

Empirical Likelihood for Non-Smooth Criterion Functions

ELISA M. MOLANES LOPEZ

Departamento de Estadística, Universidad Carlos III de Madrid

INGRID VAN KEILEGOM

Institute of Statistics, Université catholique de Louvain

NOËL VERAVERBEKE

Center for Statistics, Universiteit Hasselt

ABSTRACT. Suppose that X_1, \dots, X_n is a sequence of independent random vectors, identically distributed as a d -dimensional random vector X . Let $\mu \in \mathbb{R}^d$ be a parameter of interest and $\nu \in \mathbb{R}^q$ be some nuisance parameter. The unknown, true parameters (μ_0, ν_0) are uniquely determined by the system of equations $E\{g(X, \mu_0, \nu_0)\} = 0$, where $g = (g_1, \dots, g_{p+q})$ is a vector of $p+q$ functions. In this paper we develop an empirical likelihood (EL) method to do inference for the parameter μ_0 . The results in this paper are valid under very mild conditions on the vector of criterion functions g . In particular, we do not require that g_1, \dots, g_{p+q} are smooth in μ or ν . This offers the advantage that the criterion function may involve indicators, which are encountered when considering, e.g. differences of quantiles, copulas, ROC curves, to mention just a few examples. We prove the asymptotic limit of the empirical log-likelihood ratio, and carry out a small simulation study to test the performance of the proposed EL method for small samples.

Key words: confidence region, copulas, empirical likelihood, estimating equations, hypothesis testing, nuisance parameter, quantiles, ROC curve

1. Introduction and general method

Suppose that X_1, \dots, X_n is a sequence of independent random vectors, identically distributed as a d -dimensional random vector X . Let $\mu \in \mathbb{R}^d$ be a parameter of interest and $\nu \in \mathbb{R}^q$ be some nuisance parameter. The unknown, true parameters (μ_0, ν_0) are uniquely determined by the system of equations

$$E\{g(X, \mu_0, \nu_0)\} = 0, \tag{1}$$

where $g = (g_1, \dots, g_{p+q})$ is a vector of $p+q$ functions. In this paper we develop an empirical likelihood (EL) method to do inference for the parameter μ_0 . The results in this paper are valid under very mild conditions on the vector of criterion functions g . In particular, we do not require that g_1, \dots, g_{p+q} are smooth in μ or ν . This offers the advantage that the criterion function may involve indicators, which are encountered when considering, e.g. differences of quantiles, copulas, ROC curves, to mention just a few examples.

Qin & Lawless (1994) also consider the problem of developing EL theory for the parameter μ_0 . However, their results are restricted to smooth criterion functions, and hence they exclude many interesting examples. See also the remark following theorem 3.6 in Owen (2001), where some examples are given of situations that are ruled out by their result, and section 10.6 in Owen (2001), which considers in more detail the difficulties encountered when considering non-smooth estimating equations.

In this paper we will overcome this smoothness condition by using a different method of proof. In fact, our proof is based on Sherman (1993), who developed general conditions under which the maximizer of a locally quadratic criterion function is consistent and asymptotically normal. His result is valid without assuming that the criterion function is continuous. On the contrary, Qin & Lawless (1994) heavily use Taylor expansions in their proofs, for which smoothness of the functions g_1, \dots, g_{p+q} is indispensable.

Define for any $(\mu, v) \in \mathbb{R}^{p+q}$ the EL:

$$L(\mu, v) = n^n \sup \left\{ \prod_{i=1}^n p_i(\mu, v) : p_i(\mu, v) \geq 0, \sum_{i=1}^n p_i(\mu, v) = 1, \sum_{i=1}^n p_i(\mu, v) g_j(X_i, \mu, v) = 0, j = 1, \dots, p+q \right\}. \tag{2}$$

The supremum in (2) is defined to be zero when the set is empty, and exists and is unique provided that 0 belongs to the interior of the convex hull of $(g(X_1, \mu, v), \dots, g(X_n, \mu, v))$. In the latter case, the standard Lagrange multiplier method provides the optimal $p_i(\mu, v)$:

$$p_i(\mu, v) = \frac{1}{n} \left(1 + \sum_{j=1}^{p+q} \lambda_j(\mu, v) g_j(X_i, \mu, v) \right)^{-1}, \quad i = 1, \dots, n, \tag{3}$$

and also the following empirical log-likelihood ratio for μ :

$$\ell(\mu, v) = -2 \log L(\mu, v) = 2 \sum_{i=1}^n \log \left\{ 1 + \sum_{j=1}^{p+q} \lambda_j(\mu, v) g_j(X_i, \mu, v) \right\}, \tag{4}$$

where the Lagrange multipliers $\lambda_j(\mu, v) (j = 1, \dots, p+q)$ satisfy the following equations:

$$\sum_{i=1}^n \frac{g_j(X_i, \mu, v)}{1 + \sum_{k=1}^{p+q} \lambda_k(\mu, v) g_k(X_i, \mu, v)} = 0, \quad j = 1, \dots, p+q. \tag{5}$$

Now, define an estimator $\tilde{v}(\mu)$ of the nuisance parameter v by maximizing $L(\mu, v)$ over v for a fixed value of μ , or equivalently by minimizing $\ell(\mu, v)$:

$$\tilde{v}(\mu) = \operatorname{argmin}_v \ell(\mu, v), \tag{6}$$

and let $\tilde{v} = \tilde{v}(\mu_0)$. Finally, define

$$\ell(\mu) = \ell(\mu, \tilde{v}(\mu)). \tag{7}$$

The main result of this paper shows that the asymptotic distribution of $\ell(\mu_0)$ is χ_p^2 .

In a number of papers, the lack of smoothness of the criterion functions g_1, \dots, g_{p+q} has been overcome by replacing them by smooth approximations, leading to a so-called smooth EL. See, e.g. Zhou & Jing (2003) for differences of quantiles, Claeskens *et al.* (2003) for ROC curves and Chen *et al.* (2009) for copulas. However, this has the drawback that a bandwidth parameter needs to be selected, which is often a challenging problem. In this paper we do not apply any smoothing in the EL procedure, thanks to the new method of proof.

Instead of profiling out the nuisance parameter v_0 , as we have done in (6), one could also replace v_0 by a certain ‘plug-in’ estimator, different from the above profile estimator. This idea has been considered in Hjort *et al.* (2009) in a general framework (where v_0 is allowed to be a function rather than a parameter). With that approach, the limit of the empirical log-likelihood ratio is, however, not necessarily a χ_p^2 variable, but it is in general a weighted sum of χ_1^2 variables, where the weights are often unknown. The method proposed in this paper yields an unweighted χ_p^2 distribution, thanks to the way the parameter v_0 is estimated.

The paper is organized as follows. In the next section, we formulate the main result of this paper, and state the conditions under which this result is valid. We also discuss the extension of the proposed method to the case of multiple samples. In section 3, a number of specific examples are considered and the general conditions are tested on these examples. The results of a small simulation study are shown in section 4, whereas the proof of the main result and some technical lemmas are given in the Appendix.

2. Main result

As mentioned in the Introduction, the aim of this section is to show that Wilks' theorem (which says that the empirical log-likelihood ratio $\ell(\mu_0)$ converges in distribution to a χ_p^2 variable) is valid, even when the criterion functions are not smooth. This then allows to construct approximate confidence regions for the parameter of interest μ_0 .

The proof of this result relies on theorems 1 and 2 of Sherman (1993). In that paper, a general method is given for establishing the rate of convergence and the asymptotic normality of a maximization estimator that does not require differentiability of the criterion functions. The following matrix V of dimensions $(p + 2q) \times (p + 2q)$ will play an important role:

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}' & 0 \end{pmatrix}, \tag{8}$$

where

$$V_{11} = (E \{g_j(X, \mu_0, v_0)g_k(X, \mu_0, v_0)\})_{j,k=1, \dots, p+q}$$

$$V_{12} = \left(-\frac{\partial}{\partial v_k} E\{g_j(X, \mu_0, v)\} \Big|_{v=v_0} \right)_{k=1, \dots, q; j=1, \dots, p+q}.$$

We will need the following conditions:

- (C0) There exists a neighbourhood \mathcal{N} of v_0 such that $P(L(\mu_0, v) > 0 \text{ for all } v \in \mathcal{N}) \rightarrow 1$.
- (C1) The functions $g_j(x, \mu_0, v)$ ($j = 1, \dots, p + q$) are uniformly bounded in $\mathbb{R}^d \times \mathbb{R}^q$; the functions

$$E\{g_j(X, \mu_0, v)g_k(X, \mu_0, v)\}, \quad j, k = 1, \dots, p + q,$$

$$\frac{\partial}{\partial v_k} E\{g_j(X, \mu_0, v)\} \quad \text{and} \quad \frac{\partial^2}{\partial v_k \partial v_\ell} E\{g_j(X, \mu_0, v)\}, \quad k, \ell = 1, \dots, q; j = 1, \dots, p + q$$

are continuous for v in a neighbourhood of v_0 ; the function

$$E\{g(X, \mu_0, v)/[1 + \xi^t g(X, \mu_0, v)]\}$$

has continuous partial derivatives with respect to the components of v in a neighbourhood of v_0 ; the function

$$E\{g(X, \mu_0, v)g^t(X, \mu_0, v)/(1 + \xi^t g(X, \mu_0, v))^2\}$$

is uniformly continuous with respect to the components of v and ξ in a neighbourhood of v_0 and 0.

