the parameter(s) of interest, $\theta_0 = \theta_0(P) \in \Theta$, is then carried out using a p-dimensional estimating function of the form $m(Z, \theta, h)$, where, for the purposes of the present paper, h is a (possibly infinite dimensional) 'nuisance' parameter with unknown true value $h_0 = h_0(P) \in \mathcal{H}$. A slight extension would be to allow m to change with n; our results carry through immediately, but as the applications (in Section 3) do not require such generality, we do not consider this extension. In the slower than $\sqrt{n}$-rate applications (Section 4), however, we do need to allow for dependence on n.

When h_0 is known, it can replace h in the EL ratio function

$$\mathrm{EL}_n(\theta, h) = \max\Big\{\prod_{i=1}^n (nw_i) \colon \text{each } w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i m(Z_i, \theta, h) = 0\Big\},$$

leading to a confidence region $\{\theta \colon \mathrm{EL}_n(\theta, h_0) > c\}$ for θ_0. Here the constant c can be calibrated using Owen's (1990) EL theorem: if the observations are i.i.d. and $m(Z, \theta_0, h_0)$ has zero mean and a positive definite covariance matrix, then

$$-2 \log \mathrm{EL}_n(\theta_0, h_0) \to_d \chi^2_p,$$

where χ^2_p has a chi-squared distribution with p degrees of freedom.

2.1. Main result. We now establish a plug-in version of Owen's result in which the unknown h_0 is replaced by an estimator $\widehat{h}$, leading to a calibration for $\{\theta \colon \mathrm{EL}_n(\theta, \widehat{h}) > c\}$ as a confidence region for θ_0. For EL to be useful, there needs to exist a solution to the above maximisation problem given the data (even when h_0 is known), so the existence of a solution at $h = \widehat{h}$ is assumed implicitly. Apart from this, we extract the basic structure of Owen's result, and only rely on 'generic' asymptotic stability conditions, (A1)–(A3)

below, which do not require i.i.d. observations or consistency of $\widehat{h}$, although such structure may very well be needed to check the conditions in specific applications. The proof (in the Appendix) is essentially the same as Owen's, but is included here for easy reference in developing our subsequent results.

We use the following notation: $m^{\otimes 2} = mm^{\mathrm{t}}$,

$$M_n(\theta, h) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \theta, h), \quad S_n(\theta, h) = \frac{1}{n}\sum_{i=1}^{n} m^{\otimes 2}(Z_i, \theta, h).$$

For matrices $A = (a_{i,j})$ we use $|A| = \max_{i,j} |a_{i,j}|$, and for vectors a, $\|a\|$ denotes the Euclidean norm. The conditions are as follows. Here and later we use $\to_d$ and $\to_{\mathrm{pr}}$ to indicate respectively convergence in distribution and in probability, while $o_{\mathrm{pr}}(\cdot)$ and $O_{\mathrm{pr}}(\cdot)$ are stochastic order notation: $X_n = o_{\mathrm{pr}}(a_n)$ means that $X_n/a_n \to_{\mathrm{pr}} 0$ whereas $X_n = O_{\mathrm{pr}}(a_n)$ means that X_n/a_n is bounded in probability.

(A1) $n^{1/2} M_n(\theta_0, \widehat{h}) \to_d U$ where $U \sim \mathrm{N}_p(0, V_1)$ for some positive definite matrix V_1.
(A2) $S_n(\theta_0, \widehat{h}) \to_{\mathrm{pr}} V_2$ for some positive definite matrix V_2.
(A3) $\max_{1 \le i \le n} \|m(Z_i, \theta_0, \widehat{h})\| = o_{\mathrm{pr}}(n^{1/2})$.

THEOREM 2.1. *Under conditions (A1)–(A3),*

$$-2 \log \mathrm{EL}_n(\theta_0, \widehat{h}) \to_d U^{\mathrm{t}} V_2^{-1} U.$$

The limit distribution may also be expressed as $r_1\chi^2_{1,1} + \cdots + r_p\chi^2_{1,p}$, where the $\chi^2_{1,j}$s are independent chi-squared random variables with one degree of freedom and the weights $r_1, \ldots, r_p$ are the eigenvalues of $V_2^{-1}V_1$.

REMARK 2.1. When V_1 and V_2 coincide, we have the standard χ^2_p limit distribution and there is no perturbation due to plug-in. When V_1 and V_2 are not identical, the weights $r_1, \ldots, r_p$ may need to be estimated, for example via consistent estimators $\widehat{V}_1$, $\widehat{V}_2$ and computing the eigenvalues of $\widehat{V}_2^{-1}\widehat{V}_1$. It is not possible to say anything in general about estimation of V_1, which will depend on the structure of the specific application; later in this section we examine a bootstrap approach which can be applied when V_1 is difficult to estimate by other means. An estimator of V_2 is easily provided by condition (A2), with plug-in of a consistent estimator $\widehat{\theta}$ for θ_0. In the Appendix we show that $\widehat{V}_2 = S_n(\widehat{\theta}, \widehat{h})$ consistently estimates V_2 under the following two additional conditions:

(A4) For some subset $\bar{\mathcal{H}}$ of $\mathcal{H}$ such that $P\{\widehat{h} \in \bar{\mathcal{H}}\} \to 1$, and for some $\delta > 0$, the class of functions $\mathcal{F} = \{m^{\otimes 2}(\cdot, \theta, h) : \|\theta - \theta_0\| < \delta, h \in \bar{\mathcal{H}}\}$ has the Glivenko–Cantelli property, i.e.

$$\sup_{\|\theta-\theta_0\|<\delta, h\in\bar{\mathcal{H}}} \Big| \frac{1}{n}\sum_{i=1}^{n} \{m^{\otimes 2}(Z_i, \theta, h) - \mathrm{E}m^{\otimes 2}(Z, \theta, h)\} \Big| \to_{\mathrm{pr}} 0.$$

(A5) For any real sequence $\delta_n \downarrow 0$,

$$\sup_{\|\theta-\theta_0\|\le\delta_n, h\in\bar{\mathcal{H}}} |\mathrm{E}m^{\otimes 2}(Z,\theta,h) - \mathrm{E}m^{\otimes 2}(Z,\theta_0,h)| \to 0.$$

REMARK 2.2. For i.i.d. observations, with $m(Z,\theta_0,h_0)$ having zero mean (where h_0 is the true value of h) and a finite covariance matrix V_0, the multivariate central limit theorem implies

$$n^{1/2}M_n(\theta_0,h_0) \to_d \mathrm{N}(0,V_0),$$

so condition (A1) describes the perturbation of V_0 due to replacing h_0 by $\widehat{h}$. In the highly smooth case that $M(\theta_0,\widehat{h}) = o_{\mathrm{pr}}(n^{-1/2})$, where $M(\theta,h) = \mathrm{E}m(Z,\theta,h)$, it can be shown (under some additional assumptions) that there is no perturbation: $V_1 = V_0$. For instance, suppose that the class of functions $\{m(\cdot,\theta_0,h)\colon h \in \mathcal{H}\}$ is Donsker, and $\widehat{h}$ is consistent in the sense that $\rho_j(\widehat{h},h_0) \to_{\mathrm{pr}} 0$ for $j = 1,\dots,p$, where $\rho_j(h,h_0) = \mathrm{E}\{m_j(Z,\theta_0,h) - m_j(Z,\theta_0,h_0)\}^2$. Then

$$n^{1/2}M_n(\theta_0,\widehat{h}) = n^{-1/2}\sum_{i=1}^{n}\{m(Z_i,\theta_0,\widehat{h}) - M(\theta_0,\widehat{h})\} + n^{1/2}M(\theta_0,\widehat{h}) \to_d \mathrm{N}_p(0,V_0),$$

so $V_1 = V_0$, where empirical process theory is used to obtain weak convergence of the first term, cf. van der Vaart (1998, p. 280). However, $M(\theta_0,\widehat{h}) = o_{\mathrm{pr}}(n^{-1/2})$ is a strong condition, so we have avoided using it in favour of the less restrictive condition (A1), which is flexible enough to be checked within the context of the examples considered in the next section.

REMARK 2.3. The assumption of a normal limit in (A1) is not crucial, although the limit distribution of the likelihood ratio statistic then takes on a more complicated form and simulation may be needed to calibrate the confidence region. In fact, we could replace (A1) by

(A1′) $n^{1/2}M_n(\theta_0,\widehat{h}) \to_d U$ for some p-dimensional continuous random vector U.

REMARK 2.4. Kitamura (1997) introduces blockwise EL with estimating functions, without plug-in, in models having weakly dependent stationary observations. The maximum EL estimator under blocking is shown to have greater efficiency than the standard maximum EL estimator, but the blockwise approach has not been extended to allow plug-in. Standard EL (with plug-in), however, can still provide accurate confidence sets under dependent observations, for according to Theorem 2.1 the limiting distribution of the standard EL statistic, while not chi-square, is of a tractable form. If m does not depend on n and there is no plug-in, conditions (A1) and (A2) can be checked by central limit theorems and ergodic theorems for weakly dependent sequences. Condition (A3) holds provided

$\mathrm{E}\|m(Z,\theta_0)\|^2 < \infty$ by a Borel–Cantelli argument (cf. Owen, 2001, Lemma 11.2). For a one-dimensional estimating function $m(Z,\theta)$ ($p=1$) such that $\mathrm{E}m(Z,\theta)=0$, the limiting distribution of the EL statistic is $r\chi^2_1$, where $r = \sum_{i=1}^{\infty} \mathrm{Cov}\{m(Z_1,\theta), m(Z_i,\theta)\}/\mathrm{Var}\{m(Z,\theta)\}$ could be estimated easily.

REMARK 2.5. If the nuisance parameter has finite dimension q, the estimating function is smooth, and $\widehat{h} = \widehat{h}_\theta$ maximises $\mathrm{EL}_n(\theta,h)$ over h (i.e. $\mathrm{EL}_n(\theta,\widehat{h}_\theta)$ is a profile EL), it is known that $-2\log \mathrm{EL}_n(\theta_0,\widehat{h}_{\theta_0})$ has a χ^2_q limiting distribution (see Owen, 2001, page 55). This differs from the limiting distribution of Theorem 2.1, but V_1 or V_2 would be singular in this case, and our result does not apply.

2.2. Bootstrap calibration. As mentioned above, the estimation of V_1 can be difficult in certain situations and, more seriously, U may not be normally distributed, in which case a bootstrap calibration is desirable. The procedure developed below consists in replacing U (in the distribution of $U^{\mathrm{t}}V_2^{-1}U$) by a consistent bootstrap estimator, and consistently estimating V_2.

Let $\{Z_1^*,\ldots,Z_n^*\}$ be drawn randomly with replacement from $\{Z_1,\ldots,Z_n\}$, define $M_n^*(\theta,h) = n^{-1}\sum_{i=1}^n m(Z_i^*,\theta,h)$ for each θ, and h and let $\widehat{h}^*$ be the same estimator as $\widehat{h}$ but based on the bootstrap data. Also, let $\widehat{\theta}$ be a consistent estimator of θ_0, and $\widehat{V}_2 = S_n(\widehat{\theta},\widehat{h})$.

We use the abbreviated notation $\Delta_n = M_n - M$, as a function of (θ,h), and Δ_n^* denotes the bootstrap version of Δ_n (here and in the sequel we define the bootstrap version of any statistic as the expression obtained by replacing M, M_n, θ_0, h_0 and $\widehat{h}$ by $M_n, M_n^*, \widehat{\theta}, \widehat{h}$ and $\widehat{h}^*$, respectively). Let $\|\cdot\|_{\mathcal{H}}$ denote a semi-norm on $\mathcal{H}$. Also let $\Phi_n = n^{1/2}\{\Delta_n(\theta_0,h_0) + \Gamma(\theta_0,h_0)[\widehat{h}-h_0]\}$, where $\Gamma(\theta_0,h_0)[\widehat{h}-h_0]$ is the Gâteaux derivative of $M(\theta_0,h_0)$ in the direction $\widehat{h}-h_0$ (see e.g. Bickel, Klaassen, Ritov and Wellner, 1993, page 453). The bootstrap analogue of Φ_n is denoted by Φ_n^*. Finally, let P^* denote the bootstrap distribution conditional on the data. The following conditions are needed to formulate the validity of the bootstrap approximation:

(B1) $\sup_{t\in\mathcal{R}^p} |P^*\{\Phi_n^* \le t\} - P\{\Phi_n \le t\}| \to_{\mathrm{pr}} 0$.
(B2) $\sup_{\|\theta-\theta_0\|\le\delta_n, \|h-h_0\|_{\mathcal{H}}\le\delta_n} \|\Delta_n(\theta,h) - \Delta_n(\theta_0,h_0)\| = o_{\mathrm{pr}}(n^{-1/2})$ for all $\delta_n \downarrow 0$.
(B3) $\|M(\theta_0,\widehat{h}) - M(\theta_0,h_0) - \Gamma(\theta_0,h_0)[\widehat{h}-h_0]\| \le c\|\widehat{h}-h_0\|^2_{\mathcal{H}}$ for some $c>0$.
(B4) $\|\widehat{h}-h_0\|_{\mathcal{H}} = o_{\mathrm{pr}}(n^{-1/4})$.
(B5) The bootstrap analogues of conditions (B2)–(B4) hold pr-a.s.

