Ann Inst Stat Math (2016) 68:105–133
DOI 10.1007/s10463-014-0491-8

Fourier methods for model selection

M. D. Jiménez-Gamero · A. Batsidis · M. V. Alba-Fernández

Received: 19 December 2013 / Revised: 9 August 2014 / Published online: 9 October 2014

Abstract A test approach to the model selection problem based on characteristic functions (CFs) is proposed. The scheme is close to that proposed by Vuong (Econometrica 57:257–306, 1989), which is based on comparing estimates of the Kullback–Leibler distance between each candidate model and the true population. Other discrepancy measures could be used. This is specially appealing in cases where the likelihood of a model cannot be calculated or even, if it has a closed expression, it is either not easily tractable or not regular enough. In this work, the closeness is measured by means of a distance based on the CFs. As a prerequisite, some asymptotic properties of the minimum integrated squared error estimators are studied. From these properties, consistent tests for model selection based on CFs are given for separate, overlapping and nested models. Several examples illustrate the application of the proposed methods.

Keywords Empirical characteristic function · Model selection · Misspecified models

1 Introduction

The main purpose of this paper is to propose new tests for the model selection problem that can be described as follows. Given a sample from an unknown population and two

M. D. Jiménez-Gamero
Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Avda. Reina Mercedes, s.n., 41012 Sevilla, Spain

A. Batsidis
Department of Mathematics, University of Ioannina, 45110 Ioannina, Greece

M. V. Alba-Fernández (✉)
Departamento de Estadística e Investigación Operativa, Universidad de Jaén, Campus de Las Lagunillas, Edificio Ciencias Experimentales y de la Salud, 23071 Jaén, Spain
e-mail: mvalba@ujaen.es

possibly misspecified parametric models, $\mathcal{F}$ and $\mathcal{G}$, which may be separate, overlapping or nested, the problem of model selection consists in testing if the two competing models are equally close to the true population against the hypothesis that one model is closer than the other. Vuong (1989) (see also Linhart 1988; Kishino and Hasegawa 1989) has proposed tests for this problem that are based on the likelihood ratio statistic, which estimates the difference of the Kullback–Leibler distance between each candidate model and the true distribution. Although this approach is good and well founded, some alternative procedures have been proposed. For example, it may happen that, even if the data come from a continuous population, the available data consist of the number of observations in certain intervals, a partition of the space where the original data take values. In this case, because the Pearson Chi-square statistic is widely used for this kind of data, it seems natural to measure the discrepancy between the true population and the competing models by means of some Chi-square type of distance. This approach was studied in Vuong and Wang (1993). Since the Pearson Chi-square statistic is a member of the class of ϕ-divergence statistics and also of the class of K_ϕ-divergence statistics (see, for example, Pardo 2006), Jiménez-Gamero et al. (2011, 2014) have studied the model selection problem by using these two classes of statistics for non-overlapping models.

Thus, other discrepancy measures could be used to measure the closeness between each competing model and the true population model. This is specially appealing in cases where the likelihood of a model cannot be calculated. A typical example is the case of some stable distributions, since it is only in a few instances that convenient expressions for densities can be found. Even if the likelihood has a closed expression, it may be either not easily tractable or not regular enough, in the sense that it does not satisfy the regularity conditions in Vuong (1989). This is the case of the Laplace distribution with location and scale parameters. A common feature of these examples is that in each case, the characteristic function (CF) has a quite regular closed simple expression. Therefore, in these cases it is more convenient to measure the closeness between each competing model and the true population model by means of a distance based on the CFs (see Meintanis 2005; Matsui and Takemura 2008, for testing problems in stable distributions).

A problem intimately related to that of model selection is that of testing for two separate families of distributions (see Cox 1961, 1962; White 1982b). A main difference is that, while the latter assumes that one of the models is true and the objective is to select the correct model, the former does not assume it and the objective is to select the model which, according to some discrepancy measure, is closest to the true population distribution. The Cox approach is based on comparing the observed difference of log likelihoods with an estimate of that to be expected under the null hypothesis. Since this approach is based on likelihoods, the same arguments given above can be applied in favor of using other discrepancy measures. In this line, Feigin and Heathcote (1976) have proposed using the empirical characteristic function (ECF). The proposed technique employs either the real or the imaginary part of the ECF evaluated at a single point. To avoid working with either the real or the imaginary part, Epps et al. (1982) proposed using the moment-generating function, but again evaluated at a single point. A weak point of these two papers is that the ECF and the empirical moment-generating function, respectively, are evaluated at a single point, which implies choosing it and

losing the information given by the rest of the points. While the CF exists for all distributions, the moment-generating function may not exist, so we prefer to work with the CF.

To measure the closeness between two populations defined on $\mathbb{R}^d$, for some $d \in \mathbb{N}$, with CFs $c_1(t)$ and $c_2(t)$, $t \in \mathbb{R}^d$, we consider the following discrepancy measure:

$$D^2(c_1, c_2) = \int |c_1(t) - c_2(t)|^2 \mathrm{d}W(t), \tag{1}$$

where for any complex number, $z = a + \mathrm{i}b$, with $\mathrm{i} = \sqrt{-1}$, $|z|^2 = a^2 + b^2$, an unspecified integral denotes integration over the whole space $\mathbb{R}^d$, and $W(t)$ denotes a nondecreasing weight function whose total variation can, without loss of generality, be taken as unity. Since $|c_1(t) - c_2(t)|^2 \leq 4$, the presence of $\mathrm{d}W(t)$ in the expression of $D^2(c_1, c_2)$ renders the integral in (1) finite. Observe that if

$$\mathrm{d}W(t) = w(t)\mathrm{d}t, \quad \text{with } w(t) > 0, \ \forall t \in \mathbb{R}^d, \tag{2}$$

then $D(c_1, c_2)$ is a true distance between distributions; otherwise, $c_1 = c_2$ implies that $D(c_1, c_2) = 0$, but the contrary is not true in general (see Feller 1971). Observe also that if $\mathrm{d}W(t) = w(t)\mathrm{d}t$, we can assume that w satisfies

$$w(t) = w(-t), \quad \forall t \in \mathbb{R}^d, \tag{3}$$

because otherwise by defining $w_1(t) = 0.5\{w(t) + w(-t)\}$, which satisfies (3), we have

$$\int |c_1(t) - c_2(t)|^2 w(t)\mathrm{d}t = \int |c_1(t) - c_2(t)|^2 w_1(t)\mathrm{d}t.$$

Therefore, from now on, whenever $\mathrm{d}W(t) = w(t)\mathrm{d}t$, we will assume that (3) holds.

Roughly speaking, the method proposed in this paper consists in choosing that model minimizing the discrepancy measure (1) between an estimator of the population CF and an estimator of the model CF. To estimate the population CF we consider the ECF, and an estimator of the model CF is obtained by replacing the unknown parameters by suitable estimators. Specifically, the unknown parameters will be estimated by their minimum integrated squared error (ISE) estimators, which minimize the discrepancy measure (1) between the model and the ECF associated with the data. Some asymptotic properties of these estimators have been studied in Heathcote (1977) and Csörgő (1981) when the model is assumed to be correctly specified. For our objectives, we also need to know some properties of these estimators when the model is misspecified. This study is done in Sect. 2, where we give sufficient conditions for the strong consistency and asymptotic normality of these estimators.

Given two parametric models, $\mathcal{F}$ and $\mathcal{G}$, which may be separate, overlapping or nested, we propose tests of the null hypothesis that both models are equivalent, in the sense that the distance D defined in (1) between the population and each model is the same, against that one of the models is closer than the other to the population generating the observed data. Motivated by some results in Sect. 2, the test statistic

is a sample version of the difference of the distances between the population and each competing model. The problem is the same as the one studied in Vuong (1989), but since the distances considered, and thus their estimators, are rather different, the required assumptions and the proofs of the results also differ. These tests and some properties are presented in Sect. 3. A CF analog of the Cox approach for separate models is developed in Sect. 4.

In Sects. 3 and 4, it is assumed that the unknown parameters are estimated by their ISE estimators. It is natural to wonder what happen if other estimators are used. This topic is studied in Sect. 5, where we will see that some asymptotic results may change.

Section 6 gives two examples where neither Vuong nor Cox approaches can be applied because the competing families are not regular, in the sense that the assumptions in Vuong (1989) and White (1982b) do not hold. In contrast, these families satisfy the assumptions required by the methods proposed in this paper. The finite sample performance of the proposed procedures is numerically investigated in each example. Section 6 also presents an example where both approaches apply. Section 7 provides the conclusions to the article. All proofs are sketched in the last section.

Before ending this section, we introduce some notation: all limits in this paper are taken when $n \to \infty$; $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution; $\xrightarrow{P}$ denotes convergence in probability; $\xrightarrow{\text{a.s.}}$ denotes the almost sure convergence; if $R \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, then R° denotes the interior of R; if $x \in \mathbb{R}^d$, then $|x|$ denotes the Euclidean norm; the same symbol is used to denote the modulus of a complex number; to simplify notation, all 0s appearing in the paper represent vectors or matrices of the appropriate dimension; P_0 denotes the probability under the null hypothesis, P_* denotes the conditional probability, given the data.

2 Minimum ISE estimators

As a prerequisite to the model selection problem based on CFs, in this section we study some asymptotic properties of the minimum ISE estimators. The strong consistency and asymptotic normality of these estimators have been proved in Heathcote (1977) under the assumption that the parametric model is correctly specified. The aim of this section is to study the limit and the asymptotic normality of these estimators when such assumption is dropped.

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (IID) random vectors from a population X taking values in $\mathbb{R}^d$ with CF $c(t)$ and cumulative distribution function (CDF) F. Let $\mathcal{F}$ be a family of distributions so that each member in this family has CF $c(t;\theta)$ and CDF $F(x;\theta)$, for some finite dimensional parameter θ; in other words, we can write $\mathcal{F} = \{c(t;\theta);\ \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$, for some $k \in \mathbb{N}$. Equivalently, we also write $\mathcal{F} = \{F(t;\theta);\ \theta \in \Theta\}$. We assume that the elements in $\mathcal{F}$ are identifiable, where by identifiable we mean $c(t;\theta_1) \neq c(t;\theta_2)$, in the sense that $\sup_t |c(t;\theta_1) - c(t;\theta_2)| > 0$, whenever $\theta_1 \neq \theta_2$. If $c(t) \in \mathcal{F}$, Heathcote (1977) proposed estimating θ by means of $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$, so that

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} I_n(\theta),$$

where

$$I_n(\theta) = \int |c_n(t) - c(t;\theta)|^2 \mathrm{d}W(t),$$

$c_n(t)$ stands for the ECF of the sample,

$$c_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{i}t'X_j),$$

and the prime denotes transpose. The statistic $\hat{\theta}_n$ is called the (minimum) ISE estimator of θ.

Now, if the assumption $c(t) \in \mathcal{F}$ is dropped, we define $D(c(t), \mathcal{F}) = \inf_{\theta\in\Theta} D(c(t), c(t,\theta))$ and the projection of $c(t)$ on $\mathcal{F}$ as $c(t;\theta_*)$, where $\theta_* \in \Theta$ is such that

$$\theta_* = \arg\min_{\theta\in\Theta} D^2(c(t), c(t;\theta)).$$

Since the elements in $\mathcal{F}$ are identifiable, if $c(t) \in \mathcal{F}$, that is, if $c(t) = c(t;\theta)$, for some $\theta \in \Theta$, then $\theta_* = \theta$; otherwise, $c(t;\theta_*)$ is the element in $\mathcal{F}$ closest to $c(t)$. Note that θ_* may not exist or, if it exists, it may not be unique. Along the manuscript we will assume the following.

Assumption 1 $D^2(c(t), c(t;\theta))$ has a unique minimum at $\theta_* \in \Theta$.

Assumption 1 is commonly used in papers dealing with projections, in the sense of handling parameters minimizing some kind of distance or discrepancy measure between a population and a parametric family of distributions. For example, it is the analog of Assumption A3(b) in White (1982a), Assumption A.9 in Vuong and Wang (1993), Assumption 30 in Lindsay (1994) and Assumption (C.1) in Broniatowski and Keziou (2009), just to cite a few.

Note that, in general, θ_* will depend on W. Thus, to be rigorous, we should denote it as $\theta_*(W)$. Nevertheless, to keep the notation as simple as possible, we will just write θ_* for $\theta_*(W)$.

Theorem 1 *Suppose that Assumption 1 holds, then* $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_*$.

In practice, if $I_n(\theta)$ can be differentiated, to calculate $\hat{\theta}_n$ we solve in θ the equation

$$\frac{\partial}{\partial\theta} I_n(\theta) = 0. \tag{4}$$

Theorem 1 asserts that, if $\theta_* \in \Theta^\circ$, then there exists a root of (4) converging a.s. to θ_*.