- (C2) The matrix V_{11} in (8) is positive definite.
- (C3) $\tilde{v} = \text{argmin}_v \ell(\mu_0, v)$ converges in probability to v_0 .

$$(C4) \quad n^{-1} \sum_{i=1}^n [g_j(X_i, \mu_0, v) - E\{g_j(X, \mu_0, v)\}] = O_P(n^{-1/2})$$

uniformly in v in an $o(1)$ -neighbourhood of v_0 ($j = 1, \dots, p + q$).

$$(C5) \quad n^{-1} \sum_{i=1}^n [g_j(X_i, \mu_0, \nu) g_k(X_i, \mu_0, \nu) - E\{g_j(X, \mu_0, \nu) g_k(X, \mu_0, \nu)\}] = o_P(1)$$

uniformly in ν in an $o(1)$ -neighbourhood of ν_0 ($j, k = 1, \dots, p + q$).

$$(C6) \quad n^{-1} \sum_{i=1}^n [g_j(X_i, \mu_0, \nu) - E\{g_j(X, \mu_0, \nu)\} - g_j(X_i, \mu_0, \nu_0) + E\{g_j(X, \mu_0, \nu_0)\}] = o_P(n^{-1/2})$$

uniformly in ν for $\nu - \nu_0 = O(n^{-1/2})$ ($j = 1, \dots, p + q$).

Condition (C0) is equivalent to requiring that the probability that the zero-vector belongs to the interior of the convex hull of $(g(X_1, \mu_0, \nu), \dots, g(X_n, \mu_0, \nu))$ for all $\nu \in \mathcal{N}$ converges to one, and is needed to ensure that the log-likelihood $\ell(\mu_0, \nu)$ can be written as in (4). Note that in (C1) we only impose smoothness conditions on $E\{g(X, \mu_0, \nu)\}$ and not on $g(X, \mu_0, \nu)$ itself. Hence, we are able to handle non-smooth criterion functions, like indicators. The matrix in (C2) is by construction positive semidefinite. All we ask is that it is also positive definite, which is a very mild assumption. To prove condition (C3), use can be made of, e.g. theorem 5.7 in Van der Vaart (1998, p. 45). Finally, conditions (C4) and (C5) are standard uniform consistency results, while (C6) is a Bahadur-type modulus of continuity result. They can be easily proved or found in the literature for particular choices of the criterion functions g_j . See also section 3, where we check the above conditions in a few particular examples of the general method.

Theorem 1

Assume (C0)–(C6) Then,

$$\ell(\mu_0) = \ell(\mu_0, \tilde{\nu}(\mu_0)) = -2 \log L(\mu_0, \tilde{\nu}(\mu_0)) \xrightarrow{d} \chi_p^2.$$

In the special case where g_1, \dots, g_{p+q} are smooth functions, this result has been shown in corollary 4 in Qin & Lawless (1994), using a different method of proof.

Remark 1 (EL for two samples). The situation above can also be extended to the multi-sample situation. For simplicity we describe here the two-sample case. Suppose X_1, \dots, X_{n_1} is a sample from a d_1 -dimensional random vector X , and Y_1, \dots, Y_{n_2} is an independent sample from a d_2 -dimensional random vector Y . Suppose that the parameters (μ_0, ν_0) are uniquely determined by the equations

$$E\{g(X, \mu_0, \nu_0)\} = 0 \quad \text{and} \quad E\{h(Y, \mu_0, \nu_0)\} = 0, \tag{9}$$

where $g = (g_1, \dots, g_{r_1})$, $h = (h_1, \dots, h_{r_2})$ and $r_1 + r_2 = p + q$. The EL ratio for μ is now

$$L(\mu, \nu) = n_1^{n_1} n_2^{n_2} \sup \left\{ \prod_{i=1}^{n_1} p_i(\mu, \nu) \prod_{j=1}^{n_2} q_j(\mu, \nu) : p_i(\mu, \nu) \geq 0, q_j(\mu, \nu) \geq 0, \sum_{i=1}^{n_1} p_i(\mu, \nu) = 1, \right. \\ \left. \sum_{j=1}^{n_2} q_j(\mu, \nu) = 1, \sum_{i=1}^{n_1} p_i(\mu, \nu) g_j(X_i, \mu, \nu) = 0, \sum_{i=1}^{n_2} q_i(\mu, \nu) h_j(Y_i, \mu, \nu) = 0 \right\}.$$

The empirical log-likelihood for μ is again defined as

$$\ell(\mu, \nu) = -2 \log L(\mu, \nu) \\ = 2 \sum_{i=1}^{n_1} \log \left\{ 1 + \sum_{j=1}^{r_1} \lambda_j(\mu, \nu) g_j(X_i, \mu, \nu) \right\} + 2 \sum_{i=1}^{n_2} \log \left\{ 1 + \sum_{j=1}^{r_2} \kappa_j(\mu, \nu) h_j(Y_i, \mu, \nu) \right\},$$

where the Lagrange multipliers $\lambda_j(\mu, v)$ ($j=1, \dots, r_1$) and $\kappa_j(\mu, v)$ ($j=1, \dots, r_2$) satisfy equations analogous to the one in (5). The definitions of $\tilde{v} = \tilde{v}(\mu_0)$ and $\ell(\mu) = \ell(\mu, \tilde{v}(\mu))$ are the same as in (6) and (7). The analogue of the $(p + 2q) \times (p + 2q)$ matrix V in (8) now becomes

$$\begin{pmatrix} V_{11} & 0 & V_{13} \\ 0 & V_{22} & V_{23} \\ V'_{13} & V'_{23} & 0 \end{pmatrix},$$

with

$$\begin{aligned} V_{11} &= \left(E\{g_j(X, \mu_0, v_0)g_k(X, \mu_0, v_0)\} \right)_{j,k=1, \dots, r_1} \\ V_{22} &= \left(E\{h_j(Y, \mu_0, v_0)h_k(Y, \mu_0, v_0)\} \right)_{j,k=1, \dots, r_2} \\ V_{13} &= \left(-\frac{\partial}{\partial v_k} E\{g_j(X, \mu_0, v)\} \Big|_{v=v_0} \right)_{j=1, \dots, r_1; k=1, \dots, q} \\ V_{23} &= \left(-\frac{\partial}{\partial v_k} E\{h_j(Y, \mu_0, v)\} \Big|_{v=v_0} \right)_{j=1, \dots, r_2; k=1, \dots, q} \end{aligned}$$

Very similar to the one-sample case, one can prove that $\ell(\mu_0)$ is asymptotically χ^2_p .

The required conditions are completely similar to (C0)–(C6). For (C1) and (C4)–(C6), we need to impose the parallel conditions on the functions g_j and h_j , whereas for (C2) positive definiteness of V_{11} and V_{22} is required. We also need to impose the usual asymptotic balance condition on the sample sizes n_1 and n_2 .

Remark 2 (Tests and confidence regions). An approximate $100(1 - \alpha)\%$ EL confidence region for μ_0 is obtained by the following subset of \mathbb{R}^p :

$$\left\{ \mu : \ell(\mu) \leq \chi^2_{p, 1-\alpha} \right\} = \left\{ \mu : L(\mu, \tilde{v}(\mu)) \geq \exp\left(-\frac{1}{2}\chi^2_{p, 1-\alpha}\right) \right\},$$

where $\chi^2_{p, 1-\alpha}$ is the $(1 - \alpha)$ quantile of the χ^2_p distribution. Similarly, a level- α test for the null hypothesis $H_0 : \mu = \mu_0$ will reject H_0 if $\ell(\mu_0) > \chi^2_{p, 1-\alpha}$.

3. Examples of the general method

We consider four examples of the general theory in more detail. In each of these examples, the criterion function involves indicators, which could not be dealt with so far in the literature on EL methods without using smoothing techniques.

