INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0812

EMPIRICAL LIKELIHOOD FOR NON-SMOOTH CRITERION FUNCTIONS

MOLANES LOPEZ, E.M., VAN KEILEGOM, I. and N. VERAVERBEKE

Empirical Likelihood for Non-Smooth Criterion Functions

Elisa M. Molanes Lopez[1] Ingrid Van Keilegom[2]
Noël Veraverbeke[3]

March 5, 2008

Abstract

Suppose that $X_1, \ldots, X_n$ is a sequence of independent random vectors, identically distributed as a d-dimensional random vector X. Let $\mu \in I\!R^p$ be a parameter of interest and $\nu \in I\!R^q$ 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 $g = (g_1, \ldots, g_{p+q})$ is a vector of $p + q$ functions. In this paper we develop an empirical likelihood 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 $g_1, \ldots, g_{p+q}$ 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 empirical likelihood method for small samples.

Key words: Confidence region; copulas; empirical likelihood; estimating equations; hypothesis testing; nuisance parameter; quantiles; ROC curve.

[1] Departamento de Estadística, Universidad Carlos III de Madrid, Avda. de la Universidad 30, 28911 Leganés, Spain, E-mail: emolanes@est-econ.uc3m.es

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

[3] Center for Statistics, Universiteit Hasselt, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium, E-mail: noel.veraverbeke@uhasselt.be

1 Introduction and general method

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

$$E\{g(X, \mu_0, \nu_0)\} = 0, \tag{1.1}$$

where $g = (g_1, \ldots, g_{p+q})$ is a vector of $p+q$ functions. In this paper we develop an empirical likelihood 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 $g_1, \ldots, g_{p+q}$ 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 and Lawless (1994) also consider the problem of developing empirical likelihood (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 and Lawless (1994) heavily use Taylor expansions in their proofs, for which smoothness of the functions $g_1, \ldots, g_{p+q}$ is indispensable.

Define for any $(\mu, \nu) \in I\!R^{p+q}$ the empirical likelihood:

$$L(\mu, \nu) = n^n \sup\Big\{ \prod_{i=1}^n p_i(\nu) : p_i(\nu) \geq 0, \sum_{i=1}^n p_i(\nu) = 1 \Big\}, \tag{1.2}$$

subject to the constraint

$$\sum_{i=1}^n p_i(\nu) g_j(X_i, \mu, \nu) = 0 \quad (j = 1, \ldots, p+q).$$

The supremum in (1.2) is defined to be zero when the set is empty, and exists and is unique provided that 0 belongs to the convex hull of $(g(X_1, \mu, \nu), \ldots, g(X_n, \mu, \nu))$. In the latter case, the standard Lagrange multiplier method provides the optimal $p_i(\nu)$:

$$p_i(\nu) = \frac{1}{n}\left(1 + \sum_{j=1}^{p+q} \lambda_j(\nu) g_j(X_i, \mu, \nu)\right)^{-1} \qquad (i = 1, \ldots, n), \tag{1.3}$$

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

$$\ell(\mu, \nu) = -2 \log L(\mu, \nu) = 2 \sum_{i=1}^{n} \log\Big\{1 + \sum_{j=1}^{p+q} \lambda_j(\nu) g_j(X_i, \mu, \nu)\Big\}, \tag{1.4}$$

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

$$\sum_{i=1}^{n} \frac{g_j(X_i, \mu, \nu)}{1 + \sum_{k=1}^{p+q} \lambda_k(\nu) g_k(X_i, \mu, \nu)} = 0 \qquad (j = 1, \ldots, p+q). \tag{1.5}$$

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

$$\tilde{\nu}(\mu) = \mathrm{argmin}_{\nu} \ell(\mu, \nu), \tag{1.6}$$

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

$$\ell(\mu) = \ell(\mu, \tilde{\nu}(\mu)). \tag{1.7}$$

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 $g_1, \ldots, g_{p+q}$ has been overcome by replacing them by smooth approximations, leading to a so-called smooth empirical likelihood. See e.g. Zhou and Jing (2003) for differences of quantiles, Claeskens, Jing, Peng and Zhou (2003) for ROC curves and Chen, Peng and Zhao (2006) 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 ν_0, as we have done in (1.6), one could also replace ν_0 by a certain 'plug-in' estimator, different from the above profile-estimator. This idea has been considered in Hjort, McKeague and Van Keilegom (2008) in a general framework (where ν_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 ν_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^2_p-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:

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^t & 0 \end{pmatrix}, \tag{2.1}$$

where

$$\begin{aligned} V_{11} &= \left(E\left\{g_j(X,\mu_0,\nu_0)g_k(X,\mu_0,\nu_0)\right\}\right)_{j,k=1,\dots,p+q} \\ V_{12} &= \left(-\frac{\partial}{\partial\nu_k}E\left\{g_j(X,\mu_0,\nu)\right\}\big|_{\nu=\nu_0}\right)_{k=1,\dots,q;j=1,\dots,p+q}. \end{aligned}$$

We will need the following conditions:

(C1) The functions $g_j(x,\mu_0,\nu)(j=1,\dots,p+q)$ are uniformly bounded in $I\!\!R^d \times I\!\!R^q$; the functions $E\{g_j(X,\mu_0,\nu)g_k(X,\mu_0,\nu)\}$ $(j,k=1,\dots,p+q)$, $\frac{\partial}{\partial\nu_k}E\{g_j(X,\mu_0,\nu)\}$ and $\frac{\partial^2}{\partial\nu_k\partial\nu_\ell}E\{g_j(X,\mu_0,\nu)\}$ $(k,\ell=1,\dots,q;j=1,\dots,p+q)$ are continuous for ν in a neighborhood of ν_0; the function $E\{g(X,\mu_0,\nu)/[1+\xi^t g(X,\mu_0,\nu)]\}$ has continuous partial derivatives with respect to the components of ν and ξ in a neighborhood of ν_0 and 0.

(C2) The matrix V_{11} in (2.1) is positive definite.