Next, to study the convergence in law of $\hat{\theta}_n$, we will need the following technical assumption about the regularity of $c(t;\theta)$ as a function of θ. Let $u(t)$ and $v(t)$ denote the real and imaginary parts of $c(t)$, that is, $c(t) = u(t) + \mathrm{i}v(t)$. Analogously, for each $c(t;\theta) \in \mathcal{F}$, we write $c(t;\theta) = u(t;\theta) + \mathrm{i}v(t;\theta)$ and $c_n(t) = u_n(t) + \mathrm{i}v_n(t)$.

Assumption 2 For W-almost all t, $u(t;\theta)$ and $v(t;\theta)$ are twice continuously differentiable on Θ_1, where $\Theta_1 \subseteq \Theta$ is an open neighborhood of θ_*. In addition, $\frac{\partial}{\partial\theta}u(t;\theta)$, $\frac{\partial}{\partial\theta}v(t;\theta)$, $\frac{\partial}{\partial\theta}u(t;\theta)\frac{\partial}{\partial\theta}u(t;\theta)'$, $\frac{\partial}{\partial\theta}u(t;\theta)\frac{\partial}{\partial\theta}v(t;\theta)'$, $\frac{\partial}{\partial\theta}v(t;\theta)\frac{\partial}{\partial\theta}v(t;\theta)'$, $\frac{\partial^2}{\partial\theta\partial\theta'}u(t;\theta)$ and $\frac{\partial^2}{\partial\theta\partial\theta'}v(t;\theta)$ are uniformly $(\forall\theta\in\Theta_1)$ bounded by W-integrable functions.

Assumption 2 implies that $I_n(\theta)$ is twice continuously differentiable on Θ_1 and that it can be differentiated under the integral sign. Let $D_1(\theta) = (D_{11}(\theta),\ldots,D_{1k}(\theta))'$, with $D_{1j}(\theta) = \frac{\partial}{\partial\theta_j}D^2(c(t),c(t;\theta))$, $1\le j\le k$, and $D_2(\theta) = (D_{2jl}(\theta))$, with $D_{2jl}(\theta) = \frac{1}{2}\frac{\partial^2}{\partial\theta_j\partial\theta_l}D^2(c(t),c(t;\theta))$, $1\le j,l\le k$.

Now, we are ready to derive the asymptotic normality of $\hat\theta_n$.

Theorem 2 *Suppose that Assumptions 1 and 2 hold, then*

$$\sqrt{n}(\hat\theta_n-\theta_*) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} D_2(\theta_*)^{-1}h(X_j;\theta_*) + o_P(1), \tag{5}$$

with $h(x;\theta) = (h_1(x;\theta),\ldots,h_k(x;\theta))'$,

$$h_j(x;\theta) = \int\{\cos(t'x)-u(t;\theta)\}\frac{\partial}{\partial\theta_j}u(t;\theta)\mathrm{d}W(t) + \int\{\sin(t'x)-v(t;\theta)\}\frac{\partial}{\partial\theta_j}v(t;\theta)\mathrm{d}W(t),$$

$1\le j\le k$, *and thus*

$$\sqrt{n}(\hat\theta_n-\theta_*)\xrightarrow{\mathcal{L}} N_k(0,\Sigma),$$

where $\Sigma = D_2(\theta_*)^{-1}A(\theta_*)D_2(\theta_*)^{-1}$ *and* $A(\theta) = E\{h(X;\theta)h(X;\theta)'\}$.

As observed after Theorem 1, the statement in Theorem 2 asserts that there exists a root of (4), $\hat\theta_n$, such that $\sqrt{n}(\hat\theta_n-\theta_*)$ converges in law to a zero mean k-variate normal distribution.

The next result gives the asymptotic behavior of $D(c_n(t),c(t;\hat\theta_n))$ as an estimator of $D(c(t),c(t;\theta_*))$.

Theorem 3 (a) *Suppose that Assumption 1 holds,* $\theta_*\in\Theta^\circ$ *and* $u(t;\theta)$ *and* $v(t;\theta)$ *are continuous as functions of* θ *for all* t, *then*

$$D(c_n(t),c(t;\hat\theta_n))\xrightarrow{\text{a.s.}} D(c(t),c(t;\theta_*)) = D(c(t),\mathcal{F}).$$

(b) *If assumptions in Theorem 2 hold and* $c(t)\in\mathcal{F}$, *then*

$$nD^2(c_n(t),c(t;\hat\theta_n))\xrightarrow{\mathcal{L}}\sum_{j=1}^{\infty}\lambda_j^A\chi_{1j}^2,$$

where $\chi^2_{11}, \chi^2_{12}, \ldots$ *are independent Chi-square variates with one degree of freedom, the set* $\{\lambda_j^A\}$ *are the eigenvalues of operator* $\mathcal{A}$ *defined on* $L_2(\mathbb{R}^d, F(\cdot;\theta)) = \{g : \mathbb{R}^d \to \mathbb{R},\ \int g(x)^2 \mathrm{d}F(x;\theta) < \infty\}$ *by*

$$\mathcal{A}\varpi(x) = \int k^c(x, y; \theta)\varpi(y)\mathrm{d}F(y;\theta),$$

with $k^c(x, y; \theta) = k(x, y; \theta) - h(x;\theta)' D_2(\theta)^{-1} h(y;\theta)$ *and*

$$k(x, y; \theta) = \int \{\cos(t'x) - u(t;\theta)\}\{\cos(t'y) - u(t;\theta)\}\mathrm{d}W(t) + \int \{\sin(t'x) - v(t;\theta)\}\{\sin(t'y) - v(t;\theta)\}\mathrm{d}W(t). \quad (6)$$

(c) *If assumptions in Theorem 2 hold,* $c(t) \notin \mathcal{F}$ *and* W *such that* $\sigma^2(\theta_*) = \mathrm{var}\{\rho(X)\} > 0$, *where* $\rho(x) = E\{k(X_1, X_2; \theta_*) \mid X_1 = x\}$, *then*

$$\sqrt{n}\left\{D^2(c_n(t), c(t;\hat{\theta}_n)) - D^2(c(t), c(t;\theta_*))\right\} \xrightarrow{\mathcal{L}} N(0, \sigma^2(\theta_*)).$$

A simple way to ensure that $\sigma^2(\theta_*) > 0$ is by taking W satisfying (2). Recall that in this case $D(c_1, c_2)$ is a true distance between distributions.

Remark 1 Let $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$ be an arbitrary estimator of θ satisfying $\tilde{\theta}_n \xrightarrow{\text{a.s.}(P)} \theta_0 \in \Theta^\circ$. If $u(t;\theta)$ and $v(t;\theta)$ are continuous as functions of θ for all t, then we also have that $D(c_n(t), c(t;\tilde{\theta}_n)) \xrightarrow{\text{a.s.}(P)} D(c(t), c(t;\theta_0))$, but $D(c(t), c(t;\theta_0)) \neq D(c(t), \mathcal{F})$ whenever $\theta_0 \neq \theta_*$.

To end this section, we deal with the case of estimating parameters of two families of distributions, $\mathcal{F}$ and $\mathcal{G}$. In this setting, we will use the same notation as before plus a subindex indicating the family. Specifically, $\mathcal{F} = \{c_F(t;\theta) = u_F(t;\theta) + \mathrm{i}v_F(t;\theta),\ \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ or, equivalently, $\mathcal{F} = \{F(t;\theta),\ \theta \in \Theta\}$; $\mathcal{G} = \{c_G(t;\gamma) = u_G(t;\gamma) + \mathrm{i}v_G(t;\gamma),\ \gamma \in \Gamma\}$, where $\Gamma \subseteq \mathbb{R}^r$, or equivalently, $\mathcal{G} = \{G(t;\gamma),\ \gamma \in \Gamma\}$; $I_n^F(\theta) = D^2(c_n(t), c_F(t;\theta))$ and $I_n^G(\gamma) = D^2(c_n(t), c_G(t;\gamma))$; we write $D_{1F}(\theta)$ and $D_{1G}(\gamma)$ for the vectors of the first derivatives of $D^2(c(t), c_F(t;\theta))$ and $D^2(c(t), c_G(t;\gamma))$, respectively; analogously, we write $D_{2F}(\theta)$ and $D_{2G}(\gamma)$ for 0.5 times the Hessian matrices of $D^2(c(t), c_F(t;\theta))$ and $D^2(c(t), c_G(t;\gamma))$, respectively; we denote by $h_F(x;\theta)$ and $h_G(x;\gamma)$ the vector h appearing in Theorem 2 for the families $\mathcal{F}$ and $\mathcal{G}$, respectively; analogously, we write $A_F(\theta)$ and $A_G(\gamma)$ for the matrix A appearing in Theorem 2 for the families $\mathcal{F}$ and $\mathcal{G}$, respectively; let $B_{FG}(\theta,\gamma)$ denote the $k \times r$-matrix $E\{h_F(X;\theta)\, h_G(X;\gamma)'\}$; and finally, let

$$D_{2FG}(\theta,\gamma) = \begin{pmatrix} D_{2F}(\theta) & 0 \\ 0 & D_{2G}(\gamma) \end{pmatrix}, \quad A_{FG}(\theta,\gamma) = \begin{pmatrix} A_F(\theta) & B_{FG}(\theta,\gamma) \\ B_{FG}(\theta,\gamma)' & A_G(\gamma) \end{pmatrix}.$$

Corollary 1 (a) *Suppose that the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 3(a), then*

$$m_F(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} -D^2(c_F(t;\theta_*), c_G(t;\gamma_*)),$$

where $m_F(\theta,\gamma) = E_\theta\{D^2(c_n(t), c_F(t;\theta)) - D^2(c_n(t), c_G(t;\gamma))\}$ and E_θ denotes expectation assuming that the data have CF $c_F(t;\theta)$.

(b) *Suppose that the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 2, then*

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_*\end{pmatrix} \xrightarrow{\mathcal{L}} N_{k+r}(0, \Sigma_{FG}(\theta_*, \gamma_*)),$$

where $\Sigma_{FG}(\theta,\gamma) = D_{2FG}(\theta,\gamma)^{-1} A_{FG}(\theta,\gamma) D_{2FG}(\theta,\gamma)^{-1}$.

Remark 2 Let $\tilde{\theta}_n = \tilde{\theta}_n(X_1,\ldots,X_n)$ and $\tilde{\gamma}_n = \tilde{\gamma}_n(X_1,\ldots,X_n)$ be arbitrary estimators of θ and γ, respectively, satisfying

$$\tilde{\theta}_n \xrightarrow{\text{a.s.}(P)} \theta_0 \in \Theta^\circ \quad \text{and} \quad \tilde{\gamma}_n \xrightarrow{\text{a.s.}(P)} \gamma_0 \in \Gamma^\circ. \tag{7}$$

If the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 3(a), then we also have that $m_F(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\text{a.s.}(P)} -D^2(c_F(t;\theta_0), c_G(t;\gamma_0))$, where $m_F(\theta,\gamma)$ is as defined in Corollary 1(a).

3 Model selection

This section deals with the CF approach to the problem studied in Vuong (1989). With this aim, we first define the problem; then we give some results that will let us provide several decision rules for separate, overlapping and nested models. Along this and next sections, we will assume that the unknown parameters are estimated by means of their ISE estimators. Later in Sect. 5, we will see that some results may change if other estimators are used. From now on, we will assume that (2) holds, so that D is a true distance between distributions.

3.1 Statement of the problem and main results

Given two possibly misspecified parametric models, $\mathcal{F}$ and $\mathcal{G}$, which may be nested, non-nested or overlapping, the problem of model selection consists in testing if the two competing models are equally close to the true distribution, against the hypothesis that one model is closer than the other. Here, the closeness is measured by means of the distance D defined in (1). Therefore, the problem is that of constructing a test for

$$H_0: \; D^2(c(t), c_F(t;\theta_*)) = D^2(c(t), c_G(t;\gamma_*))$$

against the alternatives

$$H_{1F} : D^2(c(t), c_F(t;\theta_*)) < D^2(c(t), c_G(t;\gamma_*)) \quad \text{or}$$
$$H_{1G} : D^2(c(t), c_F(t;\theta_*)) > D^2(c(t), c_G(t;\gamma_*)).$$

Such a test is of practical interest since rejection of H_0 in favor of H_{1F} (H_{1G}) would indicate that $F(x;\theta_*)$ ($G(x;\gamma_*)$) is a better approximation to the true distribution.