3.1. Difference of quantiles in the one sample problem

For $d = 1$ and X_1, \dots, X_n a random sample from X with distribution function F , we consider the difference of quantiles

$$\mu_0 = F^{-1}(p_2) - F^{-1}(p_1),$$

where $0 < p_1 < p_2 < 1$ and $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ for $0 < u < 1$. Clearly, when $p_1 = 0.25$ and $p_2 = 0.75$ we get the interquartile range. Introducing the parameter $v_0 = F^{-1}(p_1)$, we have the equations

$$\begin{cases} F(\mu_0 + v_0) - p_2 = 0 \\ F(v_0) - p_1 = 0. \end{cases}$$

So, we have (1) with $p=q=1$, $g_1(X, \mu, v)=I(X \leq v) - p_1$ and $g_2(X, \mu, v)=I(X \leq \mu + v) - p_2$. This problem has been studied by Zhou & Jing (2003) by using a smoothed EL approach, whereas Chen & Hall (1993) studied one single quantile using a smoothed EL approach.

From theorem 1 we know that $\ell(\mu_0)$ converges to a χ_1^2 distribution, provided conditions (C0)–(C6) are satisfied. First, define $\mathcal{N} = \{v: \delta < F(v) < F(\mu_0 + v) < 1 - \delta\}$ for some $\delta > 0$. Then, with probability growing to one, we have that for all $v \in \mathcal{N}$, the convex hull of $(g(X_1, \mu_0, v), \dots, g(X_n, \mu_0, v))$ is equal to the triangle in \mathbb{R}^2 with corners $(1 - p_1, 1 - p_2), (-p_1, 1 - p_2)$ and $(-p_1, -p_2)$. It is easily seen that $(0, 0)$ is inside this triangle as $p_1 < p_2$. Hence, (C0) is satisfied. Condition (C1) is satisfied if $F(x)$ is twice continuously differentiable in a neighbourhood of $x = v_0$. (C2) easily follows from the fact that $p_1 < p_2$, whereas conditions (C4)–(C6) follow from the rate of convergence and the modulus of continuity of the empirical distribution function. It remains to show the validity of (C3). For a fixed sample X_1, \dots, X_n and for v such that $L(\mu_0, v) > 0$, we can write

$$\ell(\mu_0, v) = 2 \sum_{i=1}^n \log \left\{ 1 + \sum_{j=1}^2 \lambda_j(\mu_0, v) g_j(X_i, \mu_0, v) \right\},$$

where (for $j = 1, 2$)

$$\sum_{i=1}^n \frac{g_j(X_i, \mu_0, v)}{1 + \sum_{k=1}^2 \lambda_k(\mu_0, v) g_k(X_i, \mu_0, v)} = 0.$$

Note that \tilde{v} and v_0 are the maximizers of $\Gamma_n(v)$, and $\Gamma(v)$, respectively, where $\Gamma_n(v)$ and $\Gamma(v)$ are as in (A1) and (A2). Hence, to show condition (C3), we will check the conditions of theorem 5.7 in Van der Vaart (1998), i.e. we will show that

$$\sup_v |\Gamma_n(v) - \Gamma(v)| \xrightarrow{P} 0, \tag{10}$$

$$\sup_{|v-v_0| > \epsilon} \Gamma(v) < \Gamma(v_0) \tag{11}$$

for all $\epsilon > 0$. Condition (11) is ensured by the fact that v_0 is assumed to be unique together with equation (A5), whereas for (10) we write

$$\Gamma_n(v) - \Gamma(v) = \left[-n^{-1} \sum_{i=1}^n \log(1 + \lambda(\mu_0, v)^t g(X_i, \mu_0, v)) + E\{\log(1 + \lambda(\mu_0, v)^t g(X, \mu_0, v))\} \right] + [-E\{\log(1 + \lambda(\mu_0, v)^t g(X, \mu_0, v))\} + E\{\log(1 + \zeta(\mu_0, v)^t g(X, \mu_0, v))\}].$$

The second term above is easily seen to be $o_P(1)$ by using standard arguments concerning parametric Z -estimators (see, e.g. Van der Vaart, 1998, p. 41, for the notion of Z -estimators). To show that the first term goes to zero uniformly in v , we will prove that the class

$$\mathcal{F} = \{x \rightarrow \log(1 + \eta^t g(x, \mu_0, v)) : \eta \in R, v \in \mathbb{R}\}$$

is Glivenko–Cantelli, where $R \subset \mathbb{R}^2$ is such that $(\lambda_1(\mu_0, v), \lambda_2(\mu_0, v))^t$ belongs to R for all v and for all samples for which $L(\mu_0, v) > 0$. Note that R can be taken compact. This is because $1 + \lambda(\mu_0, v)^t g(X_i, \mu_0, v)$ is strictly positive for all i and all v , and hence $(\lambda_1(\mu_0, v), \lambda_2(\mu_0, v))^t$ needs to satisfy the constraints

$$\begin{cases} 1 + \lambda_1(\mu_0, v)(1 - p_1) + \lambda_2(\mu_0, v)(1 - p_2) > 0, \\ 1 - \lambda_1(\mu_0, v)p_1 + \lambda_2(\mu_0, v)(1 - p_2) > 0, \\ 1 - \lambda_1(\mu_0, v)p_1 - \lambda_2(\mu_0, v)p_2 > 0. \end{cases}$$

The intersection of these three half planes is a triangle, and hence it is compact.

The Glivenko–Cantelli property of the class \mathcal{F} can now be easily shown by using theorem 2.7.5 in Van der Vaart & Wellner (1996), together with the monotonicity (in x) of the function $g_j(x, \mu_0, v)$ ($j=1, 2$). This shows that condition (C3) is satisfied.

As a consequence of theorem 1, we can now construct an EL confidence region for μ_0 or test hypotheses concerning the value of μ_0 .

3.2. Copula functions

Take $d=2$ and let $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$ be a random sample from $X=(X_1, X_2)$ with unknown bivariate distribution function $H(x_1, x_2)=P(X_1 \leq x_1, X_2 \leq x_2)$. According to Sklar’s theorem, see, e.g. Nelsen (1999), there exists a copula function C that links the bivariate H to the marginals F_1 of X_1 and F_2 of X_2 via the formula $H(x_1, x_2)=C(F_1(x_1), F_2(x_2))$. The copula function C is itself a bivariate distribution on the unit square with uniform marginals. Moreover, if F_1 and F_2 are continuous, C is unique and given by $C(u_1, u_2)=H(F_1^{-1}(u_1), F_2^{-1}(u_2))$, $0 \leq u_1, u_2 \leq 1$. We want to do inference on the value of the unknown copula function C at fixed $0 \leq u_1, u_2 \leq 1$. Let

$$\mu_0 = C(u_1, u_2).$$

Introducing the parameters $v_{01}=F_1^{-1}(u_1)$ and $v_{02}=F_2^{-1}(u_2)$, we have the equations

$$\begin{cases} H(v_{01}, v_{02}) - \mu_0 = 0 \\ F_1(v_{01}) - u_1 = 0 \\ F_2(v_{02}) - u_2 = 0. \end{cases}$$

This is of the form (1) with

$$p=1, \quad q=2, \quad g_1(X, \mu, v)=I(X_1 \leq v_1, X_2 \leq v_2) - \mu, \\ g_2(X, \mu, v)=I(X_1 \leq v_1) - u_1 \quad \text{and} \quad g_3(X, \mu, v)=I(X_2 \leq v_2) - u_2.$$

This example has been studied using a smoothed EL by Chen *et al.* (2009).

The verification of conditions (C0)–(C6) can be carried out in much the same way as in the previous example. Note that we now have three estimating equations instead of two, and some arguments (especially the geometric arguments) are therefore somewhat more technical than in the previous example. The main reasoning is, however, the same and details are therefore omitted.

In the last two examples we consider the context of two samples. The verification of the conditions is very analogous to the first example, as the criterion functions are again based on indicators. The same method of proof as in the first example can therefore be followed.

3.3. Difference of quantiles in the two sample problem

For independent random samples X_1, \dots, X_{n_1} from X with distribution function F_1 and Y_1, \dots, Y_{n_2} from Y with distribution function F_2 , we consider

$$\mu_0 = F_2^{-1}(t) - F_1^{-1}(t),$$

where $0 < t < 1$. Introducing the parameter $v_0 = F_1^{-1}(t)$, we have the equations

$$\begin{cases} F_1(v_0) - t = 0 \\ F_2(\mu_0 + v_0) - t = 0. \end{cases}$$

This is of the form (9) with $g(X, \mu, v)=I(X \leq v) - t$ and $h(Y, \mu, v)=I(Y \leq \mu + v) - t$.

3.4. ROC curves

In the situation of section 3.3, we consider

$$\mu_0 = 1 - F_1(F_2^{-1}(1 - t)),$$

where $0 < t < 1$, which is the receiver operating characteristic (ROC) curve, evaluated in the point t . For a nice introduction on ROC curves, see, e.g. Pepe (2003). Introducing the parameter $v_0 = F_2^{-1}(1 - t)$ leads to the equations

$$\begin{cases} F_1(v_0) - (1 - \mu_0) = 0 \\ F_2(v_0) - (1 - t) = 0, \end{cases}$$

which is again of the form (9). A smoothed EL approach to this problem has been considered in Claeskens, *et al.* (2003) for completely observed data and in Molanes Lopez *et al.* (2008) (unpublished data) for censored and truncated data.