(C3) $\tilde{\nu}$ converges in probability to ν_0.

(C4) $n^{-1}\sum_{i=1}^{n}[g_j(X_i,\mu_0,\nu) - E\{g_j(X,\mu_0,\nu)\}] = O_P(n^{-1/2})$, uniformly in ν in a $o(1)$-neighborhood of ν_0 ($j = 1,\ldots,p+q$).

(C5) $n^{-1}\sum_{i=1}^{n}[g_j(X_i,\mu_0,\nu)g_k(X_i,\mu_0,\nu) - E\{g_j(X,\mu_0,\nu)g_k(X,\mu_0,\nu)\}] = o_P(1)$, uniformly in ν in a $o(1)$-neighborhood of ν_0 ($j,k = 1,\ldots,p+q$).

(C6) $n^{-1}\sum_{i=1}^{n}[g_j(X_i,\mu_0,\nu) - E\{g_j(X,\mu_0,\nu)\} - g_j(X_i,\mu_0,\nu_0) + E\{g_j(X,\mu_0,\nu_0)\}] = o_P(n^{-1/2})$, uniformly in ν for $\nu - \nu_0 = O(n^{-1/2})$ ($j = 1,\ldots,p+q$).

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)-(C6) are standard uniform consistency and modulus of continuity results, that 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 2.1 *Assume (C1)-(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 $g_1,\ldots,g_{p+q}$ are smooth functions, this result has been shown in Corollary 4 in Qin and Lawless (1994), using a different method of proof.

Remark 2.1 [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 $X_1,\ldots,X_{n_1}$ is a sample from a d_1-dimensional random vector X, and $Y_1,\ldots,Y_{n_2}$ 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

$$E\{g(X,\mu_0,\nu_0)\} = 0, \quad \text{and} \quad E\{h(Y,\mu_0,\nu_0)\} = 0, \tag{2.2}$$

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

$$L(\mu, \nu) = n_1^{n_1} n_2^{n_2} \sup \Big\{ \prod_{i=1}^{n_1} p_i(\nu) \prod_{j=1}^{n_2} q_j(\nu) : p_i(\nu) \geq 0, q_j(\nu) \geq 0, \sum_{i=1}^{n_1} p_i(\nu) = 1, \sum_{j=1}^{n_2} q_j(\nu) = 1 \Big\},$$

subject to the restrictions

$$\sum_{i=1}^{n_1} p_i(\nu) g_j(X_i, \mu, \nu) = 0, \quad \text{and} \quad \sum_{i=1}^{n_2} q_i(\nu) h_j(Y_i, \mu, \nu) = 0.$$

The empirical log-likelihood for μ is again defined as

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

where the Lagrange multipliers $\lambda_j(\nu)$ $(j = 1, \dots, r_1)$ and $\kappa_j(\nu)$ $(j = 1, \dots, r_2)$ satisfy equations analogous to the one in (1.5). The definitions of $\tilde{\nu} = \tilde{\nu}(\mu_0)$ and $\ell(\mu) = \ell(\mu, \tilde{\nu}(\mu))$ are the same as in (1.6) and (1.7). The analogue of the $(p + 2q) \times (p + 2q)$ matrix V in (2.1) now becomes

$$\begin{pmatrix} V_{11} & 0 & V_{13} \\ 0 & V_{22} & V_{23} \\ V_{13}^t & V_{23}^t & 0 \end{pmatrix},$$

with

$$\begin{aligned} V_{11} &= \Big(E\{g_j(X, \mu_0, \nu_0) g_k(X, \mu_0, \nu_0)\} \Big)_{j,k=1,\dots,r_1} \\ V_{22} &= \Big(E\{h_j(Y, \mu_0, \nu_0) h_k(Y, \mu_0, \nu_0)\} \Big)_{j,k=1,\dots,r_2} \\ V_{13} &= \Big(-\frac{\partial}{\partial \nu_k} E\{g_j(X, \mu_0, \nu)\}|_{\nu=\nu_0} \Big)_{j=1,\dots,r_1; k=1,\dots,q} \\ V_{23} &= \Big(-\frac{\partial}{\partial \nu_k} E\{h_j(Y, \mu_0, \nu)\}|_{\nu=\nu_0} \Big)_{j=1,\dots,r_2; k=1,\dots,q}. \end{aligned}$$

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

The required conditions are completely similar to (C1)-(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 2.2 [**Tests and confidence regions**] An approximate $100(1-\alpha)\%$ empirical likelihood confidence region for μ_0 is obtained by the following subset of $I\!R^p$:

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

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

In this section 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 $X_1, \ldots, X_n$ 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 $\nu_0 = F^{-1}(p_1)$, we have the equations

$$\begin{cases} F(\mu_0 + \nu_0) - p_2 = 0 \\ F(\nu_0) - p_1 = 0. \end{cases}$$

So, we have (1.1) with $p = q = 1$, $g_1(X, \mu, \nu) = I(X \leq \nu) - p_1$ and $g_2(X, \mu, \nu) = I(X \leq \mu + \nu) - p_2$. This problem has been studied by Zhou and Jing (2003) by using a smoothed empirical likelihood approach, whereas Chen and Hall (1993) studied one single quantile using a smoothed EL approach.