THEOREM 2.2. *Under conditions (A1′), (A2)–(A5) and (B1)–(B5),*

$$\sup_{t\ge 0}\Big| P^*\{n[M_n^*(\widehat{\theta},\widehat{h}^*) - M_n(\widehat{\theta},\widehat{h})]^{\mathrm{t}}\widehat{V}_2^{-1}[M_n^*(\widehat{\theta},\widehat{h}^*) - M_n(\widehat{\theta},\widehat{h})] \le t\} - P\{-2\log \mathrm{EL}_n(\theta_0,\widehat{h}) \le t\}\Big| \to_{\mathrm{pr}} 0.$$

REMARK 2.6. When $\widehat{\theta}$ is defined as the minimiser of $\|M_n(\theta, \widehat{h})\|$, sufficient conditions for $\widehat{\theta}$ to be consistent can be found in Theorem 1 in Chen, Linton and Van Keilegom (2003). In order to verify condition (B2) in the case of i.i.d. observations, it suffices by Corollary 2.3.12 in van der Vaart and Wellner (1996) to show that the class $\{m(\cdot, \theta, h)\colon \theta \in \Theta, h \in \mathcal{H}\}$ is Donsker, and that

$$\mathrm{Var}\{m(Z, \theta, h) - m(Z, \theta_0, h_0)\} \le K_1\|\theta - \theta_0\| + K_2\|h - h_0\|_{\mathcal{H}} + \varepsilon_n$$

for some $K_1, K_2 \ge 0$, and for some $\varepsilon_n \downarrow 0$. The bootstrap analogue of (B2) then follows from Giné and Zinn (1990), provided

$$\mathrm{Var}^*\{m(Z^*, \theta, h) - m(Z^*, \widehat{\theta}, \widehat{h})\} \le K_1'\|\theta - \widehat{\theta}\| + K_2'\|h - \widehat{h}\|_{\mathcal{H}} + \varepsilon_n'$$

for some $K_1', K_2' = O(1)$ a.s. and for some $\varepsilon_n' = o(1)$ a.s. Finally, condition (B3) and its bootstrap version can often be verified by using a two-term Taylor expansion of $M(\theta_0, \widehat{h})$ and of $M(\widehat{\theta}, \widehat{h}^*)$ around h_0 and $\widehat{h}$, respectively.

2.3. Simultaneous confidence bands. We now briefly discuss an extension of our approach that may be useful for obtaining a simultaneous confidence band for a function $t \mapsto \theta_0(t)$ defined on an interval $\mathcal{T}$. Given an estimating function $m(Z, \theta(t), h, t)$ of the previous form, but now also depending on t, the earlier definitions extend in the obvious way. It can be shown that the process $-2\log \mathrm{EL}_n(\theta_0(t), \widehat{h}, t)$ has a weak limit of the form $U(t)^{\mathrm{t}}V_2(t)^{-1}U(t)$ under the following conditions:

(A1*) $n^{1/2}M_n(\theta_0(t), \widehat{h}, t)$ converges weakly to a process $U(t)$.

(A2*) $\sup_{t\in\mathcal{T}} |S_n(\theta_0(t), \widehat{h}, t) - V_2(t)| \to_{\mathrm{pr}} 0$ for a matrix-valued function $V_2(t)$, and the eigenvalues of $V_2(t)$ are uniformly bounded away from zero and infinity.

(A3*) $\sup_{1\le i\le n, t\in\mathcal{T}} \|m(Z_i, \theta_0(t), \widehat{h}, t)\| = o_{\mathrm{pr}}(n^{1/2})$.

The simultaneous confidence band would need to be calibrated from the quantiles of $\sup_{t\in\mathcal{T}} |\widehat{U}(t)^{\mathrm{t}}\widehat{V}_2(t)^{-1}\widehat{U}(t)|$, where $\widehat{V}_2(t)$ is a uniformly consistent estimator of $V_2(t)$ and $\widehat{U}(t)$ is an estimated version of the process $U(t)$.

3. Applications of the plug-in theory

This section gives four illustrations of the preceding plug-in theory. The first uses parametric plug-in for a nonparametric estimand while the three others effectively use nonparametric plug-in to solve nonparametric empirical likelihood problems.

3.1. Symmetric distribution functions. Let F be a continuous distribution function that is symmetric about an unknown location a, so $F(x) = 1 - F(2a - x)$ for all x. Consider estimation of $\theta_0 = F(x)$ at a fixed x from n i.i.d. observations from F. The

estimating function has $p = 2$ components (the first being the usual estimating function and the second making use of the symmetry assumption):

$$m(X, \theta, a) = \begin{pmatrix} 1\{X \le x\} - \theta \\ 1\{X > 2a - x\} - \theta \end{pmatrix}.$$

With plug-in of the sample median $\widehat{a}$ in place of a, and provided $0 < \theta_0 < 1$, Theorem 2.1 gives

$$-2 \log \mathrm{EL}_n(\theta_0, \widehat{a}) \to_d \chi^2_2,$$

which is the same limit as in the case a is known. Condition (A1) can be checked using a Skorohod construction (cf. Li and Doss, 1993, p. 788), and

$$V_1 = V_2 = \begin{pmatrix} \theta_0(1-\theta_0) & -\theta_0^2 \\ -\theta_0^2 & \theta_0(1-\theta_0) \end{pmatrix}.$$

3.2. Integral of squared densities. Let $X_1, \ldots, X_n$ be i.i.d. from an unknown density f_0. The quantity $\theta_0 = \int f_0^2 \, \mathrm{d}x$ is of interest for various problems related to non-parametric density estimation. The limit distribution of the Hodges–Lehmann estimator of location has variance proportional to $1/\theta_0^2$, see Lehmann (1983, page 383). Similarly, the power of the Wilcoxon rank test is essentially determined by the size of θ, see Lehmann (1975, page 72).

To see how our extended EL machinery can be used to make inference about this parameter, study $m(X, \theta, f) = f(X) - \theta$, for which $\mathrm{E}m(X, \theta_0, f_0) = 0$. We employ a kernel density estimator $\widehat{f}(x) = n^{-1} \sum_{i=1}^n k_b(X_i - x)$, where $k_b = k(\cdot/b)/b$ is a scaled version of a symmetric and bounded kernel function k using bandwidth $b = b_n$, and wish to use the plug-in likelihood $\mathrm{EL}_n(\theta, \widehat{f})$. (For discussion of methods for deciding on good kernel bandwidths, when the specific purpose is precise estimation of θ, see Schweder, 1975.) For this we must go through conditions (A1)–(A3) of Theorem 2.1. Define

$$V = \int (f_0 - \theta_0)^2 f_0 \, \mathrm{d}x = \int f_0^3 \, \mathrm{d}x - \Big(\int f_0^2 \, \mathrm{d}x \Big)^2;$$

this is the variance of the limit distribution of $n^{1/2} M_n(\theta_0, f_0)$. To check (A2) first, write

$$S_n(\theta_0, \widehat{f}) = n^{-1} \sum_{i=1}^n \{\widehat{f}(X_i) - \theta_0\}^2 = \int \widehat{f}^2 \, \mathrm{d}F_n - 2\theta_0 \widehat{\theta} + \theta_0^2,$$

in terms of the empirical distribution function F_n and $\widehat{\theta} = n^{-1} \sum_{i=1}^n \widehat{f}(X_i) = \int \widehat{f} \, \mathrm{d}F_n$. One may now prove that $\int \widehat{f} \, \mathrm{d}F_n$ and $\int \widehat{f}^2 \, \mathrm{d}F_n$ have the required limits in probability $\int f_0^2 \, \mathrm{d}x$ and $\int f_0^3 \, \mathrm{d}x$, provided $b \to 0$ and $nb \to \infty$. This leads to $S_n(\theta_0, \widehat{f}) \to_{\mathrm{pr}} V$ and verifies (A2).

It is a little more laborious task to go through (A1), which also demands a more precise study of

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(X_i) = \frac{1}{n^2}\sum_{i,j} k_b(X_i - X_j) = \frac{k(0)}{nb} + \frac{n-1}{n}\widehat{g}.$$

Here $\widehat{g} = \widehat{g}(0)$, where $\widehat{g}(y) = \binom{n}{2}^{-1}\sum_{i<j}\bar{k}_b(Y_{i,j}, y)$ is a natural kernel estimator of the difference density $g(y) = \int f(y+x)f(x)\,\mathrm{d}x$ of $Y_{i,j} = X_i - X_j$; here $\bar{k}_b(Y_{i,j}, y) = \frac{1}{2}\{k_b(Y_{i,j} - y) + k_b(Y_{i,j} + y)\}$. Hjort (1999, Section 7) shows that $\widehat{g}(y)$ has mean value $g(y) + \frac{1}{2}b^2 g''(y)\int u^2 k(u)\,\mathrm{d}u + o(b^2)$, with variance $(4/n)\{g^*(y) - g(y)^2\}$ plus smaller order terms, where $g^*(y) = (1/4)\{\bar{g}(y,y) + \bar{g}(y,-y) + \bar{g}(-y,y) + \bar{g}(-y,-y)\}$ and $\bar{g}(y_1,y_2)$ is the simultaneous density of two related differences $(X_2 - X_1, X_3 - X_1)$. Thus $n^{1/2}M_n(\theta_0, \widehat{f}) = n^{1/2}(\widehat{\theta} - \theta_0)$ has mean of order $O(1/(n^{1/2}b) + n^{1/2}b^2)$ and variance going to $4V$. This, in conjunction with the asymptotic theory of U-statistics, verifies (A1), under the conditions $n^{1/2}b \to \infty$ and $n^{1/2}b^2 \to 0$. (If $b = b_0 n^{-\alpha}$, we need $\frac{1}{4} < \alpha < \frac{1}{2}$.) Finally, for (A3), note that $\widehat{f}(x) \le b^{-1}k_{\max}$ for all x, where $k_{\max}$ is the maximum of $k(u)$. Hence $\max_{i\le n}|\widehat{f}(X_i) - \theta_0|$ is bounded by $b^{-1}k_{\max} + \theta_0$, which implies (A3), provided only that $n^{1/2}b \to \infty$. We may conclude that $-2\log \mathrm{EL}_n(\theta_0, \widehat{f}) \to_d 4\chi_1^2$.

3.3. Functionals of survival distributions. Wang and Jing (2001) developed a plug-in version of EL for a class of functionals of a survival function (including its mean) in the presence of censoring. Denote the survival and censoring distribution functions by F and G, respectively. The parameter of interest here is a general linear functional of F, say

$$\theta = \theta(F) = \int_0^\infty \xi(t)\,\mathrm{d}F(t),$$

where $\xi(t)$ is some nonnegative measurable function for which the mean $\theta(F)$ is finite.

The estimating function m implicit in the approach of Wang and Jing is given by

$$m(Z, \delta, \theta, G) = \frac{\xi(Z)\delta}{1 - G(Z)} - \theta,$$

where $Z = \min(X, Y)$, $\delta = I\{X < Y\}$, $Y \sim G$, and X and Y are independent. It is easily shown that $\mathrm{E}m(Z, \delta, \theta, G) = 0$, which is the basic identity underlying the interpretation of the Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average; see Satten and Datta (2001) for discussion and references.

The censoring distribution function G plays the role of h in our framework, and we set $\widehat{h}(t) = \widehat{G}_n(t \wedge Z_{(n)})$, where $\widehat{G}_n$ is the Kaplan–Meier estimator of G. We refer to Wang and Jing (2001) for the assumptions. The conditions (A1)–(A3) needed to apply our Theorem 2.1 are now checked by referring to various parts of the proof of their Theorem 2.1. (A1) follows from their lemma on page 524, with V_1 being the asymptotic

variance of $\widehat{\theta} = \theta(\widehat{F}_n)$, where $\widehat{F}_n$ is the Kaplan–Meier estimator of F. For (A2), note that $\max_{i \le n} |m(Z_i, \delta_i, \theta_0, G) - m(Z_i, \delta_i, \theta_0, \widehat{G}_n)| = o_{\mathrm{pr}}(n^{1/2})$, cf. Wang and Jing (op. cit., page 525), which implies

$$S_n(\theta_0, \widehat{h}) = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \delta_i, \theta_0, \widehat{G}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \delta_i, \theta_0, G)^2 + o_{\mathrm{pr}}(1) \to_{\mathrm{pr}} V_2,$$

where $V_2 < \infty$ by their condition (C3). Condition (A3) is their displayed inequality immediately before (4.5).

It remains to provide consistent estimators of V_1 and V_2, and we do this along the lines of Remark 2.1. Stute's (1996) jackknife estimator can be used for $\widehat{V}_1$. Under conditions (A4)–(A5), we have that $\widehat{V}_2 = S_n(\widehat{\theta}, \widehat{h})$ consistently estimates V_2. To check (A4), assume that $G(\tau_H-) < 1$, i.e. $P\{Y \ge \tau_H\} > 0$. Let $G(\tau_H-) < c < 1$. Specify $\bar{\mathcal{H}}$ as the class of increasing nonnegative functions h such that $h(\tau_H-) < c$ and $h(t) = h(\tau_H)$ for $t \ge \tau_H$. Now,

$$\begin{aligned} \sup_{0 \le t < \tau_H} |\widehat{h}(t) - G(t)| &\le \sup_{0 \le t < \tau_H} |\widehat{h}(t) - G(t \wedge Z_{(n)})| + \sup_{0 \le t < \tau_H} |G(t \wedge Z_{(n)}) - G(t)| \\ &= \sup_{0 \le t \le Z_{(n)}} |\widehat{G}_n(t) - G(t)| + \sup_{Z_{(n)} < t < \tau_H} |G(Z_{(n)}) - G(t)| \to_{\mathrm{pr}} 0, \end{aligned}$$

by uniform consistency of $\widehat{G}_n$ on the interval $[0, Z_{(n)}]$, see Wang (1987). Thus $P\{\widehat{h} \in \bar{\mathcal{H}}\} = P\{\widehat{h}(\tau_H-) < c\} \to 1$. The class $\{1/(1-h) \colon h \in \bar{\mathcal{H}}\}$ is contained in the class of all monotone functions into $[0, 1/(1-c)]$, which is Glivenko–Cantelli, see van der Vaart and Wellner (1996, p. 149). Thus, using the preservation property of Glivenko–Cantelli classes under a continuous function, see van der Vaart and Wellner (2000), it follows that $\mathcal{F}$ is Glivenko–Cantelli. Condition (A5) follows from

$$\begin{aligned} \mathrm{E}|m^2(Z, \theta, h) - m^2(Z, \theta_0, h)| &\le \mathrm{E}\left(|m(Z, \theta, h) - m(Z, \theta_0, h)|\, |m(Z, \theta, h) + m(Z, \theta_0, h)|\right) \\ &\le \|\theta - \theta_0\| \{\|\theta + \theta_0\| + 2\,\mathrm{E}|\xi(Z)|/(1-c)\} \end{aligned}$$

for $h \in \bar{\mathcal{H}}$.