The quantity $\mu_{FG}(\theta_*, \gamma_*) = D^2(c(t), c_F(t;\theta_*)) - D^2(c(t), c_G(t;\gamma_*))$ is unknown, but from Theorem 3(a), it can be consistently estimated through $T(\hat{\theta}_n, \hat{\gamma}_n)$, where

$$T(\theta, \gamma) = D^2(c_n(t), c_F(t;\theta)) - D^2(c_n(t), c_G(t;\gamma)). \tag{8}$$

This difference converges to 0 under the null hypothesis H_0, but it converges to a strictly negative or positive constant under alternatives. Thus, the null hypothesis H_0 should be rejected for "large" or "small" values of $T(\hat{\theta}_n, \hat{\gamma}_n)$. To decide what is "large" or "small", we must calculate the null distribution of $T(\hat{\theta}_n, \hat{\gamma}_n)$, or at least a consistent approximation to it. Since the exact null distribution of $T(\hat{\theta}_n, \hat{\gamma}_n)$ is clearly unknown, we approximate it through its asymptotic null distribution. With this aim, we first observe that

$$T(\theta, \gamma) = \frac{1}{n}\sum_{j=1}^{n} \xi(X_j, \theta, \gamma),$$

where $\xi(x, \theta, \gamma) = \int \{u_G(t;\gamma) - u_F(t;\theta)\}\{2\cos(t'x) - u_G(t;\gamma) - u_F(t;\theta)\}w(t)\mathrm{d}t + \int \{v_G(t;\gamma) - v_F(t;\theta)\}\{2\sin(t'x) - v_G(t;\gamma) - v_F(t;\theta)\}w(t)\mathrm{d}t$.

Theorem 4 *Suppose that the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 2.*

(a) *If $c_F(t;\theta_*) = c_G(t;\gamma_*)$, then $nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} T_1 = \sum_{j=1}^{k+r} \lambda_j \chi^2_{1j}$, where $\chi^2_{11}, \chi^2_{12}, \ldots$, are independent Chi-square variates with one degree of freedom and the set $\{\lambda_j\}$ are the eigenvalues of the matrix*

$$S_{FG}(\theta_*, \gamma_*) = \begin{pmatrix} -D_{2F}(\theta_*) & 0 \\ 0 & D_{2G}(\gamma_*) \end{pmatrix} \Sigma_{FG}(\theta_*, \gamma_*).$$

(b) *If $c_F(t;\theta_*) = c_G(t;\gamma_*)$ then $\sup_x \left| P_0\{nT(\hat{\theta}_n, \hat{\gamma}_n) \le x\} - P_*(\hat{T}_1 \le x) \right| \xrightarrow{\text{a.s.}} 0$, where $\hat{T}_1 = \sum_{j=1}^{k+r} \hat{\lambda}_j \chi^2_{1j}$ and $\{\hat{\lambda}_j\}$ are the eigenvalues of the matrix $\hat{S}_{FG}(\hat{\theta}_n, \hat{\gamma}_n)$, having the same structure as $S_{FG}(\theta, \gamma)$ with $A_F(\theta)$, $A_G(\theta)$, $B_{FG}(\theta, \gamma)$, $D_{2F}(\theta)$ and $D_{2G}(\gamma)$ replaced by $\hat{A}_F(\theta) = \frac{1}{n}\sum_{j=1}^{n} h_F(X_j;\theta)h_F(X_j;\theta)'$, $\hat{A}_G(\gamma) = \frac{1}{n}\sum_{j=1}^{n} h_G(X_j;\gamma)h_G(X_j;\gamma)'$, $\hat{B}_{FG}(\theta, \gamma) = \frac{1}{n}\sum_{j=1}^{n} h_F(X_j;\theta)\, h_G(X_j;\gamma)'$, $\hat{D}_{2F}(\theta) = \frac{1}{2}\frac{\partial^2}{\partial\theta\partial\theta'} I_n^F(\theta)$ and $\hat{D}_{2G}(\gamma) = \frac{1}{2}\frac{\partial^2}{\partial\gamma\partial\gamma'} I_n^G(\gamma)$, respectively. Moreover, if H_{1F} holds then $P_*\{\hat{T}_1 > nT(\hat{\theta}_n, \hat{\gamma}_n)\} \xrightarrow{\text{a.s.}} 1$, while if H_{1G} holds then $P_*\{\hat{T}_1 < nT(\hat{\theta}_n, \hat{\gamma}_n)\} \xrightarrow{\text{a.s.}} 1$.*

(c) *If* $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$, *then* $\sqrt{n}\left\{T(\hat{\theta}_n,\hat{\gamma}_n) - \mu_{FG}(\theta_*,\gamma_*)\right\} \xrightarrow{\mathcal{L}} N(0,\sigma^2_{FG}(\theta_*,\gamma_*))$, *with* $\sigma^2_{FG}(\theta,\gamma) = \text{var}\{\xi(X,\theta,\gamma)\} > 0$.

Remark 3 Observe that, in spite of using rather different discrepancy measures between populations, the results in Theorem 4 are quite similar to those in Theorem 3.3 in Vuong (1989), in the sense that the limit distributions are of the same type, that is, a linear combination of Chi-square variates in case (a), and a zero mean normal distribution otherwise.

Theorem 4 says that the limiting distribution of $T(\hat{\theta}_n,\hat{\gamma}_n)$ depends on whether or not $c_F(t;\theta_*) = c_G(t;\gamma_*)$. Therefore, it is important to know if such equality holds. The result in the next theorem, which is similar to that in Lemma 4.1 in Vuong (1989), will be useful in this respect.

Theorem 5 *Let* $\sigma^2_{FG}(\theta,\gamma)$ *be as defined in Theorem 4. Then,* $\sigma^2_{FG}(\theta,\gamma) = 0 \iff c_F(t;\theta) = c_G(t;\gamma)$, $\forall t$.

Therefore, testing for $c_F(t;\theta_*) = c_G(t;\gamma_*)$ versus $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$ is equivalent to testing for

$$H_{0\sigma}: \ \sigma^2_{FG}(\theta_*,\gamma_*) = 0,$$

versus

$$H_{1\sigma}: \ \sigma^2_{FG}(\theta_*,\gamma_*) > 0.$$

With this aim, taking into account that $\sigma^2_{FG}(\theta,\gamma) = \text{var}\{\xi(X,\theta,\gamma)\}$, we estimate $\sigma^2_{FG}(\theta_*,\gamma_*)$ by means of $\hat{\sigma}^2_{FG}(\hat{\theta}_n,\hat{\gamma}_n)$, with

$$\hat{\sigma}^2_{FG}(\theta,\gamma) = \frac{1}{n}\sum_{j=1}^{n}\xi^2(X_j,\theta,\gamma) - \left\{\frac{1}{n}\sum_{j=1}^{n}\xi(X_j,\theta,\gamma)\right\}^2. \tag{9}$$

This estimator satisfies the following.

Theorem 6 (a) *If the families* $\mathcal{F}$ *and* $\mathcal{G}$ *both satisfy Assumption 1,* $\theta_* \in \Theta^\circ$, $\gamma_* \in \Gamma^\circ$, $u_F(t;\theta)$ *and* $v_F(t;\theta)$ *are continuous functions of* θ *for each* t *and* $u_G(t;\gamma)$ *and* $v_G(t;\gamma)$ *are continuous functions of* γ *for each* t, *then* $\hat{\sigma}^2_{FG}(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{\text{a.s.}} \sigma^2_{FG}(\theta_*,\gamma_*)$.

(b) *If the families* $\mathcal{F}$ *and* $\mathcal{G}$ *both satisfy the assumptions in Theorem 2 and* $\sigma^2_{FG}(\theta_*,\gamma_*) = 0$, *then* $T_\sigma = 0.25n\hat{\sigma}^2_{FG}(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{\mathcal{L}} \sum_{j=1}^{k+r}\lambda^\sigma_j\chi^2_{1j}$, *where* $\chi^2_{11},\chi^2_{12},\ldots,$ *are independent Chi-square variates with one degree of freedom and the set* $\{\lambda^\sigma_j\}$ *are the eigenvalues of the matrix* $A_{FG}(\theta_*,\gamma_*)\Sigma_{FG}(\theta_*,\gamma_*)$.

(c) *Suppose that the families* $\mathcal{F}$ *and* $\mathcal{G}$ *both satisfy the assumptions in Theorem 2. If* $H_{0\sigma}$ *is true, then* $\sup_x \left|P_{0\sigma}(T_\sigma \leq x) - P_*(\hat{T}_\sigma \leq x)\right| \xrightarrow{\text{a.s.}} 0$; *otherwise,* $P_*(\hat{T}_\sigma \leq T_\sigma) \xrightarrow{\text{a.s.}} 1$, *where* $\hat{T}_\sigma = \sum_{j=1}^{k+r}\hat{\lambda}^\sigma_j\chi^2_{1j}$ *and* $\{\hat{\lambda}^\sigma_j\}$ *are the eigenvalues of the matrix*

$\hat{A}_{FG}(\hat{\theta}_n, \hat{\gamma}_n)\hat{\Sigma}_{FG}(\hat{\theta}_n, \hat{\gamma}_n)$, having the same structure as $A_{FG}(\theta, \gamma)\Sigma_{FG}(\theta, \gamma)$ with $A_F(\theta)$, $A_G(\theta)$, $B_{FG}(\theta, \gamma)$, $D_{2F}(\theta)$ and $D_{2G}(\gamma)$ replaced by $\hat{A}_F(\theta)$, $\hat{A}_G(\gamma)$, $\hat{B}_{FG}(\theta, \gamma)$, $\hat{D}_{2F}(\theta)$ and $\hat{D}_{2G}(\gamma)$, respectively.

As a consequence of Theorem 6, for testing $H_{0\sigma}$ versus $H_{1\sigma}$, the test that rejects $H_{0\sigma}$ when $T_\sigma \geq \hat{t}_{\sigma,1-\alpha}$, where $\hat{t}_{\sigma,1-\alpha}$ is such that $P_*(\hat{T}_\sigma \leq \hat{t}_{\sigma,1-\alpha}) = 1-\alpha$, for some $\alpha \in (0,1)$, is consistent against fixed alternatives. It rejects the null hypothesis with probability tending to 1 when the null hypothesis is false and it is also asymptotically correct in the sense that it asymptotically has the desired level α. Note that, since the matrices $A_{FG}(\theta_*, \gamma_*)$, $\Sigma_{FG}(\theta_*, \gamma_*)$ are unknown, we cannot employ the asymptotic null distribution of T_σ for testing $H_{0\sigma}$, but a consistent estimator of it. A further practical problem is the calculation of $\hat{t}_{\sigma,1-\alpha}$, since the distribution of a linear combination of χ^2 variates is, in general, unknown. To overcome it, the conditional distribution of $\hat{T}_\sigma$, given the data, can be approximated either by simulation or by some numerical method (see for example Kotz et al. 1967; Castaño-Martínez and López-Blázquez 2005).

If we can assume that $c_F(t; \theta_*) = c_G(t; \gamma_*)$, then reasoning analogously, it follows from Theorem 4 that the test that rejects H_0 when $nT(\hat{\theta}_n, \hat{\gamma}_n) \leq \hat{t}_{\alpha_1}$ or $nT(\hat{\theta}_n, \hat{\gamma}_n) \geq \hat{t}_{1-\alpha_2}$, for any $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = \alpha$, where $\hat{t}_\beta$ is such that $P_*(\hat{T}_1 \leq \hat{t}_\beta) = \beta$, is asymptotically correct.

Now, we are ready to test for H_0 versus H_{1F} or H_{1G}. Recall that the models $\mathcal{F}$ and $\mathcal{G}$ can be nested, nonnested or overlapping. These three cases will be separately studied.

3.2 Nonnested or separate models

Two models, $\mathcal{F}$ and $\mathcal{G}$, are said to be nonnested or separate if $\mathcal{F} \cap \mathcal{G} = \emptyset$. In this case, we always have that $c_F(t; \theta_*) \neq c_G(t; \gamma_*)$. Therefore, as an immediate consequence of Theorems 3, 4 and 6, we have

(i) Under H_0, $D = \sqrt{n}T(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\sigma}_{FG}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} N(0,1)$.
(ii) Under H_{1F}, $D \xrightarrow{\text{a.s.}} -\infty$.
(iii) Under H_{1G}, $D \xrightarrow{\text{a.s.}} \infty$.

Thus, for fixed $\alpha \in (0,1)$, the decision rule is: if $D < -Z_{1-\alpha/2}$ then select model $\mathcal{F}$; if $D > Z_{1-\alpha/2}$ then select model $\mathcal{G}$; otherwise conclude that there is not sufficient evidence to discriminate between the competing models $\mathcal{F}$ and $\mathcal{G}$, where $\Phi(Z_{1-\alpha/2}) = 1-\alpha/2$, Φ being the CDF of a univariate standard normal distribution, $N(0,1)$. From statements (i)–(iii) above, it follows that this test is asymptotically correct and consistent, in the sense that it chooses the model $\mathcal{F}(\mathcal{G})$ with probability tending to 1 when it is closer than $\mathcal{G}(\mathcal{F})$ to the true population.