From Theorem 2.1 we know that $\ell(\mu_0)$ converges to a χ^2_1-distribution, provided conditions (C1)-(C6) are satisfied. Condition (C1) is satisfied if $F(x)$ is twice continuously differentiable in a neighborhood of $x = \nu_0$. (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). First, note that for $0 < F(\nu) < F(\mu_0 + \nu) < 1$ and for n large enough, the convex hull of $(g(X_1, \mu_0, \nu), \dots, g(X_n, \mu_0, \nu))$ is equal to the triangle in $I\!R^2$ 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 since $p_1 < p_2$. Hence, for any $0 < F(\nu) < F(\mu_0 + \nu) < 1$, we can write the empirical log-likelihood ratio as in (1.4), i.e. $\ell(\mu, \nu) = 2\sum_{i=1}^n \log\{1 + \sum_{j=1}^2 \lambda_j(\nu) g_j(X_i, \mu, \nu)\}$, where (for $j = 1, 2$)

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

Now, note that $\tilde{\nu}$ and ν_0 are the maximizers of $\Gamma_n(\nu)$ and $\Gamma(\nu)$ respectively, where $\Gamma_n(\nu)$ and $\Gamma(\nu)$ are as in (A.8) and (A.9). 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

$$\sup_{\nu} |\Gamma_n(\nu) - \Gamma(\nu)| \xrightarrow{P} 0, \tag{3.1}$$

$$\sup_{|\nu - \nu_0| > \varepsilon} \Gamma(\nu) < \Gamma(\nu_0) \tag{3.2}$$

for all $\varepsilon > 0$. Condition (3.2) is ensured by the fact that ν_0 is assumed to be unique, whereas for (3.1) we write

$$\begin{aligned} &\Gamma_n(\nu) - \Gamma(\nu) \\ &= \Big[-n^{-1} \sum_{i=1}^n \log(1 + \lambda(\nu)^t g(X_i, \mu_0, \nu)) + E\{\log(1 + \lambda(\nu)^t g(X, \mu_0, \nu))\} \Big] \\ &\quad + \Big[-E\{\log(1 + \lambda(\nu)^t g(X, \mu_0, \nu))\} + E\{\log(1 + \xi(\nu)^t g(X, \mu_0, \nu))\} \Big]. \end{aligned}$$

The second term above is easily seen to be $o_P(1)$ by using standard arguments concerning parametric Z-estimators. To show that the first term goes to zero uniformly in ν, we will prove that the class

$$\mathcal{F} = \Big\{ x \to \log(1 + \eta^t g(x, \mu_0, \nu)) : \eta \in R, 0 < \nu < 1 \Big\}$$

is Glivenko-Cantelli, where $R \subset I\!R^2$ is such that $(\lambda_1(\nu), \lambda_2(\nu))^t$ belongs to R for all $0 < \nu < 1$ and for almost all samples $X_1, \dots, X_n$. Note that R can be taken compact. This is because $1 + \lambda(\nu)^t g(X_i, \mu_0, \nu)$ is strictly positive for all i and all ν, and hence $(\lambda_1(\nu), \lambda_2(\nu))^t$ needs to satisfy the constraints

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

The intersection of these three halfplanes 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 and Wellner (1996), together with the monotonicity (in x) of the function $g_j(x, \mu_0, \nu)$ $(j = 1, 2)$. This shows that condition (C3) is satisfied.

As a consequence of Theorem 2.1, 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 $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ be a random sample from $X = (X_1, X_2)$ with 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 $C(u_1, u_2) = H(F_1^{-1}(u_1), F_2^{-1}(u_2))$, $0 \leq u_1, u_2 \leq 1$. Let

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

for fixed $0 \leq u_1, u_2 \leq 1$. Introducing the parameters $\nu_{01} = F_1^{-1}(u_1)$ and $\nu_{02} = F_2^{-1}(u_2)$, we have the equations

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

This is of the form (1.1) with $p = 1$, $q = 2$, $g_1(X, \mu, \nu) = I(X_1 \leq \nu_1, X_2 \leq \nu_2) - \mu$, $g_2(X, \mu, \nu) = I(X_1 \leq \nu_1) - u_1$, and $g_3(X, \mu, \nu) = I(X_2 \leq \nu_2) - u_2$. This example has been studied using a smoothed empirical likelihood by Chen, Peng and Zhao (2006).

The verification of conditions (C1)-(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, since 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 $X_1, \ldots, X_{n_1}$ from X with distribution function F_1 and $Y_1, \ldots, Y_{n_2}$ 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 $\nu_0 = F_1^{-1}(t)$, we have the equations

$$\begin{cases} F_1(\nu_0) - t = 0 \\ F_2(\mu_0 + \nu_0) - t = 0. \end{cases}$$

This is of the form (2.2) with $g(X, \mu, \nu) = I(X \leq \nu) - t$ and $h(Y, \mu, \nu) = I(Y \leq \mu + \nu) - t$.

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 curve (ROC curve), evaluated in the point t. For a nice introduction on ROC curves see e.g. Pepe (2003). Introducing the parameter $\nu_0 = F_2^{-1}(1 - t)$ leads to the equations

$$\begin{cases} F_1(\nu_0) - (1 - \mu_0) = 0 \\ F_2(\nu_0) - (1 - t) = 0, \end{cases}$$

which is again of the form (2.2). A smoothed empirical likelihood approach to this problem has been considered in Claeskens, Jing, Peng and Zhou (2003) for completely observed data and in Cao, Molanes Lopez and Van Keilegom (2008) 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 behavior of our empirical likelihood method in this particular case and compare it with the smoothed empirical likelihood approach proposed by Chen, Peng and Zhao (2006).

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 $H_0 : C(u_1, u_2) = \mu$ can not be rejected. According to Remark 2.2, an approximate $100(1-\alpha)\%$ confidence region for μ_0 is given by

$$I_{1-\alpha}(u_1, u_2) = \left\{\mu : \ell(\mu) \le \chi^2_{1,1-\alpha}\right\}. \tag{4.1}$$