4. Simulations

For the sake of brevity, in this section, we only consider the example introduced in section 3.2, regarding copula functions. Through a simulation study, we check the behaviour of our EL method in this particular case and compare it with the smoothed EL approach proposed by Chen *et al.* (2009), hereafter abbreviated by CPZ.

A $100(1 - \alpha)\%$ confidence region for $\mu_0 = C(u_1, u_2)$ includes all those values of μ for which the null hypothesis $H_0 : C(u_1, u_2) = \mu$ cannot be rejected. According to remark 2, an approximate $100(1 - \alpha)\%$ confidence region for μ_0 is given by

$$I_{1-\alpha}(u_1, u_2) = \{ \mu : \ell(\mu) \leq \chi_{1, 1-\alpha}^2 \}. \tag{12}$$

We draw 1000 samples of size n from the mixture copula, $C(u_1, u_2; \gamma, \theta_1, \theta_2)$, given by:

$$C(u_1, u_2; \gamma, \theta_1, \theta_2) = \gamma \left\{ u_1^{-\theta_1} + u_2^{-\theta_1} - 1 \right\}^{1/\theta_1} + (1 - \gamma) \exp \{ -((-\log u_1)^{\theta_2} + (-\log u_2)^{\theta_2})^{1/\theta_2} \}, \tag{13}$$

where the marginals are standard normal distributions and the parameters θ_1, θ_2 and γ are such that $\theta_1 > 0, \theta_2 > 1$ and $\gamma \in [0, 1]$. We take the above mixture copula with $\theta_1 = 2$ and $\theta_2 = 3$, as previously considered by CPZ. When $\gamma = 1$ the mixture copula in (13) becomes a Clayton copula with parameter θ_1 and when $\gamma = 0$ it becomes a Gumbel–Hougaard copula with parameter θ_2 . The parameter γ denotes the mixing probability of these two copulas in the mixture.

The selection of observations from a given copula has been carried out based on a general approach, which is outlined in, e.g. Embrechts *et al.* (2003). This general method entails solving an equation which, in the particular case of a Gumbel–Hougaard copula, does not have an analytical solution. Although a numerical algorithm can in principle be used to solve this equation, this approach turns out to be very time consuming, given the large number of times the algorithm needs to be applied. For this reason, drawing from a Gumbel–Hougaard copula has been performed using an alternative algorithm proposed by Marshall & Olkin (1988), based on a mixture of powers.

In order to check the performance of our method, a Monte Carlo approximation of the coverage probability of (12) is obtained under different scenarios. For every trial, we first obtain the value of $\ell(\mu_0)$ by solving the optimization problem in (14)–(16):

$$\min_v \ell(\mu_0, v) \tag{14}$$

subject to

$$\sum_{i=1}^n \frac{g_j(X_i, \mu_0, v)}{1 + \sum_{k=1}^3 \lambda_k(\mu_0, v)g_k(X_i, \mu_0, v)} = 0, \quad \text{for } j = 1, 2, 3, \tag{15}$$

$$1 + \sum_{k=1}^3 \lambda_k(\mu_0, v)g_k(X_i, \mu_0, v) > 0, \quad \text{for } i = 1, \dots, n, \tag{16}$$

where $g_k(X_i, \mu_0, v)$ ($k = 1, 2, 3$ and $i = 1, \dots, n$) are given in section 3.2. Once we obtain $\ell(\mu_0)$ we check if μ_0 falls in the confidence region given in (12) by checking whether $\ell(\mu_0) \leq \chi_{1,1-\alpha}^2$. The proportion of times that μ_0 falls in (12) gives us a Monte Carlo approximation of the coverage probability of (12).

Note that the constraints in (16) must be imposed to exclude any $\lambda(\mu_0, v) = (\lambda_1(\mu_0, v), \lambda_2(\mu_0, v), \lambda_3(\mu_0, v))$ for which some $p_i(\mu_0, v) \leq 0$ [(see (3)]. Following the ideas presented in Owen (2001, p. 62 and 235), the constrained optimization problem in (14)–(16) is equivalent to

$$\min_v \ell(\mu_0, v), \tag{17}$$

subject to

$$\sum_{i=1}^n \log_*^{(1)} \left\{ 1 + \sum_{k=1}^3 \lambda_k(\mu_0, v)g_k(X_i, \mu_0, v) \right\} g_j(X_i, \mu_0, v) = 0, \quad \text{for } j = 1, 2, 3, \tag{18}$$

where

Table 1. Empirical coverage probabilities for the empirical likelihood-based confidence region in (12) with sample sizes $n=200, 400$ and points (u_1, u_2) on the unit square diagonal. The first line corresponds to the new method, the second line to the smoothed empirical likelihood method of Chen et al. (2009)

(γ, u_1, u_2)	$n=200$		$n=400$	
	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$
(0.0, 0.25, 0.25)	0.893	0.940	0.886	0.944
	0.923	0.957	0.902	0.944
(0.0, 0.50, 0.50)	0.906	0.954	0.891	0.954
	0.889	0.940	0.896	0.948
(0.0, 0.75, 0.75)	0.844	0.934	0.862	0.936
	0.897	0.951	0.890	0.939
(0.5, 0.25, 0.25)	0.900	0.930	0.889	0.940
	0.919	0.970	0.895	0.949
(0.5, 0.50, 0.50)	0.907	0.943	0.900	0.936
	0.908	0.961	0.846	0.908
(0.5, 0.75, 0.75)	0.897	0.952	0.878	0.942
	0.876	0.932	0.869	0.924
(1.0, 0.25, 0.25)	0.870	0.915	0.894	0.935
	0.922	0.964	0.901	0.954
(1.0, 0.50, 0.50)	0.877	0.931	0.885	0.938
	0.898	0.952	0.893	0.948
(1.0, 0.75, 0.75)	0.902	0.952	0.858	0.919
	0.825	0.910	0.762	0.850

$$\log_*(z) = \begin{cases} \log(z), & \text{if } z \geq \frac{1}{n}, \\ \log(\frac{1}{n}) - 1.5 + 2nz - \frac{(nz)^2}{2}, & \text{if } z \leq \frac{1}{n}, \end{cases}$$

and $\log_*^{(1)}(z) = (\partial/\partial z)(\log_*(z))$. With this new formulation of the problem, the inequality constraints in (16) have been ruled out.

Table 1 shows the coverage probabilities of our method and those reported in CPZ, which makes our results directly comparable with theirs. The expected standard deviation of each number in the table due to Monte Carlo is in general less than 0.01. The bandwidth parameter in their method is selected using a cross validation procedure (see their paper for a detailed description of the procedure). From this comparison, we conclude that the behaviour of our method, which has the advantage of avoiding a bandwidth selection problem, is in general at least as good as the behaviour of the smoothed EL approach of CPZ. A closer inspection of the table reveals that CPZ’s method has a coverage probability that is in general closer to the target than our method, but, on the other hand, CPZ’s method is slightly more variable. Moreover, as CPZ indicate, the bandwidth in their procedure has a non-negligible impact on the coverage probability, and the choice of the optimal bandwidth in terms of coverage probability remains an open problem, both theoretically and practically. Our method on the contrary does not depend on a bandwidth, and hence it does not share this drawback.

In Table 2 we show the coverage probabilities obtained with our method for other sample sizes, other values for the parameter γ in (13) and for points (u_1, u_2) falling outside of the unit square diagonal. CPZ did not consider this setting in their simulation study. However, from the experience we gained during the implementation of our methodology, we observed that achieving convergence is more challenging when dealing with non-diagonal points. The table shows that the empirical coverage probabilities are close to their nominal values and that the results improve when the sample size increases.