3.3 Overlapping models

Two models, $\mathcal{F}$ and $\mathcal{G}$, are said to be overlapping if $\mathcal{F} \cap \mathcal{G} \neq \emptyset$, $\mathcal{F} \not\subseteq \mathcal{G}$ and $\mathcal{G} \not\subseteq \mathcal{F}$. In this case, since $\mathcal{F} \cap \mathcal{G} \neq \emptyset$, it may happen that $c_F(t; \theta_*) = c_G(t; \gamma_*)$. Therefore, we

must first test $H_{0\sigma}$ versus $H_{1\sigma}$. Since $H_{0\sigma}$ is included in H_0, if $H_{0\sigma}$ cannot be rejected, then it is concluded that we cannot discriminate between the competing models. If $H_{0\sigma}$ is rejected, then H_0 can still be true. In this case, statements (i)–(iii) in Sect. 3.2 also apply.

Thus, for fixed $\alpha_1, \alpha_2 \in (0, 1)$, the decision rule is: if $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$ and $D < -Z_{1-\alpha_2/2}$, then select model $\mathcal{F}$; if $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$ and $D > Z_{1-\alpha_2/2}$, then select model $\mathcal{G}$; otherwise, conclude that there is not sufficient evidence to discriminate between the competing models $\mathcal{F}$ and $\mathcal{G}$.

Reasoning as in Vuong (1989), the sequential procedure described above has a significance level which is asymptotically bounded above by the maximum of the asymptotic significance levels α_1 and α_2. The procedure is also consistent in the sense explained in Sect. 3.2.

The above sequential procedure is needed for investigating if $c_F(t; \theta_*) = c_G(t; \gamma_*)$, but if we know that at least one of the models is correctly specified, such procedure could be shortened because of the following property, which is analogous to Lemma 6.2 in Vuong (1989).

Lemma 1 *If the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy Assumption 1, are overlapping and at least one model is correctly specified, then the following statements are equivalent:*

(a) $F \in \mathcal{F} \cap \mathcal{G}$.
(b) $c_F(t; \theta_*) = c_G(t; \gamma_*)$.
(c) $D^2(c(t), c_F(t; \theta_*)) = D^2(c(t), c_G(t; \gamma_*))$.

Therefore, if at least one model is correctly specified, H_0 is equivalent to $c_F(t; \theta_*) = c_G(t; \gamma_*)$. Hence, as an immediate consequence of Theorems 3 and 4, we have that

(iv) Under H_0, $nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} T_1$.
(v) Under H_{1F}, $nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} -\infty$.
(vi) Under H_{1G}, $nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} \infty$.

Thus, for fixed $\alpha = \alpha_1 + \alpha_2 \in (0, 1)$, with $\alpha_1, \alpha_2 > 0$, when the families are overlapping and at least one model is correctly specified, the decision rule is: if $nT(\hat{\theta}_n, \hat{\gamma}_n) < \hat{t}_{\alpha_1}$, then select model $\mathcal{F}$; if $nT(\hat{\theta}_n, \hat{\gamma}_n) > \hat{t}_{1-\alpha_2}$, then select model $\mathcal{G}$; otherwise, conclude that there is not sufficient evidence to discriminate between the competing models $\mathcal{F}$ and $\mathcal{G}$. From statements (iv)–(vi) above, it follows that this test is asymptotically correct and consistent.

3.4 Nested models

The model $\mathcal{G}$ is said to be nested in model $\mathcal{F}$ if $\mathcal{G} \subset \mathcal{F}$. Note that in this case model $\mathcal{G}$ can never give a better fit to the data than $\mathcal{F}$; therefore, we only have one alternative, H_0 versus H_{1F}. To study this case, we make the following assumption on the parametrization of these families.

Assumption 3 There exists a function $\phi : \Gamma \to \Theta$ such that for any $\gamma \in \Gamma$, $c_G(t; \gamma) = c_F(t; \phi(\gamma))$.

The following result is analogous to Lemma 7.1 in Vuong (1989).

Lemma 2 *If the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy Assumption 1, $\mathcal{G}$ is nested in $\mathcal{F}$ and Assumption 3 holds, then the following statements are equivalent:*

(a) $\theta_* = \phi(\gamma_*)$.
(b) $\theta_* \in \phi(\Gamma)$.
(c) $c_F(t;\theta_*) = c_G(t;\gamma_*)$.
(d) $D^2(c(t), c_F(t;\theta_*)) = D^2(c(t), c_G(t;\gamma_*))$.

Therefore, if $\mathcal{G}$ is nested in $\mathcal{F}$, then H_0 is equivalent to $c_F(t;\theta_*) = c_G(t;\gamma_*)$. Hence, statements (iv) and (v) in Sect. 3.3 apply. Thus, for fixed $\alpha \in (0,1)$, the decision rule is: if $nT(\hat{\theta}_n, \hat{\gamma}_n) < \hat{t}_\alpha$, then select model $\mathcal{F}$; otherwise, conclude that there is not sufficient evidence to discriminate between the competing models $\mathcal{F}$ and $\mathcal{G}$. This test is asymptotically correct and consistent.

In the above paragraph, the null distribution of the test statistic $nT(\hat{\theta}_n, \hat{\gamma}_n)$ is estimated as in Theorem 4(b). Nevertheless, in the nested case, calculations can be simplified because, as the next lemma shows, the nonzero eigenvalues of $S_{FG}(\theta_*, \gamma_*)$ coincide with those of $S_{1FG}(\theta_*, \gamma_*)$ below, which has lower dimensions than $S_{FG}(\theta_*, \gamma_*)$,

$$S_{1FG}(\theta_*, \gamma_*) = A_F(\theta_*)\frac{\partial}{\partial \gamma'}\phi(\gamma_*)D_{2G}(\gamma_*)^{-1}\frac{\partial}{\partial \gamma}\phi'(\gamma_*) - A_F(\theta_*)D_{2F}(\theta_*)^{-1},$$

and, therefore, under the null hypothesis $nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} \sum_{j=1}^{k} \lambda_j \chi^2_{1j}$, where the set $\{\lambda_j\}$ are the eigenvalues of the matrix $S_{1FG}(\theta_*, \gamma_*)$.

Lemma 3 *If the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 2, $\mathcal{G}$ is nested in $\mathcal{F}$, Assumption 3 holds and the function ϕ in Assumption 3 is twice continuously differentiable in an open neighborhood of γ_*, then*

(a) *$S_{FG}(\theta_*, \gamma_*)$ and $S_{1FG}(\theta_*, \gamma_*)$ have the same nonzero eigenvalues.*
(b) *If H_0 is true, then $\sup_x \left| P_0\{nT(\hat{\theta}_n, \hat{\gamma}_n) \le x\} - P_*(\hat{T}_2 \le x) \right| \xrightarrow{\text{a.s.}} 0$; otherwise, $P_*\{\hat{T}_2 > nT(\hat{\theta}_n, \hat{\gamma}_n)\} \xrightarrow{\text{a.s.}} 1$, where $\hat{T}_2 = \sum_{j=1}^{k} \hat{\lambda}_{1j}\chi^2_{1j}$ and $\{\hat{\lambda}_{1j}\}$ are the eigenvalues of the matrix $\hat{S}_{1FG}(\hat{\theta}_n, \hat{\gamma}_n)$, having the same structure as $S_{1FG}(\theta, \gamma)$ with $A_F(\theta)$, $D_{2F}(\theta)$ and $D_{2G}(\gamma)$ replaced by $\hat{A}_F(\theta)$, $\hat{D}_{2F}(\theta)$ and $\hat{D}_{2G}(\gamma)$, respectively.*

As an immediate consequence of Lemma 3, in the above rule. we can replace $nT(\hat{\theta}_n, \hat{\gamma}_n) < \hat{t}_\alpha$ by $nT(\hat{\theta}_n, \hat{\gamma}_n) < \hat{t}_{2,\alpha}$, where $\hat{t}_{2,\alpha}$ is such that $P_*(\hat{T}_2 \le \hat{t}_{2,\alpha}) = \alpha$.

Note that if model $\mathcal{F}$ is correctly specified and Assumption 3 holds, from Lemma 2 it follows that the problem of testing H_0 versus H_{1F} is equivalent to the classical parametric problem of testing $H_{0,\theta} : \theta \in \phi(\Gamma)$ versus $H_{1,\theta} : \theta \notin \phi(\Gamma)$. In this setting and when H_0 is true, the test in Vuong (1989) is asymptotically distribution free. Specifically, it asymptotically has a χ^2_{k-r} distribution. Routine calculations show that if $A_F(\theta_*)D_{2F}(\theta_*)^{-1} = I_k$, then this good result is also true for the approach studied in this work, but unfortunately such an equality does not hold.

Finally, observe that in the nested case, testing H_0 versus H_{1F} is equivalent to testing $H_{0\sigma}$ versus $H_{1\sigma}$. Therefore, we can also consider the following rule: if $T_\sigma \geq \hat{t}_{\sigma,1-\alpha}$, then select model $\mathcal{F}$; otherwise, conclude that there is not sufficient evidence to discriminate between the competing models $\mathcal{F}$ and $\mathcal{G}$. From Theorem 6 and Lemma 2, this test is asymptotically correct and consistent.

4 Cox approach for testing two separate families

Along this section, we will assume that $\mathcal{F}$ and $\mathcal{G}$ are two separate families. For testing

$$H_F: \mathcal{F} \text{ is correctly specified} \quad \text{versus} \quad H_G: \mathcal{G} \text{ is correctly specified}, \tag{10}$$

we consider the test statistic $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)$, where

$$T_{\text{CoxF}}(\theta, \gamma) = T(\theta, \gamma) - m_F(\theta, \gamma), \tag{11}$$

with $T(\theta, \gamma)$ and $m_F(\theta, \gamma)$ as defined in (8) and Corollary 1, respectively. Observe that $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)$ is a CF version of the Cox statistic for the testing problem (10) obtained by replacing the Kullback–Leibler distance by D^2. From Theorem 3(a) and Corollary 1(a),

$$\begin{aligned} T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} & D^2(c(t), c_F(t; \theta_*)) - D^2(c(t), c_G(t; \gamma_*)) \\ & + D^2(c_F(t; \theta_*), c_G(t; \gamma_*)). \end{aligned}$$

As a consequence, if H_F is true then $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} 0$, while if H_G is true then $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\text{a.s.}} 2D^2(c_F(t; \theta_*), c_G(t; \gamma_*)) > 0$. Thus, the hypothesis H_F should be rejected in favor of H_G for "large" values of $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)$. To decide what is "large" we must calculate the null distribution of $T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)$. Since the exact null distribution of $T(\hat{\theta}_n, \hat{\gamma}_n)$ is unknown, we approximate it through its asymptotic null distribution. Let var_θ and cov_θ denote the variance and covariance, respectively, when the data have CF $c_F(t; \theta)$.

Theorem 7 *If the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy the assumptions in Theorem 2, then under H_F*

$$\sqrt{n} T_{\text{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{1FG}(\theta_*, \gamma_*)),$$

where

$$\sigma^2_{1FG}(\theta, \gamma) = \delta' \begin{pmatrix} \text{var}_\theta\{\xi(X, \theta, \gamma)\} & \text{cov}_\theta\{\xi(X, \theta, \gamma), H_F(X, \theta)\}' \\ \text{cov}_\theta\{\xi(X, \theta, \gamma), H_F(X, \theta)\} & \text{var}_\theta\{H_F(X, \theta)\} \end{pmatrix} \delta,$$

$H_F(X; \theta) = D_{2F}(\theta)^{-1} h_F(X; \theta)$, $\delta' = (1, \psi_F(\theta, \gamma)')$ *and* $\psi_F(\theta, \gamma) = \frac{\partial}{\partial \theta} D^2(c_F(t; \theta), c_G(t; \gamma))$.