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

$$C(u_1, u_2; \lambda, \theta_1, \theta_2) = \lambda \left\{u_1^{-\theta_1} + u_2^{-\theta_1} - 1\right\}^{\frac{1}{\theta_1}} + (1-\lambda) \exp\left\{-((-\log u_1)^{\theta_2} + (-\log u_2)^{\theta_2})^{\frac{1}{\theta_2}}\right\}, \tag{4.2}$$

where the marginals are standard normal distributions and the parameters θ_1, θ_2 and λ are such that $\theta_1 > 0$, $\theta_2 > 1$ and $\lambda \in [0, 1]$. The above mixture copula has been previously considered by Chen, Peng and Zhao (2006). When $\lambda = 1$ the mixture copula in (4.2) becomes a Clayton copula with parameter θ_1 and when $\lambda = 0$ it becomes a Gumbel-Hougaard copula with parameter θ_2. The parameter λ 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, Lindskog and McNeil (2001). 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 done using an alternative algorithm proposed by Marshall and 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 (4.1) is obtained under different scenarios. For every trial, we first obtain the value of $\ell(\mu_0)$ by solving the optimization problem in (4.3)-(4.5) below:

$$\min_{\nu} \ell(\mu_0, \nu) \tag{4.3}$$

subject to

$$\sum_{i=1}^{n} \frac{g_j(X_i, \mu_0, \nu)}{1 + \sum_{k=1}^{3} \lambda_k(\nu) g_k(X_i, \mu_0, \nu)} = 0, \text{ for } j = 1, 2, 3, \tag{4.4}$$

$$1 + \sum_{k=1}^{3} \lambda_k(\nu) g_k(X_i, \mu_0, \nu) > 0, \text{ for } i = 1, \dots, n, \tag{4.5}$$

where $g_k(X_i, \mu_0, \nu)$ ($k = 1, 2, 3$ and $i = 1, \ldots, n$) are given in Section 3.2 above. Once we obtain $\ell(\mu_0)$ we check if μ_0 falls in the confidence region given in (4.1) by checking whether $\ell(\mu_0) \leq \chi^2_{1,1-\alpha}$. The proportion of times that μ_0 falls in (4.1) gives us a Monte Carlo approximation of the coverage probability of (4.1).

Note that the constraints in (4.5) must be imposed to exclude any $\lambda(\nu) = (\lambda_1(\nu), \lambda_2(\nu), \lambda_3(\nu))$ for which some $p_i(\nu) \leq 0$ (see (1.3)). Following the ideas presented in Owen (2001), the constrained optimization problem in (4.3)-(4.5) is equivalent to

$$\min_{\nu} \ell(\mu_0, \nu), \tag{4.6}$$

subject to

$$\sum_{i=1}^{n} \log_{\star}^{(1)} \{1 + \sum_{k=1}^{3} \lambda_k(\nu) g_k(X_i, \mu_0, \nu)\} g_j(X_i, \mu_0, \nu) \quad = \quad 0, \text{ for } j = 1, 2, 3, \tag{4.7}$$

where

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

and $\log_{\star}^{(1)}(z) = \frac{\partial}{\partial z}(\log_{\star}(z))$. With this new formulation of the problem, the inequality constraints in (4.5) have been ruled out.

Put Table 1 about here

Table 1 shows the coverage probabilities of our method and those reported in Chen, Peng and Zhao (2006), which makes our results directly comparable with theirs. 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 behavior of our method, which has the advantage of avoiding a bandwidth selection problem, is in general at least as good as the behavior of the smoothed empirical likelihood approach of Chen, Peng and Zhao (2006). Moreover, as Chen, Peng and Zhao (2006) 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 parameters λ, θ_1 and θ_2 defining (4.2), and for points (u_1, u_2) falling outside of the unit square diagonal. Chen, Peng and Zhao (2006) did not consider this setting in their simulation study. The table shows that the empirical coverage probabilities are close to their nominal values and that the results improve when the sample size increases.

Put Table 2 about here

Note that the fact that we do not introduce smoothing and avoid a delicate bandwidth selection problem, inherent to Chen, Peng and Zhao (2006)'s method, entails on the other hand that our methodology is more complex to program, because no derivatives can be taken.

In the implementation of our method, we use a combination of algorithms. For every trial we first try to solve the corresponding optimization problem specified in (4.6)-(4.7) by using a Matlab function that solves non-linear constrained optimization problems (*fmincon*). If this algorithm fails to find the solution to (4.6)-(4.7), we then try to use a modification of a basic generating set search (GSS) algorithm for unconstrained optimization, proposed by Frimannslund and Steihaug (2007). This algorithm is derivative-free and tries to incorporate curvature information about the objective function as the search progresses, in such a way that the search directions of a basic GSS algorithm are adapted to the local topography. In order to use this methodology and solve the constrained optimization problem in (4.6)-(4.7), we consider that the evaluation of the objective function in (4.6) at a given ν goes through previously finding the solution, $\lambda(\nu)$, to the nonlinear system of equations defined in (4.7). The Matlab function *fsolve* has been used to solve these non-linear equations. If this algorithm does not converge neither, then we use a crude grid search. We solve (4.6)-(4.7) by computing (4.6) at an equally spaced grid of points, $\nu = (\nu_1, \nu_2)$, 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 $j = 1, 2$. For every point ν in the grid, we first use *fsolve* to find the solution to (4.7) and then we evaluate $\ell(\mu_0, \nu)$.

Appendix: Proofs

In this appendix the proofs will be given of the main theorem and of several lemmas. Since 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. Since we know from condition (C3) that $\tilde{\nu}$ converges in probability to ν_0, we can restrict attention in what follows to a $o(1)$-neighborhood of ν_0. In that case, the empirical log-likelihood ratio can be written in the form (1.4) for n large, and we therefore define:

$$\Gamma_n(\nu) = -n^{-1} \sum_{i=1}^{n} \log(1 + \lambda(\nu)^t g(X_i, \mu_0, \nu)) \tag{A.8}$$

$$\Gamma(\nu) = -E\{\log(1 + \xi(\nu)^t g(X, \mu_0, \nu))\}, \tag{A.9}$$

where $\xi(\nu) = (\xi_1(\nu), \ldots, \xi_{p+q}(\nu))$ satisfies

$$E\Big\{\frac{g(X, \mu_0, \nu)}{1 + \xi(\nu)^t g(X, \mu_0, \nu)}\Big\} = 0. \tag{A.10}$$

Note that $\xi(\nu)$ exists and is unique for ν in a neighborhood of ν_0. This follows from the implicit function theorem (see e.g. Theorem 13.7 p. 374 in Apostol (1974)), together with condition (C1).