3.4. Error distributions in nonparametric regression. Consider the model $Y = \mu(X) + \varepsilon$, where X and ε are independent, ε has unknown distribution function F_ε, and $\mu(\cdot)$ is an unknown regression function. We now use our approach with bootstrap calibration to construct an EL confidence interval for $\theta_0 = F_\varepsilon(y)$, at a fixed point y. The same assumptions as in Akritas and Van Keilegom (2001) are imposed. In particular, F_ε is assumed to be continuous, $\mu(\cdot)$ is smooth and X is bounded. For simplicity we restrict X to $(0, 1)$.

Consider the Nadaraya–Watson estimator $\widehat{\mu}(x) = \sum_{i=1}^{n} W_{n,i}(x; b_n) Y_i$, where

$$W_{n,i}(x; b_n) = k_{b,x}(X_i) / \sum_{j=1}^{n} k_{b,x}(X_j),$$

in terms of a kernel function k and scaled versions $k_{b,x}(u) = b^{-1}k((u-x)/b)$ thereof, with $b = b_n = b_0 n^{-2/7}$ a bandwidth sequence (other choices of the bandwidth are possible). The estimating function is $m(X, Y, \theta, \mu) = I\{Y - \mu(X) \le y\} - \theta$. We now check the conditions of Theorem 2.1. (A1) follows from the asymptotic normality of the estimator $\widehat{\theta} = n^{-1}\sum_{i=1}^{n} I\{\widehat{\varepsilon}_i \le y\}$ (with $\widehat{\varepsilon}_i = Y_i - \widehat{\mu}(X_i)$), given by Theorem 2 in Akritas and Van Keilegom (2001): $n^{1/2}\{\widehat{F}_\varepsilon(y) - F_\varepsilon(y)\} = n^{1/2}M_n(\theta_0, \widehat{\mu}) \to_d N(0, V_1)$ where V_1 is defined in their paper. Condition (A2) holds with $V_2 = \theta_0(1-\theta_0)$, provided $0 < \theta_0 < 1$. Also (A3) holds, since the function m is uniformly bounded by 1.

It remains to estimate V_1 and V_2. Note that $\widehat{V}_2 = \widehat{\theta}(1-\widehat{\theta})$ consistently estimates V_2. However, V_1 is harder to estimate. A plug-in type estimator can be obtained by making use of the estimator of the error density in Van Keilegom and Veraverbeke (2002). Since this approach requires the selection of a new bandwidth, we prefer to use the bootstrap approach. We now check the conditions of Theorem 2.2. For (A4), set $\delta > 0$ and define

$$C^{1+\delta}(0,1) = \{\text{differentiable } f\colon (0,1) \to \mathcal{R}, \text{such that } \|f\|_{1+\delta} \le 1\},$$

where

$$\|f\|_{1+\delta} = \max\{\|f\|_\infty, \|f'\|_\infty\} + \sup_{x,y} \frac{|f'(x) - f'(y)|}{|x-y|^\delta},$$

and $\|\cdot\|_\infty$ denotes the supremum norm. It follows along the lines of the proof of Lemma 1 in Akritas and Van Keilegom (2001), using for example $\delta = 1/5$, that the class

$$\begin{aligned}\{I(\varepsilon \le y + f(X)) - \theta\colon f \in C^{1+\delta}(0,1), \theta \in [0,1]\}& \\ = \{I\{Y - h(X) \le y\} - \theta\colon h \in \bar{\mathcal{H}}, \theta \in [0,1]\}&\end{aligned}$$

is Donsker, and hence Glivenko–Cantelli, where $\bar{\mathcal{H}} = \mathcal{H} = \mu + C^{1+\delta}(0,1)$, and $\bar{\mathcal{H}}$ is endowed with the supremum norm. As a consequence, the class $\mathcal{F}$ in (A4) is also Glivenko–Cantelli. Moreover, $P\{\widehat{\mu} \in \bar{\mathcal{H}}\} \to 1$ by Propositions 3–5 in Akritas and Van Keilegom (2001). Condition (A5) is satisfied since for any $\delta_n \downarrow 0$,

$$\begin{aligned}\sup_{|\theta-\theta_0|\le\delta_n, h\in\bar{\mathcal{H}}} &|\mathrm{E}m^2(X,Y,\theta,h) - \mathrm{E}m^2(X,Y,\theta_0,h)| \\ &\le \delta_n \sup_{|\theta-\theta_0|\le\delta_n, h\in\bar{\mathcal{H}}} \mathrm{E}|2I\{Y - h(X) \le y\} - \theta - \theta_0| \to 0.\end{aligned}$$

Next, let us calculate $\Gamma(\theta, h)[\bar{h} - h]$ for any $h, \bar{h} \in \mathcal{H}$:

$$\begin{aligned}&\Gamma(\theta,h)[\bar{h}-h] \\ &= \lim_{\tau\to 0}\{M(\theta, h + \tau(\bar{h}-h)) - M(\theta,h)\}/\tau \\ &= \lim_{\tau\to 0} \tau^{-1}\int [F_{Y\,|\,x}(y + h(x) + \tau(\bar{h}(x) - h(x))) - F_{Y\,|\,x}(y + h(x))]\,\mathrm{d}F_X(x) \\ &= \int f_{Y\,|\,x}(y + h(x))(\bar{h}(x) - h(x))\,\mathrm{d}F_X(x),\end{aligned}$$

where $F_{Y\,|\,x}$ and $f_{Y\,|\,x}$ are the distribution and density function of Y given $X = x$, and F_X is the distribution function of X. Consequently,

$$\begin{aligned}\Phi_n &= n^{1/2}\Big[n^{-1}\sum_{i=1}^{n} I\{Y_i - \mu(X_i) \le y\} - \theta_0 \\ &\quad + n^{-1}\int f_{Y\,|\,x}(y+\mu(x))\sum_{i=1}^{n}(k_{b,x}(X_i)Y_i - \mathrm{E}\{k_{b,x}(X)Y\})\,\mathrm{d}x\Big] + o_{\mathrm{pr}}(1) \\ &= n^{1/2}\Big[n^{-1}\sum_{i=1}^{n} I\{Y_i - \mu(X_i) \le y\} - \theta_0\Big] \\ &\quad + n^{1/2}\Big[n^{-1}\sum_{i=1}^{n} f_{Y\,|\,X_i}(y+\mu(X_i))Y_i - \mathrm{E}[f_{Y\,|\,X}(y+\mu(X))Y]\Big] + o_{\mathrm{pr}}(1).\end{aligned} \tag{3.1}$$

In a similar way, we obtain

$$\begin{aligned}\Phi_n^* &= n^{1/2}\Big[n^{-1}\sum_{i=1}^{n} I\{Y_i^* - \widehat{\mu}(X_i^*) \le y\} - n^{-1}\sum_{i=1}^{n} I\{Y_i - \widehat{\mu}(X_i) \le y\}\Big] \\ &\quad + n^{1/2}\Big[n^{-1}\sum_{i=1}^{n} f_{Y\,|\,X_i^*}(y+\widehat{\mu}(X_i^*))Y_i^* - \mathrm{E}^*[f_{Y\,|\,X^*}(y+\widehat{\mu}(X^*))Y^*]\Big] + o_{P^*}(1).\end{aligned} \tag{3.2}$$

Both (3.1) and (3.2) converge to zero-mean normal random variables (use e.g. the Lindeberg condition to show the convergence of (3.2)). We next show that the asymptotic variance of (3.2) converges in probability to the asymptotic variance of (3.1). To show this we restrict attention to the first term of (3.1) and (3.2) (the convergence of the variance of the second term and of the covariance between the two terms can be established in a similar way). Note that the variance of the first term of (3.1) respectively (3.2) equals $\theta_0(1-\theta_0)$ respectively $n^{-1}\sum_{i=1}^n I\{Y_i - \widehat{\mu}(X_i) \le y\}[1 - n^{-1}\sum_{i=1}^n I\{Y_i - \widehat{\mu}(X_i) \le y\}]$. Since it follows from Lemma 1 in Akritas and Van Keilegom (2001) that

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n} I\{Y_i - \widehat{\mu}(X_i) \le y\} \\ &= \theta_0 + \sum_{i=1}^{n}[I\{Y_i - \mu(X_i) \le y\} - \theta_0] + P\{Y - \widehat{\mu}(X) \le y \,|\, \widehat{\mu}\} - \theta_0 + o_{\mathrm{pr}}(n^{-1/2}) \\ &= \theta_0 + o_{\mathrm{pr}}(1),\end{aligned}$$

the result follows. Hence, (B1) is satisfied. For (B2) it suffices by Remark 2.6 to show that the class $\{I\{Y - h(X) \le y\} - \theta : 0 \le \theta \le 1, h \in \bar{\mathcal{H}}\}$ is Donsker, which we have already established before, and that

$$\mathrm{Var}\big[I\{Y - h(X) \le y\} - I\{Y - \mu(X) \le y\} - \theta + \theta_0\big] \le K_1|\theta - \theta_0| + K_2\|h - \mu\|_\infty$$

for some $K_1, K_2 \geq 0$. A similar derivation can be given for the bootstrap analogue of (B2). Next write

$$
\begin{aligned}
&|M(\theta_0, \widehat{\mu}) - \Gamma(\theta_0, \mu)[\widehat{\mu} - \mu]| \\
&= \Big| P\{Y - \widehat{\mu}(X) \leq y\} - \theta_0 - \int f_{Y\,|\,x}(y + \mu(x))\{\widehat{\mu}(x) - \mu(x)\}\,\mathrm{d}F_X(x) \Big| \\
&= |\int [F_{Y\,|\,x}(y + \widehat{\mu}(x)) - F_{Y\,|\,x}(y + \mu(x)) - f_{Y\,|\,x}(y + \mu(x))(\widehat{\mu}(x) - \mu(x))]\,\mathrm{d}F_X(x)| \\
&= \tfrac{1}{2}|\int f'_{Y\,|\,x}(y + \xi(x))(\widehat{\mu}(x) - \mu(x))^2\,\mathrm{d}F_X(x)| \leq K \sup_x |\widehat{\mu}(x) - \mu(x)|^2,
\end{aligned}
$$

for some $\xi(x)$ between $\mu(x)$ and $\widehat{\mu}(x)$, and for some positive K. This shows that (B3) holds. In a similar way, the bootstrap version of (B3) can be shown to hold. Finally, condition (B4) follows from e.g. Härdle, Janssen and Serfling (1988), and its bootstrap version can be established in a very similar way. It now follows that a $100(1-\alpha)\%$ confidence interval for $F_\varepsilon(y)$ is given by $\{\theta\colon -2\log \mathrm{EL}_n(\theta, \widehat{\mu}) \geq e^*_{1-\alpha}\}$, where $e^*_{1-\alpha}$ is the $100(1-\alpha)\%$ percentile of the distribution of

$$
n\Big[n^{-1}\sum_{i=1}^n I\{Y_i^* - \widehat{\mu}^*(X_i^*) \leq y\} - \widehat{\theta}\Big]^2 \Big/ \{\widehat{\theta}(1-\widehat{\theta})\}.
$$

4. Plug-in empirical likelihood for slower than root-n-convergence

In this section we develop a version of our main result (Theorem 2.1) when the rate of convergence of the sample mean of the estimating function is slower than the standard $\sqrt{n}$-rate, i.e. of the form n^α for some α in $(0, \frac{1}{2})$. Such rates are common in smoothing problems, inverse problems, and in settings where the observations have long-range dependence (as opposed to the weak dependence discussed in Remark 2.4). We examine in detail an application to density estimation, and survival function estimation for current status data, the latter being a classic example of cube-root asymptotics ($\alpha = 1/3$). In these applications we need the estimating function m to change with n. A related result (without plug-in or explicit rates) is the triangular array EL theorem of Owen (2001, page 85).

4.1. Main result. We need the following asymptotic stability conditions:

(C1) $n^\alpha M_n(\theta_0, \widehat{h}) \to_d U \sim \mathrm{N}_p(0, V_1)$ for some positive definite matrix V_1.

(C2) $n^{2\alpha-1} S_n(\theta_0, \widehat{h}) \to_{\mathrm{pr}} V_2$ for some positive definite matrix V_2.

(C3) $\max_{1\leq i\leq n} \|m_n(Z_i, \theta_0, \widehat{h})\| = O_{\mathrm{pr}}(n^\alpha)$.

THEOREM 4.1. *Assume conditions (C1)–(C3) hold for an appropriate α in $(0, \frac{1}{2})$. Then the statement of Theorem 2.1 is in force.*

REMARK 4.1. Conditions (C1), (C2) are natural extensions of (A1), (A2). However, (C3) only requires $O_{\mathrm{pr}}(n^\alpha)$ rather than $o_{\mathrm{pr}}(n^\alpha)$, so it is not an extension of (A3). This is an important departure from the $\sqrt{n}$-rate result, and turns out to be crucial for the application to current status data.