For the result in Theorem 7 to be useful to approximate the distribution of $T_{\mathrm{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)$ under H_F, we need a consistent estimator of σ^2_{1FG}. We will consider $\hat{\sigma}^2_{1FG}(\hat{\theta}_n, \hat{\gamma}_n)$, where

$$\hat{\sigma}^2_{1FG}(\theta, \gamma) = \delta' \begin{pmatrix} \widehat{\mathrm{var}}\{\xi(X, \theta, \gamma)\} & \widehat{\mathrm{cov}}\{\xi(X, \theta, \gamma), \hat{H}_F(X, \theta)\}' \\ \widehat{\mathrm{cov}}\{\xi(X, \theta, \gamma), \hat{H}_F(X, \theta)\} & \widehat{\mathrm{var}}\{\hat{H}_F(X, \theta)\} \end{pmatrix} \delta,$$

$$\widehat{\mathrm{var}}\{\xi(X, \theta, \gamma)\} = \hat{\sigma}^2_{FG}(\theta, \gamma),\ \hat{H}_F(x; \theta) = \hat{D}_{2F}(\theta)^{-1} h_F(x; \theta),$$

$$\begin{aligned} \widehat{\mathrm{cov}}\{\xi(X, \theta, \gamma), \hat{H}_F(X, \theta)\} &= \frac{1}{n}\sum_{j=1}^{n} \xi(X_j, \theta, \gamma)\hat{H}_F(X_j, \theta) \\ &\quad - \frac{1}{n}\sum_{j=1}^{n} \xi(X_j, \theta, \gamma)\frac{1}{n}\sum_{j=1}^{n} \hat{H}_F(X_j, \theta), \\ \widehat{\mathrm{var}}\{\hat{H}_F(X, \theta)\} &= \frac{1}{n}\sum_{j=1}^{n} \hat{H}_F(X_j, \theta)\hat{H}_F(X_j, \theta)' \\ &\quad - \frac{1}{n}\sum_{j=1}^{n} \hat{H}_F(X_j, \theta)\frac{1}{n}\sum_{j=1}^{n} \hat{H}_F(X_j, \theta)'. \end{aligned}$$

It can be easily checked that, under assumptions in Theorem 7, $\hat{\sigma}^2_{1FG}(\hat{\theta}_n, \hat{\gamma}_n)$ is a consistent estimator of σ^2_{1FG}. Therefore, if H_F is true then $D_{\mathrm{CoxF}} \xrightarrow{\mathcal{L}} N(0, 1)$, while if H_G is true then $D_{\mathrm{CoxF}} \xrightarrow{\text{a.s.}} \infty$, where $D_{\mathrm{CoxF}} = \sqrt{n}T_{\mathrm{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\sigma}_{1FG}(\hat{\theta}_n, \hat{\gamma}_n)$. Thus, the test that rejects H_F in favor of H_G when $D_{\mathrm{CoxF}} > Z_{1-\alpha}$ has asymptotically level α and is consistent.

If the roles of H_F and H_G as null and alternative hypotheses are interchanged, a test statistic $D_{\mathrm{CoxG}} = \sqrt{n}T_{\mathrm{CoxG}}(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\sigma}_{1GF}(\hat{\theta}_n, \hat{\gamma}_n)$ is obtained. Observe that $T_{\mathrm{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) = -T_{\mathrm{CoxG}}(\hat{\theta}_n, \hat{\gamma}_n) - 2m_F(\hat{\theta}_n, \hat{\gamma}_n)$, but since $\hat{\sigma}_{1FG}(\hat{\theta}_n, \hat{\gamma}_n)$ and $\hat{\sigma}_{1GF}(\hat{\theta}_n, \hat{\gamma}_n)$ are not related, the statistics D_{CoxF} and D_{CoxG} are different functions of the observations.

Following the original Cox approach, in practice we test the null hypothesis H_F versus the alternative H_G based on the test statistic D_{CoxF} and vice versa, that is, H_G versus the alternative H_F based on the test statistic D_{CoxG}. The decision rule is: reject both H_F and H_G if $D_{\mathrm{CoxF}} > Z_{1-\alpha}$ and $D_{\mathrm{CoxG}} > Z_{1-\alpha}$; reject neither H_F nor H_G if $D_{\mathrm{CoxF}} < Z_{1-\alpha}$ and $D_{\mathrm{CoxG}} < Z_{1-\alpha}$; reject H_F, but not H_G if $D_{\mathrm{CoxF}} > Z_{1-\alpha}$ and $D_{\mathrm{CoxG}} < Z_{1-\alpha}$; and reject H_G, but not H_F if $D_{\mathrm{CoxF}} < Z_{1-\alpha}$ and $D_{\mathrm{CoxG}} > Z_{1-\alpha}$.

5 On the use of other point estimators

In Sects. 3 and 4, we have estimated the parameters θ and γ by means of their ISE estimators because, according to the definition of θ_* and γ_*, the ISE estimators are their natural estimators. Nevertheless, motivated by the fact that in certain settings the

calculation of the ISE estimators can be time consuming (see for example Matsui and Takemura 2005), other estimators could be used. In such a case, although the proposed methods could be applied, some asymptotic properties may differ while, as seen in Remarks 1 and 2, others continue to be true, whenever the estimators satisfy certain assumptions. The aim of this section is to rewrite the results in Sects. 3.1 and 4 when instead of $\hat{\theta}_n$ and $\hat{\gamma}_n$, arbitrary estimators, say $\tilde{\theta}_n$ and $\tilde{\gamma}_n$, are employed. The results in this section will be stated without proofs, since they follow quite similar steps to those of the results in Sects. 3 and 4. In contrast to Sects. 3 and 4, where in addition to deriving the asymptotic distribution of certain statistics, we also provided estimators of such asymptotic distributions whenever they were unknown, to save space here we will not deal with the estimation, which could be done along the same lines. The decision rules given in Sects. 3.2–3.4 when θ and γ are estimated by means of $\tilde{\theta}_n$ and $\tilde{\gamma}_n$, respectively, will vary in an obvious way.

Assume that $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$ and $\tilde{\gamma}_n = \tilde{\gamma}_n(X_1, \ldots, X_n)$ are arbitrary estimators of θ and γ, respectively, satisfying (7). We will also assume that

$$(E\{h_F(X;\theta_0)\}', E\{h_G(X;\gamma_0)\}') \neq 0, \tag{12}$$

since otherwise the estimators $\tilde{\theta}_n$ and $\tilde{\gamma}_n$ are (asymptotically) equivalent to the ISE estimators. The following result is the analog of Theorem 4.

Theorem 8 *Suppose that $\tilde{\theta}_n$ and $\tilde{\gamma}_n$ satisfy (7) and (12), the families $\mathcal{F}$ and $\mathcal{G}$ both satisfy Assumption 2 with θ_* (γ_*) replaced by θ_0 (γ_0) and*

$$n^{\tau}\left(\tilde{\theta}_n' - \theta_0', \tilde{\gamma}_n' - \gamma_0'\right)' \xrightarrow{\mathcal{L}} Y_0 \tag{13}$$

for some $\tau > 0$ and for some random vector Y_0. Let $d_0' = (-2E\{h_F(X;\theta_0)\}', 2E\{h_G(X;\gamma_0)\}')$.

(a) *If $c_F(t;\theta_0) = c_G(t;\gamma_0)$, then $n^{\tau}T(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\mathcal{L}} d_0'Y_0$.*
(b) *Assume that $c_F(t;\theta_0) \neq c_G(t;\gamma_0)$.*
(b.1) *If $\tau < 1/2$, then $n^{\tau}\left\{T(\tilde{\theta}_n, \tilde{\gamma}_n) - \mu_{FG}(\theta_0, \gamma_0)\right\} \xrightarrow{\mathcal{L}} d_0'Y_0$.*
(b.2) *If $\tau = 1/2$ and*

$$n^{1/2}\left(n^{-1}\sum_{j=1}^{n}\xi(X_j, \theta_0, \gamma_0) - \mu_{FG}(\theta_0, \gamma_0), \tilde{\theta}_n' - \theta_0', \tilde{\gamma}_n' - \gamma_0'\right)' \xrightarrow{\mathcal{L}} Y_1, \tag{14}$$

for some random vector Y_1, then $n^{1/2}\{T(\tilde{\theta}_n, \tilde{\gamma}_n) - \mu_{FG}(\theta_0, \gamma_0)\} \xrightarrow{\mathcal{L}} d_1'Y_1$, where $d_1' = (1, d_0')$.

(b.3) *If $\tau > 1/2$, then $n^{1/2}\{T(\tilde{\theta}_n, \tilde{\gamma}_n) - \mu_{FG}(\theta_0, \gamma_0)\} \xrightarrow{\mathcal{L}} N(0, \sigma_{FG}^2(\theta_0, \gamma_0))$, where $\sigma_{FG}^2(\theta_0, \gamma_0)$ is as defined in Theorem 4(c).*

From the above result, we observe that except when $c_F(t;\theta_0) \neq c_G(t;\gamma_0)$ and $\tau > 1/2$, the choice of other estimators, different from the ISE, makes the asymptotic distributions of $T(\hat{\theta}_n, \hat{\gamma}_n)$ and $T(\tilde{\theta}_n, \tilde{\gamma}_n)$ differ.

The asymptotic null distribution depends on whether $c_F(t;\theta_0) = c_G(t;\gamma_0)$ and on the value of τ. Motivated by Theorem 5, when $\tau \geq 1/2$, which is the usual setting, testing for $c_F(t;\theta_0) = c_G(t;\gamma_0)$ versus $c_F(t;\theta_0) \neq c_G(t;\gamma_0)$ is equivalent to testing for

$$\tilde{H}_{0\sigma} : \sigma^2_{FG}(\theta_0, \gamma_0) = 0,$$

versus

$$\tilde{H}_{1\sigma} : \sigma^2_{FG}(\theta_0, \gamma_0) > 0.$$

With this aim, taking into account that $\sigma^2_{FG}(\theta,\gamma) = \text{var}\{\xi(X,\theta,\gamma)\}$, we estimate $\sigma^2_{FG}(\theta_0,\gamma_0)$ by means of $\hat{\sigma}^2_{FG}(\tilde{\theta}_n, \tilde{\gamma}_n)$, with $\hat{\sigma}^2_{FG}(\theta,\gamma)$ as defined in (9). The following result is the analog of Theorem 6 and it gives some properties of $\hat{\sigma}^2_{FG}(\tilde{\theta}_n, \tilde{\gamma}_n)$.

Theorem 9 (a) *If $\tilde{\theta}_n$ and $\tilde{\gamma}_n$ satisfy (7), $u_F(t;\theta)$ and $v_F(t;\theta)$ are continuous functions of θ for each t, and $u_G(t;\gamma)$ and $v_G(t;\gamma)$ are continuous functions of γ for each t, then $\hat{\sigma}^2_{FG}(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\text{a.s.}(P)} \sigma^2_{FG}(\theta_0,\gamma_0)$.*

(b) *Suppose that the assumptions in Theorem 8 hold. Then $0.25n\hat{\sigma}^2_{FG}(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\mathcal{L}} Y_0' M Y_0$, where M is the variance matrix of the random vector $(h_F(X;\theta_0)', h_G(X;\gamma_0)')'$, and Y_0 is as defined in (13).*

The last result of this section is the analog of Theorem 7 and it gives the asymptotic null distribution of the CF version of Cox tests statistic for testing H_F vs H_G when arbitrary estimators $\tilde{\theta}_n$ and $\tilde{\gamma}_n$ are used, that is, for the test statistic $T_{\text{CoxF}}(\tilde{\theta}_n, \tilde{\gamma}_n)$, with $T_{\text{CoxF}}(\theta,\gamma)$ as defined in (11).

Theorem 10 *Suppose that the assumptions in Theorem 8 hold. Let $Y_{0,\theta}$ denote the first k components of Y_0 giving the marginal asymptotic distribution of $n^\tau(\tilde{\theta}_n - \theta_0)$, where Y_0 is as defined in (13). Under H_F, we have:*

(a) *If $\tau < 1/2$, then $n^\tau T_{\text{CoxF}}(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\mathcal{L}} -\delta_0' Y_{0,\theta}$, where $\delta_0 = \frac{\partial}{\partial\theta} m_F(\theta_0,\gamma_0)$.*

(b) *If $\tau = 1/2$ and (14) holds for some random vector Y_1, then $n^{1/2} T_{\text{CoxF}}(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\mathcal{L}} \delta_1' Y_1$, where $\delta_1' = (1, \delta_0', 0_r')$, $0_r = (0,\ldots,0)' \in \mathbb{R}^r$.*

(c) *If $\tau > 1/2$, then $n^{1/2} T_{\text{CoxF}}(\tilde{\theta}_n, \tilde{\gamma}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{FG}(\theta_0,\gamma_0))$, where $\sigma^2_{FG}(\theta_0,\gamma_0)$ is as defined in Theorem 4(c).*

Observe that when $\tau > 1/2$, then $n^{1/2}\{T(\tilde{\theta}_n, \tilde{\gamma}_n) - \mu_{FG}(\theta_0,\gamma_0)\}$ and $n^{1/2} T_{\text{CoxF}}(\tilde{\theta}_n, \tilde{\gamma}_n)$ both have the same asymptotic distribution, which does not depend on Y_0.