We start with a preliminary lemma concerning the Lagrange multiplier $\lambda(\nu)$.

Lemma A.1 *Under (C1)-(C3),*

$$\tilde{\nu} = \arg\max_{\nu} \Gamma_n(\nu), \tag{A.11}$$

$$\nu_0 = \arg\max_{\nu} \Gamma(\nu). \tag{A.12}$$

Proof. (A.11) follows from the fact that

$$\max_{\nu} \Gamma_n(\nu) = -\frac{1}{2} n^{-1} \min_{\nu} \ell(\mu_0, \nu) = -\frac{1}{2} n^{-1} \ell(\mu_0, \tilde{\nu}) = \Gamma_n(\tilde{\nu}).$$

For (A.12), note that $\Gamma(\nu_0) = 0$ since $\xi(\nu_0) = 0$, and that for any $\nu \neq \nu_0$,

$$\Gamma(\nu) = -\xi(\nu)^t E\Big\{\frac{g(X, \mu_0, \nu)}{1 + \xi(\nu)^t g(X, \mu_0, \nu)}\Big\} - \frac{1}{2} E\Big\{\frac{(\xi(\nu)^t g(X, \mu_0, \nu))^2}{(1 + \alpha(\nu)^t g(X, \mu_0, \nu))^2}\Big\},$$

for some $\alpha(\nu)$ on the line segment between 0 and $\xi(\nu)$. The first term above equals 0, whereas the second one is strictly negative. Hence ν_0 is a maximizer of $\Gamma(\nu)$. $\square$

Lemma A.2 *Under (C1), (C2), (C4), (C5), we have*

$$\lambda(\nu) = O_P(n^{-1/2}) + O_P(\|\nu - \nu_0\|), \tag{A.13}$$

$$\lambda(\nu) - \xi(\nu) = O_P(n^{-1/2}) \tag{A.14}$$

uniformly for all ν in a $o(1)$-neighborhood of ν_0, and

$$\lambda(\nu) = V_{11}^{-1}(\nu) n^{-1} \sum_{i=1}^{n} g(X_i, \mu_0, \nu) + o_P(n^{-1/2}), \tag{A.15}$$

uniformly for all ν in a $O(n^{-1/2})$-neighborhood of ν_0, where $V_{11}(\nu) = (E\{g_j(X, \mu_0, \nu)$ $g_k(X, \mu_0, \nu)\})_{j,k=1,\ldots,p+q}$.

Proof. First note that $V_{11}(\nu)$ is positive definite for ν in a neighborhood of ν_0 because of conditions (C1) and (C2). The proof of (A.13) and (A.15) follows along the same lines as the proof of e.g. Theorem 3.2 in Owen (2001) (page 219). The proof of (A.14) follows using standard arguments concerning parametric Z-estimators. $\square$

Lemma A.3 *Under (C1)-(C2), there exists a neighborhood N of ν_0 and a constant $K > 0$ for which*

$$\Gamma(\nu) \leq -K\|\nu - \nu_0\|^2$$

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

Proof. By Taylor expansion and using that $\Gamma(\nu_0)$ and the partial derivatives of first order are zero, we have

$$\Gamma(\nu) = \frac{1}{2}(\nu - \nu_0)^t F(\nu - \nu_0),$$

where $F = \frac{\partial^2}{\partial\nu\partial\nu^t}\Gamma(\nu)|_{\nu=\eta}$ and η is on the line segment between ν and ν_0. Since $\frac{\partial^2}{\partial\nu\partial\nu^t}\Gamma(\nu)|_{\nu=\nu_0}$ is negative definite, we also have that F is negative definite if η is close to ν_0. It follows that the matrix F has the following representation: $F = P^{-1}\Lambda P$, where $P^{-1} = P^t$ and Λ is a diagonal matrix with the eigenvalues λ_i on the diagonal. These eigenvalues λ_i are strictly negative. Then,

$$\begin{aligned}
(\nu - \nu_0)^t F(\nu - \nu_0) &= (\nu - \nu_0)^t P^{-1}\Lambda P(\nu - \nu_0) \\
&= (P(\nu - \nu_0))^t \Lambda(P(\nu - \nu_0)) \\
&\leq -K\|P(\nu - \nu_0)\|^2 \\
&= -K(\nu - \nu_0)^t P^t P(\nu - \nu_0) \\
&= -K\|\nu - \nu_0\|^2 \qquad \text{with } K = \min_i(-\lambda_i).
\end{aligned}$$

$\square$

Lemma A.4 *Under (C1), (C4), (C5), we have*

$$\Gamma_n(\nu) = \Gamma(\nu) + O_P(n^{-1/2}\|\nu - \nu_0\|) + o_P(\|\nu - \nu_0\|^2) + O_P(n^{-1})$$

uniformly in ν, for $\nu - \nu_0 = o(1)$.

Proof. First note that

$$
\begin{aligned}
&E\{\log(1+\lambda(\nu)^t g(X,\mu_0,\nu))\} - E\{\log(1+\xi(\nu)^t g(X,\mu_0,\nu))\} \\
&= \{\lambda(\nu)-\xi(\nu)\}^t E\Big\{\frac{g(X,\mu_0,\nu)}{1+\xi(\nu)^t g(X,\mu_0,\nu)}\Big\} \\
&\quad + \frac{1}{2}\{\lambda(\nu)-\xi(\nu)\}^t E\Big\{\frac{g(X,\mu_0,\nu)g^t(X,\mu_0,\nu)}{(1+\eta(\nu)^t g(X,\mu_0,\nu))^2}\Big\}\{\lambda(\nu)-\xi(\nu)\},
\end{aligned}
$$

for some $\eta(\nu)$ on the line segment between $\lambda(\nu)$ and $\xi(\nu)$. The first term above equals zero by equation (A.10), while the second one is $O_P(n^{-1})$ by Lemma A.2.