REMARK 4.2. As with (A1), the assumption of a normal limit in (C1) is not crucial. In the application to current status data, however, we do have a normal limit and $V_1 = V_2$, so it is unnecessary to estimate either V_1 or V_2. In general it will be necessary to estimate both matrices for the result to be useful. Along the lines of Remark 2.1, it can be shown that $\widehat{V}_2 = n^{2\alpha-1} S_n(\widehat{\theta}, \widehat{h})$ consistently estimates V_2 if $\widehat{\theta}$ consistently estimates θ_0 and the following two additional conditions hold:

(C4) For some subset $\bar{\mathcal{H}}$ of $\mathcal{H}$ such that $P\{\widehat{h} \in \bar{\mathcal{H}}\} \to 1$,

$$\sup_{\|\theta-\theta_0\|<\delta, h\in\bar{\mathcal{H}}} \Big| n^{2(\alpha-1)} \sum_{i=1}^{n} \{m_n^{\otimes 2}(Z_i, \theta, h) - \mathrm{E} m_n^{\otimes 2}(Z, \theta, h)\} \Big| \to_{\mathrm{pr}} 0.$$

(C5) For any real sequence $\delta_n \downarrow 0$,

$$n^{2\alpha-1} \sup_{\|\theta-\theta_0\|\le\delta_n, h\in\bar{\mathcal{H}}} |\mathrm{E} m_n^{\otimes 2}(Z, \theta, h) - \mathrm{E} m_n^{\otimes 2}(Z, \theta_0, h)| \to 0.$$

4.2. Density estimation. Let $X_1, \ldots, X_n$ be i.i.d. from an unknown density f_0, and suppose we are interested in estimating $\theta_0 = f_0(t)$, for t fixed. We do this using the kernel density estimator $\widehat{f}_n(t) = n^{-1} \sum_{i=1}^{n} k_b(X_i - t)$, where $k_b(u) = b^{-1} k(b^{-1}u)$ is a b-scaled version of a symmetric, bounded kernel function k, supported on $[-1, 1]$. We choose here to employ bandwidths $b = b_n = b_0 n^{-\beta}$, where $0 < \beta < 1/5$. The $\beta = 1/5$ rate is optimal for estimating $f_0(t)$, in the sense of minimising the asymptotic mean squared error, but as we here aim at constructing confidence intervals, an undersmoothing rate of $\beta < 1/5$ is preferable. Hall and Owen (1993) constructed EL confidence bands for f_0, and Chen (1996) showed that the pointwise EL confidence intervals (with and without Bartlett correction) are more accurate than those based on the bootstrap.

Following these authors, we use the sequence of estimating functions $m_n(x, \theta) = k_b(x - t) - \theta$, which does not involve plug-in, and show that our main result yields the analogue of Wilks's theorem. We now check the conditions of Theorem 4.1 using $\alpha = \frac{1}{2}(1 - \beta)$. Condition (C1) can be checked under mild conditions on the density, as it follows from standard asymptotic theory for kernel density estimators that

$$n^{\alpha} M_n(\theta_0) = b_0^{-1/2} (nb)^{1/2} \{\widehat{f}_n(t) - f_0(t)\} \to_d \mathrm{N}(0, V_1),$$

where

$$V_1 = b_0^{-1} f_0(t) R(k) \quad \text{and} \quad R(k) = \int k(u)^2 \, \mathrm{d}u. \tag{4.1}$$

For (C2),

$$\begin{aligned} n^{2\alpha-1} S_n(\theta_0) &= b_0^{-1} \frac{b}{n} \sum_{i=1}^{n} \{k_b(X_i - t) - \theta_0\}^2 \\ &= b_0^{-1} \frac{1}{nb} \sum_{i=1}^{n} k((X_i - t)/b)^2 + O_{\mathrm{pr}}(b) \to_{\mathrm{pr}} b_0^{-1} f_0(t) R(k) = V_1. \end{aligned}$$

For (C3),

$$\max_{i\le n} |m_n(X_i,\theta_0)| = O(b^{-1}) = O(n^{\beta}) = O(n^{\alpha})$$

because k is bounded, and $\beta \le \alpha = \frac{1}{2}(1-\beta)$, which only requires $\beta \le 1/3$, whereas we assume $\beta < 1/5$.

4.3. Survival function estimation for current status data. Suppose there is a failure time of interest $T \sim F$, with survival function $S = 1-F$ and density f, but we only get to observe $Z = (C, \Delta)$, where $\Delta = 1\{T \le C\}$ and $C \sim G$ is an independent check-up time (with density g). The observations are assumed to be i.i.d.

The nonparametric maximum likelihood estimator $S_n(t)$ of $S(t)$ exists. Groeneboom (1987) showed that $n^{1/3}\{S_n(t)-S(t)\}$ converges to a non-degenerate limit law. The limit is not distribution-free, however, and is unsuitable for providing a confidence region for $S(t)$. Banerjee and Wellner (2002) found a universal limit law for the likelihood ratio statistic, leading to tractable confidence intervals. Our approach based on estimating equations offers a simpler type of EL confidence region, and extends to the setting in which T and C are conditionally independent given a covariate (although for simplicity we restrict to the case of no covariates).

First consider estimation of a smooth functional of S (such as its mean):

$$\theta_0 = \int_0^{\infty} k(u)S(u)\,\mathrm{d}u,$$

where $k\colon [0,\infty) \to \mathcal{R}$ is fixed. This parameter can be estimated at a $\sqrt{n}$-rate, there is an efficient influence curve

$$m(Z,\theta,F,g,k) = \frac{k(C)(1-\Delta)}{g(C)} - \theta - \frac{k(C)(1-F(C))}{g(C)} + \int_0^{\infty} k(u)(1-F(u))\,\mathrm{d}u,$$

and, given any preliminary estimators $\widehat{F}$ and $\widehat{g}$ of F and g, respectively, $m(Z,\theta,\widehat{F},\widehat{g},k)$ is a plug-in estimating function, which yields a consistent estimator of θ_0 when either $\widehat{F}$ or $\widehat{g}$ is consistent; see van der Laan and Robins (1998).

Now consider estimation of $\theta_0 = S(t)$. Van der Laan and van der Vaart (2000) introduced a kernel-type estimator $S_{n,b}(t)$ and showed that $n^{1/3}\{S_{n,b}(t) - S(t)\} \to_d N(0, V_1)$, for appropriate and positive V_1. Their approach is to replace k above by $k_n = k_{b,t}$, a kernel function of bandwidth $b = b_n = b_0 n^{-1/3}$ centred at t. Here $k_{b,t}(u) = k((u-t)/b)/b$ in terms of a bounded density k supported on $[-1,1]$. This yields a sequence of (plug-in) estimating functions $m_n(Z,\theta,\widehat{F},\widehat{g}) = m(Z,\theta,\widehat{F},\widehat{g},k_n)$, and the estimator is written as $S_{n,b}(t) = \mathbf{P}_n \mathrm{IC}(\widehat{F},\widehat{g},k_n)$, where $\mathbf{P}_n$ is the empirical measure of the observations, and $\mathrm{IC}(F,g,k_n)(Z) = m(Z,0,F,g,k_n)$ is the influence curve. The asymptotic variance of $S_{n,b}(t)$ is $V_1 = b_0^{-1}\sigma^2 R(k)$, where $R(k)$ is as in (4.1) and σ^2 depends on F and g, as well as on the limits g_1 and F_1 of respectively $\widehat{g}$ and $\widehat{F}$.

We adopt the same assumptions as van der Laan and van der Vaart. In particular, assume that F is differentiable at t, and g is twice continuously differentiable and bounded away from zero in a neighborhood of t. Also, $\widehat{g}$ and $\widehat{F}$ are assumed to belong to classes of functions having uniform entropy of order $(1/\epsilon)^V$, $V < 2$, with probability tending to 1, and $\widehat{g}$, or $\widehat{F}$, or both, are locally consistent at t.

Our result for estimating functions (with plug-in) under cube-root asymptotics (Theorem 4.1 with $\alpha = 1/3$) gives

$$-2 \log \mathrm{EL}_n(S(t), \widehat{F}, \widehat{g}, k_n) \to_d \chi^2_1.$$

Conditions (C1)–(C3) are easily checked by referring to van der Laan and van der Vaart's Theorem 2.1 and its proof. First note that $M_n(\theta_0, \widehat{F}, \widehat{g}) = S_{n,b}(t) - S(t)$, so (C1) holds (with V_1 given by the asymptotic variance of $S_{n,b}(t)$). Then, with $\mathbf{P}$ denoting the true distribution,

$$\begin{aligned} bS_n(\theta_0, \widehat{F}, \widehat{g}) &= b\,\mathbf{P}_n\{\mathrm{IC}(\widehat{F}, \widehat{g}, k_n) - S(t)\}^2 \\ &= b\,\mathbf{P}_n\{\mathrm{IC}(\widehat{F}, \widehat{g}, k_n) - \mathbf{P}\,\mathrm{IC}(\widehat{F}, \widehat{g}, k_n)\}^2 \\ &\quad + 2b\{S_{n,b}(t) - S(t)\}\{\mathbf{P}\,\mathrm{IC}(\widehat{F}, \widehat{g}, k_n) - S(t)\} \\ &\quad - b\,\{\mathbf{P}\,\mathrm{IC}(\widehat{F}, \widehat{g}, k_n) - S(t)\}^2. \end{aligned} \tag{4.2}$$

Along the lines of van der Laan and van der Vaart (Theorem 2.1 and the start of its proof), the last two terms are $o_{\mathrm{pr}}(b^3)$, and to handle the first term the influence function IC is split into a sum of two terms IC_1 and IC_2, where

$$\mathrm{IC}_2(F, g, k_n)(Z) = \int_0^\infty k_n(u)\{1 - F(u)\}\,\mathrm{d}u$$

does not give any contribution in the limit. In our case, IC_2 acts as a constant function (there are no covariates), so the first term in (4.2) with IC replaced by IC_2 is $O(b)$. The first term of (4.2) with IC replaced by IC_1 can be expressed as

$$bb_0^{-3/2}(b^{3/2}\mathbf{G}_n H_n) + b\,\mathbf{P}H_n, \tag{4.3}$$

where $\mathbf{G}_n = \sqrt{n}(\mathbf{P}_n - \mathbf{P})$ is the empirical process and

$$H_n(\widehat{F}, \widehat{g}, k_n)(\cdot) = \{\mathrm{IC}_1(\widehat{F}, \widehat{g}, k_n) - \mathbf{P}\mathrm{IC}_1(\widehat{F}, \widehat{g}, k_n)\}^2.$$

Applying the part of their proof that deals with IC_1, but with IC_1 replaced by H_n and $b^{3/2}k_n^2$ as the envelope functions, it can be shown that $b^{3/2}\,\mathbf{G}_n H_n$ is asymptotically tight. They also show that $b\,\mathbf{P}H_n \to_{\mathrm{pr}} \sigma^2 R(k)$, with $R(k)$ as in Section 4.2. Thus, only the second term in (4.3) gives a contribution in the limit, and we have

$$n^{-1/3} S_n(\theta_0, \widehat{F}, \widehat{g}) \to_{\mathrm{pr}} b_0^{-1}\sigma^2 R(k) = V_1,$$

establishing (C2) with $V_2 = V_1$. Finally, (C3) can be checked using their assumption that $\widehat{g}$ is asymptotically bounded away from zero in a fixed neighborhood of t. Note that $k_n \le cb_n^{-1}1_{[t-b_n,t+b_n]}$ for some constant c, so

$$\max_{1\le i\le n} |m_n(Z_i,\theta_0,\widehat{F},\widehat{g})| = O_{\rm pr}(b_n^{-1}) = O_{\rm pr}(n^{1/3}).$$

5. Empirical likelihood asymptotics with growing dimensions

The traditional empirical likelihood theory works for a fixed number of estimating functions p, or, when estimating a mean, for data having a fixed dimension d. The present section is concerned with the question of how this theory may be extended towards allowing p to increase with growing sample size. For simplicity, we assume there are no nuisance parameters. We consider the case of i.i.d. observations $Z_1,\ldots,Z_n$ from a d-dimensional distribution $F = F_n$ with mean vector $\mu_{0,n}$ and non-singular variance matrix Σ_n, so the estimating function is simply $m(Z,\mu_{0,n}) = Z - \mu_{0,n}$. Inside this triangular framework we allow $d = p = p_n$ to grow with n, and study the problem of establishing sufficient conditions under which the standard χ^2_p calibration can still be used. We shall use several steps to approximate the EL statistic

$$\mathrm{EL}_n(\mu) = \max\Big\{\prod_{i=1}^n (nw_i)\colon \text{each } w_i>0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i Z_i = \mu\Big\}, \tag{5.1}$$

and approximation results will be reached under different sets of conditions.

5.1. Approximation to a quadratic form. At the heart of the standard large-sample EL theorem lies the fact that

$$\Lambda_n = -2\log \mathrm{EL}_n(\mu_{0,n}) \quad \text{is close to} \quad Q_n = n(\bar{Z}-\mu_{0,n})^{\rm t} S_n^{-1}(\bar{Z}-\mu_{0,n}), \tag{5.2}$$

writing

$$\bar{Z} = n^{-1}\sum_{i=1}^n Z_i \quad \text{and} \quad S_n = n^{-1}\sum_{i=1}^n (Z_i-\mu_{0,n})(Z_i-\mu_{0,n})^{\rm t}.$$

In fact, under standard conditions, with fixed p, $\Lambda_n = Q_n + o_{\rm pr}(1)$, and Q_n of course tends to a χ^2_p. We aim now at investigating to what extent Λ_n remains close to Q_n, even when p grows with n.