6 Some numerical examples

As in Sect. 1, a problem with the ordinary Vuong and Cox approaches is that, in some cases, these methods cannot be applied because the required regularity assumptions are not met. This section gives two practical examples of this case: when the support of one of the competing models depends on the parameter; another is when the maximum likelihood estimator (MLE) of the parameter of one of the families does not have a limit under the competing model. In both examples, the methods proposed in this work can be applied under quite mild conditions. The finite sample performance of the proposed methods is numerically evaluated by means of some simulations.

The large sample properties of the ordinary Vuong and Cox tests based on likelihoods and their analogs based on CFs are quite similar (asymptotically correct, consistent). Although the goal of this paper is to propose alternative methods for model selection that can be applied when the ordinary ones cannot, it is also of interest to compare them in cases where both approaches can be applied. In this context, we worked several examples, in some cases the likelihood approach beats the CF approach, while in other instances the results were opposite. Moreover, we found an example where the results are rather different for different values of the parameter values in the families. A summary of the obtained results for this example is reported. In all examples we took $\alpha = 0.05$.

6.1 Example 1

Let ℓ be a completely specified probability density function (PDF), Θ denote the support of ℓ and L denote the CDF. For simplicity, we assume that $\Theta \subseteq \mathbb{R}$ is an interval. Now, we consider the family of PDFs, which is obtained by truncating ℓ to the left of $\theta \in \Theta$,

$$f(x;\theta) = \frac{\ell(x)}{1 - L(\theta)}, \quad x > \theta.$$

Clearly, the family $\mathcal{F} = \{f(x;\theta),\ \theta \in \Theta\}$ does not satisfy the regularity assumptions in Vuong (1989) (nor those in White 1982b for the Cox method).

Routine calculations show that if $\ell(x)$ and $\frac{\partial}{\partial x}\ell(x)$ are bounded and the weight function w is such that $\int |t| w(t) \mathrm{d}t < \infty$, then the family $\mathcal{F}$ satisfies Assumption 2 in this paper.

To study the finite sample performance of the CF version of the Vuong approach to model selection for the above setting, we considered the families obtained by truncating to the left at $\theta \in \Theta = \mathbb{R}$ the PDF of a standard normal distribution, family $\mathcal{F}$, and the PDF of a Laplace distribution with mean 0 and variance 2, family $\mathcal{G}$. Figure 1 graphs the truncated PDFs for $\theta = 0, 1, 2$. In this example, the parameter is the same for the two competing families. It was estimated through $\tilde{\theta} = X_{(1)}$, the minimum of the sample. If the population PDF has a bounded derivative, then $n(X_{(1)} - \theta)$ converges in law to a negative exponential random variate. Therefore, for this estimator Theorem 8(b.3) holds. We generated 10,000 samples of size n ($n = 50, 100, 200, 300$)

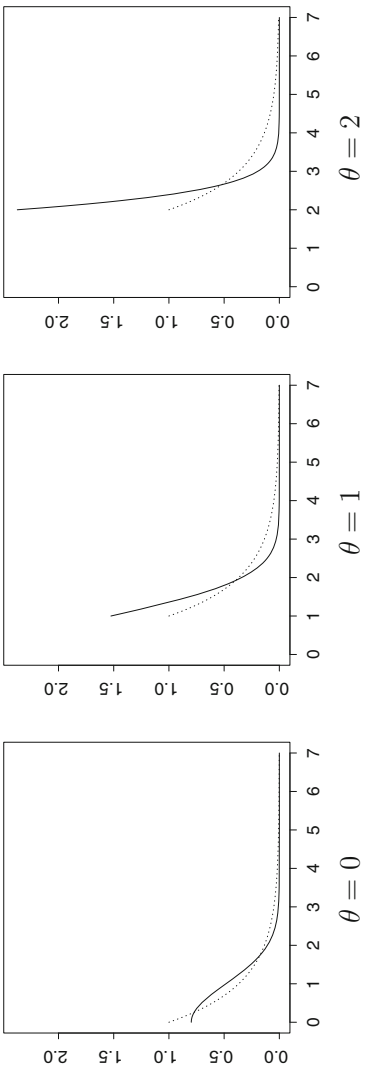

Fig. 1 Truncated Laplace PDF (*dotted line*) and truncated normal PDF (*black line*)

Table 1 Percentage of selections for the CF Vuong procedure

n	$\theta = 0$			$\theta = 1$			$\theta = 2$		
	F	G	Both	F	G	Both	F	G	Both
w is the PDF of a law $N(0, 1)$									
(a)									
50	43.09	0.00	55.91	81.82	0.00	18.18	99.09	0.00	0.91
100	63.26	0.00	36.74	97.08	0.00	2.92	99.99	0.00	0.01
200	88.08	0.00	11.92	99.96	0.00	0.04	100.00	0.00	0.00
300	96.52	0.00	3.48	100.00	0.00	0.00	100.00	0.00	0.00
(b)									
50	0.70	12.26	87.04	0.06	37.73	62.21	0.00	70.27	29.73
100	0.10	26.76	73.14	0.00	71.12	28.88	0.00	96.14	3.86
200	0.00	54.30	45.70	0.00	95.65	4.35	0.00	99.96	0.04
300	0.00	73.74	26.26	0.00	99.47	0.53	0.00	100.00	0.00
w is the PDF of a law $Ca(0, 1)$									
(a)									
50	32.38	0.04	67.58	73.92	0.00	26.08	97.28	0.00	2.72
100	52.08	0.00	47.92	94.00	0.00	6.00	99.96	0.00	0.04
200	79.14	0.00	20.86	99.94	0.00	0.06	100.00	0.00	0.00
300	92.53	0.00	7.47	100.00	0.00	0.00	100.00	0.00	0.00
(b)									
50	0.25	16.60	83.15	0.03	37.84	62.13	0.00	71.48	28.52
100	0.07	32.85	67.08	0.00	70.45	29.55	0.00	96.07	3.93
200	0.00	58.79	41.21	0.00	94.86	5.14	0.00	99.97	0.03
300	0.00	76.90	23.10	0.00	99.46	0.54	0.00	100.00	0.00

True population: (a) F, (b) G
F truncated normal, G truncated Laplace

from the families $\mathcal{F}$ and $\mathcal{G}$ and applied the CF version of Vuong approach for model selection. We considered two weight functions: the PDF of a standard normal distribution, $N(0, 1)$, and the PDF of a standard Cauchy distribution, $Ca(0, 1)$. These choices for the weight function were motivated by the ease of computation of the resulting test statistic, in the sense that most calculations can be analytically done. Table 1 displays the percentage of times that the decision is: F = choose model $\mathcal{F}$, G = choose model $\mathcal{G}$ or both = cannot discriminate between the competing models. Table 2 displays the results for the CF version of Cox approach. Looking at Fig. 1 we see that as θ increases, the PDFs of these families become more different. This fact is captured by the results in Tables 1 and 2: larger sample sizes are necessary for a perfect discrimination when $\theta = 0$, since in this case the models are rather close, while as θ increases the required sample size for a perfect discrimination decreases. As for the choice of the weight function, we see that, as expected from theory, it has little effect on the percentage of correct selections for large sample sizes.

Table 2 Percentage of selections for the CF Cox procedure

n	$\theta = 0$				$\theta = 1$				$\theta = 2$			
	F	G	Both	None	F	G	Both	None	F	G	Both	None
w is the PDF of a law $N(0, 1)$												
(a)												
50	89.68	2.49	7.79	0.04	97.35	0.08	0.00	2.57	97.99	0.00	0.00	2.01
100	96.84	0.97	0.00	2.19	97.06	0.00	0.00	2.94	97.55	0.00	0.00	2.45
200	96.29	0.00	0.00	3.71	96.55	0.00	0.00	3.45	96.39	0.00	0.00	3.61
300	96.15	0.00	0.00	3.85	96.24	0.00	0.00	3.76	96.96	0.00	0.00	3.04
(b)												
50	8.74	70.78	20.48	0.00	1.29	91.96	0.00	6.75	0.00	92.13	0.00	7.87
100	5.52	92.02	0.17	2.29	0.00	93.38	0.00	6.62	0.00	92.78	0.00	7.22
200	0.13	93.58	0.00	6.29	0.00	93.61	0.00	6.39	0.00	93.57	0.00	6.43
300	0.00	93.94	0.00	6.06	0.00	93.20	0.00	6.80	0.00	94.10	0.00	5.90
w is the PDF of a law $Ca(0, 1)$												
(a)												
50	84.56	4.48	10.91	0.05	96.95	0.20	0.00	2.85	97.48	0.00	0.00	2.52
100	95.36	1.88	0.01	2.75	96.87	0.00	0.00	3.13	97.04	0.00	0.00	2.96
200	95.28	0.01	0.00	4.71	96.18	0.00	0.00	3.82	96.14	0.00	0.00	3.86
300	95.30	0.00	0.00	4.70	96.03	0.00	0.00	3.97	96.03	0.00	0.00	3.97
(b)												
50	6.60	74.15	19.25	0.00	1.73	92.64	0.00	5.63	0.00	92.86	0.00	7.14
100	4.76	93.35	0.39	1.50	0.00	93.83	0.00	6.17	0.00	93.19	0.00	6.81
200	0.13	94.64	0.00	5.23	0.00	93.63	0.00	6.37	0.00	93.83	0.00	6.17
300	0.00	94.40	0.00	5.60	0.00	93.93	0.00	6.07	0.00	94.39	0.00	5.61

True population: (a) F, (b) G
F truncated normal, G truncated Laplace

6.2 Example 2

Let $\mathcal{F}$ be the set of normal distributions with mean 0 and variance $\theta \in \Theta = (0, \infty)$, and let $\mathcal{G}$ be the set of Cauchy distributions with location parameter 0 and scale parameter $\gamma \in \Gamma = (0, \infty)$. The MLE of θ is $\hat{\theta}_{ML} = \frac{1}{n}\sum_i X_i^2$, which clearly does not have a limit when the data come from the family $\mathcal{G}$, and therefore Vuong and Cox procedures cannot be applied. To discriminate between these families, we applied the CF versions proposed in this paper taking as weight function $w(t) = \exp(-|t|)$. As in the above example, the choice of this weight function was guided by the ease of computation. The parameters were estimated by their ISE estimators. We generated 10,000 samples of size n ($n = 50, 100, 200$) from the families $\mathcal{F}$, with $\theta = 1$, and $\mathcal{G}$, with $\gamma = 1$, and applied the CF version of Vuong and Cox approaches. Table 3 displays the obtained results. Observe that the CF version of Cox procedure gives very good results even for $n = 50$, in the sense of yielding a high percentage of correct classifications, while the CF version of Vuong procedure requires a bit larger sample sizes.

Table 3 Percentage of selections

n	True population F							True population G						
	Vuong			Cox				Vuong			Cox			
	F	G	Both	F	G	Both	None	F	G	Both	F	G	Both	None
50	72.28	0.00	27.72	95.44	0.05	0.00	4.51	0.06	33.97	65.97	3.97	92.55	0.08	3.40
100	93.70	0.00	6.30	95.34	0.00	0.00	4.66	0.00	64.59	35.41	0.02	93.37	0.00	6.61
200	99.78	0.00	0.22	95.11	0.00	0.00	4.89	0.00	92.94	7.06	0.00	94.15	0.00	5.85

F normal, G Cauchy

Table 4 Percentage of selections

	$a = 1$							$a = 0.5$						
	Vuong			Cox				Vuong			Cox			
	F	G	Both	F	G	Both	None	F	G	Both	F	G	Both	None
(a)														
ML	62.55	0.00	37.45	92.40	0.20	0.05	7.35	0.30	6.95	92.75	6.85	14.60	78.55	0.00
CF1	67.40	0.00	32.60	93.55	1.85	3.45	1.15	0.20	5.00	94.80	0.00	5.90	94.10	0.00
CF2	72.05	0.00	27.95	87.40	0.10	3.35	9.15	0.50	1.30	98.20	0.00	6.00	94.00	0.00
CF3	69.80	0.00	30.20	88.70	0.20	2.00	9.10	0.80	1.20	98.00	0.00	6.25	93.75	0.00
(b)														
ML	3.20	1.25	95.55	35.00	12.75	52.10	0.15	0.15	13.60	86.25	2.55	19.60	77.85	0.00
CF1	1.90	0.05	98.05	0.00	4.10	95.90	0.00	7.35	0.05	92.60	0.00	4.05	95.95	0.00
CF2	1.30	2.10	96.60	0.00	8.30	91.70	0.00	4.00	0.15	95.85	0.00	5.80	94.20	0.00
CF3	1.15	3.25	95.60	0.00	8.55	91.45	0.00	3.70	0.20	96.10	0.00	6.70	93.30	0.00
(c)														
ML	0.05	14.45	85.50	4.50	57.10	38.40	0.00	0.20	13.30	86.50	2.50	22.10	75.40	0.00
CF1	0.00	20.20	79.80	0.40	37.35	62.25	0.00	0.10	8.55	91.35	0.00	8.35	91.65	0.00
CF2	0.20	14.50	85.30	3.85	33.10	63.05	0.00	0.30	2.70	97.00	0.00	7.85	92.15	0.00
CF3	0.25	13.50	86.25	6.00	32.15	61.85	0.00	0.65	2.35	97.00	0.00	8.20	91.80	0.00