Hence, it suffices to calculate the order of

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^n \log(1+\lambda(\nu)^t g(X_i,\mu_0,\nu)) - E\{\log(1+\lambda(\nu)^t g(X,\mu_0,\nu))\} \\
&= n^{-1}\sum_{i=1}^n \big[\lambda(\nu)^t g(X_i,\mu_0,\nu) - E\{\lambda(\nu)^t g(X,\mu_0,\nu)\}\big] \\
&\quad -\frac{1}{2}n^{-1}\sum_{i=1}^n [(\lambda(\nu)^t g(X_i,\mu_0,\nu))^2 - E\{(\lambda(\nu)^t g(X,\mu_0,\nu))^2\}] \\
&\quad +\frac{1}{3}n^{-1}\sum_{i=1}^n \left[\frac{(\lambda(\nu)^t g(X_i,\mu_0,\nu))^3}{(1+\xi_{1i})^3} - E\left\{\frac{(\lambda(\nu)^t g(X,\mu_0,\nu))^3}{(1+\xi_2)^3}\right\}\right] \\
&= T_1 + T_2 + T_3,
\end{aligned}
$$

where

$$
|\xi_{1i}| \le |\lambda(\nu)^t g(X_i,\mu_0,\nu)| \le M\sum_{j=1}^{p+q}|\lambda_j(\nu)|,
$$

$$
|\xi_2| \le |\lambda(\nu)^t g(X,\mu_0,\nu)| \le M\sum_{j=1}^{p+q}|\lambda_j(\nu)|,
$$

and $M = \sup_{j,x,\nu}|g_j(x,\mu_0,\nu)| < \infty$ by condition (C1). For T_1 we have

$$
\begin{aligned}
|T_1| &\le \sum_{j=1}^{p+q}|\lambda_j(\nu)|\Big|n^{-1}\sum_{i=1}^n [g_j(X_i,\mu_0,\nu) - E\{g_j(X,\mu_0,\nu)\}]\Big| \\
&= O_P\Big(n^{-1/2}\sum_{j=1}^{p+q}|\lambda_j(\nu)|\Big) = O_P\Big(n^{-1/2}\|\nu-\nu_0\|\Big) + O_P(n^{-1})
\end{aligned}
$$

because of condition (C4) and Lemma A.2. Further,

$$
\begin{aligned}
|T_2| &\le \frac{1}{2}\sum_{j,k=1}^{p+q}|\lambda_j(\nu)\lambda_k(\nu)|\Big|n^{-1}\sum_{i=1}^{n}[g_j(X_i,\mu_0,\nu)g_k(X_i,\mu_0,\nu)-E\{g_j(X,\mu_0,\nu)g_k(X,\mu_0,\nu)\}]\Big| \\
&= o_P\Big(\Big(\sum_{j=1}^{p+q}|\lambda_j(\nu)|\Big)^2\Big) = o_P(\|\nu-\nu_0\|^2)+o_P(n^{-1})
\end{aligned}
$$

by using (C5). Similarly,

$$
\begin{aligned}
|T_3| \le \frac{1}{3}\sum_{j,k,\ell=1}^{p+q}|\lambda_j(\nu)\lambda_k(\nu)\lambda_\ell(\nu)|\Big|n^{-1}\sum_{i=1}^{n}\Big[&\frac{g_j(X_i,\mu_0,\nu)g_k(X_i,\mu_0,\nu)g_\ell(X_i,\mu_0,\nu)}{(1+\xi_{1i})^3} \\
&-E\Big\{\frac{g_j(X,\mu_0,\nu)g_k(X,\mu_0,\nu)g_\ell(X,\mu_0,\nu)}{(1+\xi_2)^3}\Big\}\Big]\Big|.
\end{aligned}
$$

By (C1) and since $|\xi_{1i}| \le M\|\lambda(\nu)\|$ and $|\xi_2| \le M\|\lambda(\nu)\|$, we have that, for each sequence of random variables $\{r_n\}$ with $r_n = o(1)$:

$$
\begin{aligned}
\sup_{\|\nu-\nu_0\|\le r_n/M}|T_3| &\le (\text{constant})\Big(\sum_{j=1}^{p+q}|\lambda_j(\nu)|\Big)^3\frac{1}{1-r_n} \\
&= O_P(\|\nu-\nu_0\|^3+n^{-3/2})\frac{1}{1-r_n} = o_P(\|\nu-\nu_0\|^2)+O_P(n^{-1}).
\end{aligned}
$$

□

Lemma A.5 *Under (C1), (C5) and (C6) we have*

$$
\Gamma_n(\nu) = \frac{1}{2}(\theta-\theta_0)^t V(\theta-\theta_0) - n^{-1/2}(\theta-\theta_0)^t W_n + o_P(n^{-1}) \qquad \text{(A.16)}
$$

uniformly in ν, for $\nu-\nu_0 = O(n^{-1/2})$, where $\theta = (\lambda(\nu),\nu)^t$, $\theta_0 = (0,\nu_0)^t$,

$$
W_n = \Big(n^{-1/2}\sum_{i=1}^{n}g_1(X_i,\mu_0,\nu_0),\ldots,n^{-1/2}\sum_{i=1}^{n}g_{p+q}(X_i,\mu_0,\nu_0),0_q\Big), \qquad \text{(A.17)}
$$

and where 0_q is a vector of q zeros.