Note that the fact that we do not introduce smoothing and avoid a delicate bandwidth selection problem, inherent to CPZ’s method, entails, on the other hand, that our methodology is more complex to programme, because no derivatives can be taken. We specially

Table 2. Empirical coverage probabilities for the empirical likelihood-based confidence region in (12) with sample sizes $n=200, 300$ and points (u_1, u_2) outside the unit square diagonal and such that $|u_1 - u_2| = 0.10, 0.20, 0.30$

(γ, u_1, u_2)	$n = 200$		$n = 300$	
	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$
(0.25, 0.30, 0.40)	0.909	0.966	0.895	0.948
(0.50, 0.30, 0.40)	0.926	0.948	0.915	0.958
(0.75, 0.30, 0.40)	0.886	0.948	0.900	0.948
(0.25, 0.40, 0.50)	0.895	0.949	0.894	0.947
(0.50, 0.40, 0.50)	0.907	0.939	0.914	0.959
(0.75, 0.40, 0.50)	0.879	0.945	0.865	0.949
(0.25, 0.30, 0.50)	0.901	0.947	0.901	0.963
(0.50, 0.30, 0.50)	0.898	0.957	0.864	0.946
(0.75, 0.30, 0.50)	0.895	0.923	0.879	0.923
(0.25, 0.40, 0.60)	0.856	0.958	0.870	0.944
(0.50, 0.40, 0.60)	0.884	0.958	0.907	0.947
(0.75, 0.40, 0.60)	0.869	0.949	0.878	0.942
(0.25, 0.30, 0.60)	0.935	0.942	0.811	0.971
(0.50, 0.30, 0.60)	0.958	0.962	0.918	0.922
(0.75, 0.30, 0.60)	0.830	0.955	0.874	0.954
(0.25, 0.40, 0.70)	0.923	0.952	0.888	0.964
(0.50, 0.40, 0.70)	0.841	0.965	0.896	0.956
(0.75, 0.40, 0.70)	0.895	0.922	0.879	0.951

found more difficulties for our method to achieve convergence when non-diagonal points are considered. This is the reason why in the implementation of our methodology we finally used a combination of numerical algorithms as explained below.

For every trial we first try to solve the optimization problem specified in (17) and (18) by using a MATLAB function for nonlinear constrained optimization problems (*fmincon*). If this algorithm fails to find the solution to (17) and (18), we then use a modification of a basic generating set search (GSS) algorithm for unconstrained optimization, proposed by Frimannslund & Steihaug (2007). This algorithm is derivative-free and incorporates curvature information about the objective function as the search progresses. If this second algorithm does not converge either, then we use a crude grid search. We evaluate (17) at an equally spaced grid of points, $v=(v_1, v_2)$, placed around $(F_{1n}^{-1}(u_1), F_{2n}^{-1}(u_2))$, where F_{jn}^{-1} denotes the empirical quantile function of F_j for $j=1, 2$. Note that for the implementation of these two last algorithms, the evaluation of the objective function in (17) at a given v goes through previously finding the solution, $\lambda(\mu_0, v)$, to the nonlinear system of equations defined in (18). The MATLAB function *fsolve* has been used to solve these nonlinear equations.

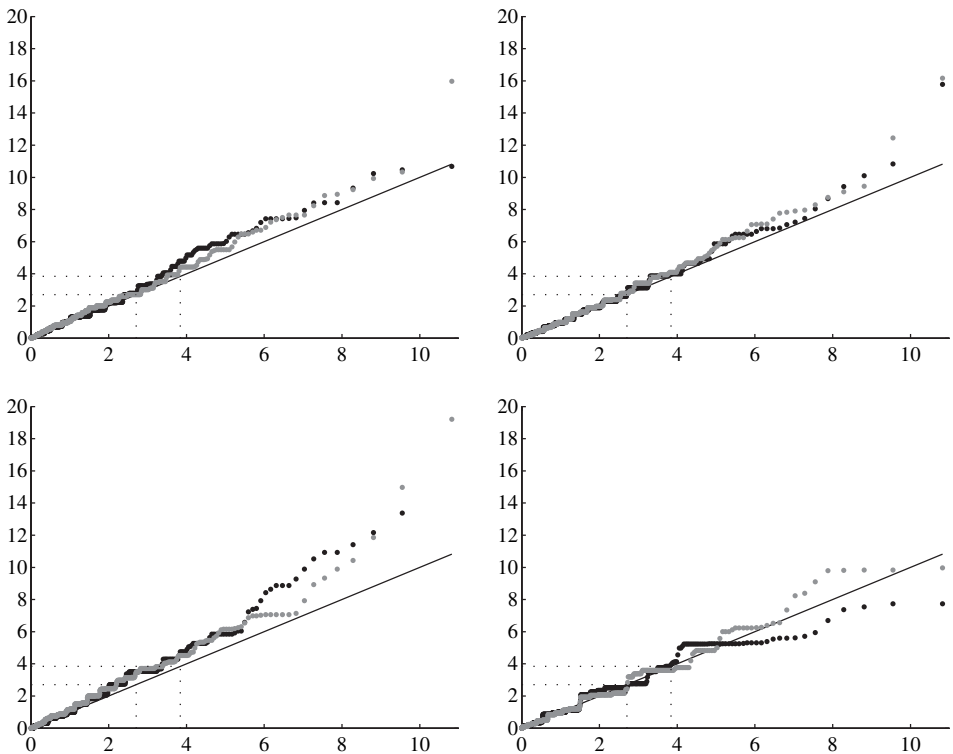


Fig. 1. A QQ plot with the c quantiles of the χ^2_1 distribution ($c \in \{0.001, 0.002, \dots, 0.999\}$) plotted against the corresponding sample quantiles obtained from 1000 values drawn from $\ell(\mu_0)$ under different settings: for $(\gamma, u_1, u_2) = (1, 0.5, 0.5)$, with $n=200$ (dark points) and $n=400$ (grey points) (left-top panel), for $(\gamma, u_1, u_2) = (0.5, 0.25, 0.25)$, with $n=200$ (dark points) and $n=400$ (grey points) (right-top panel), for $(\gamma, u_1, u_2) = (0, 0.75, 0.75)$, with $n=200$ (dark points) and $n=400$ (grey points) (left-bottom panel) and for $(\gamma, u_1, u_2) = (0.25, 0.30, 0.50)$, with $n=200$ (dark points) and $n=300$ (grey points) (right-bottom panel). The 45° line and two reference points corresponding to nominal confidence levels of 90% and 95% are also included in each plot.

In order to empirically examine the accuracy of the χ_1^2 approximation to the distribution of the log-likelihood function at the true parameter, $\ell(\mu_0)$, we draw QQ plots that compare the c quantiles of the χ_1^2 distribution for $c \in \{0.001, 0.002, \dots, 0.999\}$, with the corresponding sample quantiles obtained from 1000 values drawn from $\ell(\mu_0)$ (see Fig. 1 where different scenarios are considered). From these QQ plots we can conclude that, except for extreme upper quantiles, the distribution of $\ell(\mu_0)$ is reasonably well approximated by a χ_1^2 distribution.

Acknowledgements

The authors are grateful to two reviewers for helpful comments, that substantially improved an earlier version of this paper. The first author acknowledges support from NS project MTM2005-00429 (EU ERDF support included). The second and third authors acknowledge support from IAP research network No. P6/03 of the Belgian government (Belgian Science Policy). The research of the second author is in addition financially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC Grant agreement No. 203650.

References

- Apostol, T. (1974). *Mathematical analysis*. Addison-Wesley, Reading, MA.
- Chen, S. X. & Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21**, 1166–1181.
- Chen, J., Peng, L. & Zhao, Y. (2009). Empirical likelihood based confidence intervals for copulas. *J. Multivariate Anal.* **100**, 137–151.
- Claeskens, G., Jing, B. Y., Peng, L. & Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *Canad. J. Statist.* **31**, 173–190.
- Embrechts, P., Lindskog, F. & McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of heavy tailed distributions in finance* (ed. S. Rachev), 329–384. Elsevier, UK.
- Frimannslund, L. & Steihaug, T. (2007). A generating set search method using curvature information. *Comput. Optim. Appl.* **38**, 105–121.
- Hjort, N. L., McKeague, I. W. & Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37**, 1079–1115.
- Marshall, A. W. & Olkin, I. (1988). Families of multivariate distributions. *J. Amer. Statist. Assoc.* **83**, 834–841.
- Nelsen, R. B. (1999). *An introduction to copulas*. Springer, New York.
- Owen, A. (2001). *Empirical likelihood*. Chapman & Hall, New York.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, New York.
- Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–325.
- Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123–137.
- Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.
- Van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York.
- Zhou, W. & Jing, B.-Y. (2003). Smoothed empirical likelihood confidence intervals for the difference of quantiles. *Statist. Sinica* **13**, 83–95.

Received March 2008, in final form December 2008

Ingrid Van Keilegom, Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium.
E-mail: ingrid.vankeilegom@uclouvain.be

Appendix: Proofs

Throughout this appendix we will use the abbreviated notation $\lambda_j(v) = \lambda_j(\mu_0, v)$ ($j = 1, \dots, p + q$). We will give here the proofs of the main theorem and of several lemmas. As these will rely on theorems 1 and 2 of Sherman (1993), it is convenient to introduce some extra notation in order to bring our situation into theirs. As we know from condition (C3) that \tilde{v} converges in probability to v_0 , we can restrict our attention to what follows to an $o(1)$ neighbourhood of v_0 . In that case, the empirical log-likelihood ratio can be written in the form (A4) for n large, and we therefore define:

$$\Gamma_n(v) = -n^{-1} \sum_{i=1}^n \log(1 + \lambda(v)' g(X_i, \mu_0, v)) \tag{A1}$$

$$\Gamma(v) = -E\{\log(1 + \zeta(v)' g(X, \mu_0, v))\}, \tag{A2}$$

where $\lambda(v)$ is defined in (5) and $\zeta(v) = (\zeta_1(v), \dots, \zeta_{p+q}(v))$ satisfies

$$E \left\{ \frac{g(X, \mu_0, v)}{1 + \zeta(v)' g(X, \mu_0, v)} \right\} = 0. \tag{A3}$$

Note that $\zeta(v)$ exists, and is non-stochastic, unique and continuously differentiable for v in a neighbourhood of v_0 . This follows from the implicit function theorem (see, e.g. theorem 13.7, p. 374 in Apostol, 1974), together with condition (C1). In addition, note that $\ell(\mu_0, v) = -2n\Gamma_n(v)$, where $\ell(\mu_0, v)$ is defined in (4).