True population: (a) $N(0, 1)$, (b) $N(0, 2)$, (c) $0.5N(a, 1) + 0.5N(-a, 1)$
F normal, G equal mixture of two normals

6.3 Example 3

Let $\mathcal{F}$ be the set of normal distributions with mean 0 and variance $\theta \in \Theta = (0, \infty)$, $N(0, \theta)$, and let $\mathcal{G}$ be the set of equal mixtures of two normal populations with equal variance $\gamma \in \Gamma = (0, \infty)$ and known means a and $-a$, $0.5N(a, \gamma) + 0.5N(-a, \gamma)$. Both families satisfy the required regularity assumptions for applying the classical approach (denoted as ML) and the one proposed in this paper. To discriminate between these families we applied both approaches. For the CF methodology, we took as weight function the PDF of a normal law with mean 0 and standard deviation 1, 2 and 3 (denoted as CF1, CF2 and CF3, respectively). We carried out an experiment similar to

that described in Sect. 6.2 for several values of θ, γ and a. Table 4 displays the results for $n = 100$. Looking at this table, we observe that rather different results are obtained for different values of the parameters. For example, in cases (b) $a = 1$ and (c) $a = 0.5$, the ordinary Vuong test outperforms the one proposed in this work; the opposite is observed in cases (a) $a = 1$ and (b) $a = 0.5$; in cases (a) $a = 0.5$ and (c) $a = 1$, the results are quite similar. For Cox test, in all cases the likelihood-based method has the highest percentage of both right decisions and wrong decisions (except in case (a) $a = 1$). Figure 2 graphs the PDF and the CF of a standard normal law (black) together with the closest PDF and CF of the mixture model (dashed) with $a = 1$ and $a = 0.5$. When $a = 1$, we see that the PDFs have rather different shapes (also the CFs); when $a = 1$, they are really close. Because of this reason, the results in case (a) for $a = 1$ are better than for $a = 0.5$.

In the light of the above simulation results, at present we cannot give a general recommendation on what method to use when both apply. This point certainly deserves further research.

7 Conclusions

Two methods for the model selection problem have been proposed and studied. They are based on measuring the distance between the CF of the population generating the data and the CF in each competing model. The first method is a CF analog to that developed by Vuong (1989), while the second one is a CF version of the Cox (1961, 1962) approach for the problem of testing for two separate models. Two examples are used to illustrate that the proposed methods can be applied in settings where neither Vuong nor Cox approaches can be used.

Some generalizations of the proposed methods are possible: (a) throughout the paper we have assumed that we have only two competing models; the case of three or more competing models can be dealt by applying multiple comparison techniques as suggested in Shimodaira (1998); (b) throughout the paper we have assumed that the available data consist of IID observations; the proposed procedures can be extended to other more general settings such as regression models or dependent data; (c) throughout the paper we have assumed that the competing models are parametric; the proposed procedures can be extended to other more general models such as semiparametric models. These as well as other possible extensions constitute a field of future research.

Another open question that deserves further study is what method should be applied in cases where both approaches can be applied.

8 Proofs

Proof of Theorem 1 For each fixed $\theta \in \Theta$, $I_n(\theta)$ is a degree-2 V-statistic, $I_n(\theta) = \frac{1}{n^2}\sum_{jl} k(X_j, X_l; \theta)$, with kernel $k(x, y; \theta)$ as defined in (6). From the SLLN for V-statistics (see for example Serfling 1980), $I_n(\theta) \xrightarrow{\text{a.s.}} E\{k(X_1, X_2; \theta)\} = D^2(c(t), c(t; \theta))$. Thus, for any $\delta \neq 0$ such that $\theta_* + \delta \in \Theta$, we have that, at least for

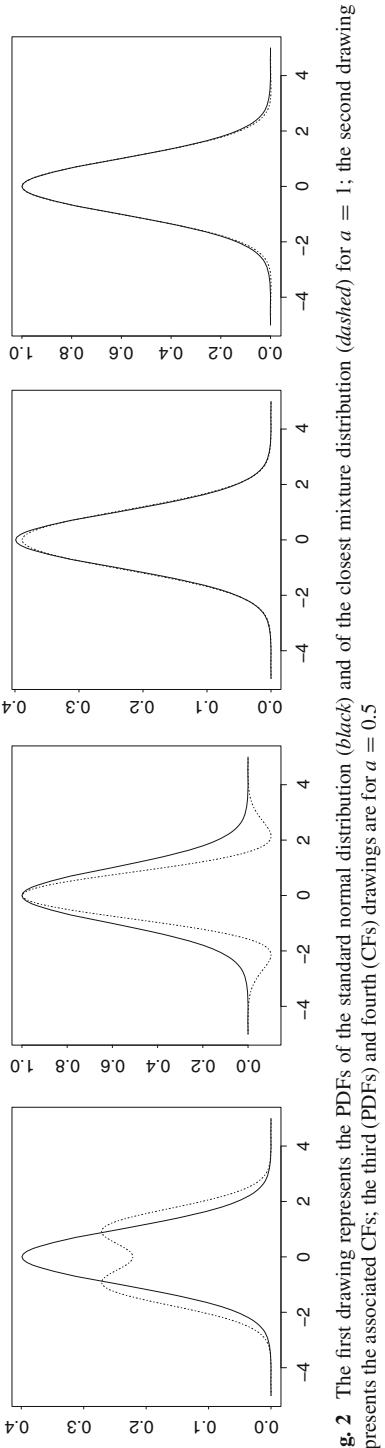

Fig. 2 The first drawing represents the PDFs of the standard normal distribution (*black*) and of the closest mixture distribution (*dashed*) for $a = 1$; the second drawing represents the associated CFs; the third (PDFs) and fourth (CFs) drawings are for $a = 0.5$

large n, $I_n(\theta_* + \delta) - I_n(\theta_*) > 0$ with probability one. Since δ is arbitrary, we conclude that $\hat{\theta}_n$ is strongly consistent for θ_*. □

Proof of Theorem 2 From Assumptions 1 and 2, $D_1(\theta_*) = 0$ and $D_2(\theta)$ is positive definite for all θ in a neighborhood of θ_*. By Taylor expansion,

$$0 = \frac{\partial}{\partial\theta} I_n(\hat{\theta}_n) = \frac{\partial}{\partial\theta} I_n(\theta_*) + \frac{\partial^2}{\partial\theta\partial\theta'} I_n(\hat{\theta}_{1n})(\hat{\theta}_n - \theta_*), \tag{15}$$

with $\hat{\theta}_{1n} = \alpha\hat{\theta}_n + (1-\alpha)\theta_*$, for some $\alpha \in (0,1)$. We have $\frac{\partial}{\partial\theta} I_n(\theta) = -2\frac{1}{n}\sum_{j=1}^n h(X_j;\theta)$, $2E\{h(X;\theta)\} = -D_1(\theta)$ and $E\{h(X;\theta)h(X;\theta)'\} = A(\theta)$. Thus, from the CLT,

$$\sqrt{n}\frac{1}{2}\frac{\partial}{\partial\theta} I_n(\theta_*) \xrightarrow{\mathcal{L}} N_k(0, A(\theta_*)). \tag{16}$$

On the other side, $E\left\{\frac{\partial^2}{\partial\theta\partial\theta'} I_n(\theta)\right\} = 2D_2(\theta)$. Since $\hat{\theta}_{1n} \xrightarrow{\text{a.s.}} \theta_*$, from Lemma 3.1 in White (1981), it follows

$$\frac{\partial^2}{\partial\theta\partial\theta'} I_n(\hat{\theta}_{1n}) \xrightarrow{\text{a.s.}} 2D_2(\theta_*). \tag{17}$$

Finally, the result follows from (15)–(17). □

Proof of Theorem 3 (a) We have

$$D^2(c_n(t), c(t;\hat{\theta}_n)) = I_n(\hat{\theta}_n) = \int \{u_n(t) - u(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t) + \int \{v_n(t) - v(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t). \tag{18}$$

For the first integral in the right hand side of (18), we have

$$\int \{u_n(t) - u(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t) = \int \{u_n(t) - u(t;\theta_*)\}^2 \mathrm{d}W(t) + \int \{u(t;\theta_*) - u(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t) + 2\int \{u_n(t) - u(t;\theta_*)\}\{u(t;\theta_*) - u(t;\hat{\theta}_n)\} \mathrm{d}W(t) := S1 + S2 + 2S3.$$

From the proof of Theorem 1, $S1 \xrightarrow{\text{a.s.}} \int \{u(t) - u(t;\theta_*)\}^2 \mathrm{d}W(t)$. To deal with $S2$, we observe the following facts: let $\varepsilon > 0$ be arbitrary but fixed, then there exists a compact set $K \subset \mathbb{R}^k$ such that $\int_K \mathrm{d}W(t) \geq 1 - \varepsilon$; let $\delta > 0$ be such that $\Theta_1 = \bar{B}(\theta_*;\delta) \subseteq \Theta$, where $\bar{B}(\theta_*;\delta) = \{x \in \mathbb{R}^k : |x - \theta_*| \leq \delta\}$; from the assumptions made, $u(t;\theta)$ is a continuous function, as a function of the pair (t,θ); thus it is a uniformly continuous function on $C = K \times \Theta_1$, which implies

that there exists a $\zeta > 0$ such that $|u(t;\theta) - u(t';\theta')| < \varepsilon$, whenever (t,θ), $(t',\theta') \in K \times \Theta_1$ and $|(t,\theta) - (t',\theta')| < \zeta$. From Theorem 1, for large enough n, we have that $\hat{\theta}_n \in \bar{B}(\theta_*;\delta)$ with probability 1. Therefore, for large enough n,

$$0 \leq \int \{u(t;\theta_*) - u(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t) \leq \varepsilon^2 \int_K \mathrm{d}W(t) + 4 \int_{K^c} \mathrm{d}W(t) \leq \varepsilon^2 + 4\varepsilon,$$

where K^c denotes the complementary of K. Since $\varepsilon > 0$ is arbitrary, this implies that $S2 \xrightarrow{\text{a.s.}} 0$. As for $S3$, taking into account that $|S3| \leq S1^{1/2} S2^{1/2}$ and $0 \leq \int \{u(t) - u(t;\theta_*)\}^2 \mathrm{d}W(t) \leq 4$, we have that $S3 \xrightarrow{\text{a.s.}} 0$. Thus,

$$\int \{u_n(t) - u(t;\hat{\theta}_n)\}^2 \mathrm{d}W(t) \xrightarrow{\text{a.s.}} \int \{u(t) - u(t;\theta_*)\}^2 \mathrm{d}W(t).$$

Proceeding analogously with the second term in the right hand side of (18), we get the result.

(b) The result follows from the results in Sections 6 and 8 of Csörgő (1981) or from Theorem 1 in Jiménez-Gamero et al. (2009).

(c) By Taylor expansion,

$$D^2(c_n(t), c(t;\hat{\theta}_n)) = I_n(\hat{\theta}_n) = I_n(\theta_*) + \frac{\partial}{\partial\theta} I_n(\hat{\theta}_{1n})'(\hat{\theta}_n - \theta_*), \tag{19}$$

with $\hat{\theta}_{1n} = \alpha\hat{\theta}_n + (1-\alpha)\theta_*$, for some $\alpha \in (0,1)$. Now, we separately study each term in the right hand side of (19).

As observed in the proof of Theorem 1, $I_n(\theta)$ is a degree-2 V-statistic with kernel $k(x,y;\theta)$, defined in (6), satisfying $|k(x,y;\theta)| \leq 8$, $\forall x, y, \theta$. We are assuming that $\mathrm{var}\{\rho(X)\} = \sigma^2(\theta_*) > 0$. Thus, from Theorem 6.4.1A in Serfling (1980),

$$\sqrt{n}\left\{I_n(\theta_*) - D^2(c(t), c(t;\theta_*))\right\} \xrightarrow{\mathcal{L}} N(0, \sigma^2(\theta_*)). \tag{20}$$

Routine calculations show that

$$\sqrt{n}\frac{\partial}{\partial\theta} I_n(\hat{\theta}_{1n})'(\hat{\theta}_n - \theta_*) \xrightarrow{P} 0. \tag{21}$$

Finally, the result follows from (19)–(21). □

Proof of Corollary 1 (a) The result follows from Theorem 1 because $m_F(\theta,\gamma) = -D^2(c_F(t;\theta), c_G(t;\gamma))$.