Proof. Throughout the proof we will use the notation $\theta_1 = \lambda(\nu)$, $\theta_2 = \nu$, $\theta_{01} = 0$ and $\theta_{02} = \nu_0$. Taylor expansion as in Lemma A.4 gives

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

where $|\xi_{1i}| \leq M \sum_{j=1}^{p+q} |\theta_{1j}|$. As in the proof of Lemma A.4 we have that $S_3 = O_P(\|\theta - \theta_0\|^3) = o_P(n^{-1})$, since $\lambda(\nu) = O_P(n^{-1/2})$ by Lemma A.2. Next consider S_1. We write

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

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

$$\begin{aligned} S_{12} = & -n^{-1} \sum_{j=1}^{p+q} \sum_{i=1}^{n} \Big[g_j(X_i, \mu_0, \theta_2) - E\{g_j(X, \mu_0, \theta_2)\} - g_j(X_i, \mu_0, \theta_{02}) + E\{g_j(X, \mu_0, \theta_{02})\} \Big] \theta_{1j} \\ & - \sum_{j=1}^{p+q} \Big[E\{g_j(X, \mu_0, \theta_2)\} - E\{g_j(X, \mu_0, \theta_{02})\} \Big] \theta_{1j}. \end{aligned} \quad \text{(A.18)}$$

From (C6) we have that the first term in (A.18) is $o_P(n^{-1/2} \sum_{j=1}^{p+q} |\theta_{1j}|) = o_P(n^{-1/2}\|\theta - \theta_0\|) = o_P(n^{-1})$. For the second term in (A.18), 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 $\frac{\partial}{\partial \theta_2} E\{g_j(X, \mu_0, \theta_2)\} \mid_{\theta_2 = \theta_{02}}$ is the vector with elements $\frac{\partial}{\partial \theta_{2k}} E\{g_j(X, \mu_0, \theta_2)\} \mid_{\theta = \theta_{02}}$. Hence,

$$S_{12} = (\theta_1 - \theta_{01})^t V_{12} (\theta_2 - \theta_{02}) + o_P(n^{-1}).$$

Now we deal with S_2. We write this term as

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

For S_{21} we write

$$S_{21} = \frac{1}{2}n^{-1}\sum_{j,k=1}^{p+q}\sum_{i=1}^{n}\Big[g_j(X_i,\mu_0,\theta_{02})g_k(X_i,\mu_0,\theta_{02})$$
$$- E\{g_j(X,\mu_0,\theta_{02})g_k(X,\mu_0,\theta_{02})\}\Big]\theta_{1j}\theta_{1k}$$
$$+\frac{1}{2}\sum_{j,k=1}^{p+q} E\{g_j(X,\mu_0,\theta_{02})g_k(X,\mu_0,\theta_{02})\}\theta_{1j}\theta_{1k}. \qquad \text{(A.19)}$$

From (C5), we have that the first term in (A.19) is $o_P((\sum_{j=1}^{p+q}|\theta_{1j}|)^2) = o_P(\|\theta-\theta_0\|^2) = o_P(n^{-1})$. The second term in (A.19) is equal to

$$\frac{1}{2}(\theta_1-\theta_{01})^t V_{11}(\theta_1-\theta_{01}).$$

The term S_{22} can be written as

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

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 (A.16). □

Lemma A.6 *Under (C1)-(C6), we have*

$$\Gamma_n(\nu) = -\frac{1}{2}(\nu-\nu_0)^t V_{12}^t V_{11}^{-1} V_{12}(\nu-\nu_0) + n^{-1/2}(\nu-\nu_0)^t V_{12}^t V_{11}^{-1} X_n$$
$$-\frac{1}{2}n^{-1}X_n^t V_{11}^{-1} X_n + o_P(n^{-1}) \qquad \text{(A.20)}$$

uniformly in ν, for $\nu-\nu_0 = O(n^{-1/2})$, and

$$\tilde{\nu} - \nu_0 = n^{-1/2}(V_{12}^t V_{11}^{-1} V_{12})^{-1} V_{12}^t V_{11}^{-1} X_n + o_P(n^{-1/2}) \qquad \text{(A.21)}$$

where $W_n = (X_n, 0_q)^t$, and V and W_n are given in (2.1) and (A.17) respectively.

Proof. We start with the first assertion. From Lemma A.2, together with conditions (C1) and (C6), we know that

$$\lambda(\nu) = V_{11}^{-1}[n^{-1/2}X_n - V_{12}(\nu-\nu_0)] + o_P(n^{-1/2})$$

uniformly for all ν in a $O(n^{-1/2})$-neighborhood of ν_0. Hence, it follows from Lemma A.5 that

$$\begin{aligned}
\Gamma_n(\nu) &= \frac{1}{2}\Big[\lambda(\nu)^t V_{11}\lambda(\nu) + 2(\nu-\nu_0)^t V_{12}^t\lambda(\nu)\Big] - n^{-1/2}(\lambda(\nu)^t, (\nu-\nu_0)^t)W_n + o_P(n^{-1}) \\
&= \frac{1}{2}\Big[\Big\{n^{-1/2}X_n^t - (\nu-\nu_0)^t V_{12}^t\Big\}V_{11}^{-1}\Big\{n^{-1/2}X_n - V_{12}(\nu-\nu_0)\Big\}\Big] \\
&\quad +(\nu-\nu_0)^t V_{12}^t V_{11}^{-1}\Big\{n^{-1/2}X_n - V_{12}(\nu-\nu_0)\Big\} \\
&\quad -n^{-1/2}\Big\{n^{-1/2}X_n^t - (\nu-\nu_0)^t V_{12}^t\Big\}V_{11}^{-1}X_n + o_P(n^{-1}) \\
&= -\frac{1}{2}(\nu-\nu_0)^t V_{12}^t V_{11}^{-1}V_{12}(\nu-\nu_0) + n^{-1/2}(\nu-\nu_0)^t V_{12}^t V_{11}^{-1}X_n \\
&\quad -\frac{1}{2}n^{-1}X_n^t V_{11}^{-1}X_n + o_P(n^{-1}). \qquad \text{(A.22)}
\end{aligned}$$

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 A.1, A.3 and A.4 imply that $\tilde{\nu} - \nu_0 = O_P(n^{-1/2})$, by applying Theorem 1 in Sherman (1993). Next, (A.22) shows that the displayed condition (4) in the statement of Theorem 2 in Sherman (1993) is satisfied, except for the term $-\frac{1}{2}n^{-1}X_n^t V_{11}^{-1}X_n$, 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, since this term does not depend on ν. It now follows from the proof of this theorem that

$$\tilde{\nu} - \nu_0 = n^{-1/2}(V_{12}^t V_{11}^{-1}V_{12})^{-1}V_{12}^t V_{11}^{-1}X_n + o_P(n^{-1/2}).$$