We start with a preliminary lemma concerning \tilde{v} and v_0 .

Lemma 1

Under (C0)–(C2), and with \tilde{v} defined in (6), we have

$$\tilde{v} = \arg \max_v \Gamma_n(v), \tag{A4}$$

$$v_0 = \arg \max_v \Gamma(v). \tag{A5}$$

Proof. Equation (A4) follows from the fact that

$$\max_v \Gamma_n(v) = -\frac{1}{2} n^{-1} \min_v \ell(\mu_0, v) = -\frac{1}{2} n^{-1} \ell(\mu_0, \tilde{v}) = \Gamma_n(\tilde{v}).$$

For (A5), note that $\Gamma(v_0) = 0$ as $\zeta(v_0) = 0$, and that for any $v \neq v_0$,

$$\Gamma(v) = -\zeta(v)' E \left\{ \frac{g(X, \mu_0, v)}{1 + \zeta(v)' g(X, \mu_0, v)} \right\} - \frac{1}{2} E \left\{ \frac{(\zeta(v)' g(X, \mu_0, v))^2}{(1 + \alpha(v)' g(X, \mu_0, v))^2} \right\},$$

for some $\alpha(v)$ on the line segment between 0 and $\zeta(v)$. The first term above equals 0, whereas the second one is strictly negative. Hence v_0 is a maximizer of $\Gamma(v)$.

Lemma 2

Under (C0)–(C2), (C4), (C5), we have

$$\lambda(v) = O_P(n^{-1/2}) + O_P(\|v - v_0\|), \tag{A6}$$

$$\lambda(v) - \zeta(v) = O_P(n^{-1/2}) \tag{A7}$$

uniformly for all v in an $o(1)$ neighbourhood of v_0 , and

$$\lambda(v) = V_{11}^{-1}(v)n^{-1} \sum_{i=1}^n g(X_i, \mu_0, v) + o_P(n^{-1/2}), \tag{A8}$$

uniformly for all v in an $O(n^{-1/2})$ neighbourhood of v_0 , where

$$V_{11}(v) = (E\{g_j(X, \mu_0, v)g_k(X, \mu_0, v)\})_{j,k=1,\dots,p+q}.$$

Proof. First note that $V_{11}(v)$ is positive definite for v in a neighbourhood of v_0 because of conditions (C1) and (C2). The proof of (A6) and (A8) follows along the same lines as the proof of, e.g. theorem 3.2 in Owen (2001, p. 219). The proof of (A7) follows using standard arguments concerning parametric Z estimators (see, e.g. Van der Vaart, 1998, p. 41, for the notion of Z -estimators).

Lemma 3

Under (C0)–(C2), there exists a neighbourhood N of v_0 and a constant $K > 0$ for which

$$\Gamma(v) \leq -K\|v - v_0\|^2$$

for all $v \in N$, where $\|\cdot\|$ is the Euclidean norm.

Proof. From the proof of lemma 1, we have

$$\Gamma(v) = -\frac{1}{2} E \left\{ \frac{(\zeta(v)^t g(X, \mu_0, v))^2}{(1 + \alpha(v)^t g(X, \mu_0, v))^2} \right\},$$

with $\alpha(v)$ on the segment between 0 and $\zeta(v)$. As from (A3) it follows that $\zeta(v)$ is continuously differentiable, we have that

$$\zeta(v) = \zeta(v_0) + \zeta'(v^*)(v - v_0) = \zeta'(v^*)(v - v_0),$$

with v^* between v_0 and v . Hence,

$$\begin{aligned} (\zeta(v)^t g(X, \mu_0, v))^2 &= (v - v_0)^t \zeta'(v^*)^t g(X, \mu_0, v) g^t(X, \mu_0, v) \zeta'(v^*)(v - v_0) \quad \text{and} \\ \Gamma(v) &= \frac{1}{2} (v - v_0)^t F(v, v^*)(v - v_0), \end{aligned}$$

where

$$F(v, v^*) = -\zeta'(v^*)^t E \left\{ \frac{g(X, \mu_0, v) g^t(X, \mu_0, v)}{(1 + \alpha(v)^t g(X, \mu_0, v))^2} \right\} \zeta'(v^*).$$

From condition (C1) it follows that $F(v, v^*)$ is uniformly continuous in a compact neighbourhood of v_0 and hence

$$\Gamma(v) \leq \frac{1}{2} (v - v_0)^t F(v_0, v_0)(v - v_0) + \delta\|v - v_0\|^2$$

for any $\delta > 0$. As $F(v_0, v_0) = -\zeta'(v_0)^t E\{g(X, \mu_0, v_0)g^t(X, \mu_0, v_0)\}\zeta'(v_0)$, we have by (C2) that the inner matrix is positive and hence that $F(v_0, v_0)$ is negative definite. This means that it has the following representation: $F(v_0, v_0) = P^{-1}\Lambda P$, where $P^{-1} = P^t$ and Λ is a diagonal matrix with the eigenvalues on the diagonal. These eigenvalues are strictly negative. Then,

$$\begin{aligned} (v - v_0)^t F(v_0)(v - v_0) &= (v - v_0)^t P^{-1}\Lambda P(v - v_0) = (P(v - v_0))^t \Lambda (P(v - v_0)) \\ &\leq -K\|P(v - v_0)\|^2 = -K(v - v_0)^t P^t P(v - v_0) = -K\|v - v_0\|^2, \end{aligned}$$

with $-K$ equal to the eigenvalue with the largest absolute value.

Lemma 4

Under (C0), (C1), (C4), (C5), we have

$$\Gamma_n(v) = \Gamma(v) + O_P(n^{-1/2} \|v - v_0\|) + o_P(\|v - v_0\|^2) + O_P(n^{-1})$$

uniformly in v , for $v - v_0 = o(1)$.

Proof. First note that

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \{ \log(1 + \lambda(v)^t g(X_i, \mu_0, v)) - \log(1 + \zeta(v)^t g(X_i, \mu_0, v)) \} \\ &= \{ \lambda(v) - \zeta(v) \}^t n^{-1} \sum_{i=1}^n \left\{ \frac{g(X_i, \mu_0, v)}{1 + \zeta(v)^t g(X_i, \mu_0, v)} \right\} \\ & \quad + \frac{1}{2} \{ \lambda(v) - \zeta(v) \}^t n^{-1} \sum_{i=1}^n \left\{ \frac{g(X_i, \mu_0, v) g'(X_i, \mu_0, v)}{(1 + \eta(v)^t g(X_i, \mu_0, v))^2} \right\} \{ \lambda(v) - \zeta(v) \}, \end{aligned}$$

for some $\eta(v)$ on the line segment between $\lambda(v)$ and $\zeta(v)$. The first term above is $O_P(n^{-1})$ by equations (A3) and (A7), while the second one is $O_P(n^{-1})$ by lemma 2 and the boundedness of the function g .