We may take $\mu_{0,n} = 0$ for simplicity of presentation. Following Owen (2001, Ch. 3), the maximising weights are characterised as $w_i = n^{-1}(1+\lambda^{\rm t}Z_i)^{-1}$ for $i = 1,\ldots,n$, where the vector $\lambda = (\lambda_1,\ldots,\lambda_p)^{\rm t}$ of Lagrange multipliers satisfies the p equations

$$g(\lambda) = n^{-1}\sum_{i=1}^n \frac{Z_i}{1+\lambda^{\rm t}Z_i} = 0. \tag{5.3}$$

Studying $\Lambda_n = 2\sum_{i=1}^n \log(1+\lambda^{\rm t} Z_i)$ in particular necessitates having control over the size of $\lambda^{\rm t} Z_i$, after which Taylor series arguments may be employed. Below we outline the main line of reasoning and results, with details again left for the Appendix.

Among the technical obstacles met when attempting to secure uniform smallness of the $\lambda^{\rm t} Z_i$s is the need to control the behaviour of the largest and smallest eigenvalues of S_n. This involves the behaviour of eigenvalues for Σ_n, as well as the closeness of S_n to Σ_n. For the following result, which is proved in our Appendix, let $S_{n,j,k}$ and $\sigma_{n,j,k}$ denote the (j,k) elements of S_n and Σ_n, and define

$$L_n = \max_{j,k\le p} |S_{n,j,k} - \sigma_{n,j,k}|.$$

LEMMA 5.1. *Assume that the $Z_{i,j}$s have finite qth order moments, for some $q \ge 4$, and let $A_n(p,q) = p^{-1}\sum_{j=1}^p {\rm E}|Z_{i,j} - \mu_{0,n,j}|^q$. Then, for a positive constant $c(q)$,*

$$P\{L_n \ge \varepsilon\} \le \frac{c(q)p^2}{\varepsilon^q n^{q/2}} A_n(p,q)^2 \quad \textit{for each positive } \varepsilon.$$

Let now $\gamma_{1,n}$ and $\gamma_{p,n}$ be the largest and smallest eigenvalues of Σ_n, and similarly $\widehat\gamma_{1,n}$ and $\widehat\gamma_{p,n}$ the largest and smallest eigenvalues of S_n. Writing $S_n = \Sigma_n + D_n$, we then have, for all unit vectors u,

$$u^{\rm t} S_n u \le u^{\rm t}\Sigma_n u + \sum_{j,k} |u_j|\,|u_k|\,|D_{n,j,k}| \le \gamma_{1,n} + pL_n,$$

and similarly $u^{\rm t} S_n u \ge \gamma_{p,n} - pL_n$, in that $\sum_{j=1}^p |u_j| \le p^{1/2}$. This shows that

$$\widehat\gamma_{1,n} \le \gamma_{1,n} + pL_n \quad \text{and} \quad \widehat\gamma_{p,n} \ge \gamma_{p,n} - pL_n.$$

In particular, the largest and smallest eigenvalues of S_n are only $o_{\rm pr}(1)$ away from those of Σ_n if $pL_n \to_{\rm pr} 0$. A sufficient condition for this is that $A_n(p,q)$ stays bounded as n grows and that $p^{2+q}/n^{q/2} \to 0$, by Lemma 5.1; for example, stable fourth order moments and $p^3/n \to 0$ secures such a closeness of eigenvalues. If higher order moments exist then the condition $p^2/n \to 0$ almost suffices. With yet further conditions being imposed, like existence of moment-generating functions or independence among components of the Z_is, even less may be required of the size of p to secure $\widehat\gamma_{1,n} - \gamma_{1,n} \to_{\rm pr} 0$ and $\widehat\gamma_{p,n} - \gamma_{p,n} \to_{\rm pr} 0$.

Our first set of conditions establish uniform smallness in probability of the variables $Y_i = \lambda^{\rm t} Z_i$. They demand that, as n grows,

(D1) for some $q \ge 4$, the sequence of ${\rm E}\|Z_i/p^{1/2}\|^q$ stays bounded;
(D2) for the q of (D1), $p^{2+4/(q-2)}/n \to 0$;
(D3) for the q of (D1), $A_n(p,q)$ of Lemma 5.1 stays bounded;
(D4) the largest and smallest eigenvalues of Σ_n stay away from infinity and zero.

LEMMA 5.2. *Under (D1) and (D2),* $B_n = \max_{i \le n} \|Z_i\| = o_{\rm pr}((n/p)^{1/2})$, *in fact,*

$$P\{B_n \ge (n/p)^{1/2}\varepsilon\} \le \frac{\mathrm{E}\|Z_i/p^{1/2}\|^q}{\varepsilon^q}\frac{p^q}{n^{q/2-1}} \quad \textit{for all positive } \varepsilon.$$

If in addition (D3) and (D4) hold, then the random λ *of (5.3) is of size* $\|\lambda\| = O_{\rm pr}((p/n)^{1/2})$.

We are now in a position to state precise versions of (5.2). We give three such. The first version assumes uniform boundedness of all $Z_{i,j}$ components, leading to quite transparent conditions for closeness of Λ_n to Q_n. The second version does not assume boundedness but spells out what is required under reasonable moments conditions. Finally the third version works with linearly transformed versions of the Z_is.

THEOREM 5.1. *Assume that all* $Z_{i,j}$ *components remain uniformly bounded and that condition (D4) holds. Then, if* $p^3/n \to 0$, $\Lambda_n = Q_n + o_{\rm pr}(p^{1/2})$. *If the more strict requirement* $p^4/n \to 0$ *holds, then* $\Lambda_n = Q_n + o_{\rm pr}(1)$.

THEOREM 5.2. *Assume conditions (D1), (D3), (D4) hold, with* $p^{3+6/(q-2)}/n \to 0$ *replacing (D2). Then* $\Lambda_n = Q_n + o_{\rm pr}(p^{1/2})$. *Under the tougher condition* $p^{4+8/(q-2)}/n \to 0$, $\Lambda_n = Q_n + o_{\rm pr}(1)$.

In situations where the Z_is have moments of all orders, the growth conditions here come close to those of the previous theorem. If data are normal, for example, then $\|Z_i\|$ is bounded by a variable of the type $c(\chi^2_p)^{1/2}$ for a suitable c, and one may show that $p^3/n \to 0$ and $p^4/n \to 0$ again suffice for the $o_{\rm pr}(p^{1/2})$ and $o_{\rm pr}(1)$ statements.

Verifying condition (D4), or the corresponding statement for the eigenvalues of S_n, which is called for in the proof of the theorem, is sometimes technically hard. A theorem of Bai and Yin (1993) works for the case of Z_i having independent components, in which case the minimal growth condition $p/n \to 0$ ensures that the full spectral distribution tends to 1, i.e. the largest and smallest eigenvalues of S_n tend to 1. See also Bai (1999) and the ensuing discussion.

While Theorems 5.1–5.2 are satisfactory for several classes of problems, there are other situations of interest where condition (D4) does not hold. For this reason we provide a parallel theorem that demands less regarding the distribution of eigenvalues. Consider $Z_i^* = \Sigma_n^{-1/2}(Z_i - \mu_{0,n})$, which have mean zero and variance matrix I_p, and let S_n^* be the empirical variance matrix of these, i.e. $S_n^* = n^{-1}\sum_{i=1}^n Z_i^*(Z_i^*)^{\rm t} = \Sigma_n^{-1/2} S_n \Sigma_n^{-1/2}$. The eigenvalues of S_n^* turn out to be sufficiently well-behaved, as we demonstrate in the Appendix while proving the following theorem.

THEOREM 5.3. *The conclusions of Theorem 5.2 continue to hold, without condition (D4), as long as conditions (D1) and (D3) hold for the transformed variables* $Z_i^* = \Sigma_n^{-1/2}(Z_i - \mu_{0,n})$.

There is similarly a (D4)-free version of Theorem 5.1, for cases when the components $Z_{i,j}^*$ are uniformly bounded as n grows.

The central point to note for Theorem 5.3 is that the empirical likelihood (5.1) is invariant with respect to the transformation that maps data Z_i to G_nZ_i, where G_n is any non-singular non-random $p \times p$ matrix. If $\mathrm{EL}_n(G_n\mu \,|\, G_n)$ is the empirical likelihood computed on the basis of $Z_i' = G_nZ_i$, for the parameter $\widetilde{\mu} = G_n\mu$, then G_n cancels out of the defining equation $\sum_{i=1}^n w_i(G_nZ_i - G_n\mu) = 0$, showing that $\mathrm{EL}_n(\widetilde{\mu} \,|\, G_n)$ is the same as $\mathrm{EL}_n(\mu)$ in (5.1), i.e. independent of G_n (and with the same maximising w_is). The same is true for the quadratic approximation of (5.2). We may in particular employ $G_n = \Sigma_n^{-1/2}$, where the resulting G_nZ_i have covariance matrix I_p. This makes it possible to prove Theorem 5.3 with arguments similar to those needed for Theorem 5.2, but under the additional simplifying assumptions that $\Sigma_n = I_p$; see the Appendix.

Yet another version of our result emerges by dividing the Z_is by $\gamma_{p,n}^{1/2}$, to avoid small eigenvalues. This gives a parallel result to that of Theorem 5.1, where the essential condition is that the ratio $\gamma_{1,n}/\gamma_{p,n}$ remains bounded. See in this connection also Owen (2001, page 86), where stability of this ratio is crucial also for some problems associated with fixed p.

5.2. Approximation to a chi-squared. In applications of EL one typically employs a χ_p^2 approximation to Λ_n of (5.2), with or without a Bartlett correction; see again Owen (2001). We show here that closeness of the distribution of Λ_n to that of a χ_p^2 is still achieved with growing p. A somewhat mild version of this statement is that

$$(\Lambda_n - p)/(2p)^{1/2} \to_d \mathrm{N}(0,1) \tag{5.4}$$

as p grows with n at appropriate rate. For statistical applications, including testing and confidence statements, the above would typically be sufficient, securing that upper quantiles for the EL distribution are reasonably close to that of the χ_p^2. We see below that (5.4) often holds, but that uniform closeness of the Λ_n and χ_p^2 distributions demand rather more.

Regarding the approximation of the distribution of Q_n to the χ_p^2, we work in two steps. The first is to study the simpler case where S_n is replaced by the real Σ_n, i.e. involving $Q_n^0 = n(\bar{Z} - \mu_{0,n})^{\mathrm{t}}\Sigma_n^{-1}(\bar{Z} - \mu_{0,n})$, while the second is to control the consequences of this simplification.

PROPOSITION 5.1. *If* $\mathrm{E}\|Z_i/p^{1/2}\|^3$ *stays bounded and* $p^5/n \to 0$, *then*

$$a_{n,p} = \max_t |P\{Q_n^0 \le t\} - P\{\chi_p^2 \le t\}|$$

goes to zero when n *grows. Secondly, provided that* $\mathrm{E}|Z_{i,j} - \mu_{0,n,j}|^6 \le E_n$ *for* $j \le p$, *where* E_n *stays bounded, the rather weaker assumption* $p/n \to 0$ *secures approximate* χ_p^2*-ness in the sense that* $(Q_n^0 - p)/(2p)^{1/2} \to_d \mathrm{N}(0,1)$.

REMARK 5.1. That the uniform bounds used above are not the best possible, though phenomenally difficult to establish and requiring sophisticated mathematical techniques,

becomes clearer in Bentkus and Götze (1996, 1997) as they are able to show that $a_{n,p} \leq k(p)/n$ for all $p \geq 6$, exploiting particularities related to the quadratic form. It is not clear how big the $k(p)$ constants are, however.

To finish the story of the closeness of Λ_n to the χ^2_p, we need to ascertain that Q_n is close enough to Q_n^0. This is a trivial matter when p is fixed or bounded, but more complicated when p grows. The condition given in the following result, which is proved in our Appendix, stems from bounds originating in Lemma 5.1. It may be softened if the $Z_{i,j}Z_{i,k}$ variables have moment-generating functions, for example.

PROPOSITION 5.2. *Suppose the conditions of Lemma 5.1 are in force, and add to these that* $p^{4+4/q}/n \to 0$. *Then* $Q_n - Q_n^0 \to_{\rm pr} 0$. *Also, if* $p^{3+4/q}/n \to 0$, *then* Q_n *is approximately a* χ^2_p *in the sense that* $(Q_n - p)/(2p)^{1/2} \to_d {\rm N}(0,1)$.

Used in conjunction with theorems of the previous subsection we see that also (5.4) holds.

REMARK 5.2. The χ^2_p approximation for Q_n and hence for $\Lambda_n = -2\log {\rm EL}_n(\mu_{0,n})$ may be improved, with suitable corrections depending on p and n. It is worthwhile investigating the distribution of Q_n for the case of normal data. Some matrix algebra shows that

$$Q_n = \frac{Q_n'}{1 + Q_n'/n}, \quad \text{where} \quad Q_n' = n(\bar Z_n - \mu_{0,n})^{\rm t}(S_n')^{-1}(\bar Z_n - \mu_{0,n}),$$

with $S_n' = n^{-1}\sum_{i=1}^n (Z_i - \bar Z_n)(Z_i - \bar Z_n)^{\rm t}$ being the usual sample covariance matrix. From classic multivariate normal theory, see e.g. Mardia, Kent and Bibby (1979, Sections 3.4–3.5),

$$Q_n' \sim \frac{n}{n-1} T^2(p, n-1) = \frac{np}{n-p} F(p, n-p),$$

in terms of Hotelling T^2 and Fisher distributions, with the appropriate parameters, if data are normal. In particular, the mean of Q_n' is $np/(n-p-2)$.

6. Applications with growing p

This section provides some examples where there is a growing number of parameters, and where the theory developed in Section 5 guarantees that the empirical likelihood methodology still is applicable.