This shows the second statement of the lemma. □

Proof of Theorem 2.1 Without loss of generality we condition on the event that $\|\tilde{\nu} - \nu_0\| \leq K$ for some $K > 0$. This is possible, since $\tilde{\nu} - \nu_0 = O_P(n^{-1/2})$ (see Lemma A.6). From Lemma A.6 it follows that, with $V_{22.1} = -V_{12}^t V_{11}^{-1}V_{12}$,

$$\begin{aligned}
\Gamma_n(\tilde{\nu}) &= -\frac{1}{2}n^{-1}X_n^t V_{11}^{-1}V_{12}V_{22.1}^{-1}V_{12}^t V_{11}^{-1}X_n - \frac{1}{2}n^{-1}X_n^t V_{11}^{-1}X_n + o_P(n^{-1}) \\
&= -\frac{1}{2}n^{-1}X_n^t V_{11}^{-1/2}DV_{11}^{-1/2}X_n + o_P(n^{-1}),
\end{aligned}$$

or equivalently,

$$\ell(\mu_0) = X_n^t V_{11}^{-1/2}DV_{11}^{-1/2}X_n + o_P(1), \qquad \text{(A.23)}$$

where $D = V_{11}^{-1/2}\{I + V_{12}V_{22.1}^{-1}V_{12}^tV_{11}^{-1}\}V_{11}^{1/2}$. Note that D can also be written as $D = V_{11}^{1/2}V^{11}V_{11}^{1/2}$, where

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

since it follows from Lemma 3 in Qin and Lawless (1994) that

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

Also, note that

$$V_{11}^{-1/2}X_n = V_{11}^{-1/2}n^{-1/2}\sum_{i=1}^{n} g(X_i, \mu_0, \nu_0) \xrightarrow{d} N(0; I),$$

so that from (A.23) it follows that $\ell(\mu_0) \xrightarrow{d} \chi_p^2$, provided we can show that

$$D \text{ is symmetric}, \tag{A.24}$$

$$D \text{ is idempotent}, \tag{A.25}$$

$$\mathrm{tr}(D) = p, \tag{A.26}$$

where $\mathrm{tr}(D)$ is the trace of the matrix D. For (A.24), we have that $D^t = D$ since it is easily seen that $V_{11}^t = V_{11}$ and $(V^{11})^t = V^{11}$. For (A.25), note that

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

since by direct calculation, it follows that

$$V^{11}V_{11}V^{11} = V_{11}^{-1}(I+V_{12}V_{22.1}^{-1}V_{12}^tV_{11}^{-1})(I+V_{12}V_{22.1}^{-1}V_{12}^tV_{11}^{-1}) = V_{11}^{-1}(I+V_{12}V_{22.1}^{-1}V_{12}^tV_{11}^{-1}) = V^{11}.$$

For (A.26), we have:

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

This finishes the proof of the theorem. $\square$

Acknowledgments. The second and third author acknowledge support from IAP research network nr. P6/03 of the Belgian government (Belgian Science Policy). The first author acknowledges support from the NS project MTM2005-00429 (EU ERDF support included).

References

Apostol, T. (1974). *Mathematical Analysis.* Addison-Wesley, Reading.

Cao, R., Molanes Lopez, E.M. and Van Keilegom, I. (2008). Empirical likelihood inference for relative distributions under left truncation and right censoring (submitted).

Chen, J., Peng, L. and Zhao, Y. (2006). Empirical likelihood based confidence intervals for copulas (submitted).

Chen, S.X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.*, **21**, 1166-1181.

Claeskens, G., Jing, B.Y., Peng, L. and Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *Canad. J. Statist.*, **31**, 173-190.

Embrechts, P., Lindskog, F. and McNeil, A. (2001). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance.* Elsevier.

Frimannslund, L. and Steihaug, T. (2007). A generating set search method using curvature information. *Comput. Optim. Appl.*, **38**, 105-121.

Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2008). Extending the scope of empirical likelihood. *Ann. Statist.* (to appear).

Marshall, A.W. and 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 and Hall, New York.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, New York.

Qin, J. and 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 Univ. Press. Cambridge.

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

Zhou, W. and Jing, B.-Y. (2003). Smoothed empirical likelihood confidence intervals for the difference of quantiles. *Statist. Sinica*, **13**, 83-95.

Table 1: Empirical coverage probabilities for the empirical likelihood based confidence region in (4.1) 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 method of Chen, Peng and Zhao (2006).

n	200		400	
(λ, u_1, u_2)	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$
(0.0,0.25,0.25)	0.893	0.940	0.886	0.944
	0.926	0.958	0.901	0.944
(0.0,0.50,0.50)	0.906	0.954	0.891	0.954
	0.889	0.933	0.894	0.946
(0.0,0.75,0.75)	0.844	0.934	0.862	0.936
	0.904	0.959	0.896	0.943
(0.5,0.25,0.25)	0.900	0.930	0.889	0.940
	0.937	0.968	0.907	0.954
(0.5,0.50,0.50)	0.907	0.943	0.900	0.936
	0.902	0.959	0.849	0.917
(0.5,0.75,0.75)	0.897	0.952	0.878	0.942
	0.914	0.958	0.870	0.924
(1.0,0.25,0.25)	0.870	0.915	0.894	0.935
	0.932	0.968	0.904	0.959
(1.0,0.50,0.50)	0.877	0.931	0.885	0.938
	0.904	0.962	0.897	0.950
(1.0,0.75,0.75)	0.902	0.952	0.858	0.919
	0.855	0.925	0.760	0.859

Table 2: Empirical coverage probabilities for the empirical likelihood based confidence region in (4.1) with sample sizes $n = 200, 300$ and points (u_1, u_2) outside the unit square diagonal and such that $|u_1 - u_2| = 0.10, 0.20, 0.30$.

n	200		300	
(λ, u_1, u_2)	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$	$I_{0.90}(u_1, u_2)$	$I_{0.95}(u_1, u_2)$
(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