Hence, it suffices to calculate the order of

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \log(1 + \zeta(v)^t g(X_i, \mu_0, v)) - E \{ \log(1 + \zeta(v)^t g(X, \mu_0, v)) \} \\ &= n^{-1} \sum_{i=1}^n [\zeta(v)^t g(X_i, \mu_0, v) - E \{ \zeta(v)^t g(X, \mu_0, v) \}] \\ & \quad - \frac{1}{2} n^{-1} \sum_{i=1}^n [(\zeta(v)^t g(X_i, \mu_0, v))^2 - E \{ (\zeta(v)^t g(X, \mu_0, v))^2 \}] \\ & \quad + \frac{1}{3} n^{-1} \sum_{i=1}^n \left[\frac{(\zeta(v)^t g(X_i, \mu_0, v))^3}{(1 + \eta_{1i})^3} - E \left\{ \frac{(\zeta(v)^t g(X, \mu_0, v))^3}{(1 + \eta_2)^3} \right\} \right] \\ &= T_1 + T_2 + T_3, \end{aligned}$$

where

$$|\eta_{1i}| \leq | \zeta(v)^t g(X_i, \mu_0, v) | \leq M \sum_{j=1}^{p+q} | \zeta_j(v) |$$

and similarly for η_2 , and where

$$M = \sup_{j,x,v} | g_j(x, \mu_0, v) | < \infty$$

by condition (C1). For T_1 we have

$$\begin{aligned} |T_1| &\leq \sum_{j=1}^{p+q} | \zeta_j(v) | \left| n^{-1} \sum_{i=1}^n [g_j(X_i, \mu_0, v) - E \{ g_j(X, \mu_0, v) \}] \right| \\ &= O_P \left(n^{-1/2} \sum_{j=1}^{p+q} | \zeta_j(v) | \right) = O_P(n^{-1/2} \|v - v_0\|) + O_P(n^{-1}) \end{aligned}$$

because of condition (C4) and lemma 2. Further,

$$\begin{aligned}
 |T_2| &\leq \frac{1}{2} \sum_{j,k=1}^{p+q} |\xi_j(v)\xi_k(v)| \left| n^{-1} \sum_{i=1}^n [g_j(X_i, \mu_0, v)g_k(X_i, \mu_0, v) - E\{g_j(X, \mu_0, v)g_k(X, \mu_0, v)\}] \right| \\
 &= o_P\left(\left(\sum_{j=1}^{p+q} |\xi_j(v)|\right)^2\right) = o_P(\|v - v_0\|^2) + o_P(n^{-1})
 \end{aligned}$$

by using (C5). Similarly,

$$\begin{aligned}
 |T_3| &\leq \frac{1}{3} \sum_{j,k,\ell=1}^{p+q} |\xi_j(v)\xi_k(v)\xi_\ell(v)| \left| n^{-1} \sum_{i=1}^n \left[\frac{g_j(X_i, \mu_0, v)g_k(X_i, \mu_0, v)g_\ell(X_i, \mu_0, v)}{(1 + \eta_{1i})^3} \right. \right. \\
 &\quad \left. \left. - E\left\{ \frac{g_j(X, \mu_0, v)g_k(X, \mu_0, v)g_\ell(X, \mu_0, v)}{(1 + \eta_2)^3} \right\} \right] \right|.
 \end{aligned}$$

By (C1) and as $|\eta_{1i}| \leq M \|\xi(v)\|$ and $|\eta_2| \leq M \|\xi(v)\|$, we have that, for each sequence of random variables $\{r_n\}$ with $r_n = o(1)$:

$$\begin{aligned}
 \sup_{\|v-v_0\| \leq r_n/M} |T_3| &\leq (\text{constant}) \left(\sum_{j=1}^{p+q} |\xi_j(v)|\right)^3 \frac{1}{1 - r_n} \\
 &= O_P(\|v - v_0\|^3 + n^{-3/2}) \frac{1}{1 - r_n} = o_P(\|v - v_0\|^2) + O_P(n^{-1}).
 \end{aligned}$$

Lemma 5

Under (C0), (C1), (C5) and (C6) we have

$$\Gamma_n(v) = \frac{1}{2}(\theta - \theta_0)' V(\theta - \theta_0) - n^{-1/2}(\theta - \theta_0)' W_n + o_P(n^{-1}) \tag{A9}$$

uniformly in v , for $v - v_0 = O(n^{-1/2})$, where $\theta = (\lambda(v), v)'$, $\theta_0 = (0, v_0)'$,

$$W_n = \left(n^{-1/2} \sum_{i=1}^n g_1(X_i, \mu_0, v_0), \dots, n^{-1/2} \sum_{i=1}^n g_{p+q}(X_i, \mu_0, v_0), 0_q \right), \tag{A10}$$

and where 0_q is a vector of q zeros.

Proof. Throughout the proof we will use the notation $\theta_1 = \lambda(v)$, $\theta_2 = v$, $\theta_{01} = 0$ and $\theta_{02} = v_0$. Taylor expansion as in lemma 4 gives

$$\begin{aligned}
 \Gamma_n(v) &= -n^{-1} \sum_{i=1}^n \log(1 + \theta_1' g(X_i, \mu_0, \theta_2)) \\
 &= -n^{-1} \sum_{i=1}^n \theta_1' g(X_i, \mu_0, \theta_2) + \frac{1}{2} n^{-1} \sum_{i=1}^n (\theta_1' g(X_i, \mu_0, \theta_2))^2 - \frac{1}{3} n^{-1} \sum_{i=1}^n \frac{(\theta_1' g(X_i, \mu_0, \theta_2))^3}{(1 + \xi_{1i})^3} \\
 &= S_1 + S_2 + S_3,
 \end{aligned}$$

where

$$|\xi_{1i}| \leq M \sum_{j=1}^{p+q} |\theta_{1j}|.$$

As in the proof of lemma 4 we have that $S_3 = O_P(\|\theta - \theta_0\|^3) = o_P(n^{-1})$, as $\lambda(v) = O_P(n^{-1/2})$ by lemma 2. Next consider S_1 . We write

$$S_1 = -n^{-1} \sum_{i=1}^n \sum_{j=1}^{p+q} g_j(X_i, \mu_0, \theta_{02}) \theta_{1j} - n^{-1} \sum_{i=1}^n \sum_{j=1}^{p+q} [g_j(X_i, \mu_0, \theta_2) - g_j(X_i, \mu_0, \theta_{02})] \theta_{1j} = S_{11} + S_{12}.$$

Note that $S_{11} = -n^{-1/2}(\theta - \theta_0)' W_n$. For S_{12} we write

$$S_{12} = -n^{-1} \sum_{j=1}^{p+q} \sum_{i=1}^n [g_j(X_i, \mu_0, \theta_2) - E\{g_j(X, \mu_0, \theta_2)\} - g_j(X_i, \mu_0, \theta_{02}) + E\{g_j(X, \mu_0, \theta_{02})\}] \theta_{1j} - \sum_{j=1}^{p+q} [E\{g_j(X, \mu_0, \theta_2)\} - E\{g_j(X, \mu_0, \theta_{02})\}] \theta_{1j}. \tag{A11}$$

From (C6) we have that the first term in (A11) is

$$o_P\left(n^{-1/2} \sum_{j=1}^{p+q} |\theta_{1j}|\right) = o_P(n^{-1/2} \|\theta - \theta_0\|) = o_P(n^{-1}).$$

For the second term in (A11), we have, using (C1), that it is equal to

$$-\sum_{j=1}^{p+q} \frac{\partial}{\partial \theta_2} E\{g_j(X, \mu_0, \theta_2)\} \Big|_{\theta_2 = \theta_{02}} (\theta_2 - \theta_{02}) \theta_{1j} + o_P(n^{-1}),$$

where

$$\frac{\partial}{\partial \theta_2} E\{g_j(X, \mu_0, \theta_2)\} \Big|_{\theta_2 = \theta_{02}}$$

is the vector with elements

$$\frac{\partial}{\partial \theta_{2k}} E\{g_j(X, \mu_0, \theta_2)\} \Big|_{\theta = \theta_{02}}.$$

Hence, $S_{12} = (\theta_1 - \theta_{01})' V_{12}(\theta_2 - \theta_{02}) + o_P(n^{-1})$. Now we deal with S_2 .

$$S_2 = \frac{1}{2} n^{-1} \sum_{i=1}^n \sum_{j,k=1}^{p+q} g_j(X_i, \mu_0, \theta_{02}) g_k(X_i, \mu_0, \theta_{02}) \theta_{1j} \theta_{1k} + \frac{1}{2} n^{-1} \sum_{i=1}^n \sum_{j,k=1}^{p+q} [g_j(X_i, \mu_0, \theta_2) g_k(X_i, \mu_0, \theta_2) - g_j(X_i, \mu_0, \theta_{02}) g_k(X_i, \mu_0, \theta_{02})] \theta_{1j} \theta_{1k} = S_{21} + S_{22}.$$

For S_{21} we write

$$S_{21} = \frac{1}{2} n^{-1} \sum_{j,k=1}^{p+q} \sum_{i=1}^n [g_j(X_i, \mu_0, \theta_{02}) g_k(X_i, \mu_0, \theta_{02}) - E\{g_j(X, \mu_0, \theta_{02}) g_k(X, \mu_0, \theta_{02})\}] \theta_{1j} \theta_{1k} + \frac{1}{2} \sum_{j,k=1}^{p+q} E\{g_j(X, \mu_0, \theta_{02}) g_k(X, \mu_0, \theta_{02})\} \theta_{1j} \theta_{1k}. \tag{A12}$$

From (C5), we have that the first term in (A12) is

$$o_P\left(\left(\sum_{j=1}^{p+q} |\theta_{1j}| \right)^2\right) = o_P(\|\theta - \theta_0\|^2) = o_P(n^{-1}).$$

The second term in (A12) is equal to $\frac{1}{2}(\theta_1 - \theta_{01})' V_{11}(\theta_1 - \theta_{01})$. The term S_{22} can be written as

$$\begin{aligned} S_{22} &= \frac{1}{2}n^{-1} \sum_{j,k=1}^{p+q} \sum_{i=1}^n [g_j(X_i, \mu_0, \theta_2)g_k(X_i, \mu_0, \theta_2) - E\{g_j(X, \mu_0, \theta_2)g_k(X, \mu_0, \theta_2)\} \\ &\quad - g_j(X_i, \mu_0, \theta_{02})g_k(X_i, \mu_0, \theta_{02}) + E\{g_j(X, \mu_0, \theta_{02})g_k(X, \mu_0, \theta_{02})\}]\theta_{1j}\theta_{1k} \\ &\quad + \frac{1}{2} \sum_{j,k=1}^{p+q} [E\{g_j(X, \mu_0, \theta_2)g_k(X, \mu_0, \theta_2)\} - E\{g_j(X, \mu_0, \theta_{02})g_k(X, \mu_0, \theta_{02})\}]\theta_{1j}\theta_{1k}. \end{aligned}$$

From (C5) it follows that the first term is $o_P(\|\theta - \theta_0\|^2) = o_P(n^{-1})$ and that the second term is $O_P(\|\theta - \theta_0\|^3) = o_P(n^{-1})$, using (C1). This shows (A9).