6.1. Many independent means. Suppose that $Z_1, \ldots, Z_n$ correspond to p independent samples $Z_{1,j}, \ldots, Z_{n,j}$, with mean $\mu_{0,j}$ and standard deviation σ_j, for $j = 1, \ldots, p$. EL may then be used to make simultaneous inference for the vector of mean parameters μ_0. Consider the normalised variable U_p with components $(Z_{i,j} - \mu_{0,j})/\sigma_j$. It has mean zero and variance matrix I_p. Results of Section 5 imply that the EL works properly, even when p grows, provided $p^3/n \to 0$ and that the U_p components stay uniformly bounded, for example. This is secured by the theorem of Bai and Yin (1993) about all eigenvalues of $\widetilde S_n^*$ coming close to 1, cf. Theorem 5.3.

Similar results may be reached in other models with a growing number of mean type parameters, possibly also in the presence of nuisance parameters. An example is analysis of variance with a large number of groups, cf. Akritas and Arnold (2000). Our theory also supports the use of EL theory when multiple comparisons between groups are made, since the covariance matrix of a collection of such differences of means is well-behaved enough to have its eigenvalues away from zero and infinity, i.e. (D4) of Theorem 5.1 will hold.

6.2. Histograms and cell probabilities. Let $X_1, \ldots, X_n$ be i.i.d. real random variables with density f. For p disjoint cells $C_1, \ldots, C_p$ let Z_i be the vector with components $Z_{i,j} = I\{X_i \in C_j\}$, for $j = 1, \ldots, p$. These have mean vector $\pi = (\pi_1, \ldots, \pi_p)^{\rm t}$, involving cell probabilities $\pi_j = \int_{C_j} f \, {\rm d}x$, and covariance matrix Σ_n, with elements $\pi_j(\delta_{j,k} - \pi_k)$. The $\bar{Z}$ vector has components $\widehat{\pi}_j$ equal to the relative frequencies in the p cells. For fixed histogram cells,

$$-2 \log {\rm EL}_n(\pi) = n(\widehat{\pi} - \pi)^{\rm t} S_n^{-1} (\widehat{\pi} - \pi) + o_{\rm pr}(1), \tag{6.1}$$

by standard EL theory. By our efforts in Section 5.1, (6.1) continues to hold also when p is allowed to grow at appropriate slow rate with n.

Determining just how slow this rate must be, in order for the remainder term in (6.1) to still go to zero in probability, turns out to be a delicate matter, and depends also on aspects of the vector of probabilities π. It turns out that the smallest eigenvalues of Σ_n go to zero with growing p, so condition (D4) of Theorems 5.1–5.2 is not met. The eigenvalues of S_n^* for the transformed variables $Z_i^* = \Sigma_n^{-1/2}(Z_i - \pi)$ are however well-behaved, if only (D3) of Theorem 5.3 holds for the transformed variables. The precise condition securing (6.1) becomes

$$A_n^*(p) = p^{-1} \sum_{j,k} \pi_k (r_{j,k} - \widetilde{r}_j)^4 \text{ bounded as } n \text{ grows},$$

where the $r_{j,k}$s are elements of $R_n = \Sigma_n^{-1/2}$ and $\widetilde{r}_j = \sum_k \pi_k r_{j,k}$. This may be seen to hold when the cell probabilities π_j do not vary too much from each other.

The theory of Section 5 applies somewhat more readily to the case of Z_i having components $I\{X_i \le t_j\}$ for specified cut points $t_1 < \cdots < t_p$. Then $\bar{Z}$ has as components the empirical distribution function $F_n(t_j)$ evaluated at these p points. Theorem 5.1 implies that $-2 \log {\rm EL}_n(F(t_1), \ldots, F(t_p))$ is close to the natural quadratic form $n(F_n(J) - F(J))^{\rm t} \Omega_{n,J}^{-1} (F_n(J) - F(J))$, where $F_n(J)$ is $F_n(t)$ evaluated at positions $J = \{t_1, \ldots, t_p\}$, and so on. This is since the eigenvalues of the variance matrix do not flee to zero for this problem.

6.3. Kernel density estimation. Let $\widehat{f}$ be a kernel density estimator based on $X_1, \ldots, X_n$, as in Section 4.2, where we take the kernel k to be bounded and symmetric with support $[-1, 1]$. We assume that the real density f is bounded and continuous. Consider $\widehat{f}(t_1), \ldots, \widehat{f}(t_p)$ at different positions. This fits in with the setup of Section 5,

with Z_i having components $Z_{i,j} = k_b(X_i - t_j)$ for $j = 1, \ldots, p$. Its mean is $\bar f(t)$ evaluated at $t_1, \ldots, t_p$, where $\bar f = k_b * f$, i.e. $\bar f(t) = \int k(u) f(t + bu)\,\mathrm{d}u = f(t) + \frac12 k_2 b^2 f''(t) + o(b^2)$, where $k_2 = \int u^2 k(u)\,\mathrm{d}u$. Also, the variance matrix of Z_i takes the form $b^{-1}\Omega_n$, where Ω_n has diagonal elements $R(k) f(t_j) - b f(t_j)^2 + O(b^2)$ and elements $-b f(t_j) f(t_l) + O(b^2)$ outside the diagonal, provided the points are at least b apart. Hence the eigenvalues of Ω_n are essentially $R(k) f(t_j)$, when b is small and the t_j points at least b apart.

We wish to conclude from this that the empirical likelihood $\mathrm{EL}_n(\bar f(t_1), \ldots, \bar f(t_p))$ has the property

$$-2 \log \mathrm{EL}_n(\bar f(t_1), \ldots, \bar f(t_p)) = nb \begin{pmatrix} \widehat f(t_1) - \bar f(t_1) \\ \vdots \\ \widehat f(t_p) - \bar f(t_p) \end{pmatrix}^{\mathrm{t}} \widehat\Omega_n^{-1} \begin{pmatrix} \widehat f(t_1) - \bar f(t_1) \\ \vdots \\ \widehat f(t_p) - \bar f(t_p) \end{pmatrix} + \delta_n, \quad (6.2)$$

where $\widehat\Omega_n$ is an empirical version of Ω_n and where δ_n is $o_{\mathrm{pr}}(1)$ or perhaps $o_{\mathrm{pr}}(p^{1/2})$, depending on growth conditions for p. Neither of Theorems 5.1–5.3 can be applied directly; for example, conditions (D1) and (D3) are not met for the Z_is. Variations of the arguments used to prove Theorem 5.3 will however suffice. The point is that the EL is invariant with respect to linear transformations of data, as explained at the end of Section 5.1. Here we work with Z_i' with components $b\{Z_{i,j} - \bar f(t_j)\}$. These are then uniformly bounded. Also, Z_i' has mean zero and variance matrix $b\Omega_n$. One may now go through arguments used to prove Theorem 5.3 and verify that (6.2) holds with $\delta_n = o_{\mathrm{pr}}(p^{1/2})$, when $p^3/n \to 0$ and $b \to 0$ at a speed where $b = O(1/p)$; this latter requirement comes from keeping the smallest eigenvalue of Ω_n away from zero. This also entails the usual $nb \to \infty$ requirement.

6.4. Growing polynomial regression. Consider the regression model

$$Y_i = \xi(X_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the pairs (X_i, ε_i) are i.i.d., with X_is coming from some density f and the ε_is having mean zero and standard deviation σ_0. The main objective is to make inference about $\xi(x)$. We do not strive for the fullest generality in this application of our theory, and are content to work with the following scenario: f is known (e.g. the uniform on the unit interval), and $\xi(x)$ may be expanded in terms of basis functions $\psi_0, \psi_1, \psi_2, \ldots$ that are orthonormal w.r.t. f, i.e. $\int f \psi_j \psi_k\,\mathrm{d}x = \delta_{j,k}$, and where we take $\psi_0 = 1$. We might for example take $\psi_j(x) = \phi_j(F(x))$ where the ϕ_js are orthogonal w.r.t. the uniform on the unit interval and F the c.d.f. of f. Hence $\xi(x) = \sum_{j=0}^\infty b_j \psi_j(x)$, where we assume that $\mathrm{E}\xi(X)^2 = \sum_{j=0}^\infty b_j^2$ is finite, and also that $\xi(x)$ is bounded.

In this setup, consider as pth order model

$$Y_i = \xi_p(X_i) + \varepsilon_i', \quad \text{with} \quad \xi_p(x) = \sum_{j=0}^p b_j \psi_j(x) = (\psi^{(p)}(x))^{\mathrm{t}} b^{(p)},$$

where the residuals are $\varepsilon_i' = \sum_{j=p+1}^{\infty} b_j\psi_j(X_i) + \varepsilon_i$ with variance $\sigma_p^2 = \sigma_0^2 + \sum_{j=p+1}^{\infty} b_j^2$; including more terms in the regression structure makes the residuals smaller in size, and vice versa. Consider $Z_i = Y_i\psi^{(p)}(X_i)$, a vector of dimension $p+1$, with mean value seen to be $b^{(p)}$. We will consider conditions under which $-2\log \mathrm{EL}_n(b^{(p)})$, based on $Z_1, \ldots, Z_n$, can be approximated by a χ^2_{p+1} distribution.

The key to verifying (5.2) lies in controlling the sizes of the eigenvalues of the covariance matrix of Z_i, which may be written

$$\Sigma_n = \mathrm{E}Y_i^2\psi^{(p)}(X_i)\psi^{(p)}(X_i)^{\mathrm{t}} - b^{(p)}(b^{(p)})^{\mathrm{t}} = \sigma_0^2 I_p + \Omega_p,$$

where I_p and Ω_p are of size $(p+1)\times(p+1)$ and where the elements of the non-negative definite Ω_p matrix are $\int \xi(x)^2\psi_j(x)\psi_k(x)f(x)\,\mathrm{d}x - b_jb_k$. The eigenvalues of Σ_n take the form $\sigma_0^2 + \phi_j$, where the ϕ_js are the eigenvalues of Ω_p, and are hence bounded downwards by σ_0^2. They are also bounded upwards, since for any unit vector u, $u^{\mathrm{t}}\Omega_p u$ is bounded by $M^2\int(u_0\psi_0 + \cdots + u_p\psi_p)^2 f\,\mathrm{d}x = M^2$, where M bounds $|\xi(x)|$.

As explained near the end of Section 1, we may now use EL methods to produce confidence regions for all of or some of the b_j parameters, to test whether some of them are equal to zero, and via profile EL methods (as in Owen, 2001, Ch. 3.4) make inference for any smooth function of the b_js, like the regression function itself at given positions.

6.5. Density estimation with orthogonal expansions. For i.i.d. data $X_1, \ldots, X_n$ from an unknown density, consider the growing class of models

$$f_p(x\,|\,a) = f_0(x)c_p(a_1,\ldots,a_p)^{-1}\exp\Big\{\sum_{j=1}^{p} a_j\psi_j(x)\Big\}.$$

Here f_0 is a 'start density', around which one models a flexible log-linear structure for deviations, the ψ_j functions are orthonormal w.r.t. f_0, i.e. $\int f_0\psi_j\psi_k\,\mathrm{d}x = \delta_{j,k}$, and c_p is the appropriate normalising constant.

Here we can carry out EL analysis for $\xi = (\xi_1,\ldots,\xi_p)^{\mathrm{t}}$, where $\xi_j = \int f\psi_j\,\mathrm{d}x$, and a growing p. This is done via the vectors $Z_i = (\psi_1(X_i),\ldots,\psi_p(X_i))^{\mathrm{t}}$. The eigenvalues of its covariance matrix will typically be well-behaved, with reasonable conditions on f, and there is stability of fourth order moments if for example the ψ_js are bounded. Thus EL theory holds for analysis of the ξ_js, if $p^3/n \to 0$. One may next transform ξ analysis to a analysis. To do this, one lets for each given ξ vector a be the maximiser of $\sum_{j=1}^{p} a_j\xi_j - \log c_p(a)$, corresponding to selecting the $f_p(x\,|\,a)$ model that minimises the Kullback–Leibler distance from f to its approximant. See in this connection Barron and Sheu (1991) for information on how quickly the best f_p comes close to any given f, as p grows.