(b) The result follows from Theorem 2.

□

Proof of Theorem 4 (a) From Taylor expansion of $D^2(c_n(t), c_F(t;\theta_*))$ around $\hat{\theta}_n$, we obtain

$$\begin{aligned} D^2(c_n(t), c_F(t;\theta_*)) &= D^2(c_n(t), c_F(t;\hat{\theta}_n)) \\ &+ (\hat{\theta}_n - \theta_*)' D_{2F}(\theta_*)(\hat{\theta}_n - \theta_*) + o_P(n^{-1}). \end{aligned} \tag{22}$$

Similarly,

$$D^2(c_n(t), c_G(t;\gamma_*)) = D^2(c_n(t), c_G(t;\hat{\gamma}_n)) + (\hat{\gamma}_n - \gamma_*)' D_{2G}(\gamma_*)(\hat{\gamma}_n - \gamma_*) + o_P(n^{-1}). \tag{23}$$

From (22) and (23),

$$nT(\hat{\theta}_n, \hat{\gamma}_n) = \sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_*\end{pmatrix}' \begin{pmatrix}-D_{2F}(\theta_*) & 0 \\ 0 & D_{2G}(\gamma_*)\end{pmatrix} \sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_*\end{pmatrix} + o_P(1).$$

Now, the result follows from the above equality and the result in Corollary 1(b).

(b) This part is a direct consequence of Lemma 3.1 in White (1981) that the eigenvalues of a matrix are a continuous function of the entries in the matrix and the Polya theorem (see for example Lemma 8.2.6 in Athreya and Lahiri 2006).

(c) Note that $E\{\xi(X,\theta,\gamma)\} = \mu_{FG}(\theta,\gamma)$ and $0 < \text{var}\{\xi(X,\theta,\gamma)\} = \sigma^2_{FG}(\theta,\gamma)$. Thus from the CLT,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\{\xi(X_j,\theta,\gamma) - \mu_{FG}(\theta,\gamma)\} \xrightarrow{\mathcal{L}} N(0, \sigma^2_{FG}(\theta,\gamma)). \tag{24}$$

Since $\frac{\partial}{\partial\theta_j}\mu_{FG}(\theta_*,\gamma_*) = \frac{\partial}{\partial\gamma_l}\mu_{FG}(\theta_*,\gamma_*) = 0$, $1 \le j \le k$, $1 \le l \le r$, (24) and Theorem 2.13 of Randles (1982) both imply that $\sqrt{n}\{T(\hat{\theta}_n, \hat{\gamma}_n) - \mu_{FG}(\theta_*,\gamma_*)\} \xrightarrow{\mathcal{L}} N(0, \sigma^2_{FG}(\theta_*,\gamma_*))$.

□

Proof of Theorem 5 If $c_F(t;\theta_*) = c_G(t;\gamma_*)$, then $\sigma^2_{FG}(\theta_*,\gamma_*) = \text{var}\{\xi(X,\theta_*,\gamma_*)\} = 0$. To show the another implication, note that $\sigma^2_{FG}(\theta_*,\gamma_*) = 0$ iff

$$\xi(x,\theta_*,\gamma_*) = \kappa_1, \quad \text{for some } \kappa_1 \in \mathbb{R},\ \forall x \in \mathbb{R}^d. \tag{25}$$

The equality in (25) can be rewritten as follows:

$$C(x) = -S(x) + \kappa_2, \tag{26}$$

for some $\kappa_2 \in \mathbb{R}$, $\forall x \in \mathbb{R}^d$, where $C(x) = \int \cos(t'x)\{u_F(t;\theta_*) - u_G(t;\gamma_*)\}w(t)\mathrm{d}t$, $S(x) = \int \sin(t'x)\{v_F(t;\theta_*) - v_G(t;\gamma_*)\}w(t)\mathrm{d}t$. Since $C(x) = C(-x)$ and $S(x) = -S(-x)$, $\forall x \in \mathbb{R}^d$, from (26) we conclude that $S(x) = 0$ and $C(x) = \kappa_2$, $\forall x \in \mathbb{R}^d$. Because $\cos(-t'x)w(-t) = \cos(t'x)w(t)$, $v_F(t;\theta) = -v_F(-t;\theta)$ and $v_G(t;\theta) = -v_G(-t;\theta)$, $\forall x, t \in \mathbb{R}^d$, $\forall\theta \in \Theta$, $\forall\gamma \in \Gamma$, we have

$$\int \cos(t'x)\{v_F(t;\theta_*) - v_G(t;\gamma_*)\}w(t)\mathrm{d}t = 0, \quad \forall x \in \mathbb{R}^d. \tag{27}$$

Now, $S(x) = 0$ and (27) are tantamount to saying that the Fourier transform of the function $\{v_F(t;\theta_*) - v_G(t;\gamma_*)\}w(t)$ is equal to 0. The uniqueness of the Fourier transform implies that $\{v_F(t;\theta_*) - v_G(t;\gamma_*)\}w(t) = 0$, $\forall t \in \mathbb{R}^d$. Since $w(t) > 0$, it follows that $v_F(t;\theta_*) = v_G(t;\gamma_*)$, $\forall t \in \mathbb{R}^d$. Proceeding analogously,

$$\int \sin(t'x)\{u_F(t;\theta_*) - u_G(t;\gamma_*)\}w(t)\mathrm{d}t = 0, \quad \forall x \in \mathbb{R}^d. \tag{28}$$

Now, $C(x) = \kappa_2$ and (28) are tantamount to saying that the Fourier transform of the function $\{u_F(t;\theta_*) - u_G(t;\gamma_*)\}w(t)$ is equal to κ_2. The Riemann–Lebesgue lemma implies that $\kappa_2 = 0$. Reasoning as before, we get $u_F(t;\theta_*) = u_G(t;\gamma_*)$, $\forall t \in \mathbb{R}^d$. This proves the result. □

Proof of Theorem 6 (a) The result follows from Theorem 1 and Lemma 3.1 in White (1981).

(b) By Taylor expansion of $\hat{\sigma}^2_{FG}(\hat{\theta}_n, \hat{\gamma}_n)$ around $(\theta_*', \gamma_*')'$, we obtain

$$n\hat{\sigma}^2_{FG}(\hat{\theta}_n, \hat{\gamma}_n) = 4\sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_*\end{pmatrix}' A_{FG}(\theta_*, \gamma_*)\sqrt{n}\begin{pmatrix}\hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_*\end{pmatrix} + o_P(1).$$

Now, the result follows from the above expression and the result in Corollary 1(b).

(c) The proof is the same as that of Theorem 4(b); so we omit it.

□

The proof of Lemma 1 is easy, and thus it is omitted. The proof of Lemma 2 is quite similar to the proof of Lemma 7.1 in Vuong (1989); so we omit it. The proof of Lemma 3(a) is quite similar to the proof of Theorem 7.2(i) in Vuong (1989); so we omit it. Lemma 3(b) is a direct consequence of part (a) and Theorem 4(a).

Proof of Theorem 7 We have $\sqrt{n}T_{\mathrm{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) = \sqrt{n}\{T(\hat{\theta}_n, \hat{\gamma}_n) - m_F(\theta_*, \gamma_*) + m_F(\theta_*, \gamma_*) - m_F(\hat{\theta}_n, \hat{\gamma}_n)\}$. Under H_F, $\mu_{FG}(\theta_*, \gamma_*) = m_F(\theta_*, \gamma_*) = -D^2(c_F(t;\theta_*), c_G(t;\gamma_*))$. From the proof of Theorem 4(c), $\sqrt{n}\{T(\hat{\theta}_n, \hat{\gamma}_n) - m_F(\theta_*, \gamma_*)\} = \frac{1}{\sqrt{n}}\sum_{j=1}^n \xi(X_j, \theta_*, \gamma_*) - \mu_{FG}(\theta_*, \gamma_*) + o_P(1)$. Routine calculations show that $\sqrt{n}\{-m_F(\hat{\theta}_n, \hat{\gamma}_n) + m_F(\theta_*, \gamma_*)\} = \psi_F(\theta_*, \gamma_*)'\frac{1}{\sqrt{n}}\sum_{j=1}^n H_F(X_j, \theta_*) + o_P(1)$. Summarizing,

$$\sqrt{n}T_{\mathrm{CoxF}}(\hat{\theta}_n, \hat{\gamma}_n) = (1,\ \psi_F(\theta_*, \gamma_*)')\frac{1}{\sqrt{n}}\sum_{j=1}^n \begin{pmatrix}\xi(X_j, \theta_*, \gamma_*) - \mu_{FG}(\theta_*, \gamma_*) \\ H_F(X_j, \theta_*)\end{pmatrix} + o_P(1).$$

The result follows from the above expression and CLT. □

Acknowledgments The authors thank the anonymous referees for their constructive comments and suggestions which helped to improve the presentation. M.D. Jiménez-Gamero and M.V. Alba-Fernández have been partially supported by the research Project UJA2013/08/01 (University of Jaén and Caja Rural Provincial of Jaén).

References

Athreya, K. B., Lahiri, S. N. (2006). *Measure theory and probability theory*. New York: Springer.

Broniatowski, M., Keziou, A. (2009). Parametric estimation and testing through divergences and the duality technique. *Journal of Multivariate Analysis*, *100*, 16–36.

Castaño-Martínez, A., López-Blázquez, F. (2005). Distribution od a sum of weighted central chi-square variables. *Communications in Statistics-Theory and Methods*, *34*, 515–524.

Cox, D. R. (1961). Tests of separate families of hypothesis. *Proceedings of the forth Berkeley symposium in mathematical statistics and probability* (pp. 105–123). Berkeley: University of California Press.

Cox, D. R. (1962). Further results on tests of separate families of hypothesis. *Journal of the Royal Statistical Society B*, *24*, 406–424.

Csörgő, S. (1981). The empirical characteristic process when parameters are estimated. In J. Gani, V. K. Rohatgi (Eds.), *Contributions to probability: A collection of papers dedicated to Eugene Lukacs* (pp. 708–723). New York: Academic Press.

Epps, T. W., Singleton, K. J., Pulley, L. B. (1982). A test of separate families of distributions based on the empirical moment generating function. *Biometrika*, *69*, 391–399.

Feigin, P. D., Heathcote, C. R. (1976). The empirical characteristic function and the Cramér–von Mises statistic. *Sankhya*, *38*, 309–325.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 2). New York: Wiley.

Heathcote, C. R. (1977). The integrated squared error estimation of parameters. *Biometrika*, *64*, 64–255.

Jiménez-Gamero, M. D., Alba-Fernández, V., Muñoz-García, J., Chalco-Cano, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions. *Computational Statistics & Data Analysis*, *53*, 3957–3971.

Jiménez-Gamero, M. D., Pino-Mejías, R., Alba-Fernández, V., Moreno-Rebollo, J. L. (2011). Minimum ϕ-divergence estimation in misspecified multinomial models. *Computational Statistics & Data Analysis*, *55*, 3365–3378.

Jiménez-Gamero, M. D., Pino-Mejías, R., Rufián-Lizana, A. (2014). Minimum K_ϕ-divergence estimators for multinomial models and applications. *Computational Statistics*, *29*, 363–401.

Kishino, H., Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, *29*, 170–179.

Kotz, S., Johnson, N. L., Boyd, D. W. (1967). Series representations of quadratic forms in normal variables. I. Central case. *The Annals of Mathematical Statistics*, *38*, 823–837.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, *22*, 1081–1114.

Linhart, H. (1988). A test whether two AIC's differ significantly. *South African Statistical Journal*, *22*, 153–161.

Matsui, M., Takemura, A. (2005). Empirical characteristic function approach to goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE. *Annals of the Institute of Mathematical Statistics*, *57*, 183–199.

Matsui, M., Takemura, A. (2008). Goodness-of-fit tests for symmetric stable distributions-Empirical characteristics function approach. *Test*, *17*(3), 546–566.

Meintanis, S. G. (2005). Consistent tests for symmetric stability with finite mean based on the empirical characteristic function. *Journal of Statistical Planning and Inference*, *128*(2), 373–380.

Pardo, L. (2006). *Statistical inference based on divergence measures*. Boca Raton: Chapman & Hall.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, *10*, 462–474.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Annals of the Institute of Mathematical Statistics*, *50*, 1–13.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 257–306.

Vuong, Q. H., Wang, W. (1993). Minimum chi-square estimation and tests for model selection. *Journal of Econometrics*, *56*, 141–168.

White, H. (1981). Misspecified nonlinear regression models. *Journal of the American Statistical Society*, *76*, 419–433.

White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

White, H. (1982b). Regularity conditions for Cox's test of non-nested hypothesis. *Journal of Econometrics*, *19*, 301–315.