Lemma 6

Under (C0)–(C6), we have

$$\begin{aligned} \Gamma_n(v) &= -\frac{1}{2}(v - v_0)' V_{12}' V_{11}^{-1} V_{12}(v - v_0) + n^{-1/2}(v - v_0)' V_{12}' V_{11}^{-1} X_n \\ &\quad - \frac{1}{2}n^{-1} X_n' V_{11}^{-1} X_n + o_P(n^{-1}) \end{aligned} \tag{A13}$$

uniformly in v , for $v - v_0 = O(n^{-1/2})$, and

$$\tilde{v} - v_0 = n^{-1/2}(V_{12}' V_{11}^{-1} V_{12})^{-1} V_{12}' V_{11}^{-1} X_n + o_P(n^{-1/2}), \tag{A14}$$

where $W_n = (X_n, 0_q)'$, and V and W_n are given in (A8) and (A10) respectively.

Proof. We start with the first assertion. From lemma 2, together with conditions (C1) and (C6), we know that $\lambda(v) = V_{11}^{-1}[n^{-1/2}X_n - V_{12}(v - v_0)] + o_P(n^{-1/2})$ uniformly for all v in an $O(n^{-1/2})$ neighbourhood of v_0 . Hence, it follows from lemma 5 that

$$\begin{aligned} \Gamma_n(v) &= \frac{1}{2} \left[\lambda(v)' V_{11} \lambda(v) + 2(v - v_0)' V_{12}' \lambda(v) \right] - n^{-1/2}(\lambda(v))' (v - v_0)' W_n + o_P(n^{-1}) \\ &= \frac{1}{2} \left\{ \left[n^{-1/2} X_n' - (v - v_0)' V_{12}' \right] V_{11}^{-1} \left[n^{-1/2} X_n - V_{12}(v - v_0) \right] \right\} \\ &\quad + (v - v_0)' V_{12}' V_{11}^{-1} \left\{ n^{-1/2} X_n - V_{12}(v - v_0) \right\} \\ &\quad - n^{-1/2} \left\{ n^{-1/2} X_n' - (v - v_0)' V_{12}' \right\} V_{11}^{-1} X_n + o_P(n^{-1}) \\ &= -\frac{1}{2}(v - v_0)' V_{12}' V_{11}^{-1} V_{12}(v - v_0) + n^{-1/2}(v - v_0)' V_{12}' V_{11}^{-1} X_n \\ &\quad - \frac{1}{2}n^{-1} X_n' V_{11}^{-1} X_n + o_P(n^{-1}). \end{aligned} \tag{A15}$$

To show the second assertion of the lemma, we will apply theorems 1 and 2 in Sherman (1993). First note that condition (C3) and lemmas 1, 3 and 4 imply that $\tilde{v} - v_0 = O_P(n^{-1/2})$, by applying theorem 1 in Sherman (1993). Next, (A15) shows that the displayed condition (4) in the statement of theorem 2 in Sherman (1993) is satisfied, except for the term $-\frac{1}{2}n^{-1} X_n' V_{11}^{-1} X_n$,

which should not be there. However, careful inspection of the proof of this theorem reveals that the result remains valid when this extra term is present, as this term does not depend on v . It now follows from the proof of this theorem that

$$\tilde{v} - v_0 = n^{-1/2}(V'_{12}V^{-1}_{11}V_{12})^{-1}V'_{12}V^{-1}_{11}X_n + o_P(n^{-1/2}).$$

This shows the second statement of the lemma.

Proof of theorem 1. Without loss of generality, we condition on the event that $\|\tilde{v} - v_0\| \leq K$ for some $K > 0$. This is possible, as $\tilde{v} - v_0 = O_P(n^{-1/2})$ (see lemma 6). From lemma 6 it follows that, with $V_{22.1} = -V'_{12}V^{-1}_{11}V_{12}$,

$$\begin{aligned} \Gamma_n(\tilde{v}) &= -\frac{1}{2}n^{-1}X'_nV^{-1}_{11}V_{12}V_{22.1}^{-1}V'_{12}V^{-1}_{11}X_n - \frac{1}{2}n^{-1}X'_nV^{-1}_{11}X_n + o_P(n^{-1}) \\ &= -\frac{1}{2}n^{-1}X'_nV^{-1/2}DV^{-1/2}X_n + o_P(n^{-1}), \end{aligned}$$

where

$$D = V^{-1/2}_{11}\{I + V_{12}V_{22.1}^{-1}V'_{12}V^{-1}_{11}\}V^{1/2}_{11},$$

or equivalently,

$$\ell(\mu_0) = X'_nV^{-1/2}_{11}DV^{-1/2}_{11}X_n + o_P(1). \tag{A16}$$

Note that D can also be written as $D = V^{1/2}_{11}V^{11}V^{1/2}_{11}$, where

$$V^{-1} = \begin{pmatrix} V^{11} & V^{12} \\ (V^{12})' & V^{22} \end{pmatrix},$$

as it follows from lemma 3 in Qin & Lawless (1994) that

$$\begin{aligned} V^{-1} &= \begin{pmatrix} I & -V^{-1}_{11}V_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} V^{-1}_{11} & 0 \\ 0 & V^{-1}_{22.1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -V'_{12}V^{-1}_{11} & I \end{pmatrix} \\ &= \begin{pmatrix} V^{-1}_{11}(I + V_{12}V_{22.1}^{-1}V'_{12}V^{-1}_{11}) & -V^{-1}_{11}V_{12}V_{22.1}^{-1} \\ -V^{-1}_{22.1}V'_{12}V^{-1}_{11} & V^{-1}_{22.1} \end{pmatrix}. \end{aligned}$$

Moreover, note that

$$V^{-1/2}_{11}X_n = V^{-1/2}_{11}n^{-1/2} \sum_{i=1}^n g(X_i, \mu_0, v_0) \xrightarrow{d} N(0; I),$$

so that from (A16) it follows that

$$\ell(\mu_0) \xrightarrow{d} \chi^2_p,$$

provided we can show that

$$D \text{ is symmetric,} \tag{A17}$$

$$D \text{ is idempotent,} \tag{A18}$$

$$\text{tr}(D) = p, \tag{A19}$$

where $\text{tr}(D)$ is the trace of the matrix D . For (A17), we have that $D' = D$ as it is easily seen that $V'_{11} = V_{11}$ and $(V^{11})' = V^{11}$. For (A18), note that

$$DD = V_{11}^{1/2} V^{11} V_{11} V^{11} V_{11}^{1/2} = V_{11}^{1/2} V^{11} V_{11}^{1/2} = D,$$

as by direct calculation, it follows that

$$\begin{aligned} V^{11} V_{11} V^{11} &= V_{11}^{-1} (I + V_{12} V_{22.1}^{-1} V_{12}' V_{11}^{-1}) (I + V_{12} V_{22.1}^{-1} V_{12}' V_{11}^{-1}) \\ &= V_{11}^{-1} (I + V_{12} V_{22.1}^{-1} V_{12}' V_{11}^{-1}) = V^{11}. \end{aligned}$$

Finally, for (A19) we have:

$$\begin{aligned} \text{tr}(D) &= \text{tr}(V_{11}^{1/2} V^{11} V_{11}^{1/2}) = \text{tr}(V_{11} V^{11}) = \text{tr}(I + V_{12} V_{22.1}^{-1} V_{12}' V_{11}^{-1}) \\ &= \text{tr}(I_{(p+q) \times (p+q)}) + \text{tr}(V_{12}' V_{11}^{-1} V_{12} V_{22.1}^{-1}) \\ &= \text{tr}(I_{(p+q) \times (p+q)}) - \text{tr}(I_{q \times q}) = p + q - q = p. \end{aligned}$$