Appendix: Proofs

PROOF OF THEOREM 2.1. The statement is established by a routine extension of the proof given in Owen (2001, p. 219) for estimating equations not depending on a nuisance function h. Write $\mathrm{EL}_n(\theta_0, \widehat{h}) = \prod_{i=1}^n (1 + \lambda^{\mathrm{t}} W_{n,i})^{-1}$, where λ is the solution of

$$\frac{1}{n} \sum_{i=1}^n \frac{W_{n,i}}{1 + Y_{n,i}} = 0, \tag{A.1}$$

$W_{n,i} = m(Z_i, \theta_0, \widehat{h})$ and $Y_{n,i} = \lambda^{\mathrm{t}} W_{n,i}$. Hence, since we can write

$$\log(1 + Y_{n,i}) = Y_{n,i} - \tfrac{1}{2} Y_{n,i}^2 + \tfrac{1}{3} \frac{Y_{n,i}^3}{(1 + \xi_{n,i})^3},$$

where $\xi_{n,i}$ lies between 0 and $Y_{n,i}$, we have

$$-2 \log \mathrm{EL}_n(\theta_0, \widehat{h}) = 2 \sum_{i=1}^n \log(1 + Y_{n,i}) = 2 \sum_{i=1}^n Y_{n,i} - \sum_{i=1}^n Y_{n,i}^2 + \frac{2}{3} \sum_{i=1}^n \frac{Y_{n,i}^3}{(1 + \xi_{n,i})^3}.$$

In order to obtain the limiting law of $-2 \log \mathrm{EL}_n(\theta_0, \widehat{h})$ we need to develop an explicit asymptotic expression for λ. Let $\lambda = \|\lambda\| u$, where u is a random unit vector. As in Owen (2001, p. 220) we write

$$\|\lambda\| \{u^{\mathrm{t}} S_n(\theta_0, \widehat{h}) u - \max_{i \le n} \|W_{n,i}\| u^{\mathrm{t}} M_n(\theta_0, \widehat{h})\} \le u^{\mathrm{t}} M_n(\theta_0, \widehat{h}).$$

From assumptions (A1) and (A3) it now follows that

$$\|\lambda\| \{u^{\mathrm{t}} S_n(\theta_0, \widehat{h}) u + o_{\mathrm{pr}}(1)\} = O_{\mathrm{pr}}(n^{-1/2})$$

and hence $\|\lambda\| = O_{\mathrm{pr}}(n^{-1/2})$, since by assumption (A2), $u^{\mathrm{t}} S_n(\theta_0, \widehat{h}) u$ is asymptotically bounded between the largest and smallest eigenvalues of V_2. Next, rewrite (A.1) as

$$0 = \frac{1}{n} \sum_{i=1}^n W_{n,i} \Big(1 - Y_{n,i} + \frac{Y_{n,i}^2}{1 + Y_{n,i}}\Big) = M_n(\theta_0, \widehat{h}) - S_n(\theta_0, \widehat{h}) \lambda + \frac{1}{n} \sum_{i=1}^n \frac{W_{n,i} Y_{n,i}^2}{1 + Y_{n,i}}.$$

The norm of the last term above is bounded above by

$$\frac{1}{n} \|\lambda\|^2 \sum_{i=1}^n \frac{\|W_{n,i}\|^3}{|1 + Y_{n,i}|} = O_{\mathrm{pr}}(n^{-1}) o_{\mathrm{pr}}(n^{1/2}) = o_{\mathrm{pr}}(n^{-1/2}),$$

from assumption (A2) and the fact that $\max_{i \le n} |Y_{n,i}| = o_{\mathrm{pr}}(1)$ by assumption (A3). Hence

$$\lambda = S(\theta_0, \widehat{h})^{-1} M_n(\theta_0, \widehat{h}) + \beta_n,$$

where $\beta_n = o_{\rm pr}(n^{-1/2})$. We may now write

$$\begin{aligned} -2\log \mathrm{EL}_n(\theta_0, \widehat{h}) &= 2n\lambda^{\rm t} M_n(\theta_0, \widehat{h}) - n\lambda^{\rm t} S_n(\theta_0, \widehat{h})\lambda + \frac{2}{3}\sum_{i=1}^{n} \frac{Y_{n,i}^3}{(1+\xi_{n,i})^3} \\ &= nM_n(\theta_0, \widehat{h})^{\rm t} S_n(\theta_0, \widehat{h})^{-1} M_n(\theta_0, \widehat{h}) - n\beta_n^{\rm t} S(\theta_0, \widehat{h})\beta_n + o_{\rm pr}(1) \\ &= nM_n(\theta_0, \widehat{h})^{\rm t} S_n(\theta_0, \widehat{h})^{-1} M_n(\theta_0, \widehat{h}) + o_{\rm pr}(1) \to_d U^{\rm t} V_2^{-1} U, \end{aligned} \tag{A.2}$$

where $U \sim \mathrm{N}(0, V_1)$, and where assumptions (A1) and (A2) are used in the last step. That the third term on the right hand side of (A.2) vanishes in probability is since its norm is bounded by

$$\frac{2}{3}\|\lambda\|^3 \sum_{i=1}^{n} \frac{\|W_{n,i}\|^3}{|1+\xi_{n,i}|^3} = O_{\rm pr}(n^{-3/2}) O_{\rm pr}(n) o_{\rm pr}(n^{1/2}) = o_{\rm pr}(1).$$

The statement of the theorem then follows using Lemma 3 of Qin and Jing (2001a) to re-express the distribution of $U^{\rm t} V_2^{-1} U$ as the weighted sum of independent χ_1^2 random variables. ■

Proof of the claim of Remark 2.1. Conditions (A4) and (A5) imply that, given any real sequence $\delta_n \downarrow 0$,

$$\sup_{\|\theta-\theta_0\| \le \delta_n,\, h \in \bar{\mathcal{H}}} \Big| \frac{1}{n} \sum_{i=1}^{n} \{ m^{\otimes 2}(Z_i, \theta, h) - m^{\otimes 2}(Z_i, \theta_0, h) \} \Big| \to_{\rm pr} 0.$$

The consistency of $\hat{\theta}$ then implies $S_n(\hat{\theta}, \widehat{h}) - S_n(\theta_0, \widehat{h}) \to_{\rm pr} 0$. Thus

$$|\widehat{V}_2 - V_2| \le |S_n(\hat{\theta}, \widehat{h}) - S_n(\theta_0, \widehat{h})| + |S_n(\theta_0, \widehat{h}) - V_2| \to_{\rm pr} 0,$$

where we have used assumption (A2) for the last term, so $\widehat{V}_2$ consistently estimates V_2. ■

Proof of Theorem 2.2. By equation (A.2), the singular value theorem applied to V_2^{-1} and $\widehat{V}_2^{-1}$, and the Cramér–Wold device, it suffices to show that $\widehat{V}_2 \to_{\rm pr} V_2$ and that

$$P^*\{ n^{1/2}[M_n^*(\widehat{\theta}, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h})] \le t \} - P\{U \le t\} = o_{\rm pr}(1).$$

The former follows from Remark 2.1, under conditions (A4) and (A5). For the latter, define, for any sequences $\alpha_n^1, \alpha_n^2 \downarrow 0$,

$$\begin{aligned} A_{n,\alpha_n} = \Big\{ &|\widehat{\theta} - \theta_0| \le \alpha_n^1,\, \sup_t |B_n(t)| \le \alpha_n^1,\, \sup_{\|\theta-\theta_0\| \le \alpha_n^1, \|h-h_0\|_{\mathcal{H}} \le \alpha_n^1} \|C_n(\theta, h)\| \le \alpha_n^2 n^{-1/2}, \\ &\|\widehat{h} - h_0\|_{\mathcal{H}} \le \alpha_n^1 n^{-1/4} \Big\}, \end{aligned}$$

where $B_n(t)$ respectively $C_n(\theta, h)$ is the expression between absolute values (norm-signs) in condition (B1) respectively (B2). Then, by conditions (B1), (B2), (B4) and the consistency of $\widehat{\theta}$, α_n^1 and α_n^2 can be chosen such that $P(A_{n,\alpha_n}) \to 1$ as n tends to infinity. Hence it suffices to establish the convergence in probability, conditionally on the event A_{n,α_n}. It now follows from condition (B5) that

$$\begin{aligned}
&\|M_n^*(\widehat{\theta}, \widehat{h}^*) - M_n^*(\widehat{\theta}, \widehat{h}) - \Gamma(\widehat{\theta}, \widehat{h})[\widehat{h}^* - \widehat{h}]\| \\
&\quad = \|M_n(\widehat{\theta}, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h}) - \Gamma(\widehat{\theta}, \widehat{h})[\widehat{h}^* - \widehat{h}]\| + o_{P^*}(n^{-1/2}) \\
&\quad \leq c\|\widehat{h}^* - \widehat{h}\|_{\mathcal{H}}^2 + o_{P^*}(n^{-1/2}) = o_{P^*}(n^{-1/2}) \text{ a.s.}
\end{aligned}$$

In a similar way it follows from (B2), (B3) and (B4) that

$$\|M_n(\theta_0, \widehat{h}) - M_n(\theta_0, h_0) - \Gamma(\theta_0, h_0)[\widehat{h} - h_0]\| = o_{\mathrm{pr}}(n^{-1/2}).$$

Hence condition (B1) implies that

$$\begin{aligned}
&n^{1/2}\{M_n^*(\widehat{\theta}, \widehat{h}^*) - M_n(\widehat{\theta}, \widehat{h})\} \\
&\quad = n^{1/2}\{M_n^*(\widehat{\theta}, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + \Gamma(\widehat{\theta}, \widehat{h})[\widehat{h}^* - \widehat{h}]\} + o_{P^*}(1) \text{ a.s.}
\end{aligned}$$

has the same limiting distribution as

$$n^{1/2}\{M_n(\theta_0, h_0) + \Gamma(\theta_0, h_0)[\widehat{h} - h_0]\} = n^{1/2} M_n(\theta_0, \widehat{h}) + o_{\mathrm{pr}}(1),$$

which by condition (A1′) converges to U. ■

Proof of Theorem 4.1. The proof is similar to Theorem 2.1, but there are some key steps where the argument needs to be modified. The first point of departure is in finding the order of the Lagrange multipliers. Now, using conditions (C1) and (C3),

$$\|\lambda\| \{u^{\mathrm{t}} S_n(\theta_0, \widehat{h}) u + O_{\mathrm{pr}}(1)\} = O_{\mathrm{pr}}(n^{-\alpha}),$$

whereas before the $O_{\mathrm{pr}}(1)$ term was asymptotically negligible. Multiplying both sides of the above display by $n^{2\alpha-1}$ and using condition (C2) we find that

$$\|\lambda\| \{u^{\mathrm{t}} V_2 u + o_{\mathrm{pr}}(1) + O_{\mathrm{pr}}(n^{2\alpha-1})\} = O_{\mathrm{pr}}(n^{\alpha-1}).$$

Then, since the term $O_{\mathrm{pr}}(n^{2\alpha-1}) = o_{\mathrm{pr}}(1)$ in the case $0 < \alpha < 1/2$, the Lagrange multipliers have order $\|\lambda\| = O_{\mathrm{pr}}(n^{\alpha-1})$. Next,

$$\beta_n = S_n(\theta_0, \widehat{h})^{-1} \|\lambda\|^2 O_{\mathrm{pr}}(n^{\alpha}) O_{\mathrm{pr}}(n^{1-2\alpha}) = O_{\mathrm{pr}}(n^{3\alpha-2}).$$

This leads to bounds on the two remainder terms in the expansion (A.2) of the likelihood ratio statistic:

$$n\beta_n^{\mathrm{t}} S(\theta_0, \widehat{h}) \beta_n = O_{\mathrm{pr}}(n^{2(2\alpha-1)}) \to_{\mathrm{pr}} 0$$

and

$$\|\lambda\|^3 \sum_{i=1}^{n} \frac{\|W_{n,i}\|^3}{|1+\xi_{n,i}|^3} = O_{\rm pr}(n^{3(\alpha-1)})O_{\rm pr}(n)O_{\rm pr}(n^{\alpha})O_{\rm pr}(n^{1-2\alpha}) = O_{\rm pr}(n^{2\alpha-1}) \to_{\rm pr} 0.$$

Hence the likelihood ratio statistic has the decomposition

$$-2\log {\rm EL}_n(\theta_0,\widehat{h}) = n^{\alpha} M_n(\theta_0,\widehat{h})^{\rm t}\{n^{2\alpha-1}S_n(\theta_0,\widehat{h})\}^{-1} n^{\alpha} M_n(\theta_0,\widehat{h}) + o_{\rm pr}(1) \to_d UV_2^{-1}U,$$

as required. ■

PROOF OF LEMMA 5.1. We may take $\mu_{0,n} = 0$ without loss of generality. For the components of the $p \times p$ matrix $D_n = S_n - \Sigma_n$ we can bound $P\{|D_{n,j,k}| \geq \varepsilon\}$ via the Markov inequality. We find

$$P\{|D_{n,j,k}| \geq \varepsilon\} \leq \frac{{\rm E}|\sqrt{n}D_{n,j,k}|^q}{(\sqrt{n}\varepsilon)^q} \leq \frac{c(q)V_{n,j,k}^{q/2}}{n^{q/2}\varepsilon^q},$$

for a constant $c(q)$, by results of von Bahr (1965). Here $V_{n,j,k} = {\rm E}(Z_{i,j}Z_{i,k})^2 - \sigma_{n,j,k}^2$ is the variance of $Z_{i,j}Z_{i,k}$. This may be further bounded by

$$V_{n,j,k} \leq ({\rm E}|Z_{i,j}|^4)^{1/2}({\rm E}|Z_{i,k}|^4)^{1/2} \leq ({\rm E}|Z_{i,j}|^q)^{2/q}({\rm E}|Z_{i,k}|^q)^{2/q}$$

for $q \geq 4$. This gives

$$P\{L_n \geq \varepsilon\} \leq \sum_{j,k} \frac{c(q){\rm E}|Z_{i,j}|^q\,{\rm E}|Z_{i,k}|^q}{n^{q/2}\varepsilon^q},$$

which is seen to imply the lemma. ■

PROOF OF LEMMA 5.2. To gauge the size of B_n we cannot appeal to arguments involving the Borel–Cantelli lemma, as Owen (2001, Ch. 11) could when analysing the fixed p situation. However,

$$P\{B_n \geq (n/p)^{1/2}\varepsilon\} \leq nP\{\|Z_i\| \geq (n/p)^{1/2}\varepsilon\} \leq n\frac{p^q}{n^{q/2}\varepsilon^q}{\rm E}\|Z_i/p^{1/2}\|^q,$$

proving the first part of the lemma.

Next write $\lambda = \|\lambda\|u$ where u has length 1. From (5.3) and $u^{\rm t}g(\lambda) = 0$ we find via some rearranging that $u^{\rm t}\bar{Z} = u^{\rm t}\widetilde{S}_n u\,\|\lambda\|$, where $\widetilde{S}_n = n^{-1}\sum_{i=1}^n Z_i Z_i^{\rm t}/(1+Y_i)$. Then $u^{\rm t}S_n u \leq u^{\rm t}\widetilde{S}_n u(1+\|\lambda\|B_n)$, which gives

$$\|\lambda\|(u^{\rm t}S_n u - u^{\rm t}\bar{Z}B_n) \leq u^{\rm t}\bar{Z}.$$

Here $|n^{1/2}u^{\rm t}\bar{Z}| \leq |n^{1/2}\bar{Z}|$, which is $O_{\rm pr}(p^{1/2})$, since its square $n\bar{Z}^{\rm t}\bar{Z}$ is seen to have mean of order p. This leads to

$$\|\lambda\| \leq \frac{n^{1/2}u^{\rm t}\bar{Z}\,n^{-1/2}}{u^{\rm t}S_n u - n^{1/2}u^{\rm t}\bar{Z}B_n n^{-1/2}} \leq \frac{O_{\rm pr}(p^{1/2})n^{-1/2}}{\widehat{\gamma}_{p,n} + O_{\rm pr}(p^{1/2})o_{\rm pr}(1/p^{1/2})},$$

in terms of the smallest eigenvalue of S_n. Accordingly, by remarks made after Lemma 5.1, $\|\lambda\| = O_{\rm pr}((p/n)^{1/2})$. ■

PROOF OF THEOREMS 5.1 AND 5.2. We have already seen that $\Lambda_n = 2\sum_{i=1}^n \log(1+Y_i)$, with control over the size of $Y_i = \lambda^{\rm t} Z_i$. Writing

$$B_n = \max_{i\le n} \|Z_i\| = (n/p)^{1/2}\varepsilon_n, \tag{A.3}$$

we have $\max_{i\le n}|Y_i| \le \|\lambda\|\, B_n = O_{\rm pr}(1)\varepsilon_n \to_{\rm pr} 0$ by Lemma 5.2. In particular, the event Ω_n that $|Y_i| \le \frac12$ for each $i \le n$ has probability going to 1. For $|y| \le \frac12$, $\log(1+y) = y - \frac12 y^2 + \frac13 y^3 h(y)$, where $|h(y)| \le 2$. This enables our writing

$$\Lambda_n = 2\sum_{i=1}^n \log(1+Y_i) = U_n + V_n,$$

under Ω_n, with $U_n = 2\sum_{i=1}^n (Y_i - \frac12 Y_i^2)$ and $V_n = 2\sum_{i=1}^n \frac13 Y_i^3 h(Y_i)$. The point is that U_n is close to the quadratic form Q_n of (5.2) while V_n will be small.

We first check the size of the last term in the (5.3) based expansion

$$0 = n^{-1}\sum_{i=1}^n Z_i\Big(1 - Y_i + \frac{Y_i^2}{1+Y_i}\Big) = \bar Z - S_n\lambda + \alpha_n,$$

which will be used to construct a sufficiently precise approximation for λ. Under Ω_n, $(1+Y_i)^{-1} \le 1 + 2|Y_i| \le 2$, and

$$\begin{aligned}\|\alpha_n\| &\le 2n^{-1}\sum_{i=1}^n Y_i^2\|Z_i\| \le 2B_n\lambda^{\rm t} S_n\lambda \\ &= O_{\rm pr}((n/p)^{1/2})\varepsilon_n O_{\rm pr}((p/n))\max\text{eigen}(S_n) = O_{\rm pr}((p/n)^{1/2})\varepsilon_n.\end{aligned}$$

We have $\lambda = S_n^{-1}(\bar Z_n + \alpha_n)$ and learn that

$$U_n = 2n\lambda^{\rm t}\bar Z - n\lambda^{\rm t} S_n\lambda = n\bar Z^{\rm t} S_n^{-1}\bar Z - n\alpha_n^{\rm t} S_n^{-1}\alpha_n = Q_n - \delta_n,$$

where $|\delta_n| \le \|n^{1/2}\alpha_n\|^2 \min\text{eigen}(S_n)$ has size $O_{\rm pr}(p)\varepsilon_n^2$. Also,

$$\begin{aligned}|V_n| &\le (4/3)\sum_{i=1}^n |Y_i|^3 \le (4/3)\|\lambda\| B_n n\lambda^{\rm t} S_n\lambda \\ &= nO_{\rm pr}((p/n)^{3/2})O_{\rm pr}((n/p)^{1/2})\varepsilon_n = O_{\rm pr}(p)\varepsilon_n.\end{aligned}$$

To sum up these efforts,

$$\Lambda_n = Q_n - \delta_n + V_n = Q_n + O_{\rm pr}(p\varepsilon_n^2) + O_{\rm pr}(p\varepsilon_n) = Q_n + O_{\rm pr}(p\varepsilon_n).$$

Theorem 5.1 deals with the case where all $|Z_{i,j}| \le M$ for some M. Then ε_n of (A.3) is at most $Mp/n^{1/2}$. The implications of $p^3/n \to 0$ and $p^4/n \to 0$ are therefore that respectively $p^{1/2}\varepsilon_n$ and $p\varepsilon_n$ go to zero in probability, proving Theorem 5.1.

For the case of Theorem 5.2, a bound for $P\{\varepsilon_n \geq \varepsilon\}$ is provided by Lemma 5.2, with implications for $P\{p^{1/2}\varepsilon_n \geq \varepsilon\}$ and $P\{p\varepsilon_n \geq \varepsilon\}$. Working through these deails one arrives at the conclusion of Theorem 5.2. ■

PROOF OF THEOREM 5.3. As argued in Section 5.1, the invariance property of the EL allows us to trace the arguments used above to prove Theorem 5.2 in the current situation, where $\Sigma_n = I_p$ and $\mu_{0,n} = 0$. The proof is seen to go through, down to the representation $\Lambda_n = Q_n - \delta_n + V_n$. Examination of the details shows that

$$|\delta_n| \leq O_{\rm pr}(p\varepsilon_n^2)\widehat{\gamma}_{1,n}^2/\widehat{\gamma}_{p,n}, \quad |V_n| \leq O_{\rm pr}(p\varepsilon_n)\widehat{\gamma}_{1,n}.$$

The point is now that the eigenvalues of S_n are secured good behaviour, in view of arguments used after Lemma 5.1 and involving $pL_n \to_{\rm pr} 0$. ■

PROOF OF PROPOSITION 5.1. For this part of the problem we may again take $\mu_{0,n} = 0$ and in addition Σ_n equal to the identity matrix I_p, to simplify matters. We need to bound $a_{n,p} = \max_t |P\{n\bar{Z}^{\rm t}\bar{Z} \leq t\} - P\{\chi_p^2 \leq t\}|$. One knows that $a_{n,p} \to 0$ for fixed p, as a consequence of the central limit theorem for $N_n = n^{1/2}\bar{X}$. The maximal approximation error will however grow with growing dimension. Bhattacharya and Ranga Rao (1976) give various uniform upper bounds for approximating the distribution of N_n with a $\mathrm{N}_p(0, I_p)$. These bounds are often quite weak in that the uniformity in question includes all the 'worst case' scenarios. When the sets for which probability approximations are sought are all convex, however, as with the quadratic form Q_n^0, sharper bounds can be established. Finessing earlier results in the literature, Götze (1991) was able to show that

$$a_{n,p} \leq b(p)\mathrm{E}\|Z_i\|^3/n^{1/2} = b(p)p^{3/2}/n^{1/2}\,\mathrm{E}\|Z_i/p^{1/2}\|^3,$$

where $b(p) = O(p)$. In fact, $b(p) \leq 158\,p + 10$ for all $p \geq 6$, with a mild constraint involved for the form of the probability mechanism behind the X_is. The first part of the proposition follows from this.

Approximations involved for the second part of the proposition can be made stronger in that a smaller class of sets is involved, namely those of the type $\{Q_n^0 \leq p + (2p)^{1/2}t\}$. The statement follows in fact from efforts of Portnoy (1988, Section 4), who used a martingale central limit theorem. ■

PROOF OF PROPOSITION 5.2. We may transform Z_i to $Z_i^* = \Sigma_n^{-1/2}(Z_i - \mu_{0,n})$ here, as in the preamble to Theorem 5.2. We therefore assume without loss of generality that $\Sigma_n = I_p$ and $\mu_{0,n} = 0$. Write now $S_n = I_p + D_n$ where D_n becomes small, as per Lemma 5.1. Hence Σ_n may be inverted to give $S_n^{-1} = I_p - D_n + D_n^2 - \cdots$, giving, on a set with probability going to 1,

$$Q_n - Q_n^0 = -n\bar{Z}^{\rm t}D_n\bar{Z} + o_{\rm pr}(1).$$

The first term may be bounded by

$$\sum_{j,k} n|\bar{Z}_{n,j}\bar{Z}_{n,k}|\,|D_{n,j,k}| \le L_n \sum_{j,k} n|\bar{Z}_{n,j}\bar{Z}_{n,k}| \le L_n \|n^{1/2}\bar{Z}_n\|^2 = O_{\rm pr}(p^2 L_n),$$

in terms of the L_n of Lemma 5.1. This goes to zero in probability if $p^{2+2q}/n^{q/2} \to 0$. The statement involving $(Q_n - p)/(2p)^{1/2}$ is proved similarly, using also Proposition 5.1, seeing that its distance to $(Q_n^0 - p)/(2p)^{1/2}$ is of size $O_{\rm pr}(p^{3/2}L_n)$. ■

Acknowledgments. The research of Ingrid Van Keilegom was supported by IAP research network grant nr. P5/24 of the Belgian government (via its Belgian Science Policy). Ian McKeague was supported by NSF Grant 0204688. The three authors are also grateful to the biostatistical working group NorEvent at the University of Oslo and its director Professor Odd Aalen for support in connection with a Research Kitchen in Oslo, where they also benefitted from discussions with Gerda Claeskens.

References

Akritas, M.G. and Arnold, S. (2000). Asymptotics for analysis of variance when the number of levels is large. *Journal of the American Statistical Association* **95**, 212–226.

Akritas, M.G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scandinavian Journal of Statistics* **28**, 549–568.

von Bahr, B. (1965). On convergence of moments in the central limit theorem. *Annals of Mathematical Statistics* **36**, 808–818.

Bai, Z.D. and Yin, Y.Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability* **21**, 1275–1294.

Bai, Z.D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review [with discussion]. *Statistica Sinica* **9**, 611–677.

Banerjee, M. and Wellner J.A. (2002). Confidence intervals for current status data. Preprint.

Barron, A.R. and Sheu, C. (1991). Approximations of density functions by sequences of exponential families. *Annals of Statistics* **19**, 1347–1369.

Bentkus, V. and Götze, F. (1996). Optimal rates of convergence in the CLT for quadratic forms. *Annals of Probability Theory* **24**, 466–490.

Bentkus, V. and Götze, F. (1997). Uniform rates of convergence in the CLT for quadratic forms in multidimensional spaces. *Probability Theory and Related Fields* **109**, 367–416.

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

Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, London.

Chen, S.L. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83**, 329–341.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591–1608.

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Annals of Probability* **18**, 851–869.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Annals of Probability* **19**, 724–739.

Groeneboom, P. (1987). Asymptotics for interval censored observations. Report 87-18. Department of Mathematics, University of Amsterdam.

Hall, P. and Owen, A. (1993). Empirical likelihood confidence bands in density estimation. *Journal of Computational Graphics and Statistics* **2**, 273–289.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals. *Annals of Statistics* **16**, 1428–1449.

Hjort, N.L. (1999). Towards semiparametric bandwidth selectors for kernel density estimators. Statistical Research Report, Department of Mathematics, University of Oslo.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Annals of Statistics* **25**, 2084–2102.

van der Laan, M.J. and Robins, J.M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association* **93**, 693–701.

van der Laan, M.J. and van der Vaart A.W. (2000). Estimating a survival distribution with current status data and high-dimensional covariates. Preprint.

Lehmann, E.L. (1975). *Nonparametrics. Statistical methods based on ranks.* Holden-Day Series in Probability and Statistics. Holden-Day, San Francisco.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

Li, G. and Doss, H. (1993). Generalized Pearson–Fisher chi-square goodness-of-fit tests, with applications to models with life history data. *Annals of Statistics* **21**, 772–797.

Li, G. and Wang, Q.-H. (2003). Empirical likelihood regression analysis for right censored data. *Statistica Sinica* **13**, 51–68.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, New York.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics* **18**, 90–120.

Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall/CRC, London.

Portnoy, S. (1986). On the central limit theorem in R^p when p$\to \infty$. *Probability Theory and Related Fields* **73**, 571–583.

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

Qin, G. and Jing, B.-Y. (2001a). Empirical likelihood for censored linear regression. *Scandinavian Journal of Statistics* **28**, 661–673.

Qin, G. and Jing, B.-Y. (2001b). Censored partial linear models and empirical likelihood. *Journal of Multivariate Analysis* **78**, 37–61.

Qin, G. and Tsao, M. (2003). Empirical likelihood inference for median regression models for censored survival data. *Journal of Multivariate Analysis* **85**, 416–430.

Satten, G.A. and Datta, S. (2001). The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *American Statistician* **55**, 207–210.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics* **2**, 113–126.

Stute, W. (1996). The jackknife estimate of variance of a Kaplan–Meier integral. *Annals of Statistics* **24**, 2679–2704.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

van der Vaart, A. and Wellner, J.A. (2000). Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes. In *High Dimensional Probability II* (Seattle, 1999), 115–133, Progr. Probab. **47**, Birkhäuser Boston, Boston, MA.

Van Keilegom, I. and Veraverbeke, N. (2002). Density and hazard estimation in censored regression models. *Bernoulli* **8**, 607–625.

Wang, J.G. (1987). A note on the uniform consistency of the Kaplan–Meier estimator. *Annals of Statistics* **15**, 1313–1316.

Wang, Q.-H. and Jing, B.-Y. (2001). Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics* **53**, 517–527.

Wang, Q.-H. and Rao, J.N.K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics* **30**, 896–924.