A spline approach to nonparametric test of hypothesis

Ronaldo Dias and Nancy L. Garcia
Departamento de Estatística.
Universidade Estadual de Campinas. São Paulo, Brasil
E-mail address: dias@ime.unicamp.br and nancy@ime.unicamp.br

Abstract

We propose a test of hypothesis for the closeness of two distributions whose test statistic is asymptotically normal. The divergent is based on the estimation procedure developed in Dias (2000) using a proxy of symmetrized Kullback-Leibler distance. Simulation results show that for mixture of normal distributions this test is more powerful than Kolmogorov-Smirnov test. As an application we compare acoustic data from several languages in order to identify rhythmic classes.

Key words: Asymptotic Theory, B-splines, Kolmogorov-Sminorv Test, Kullback-Leibler divergent.

1 Introduction

There are several situations where we have independent samples and wish to test whether they come from the same distribution. If it is possible to conjecture a parametric family for the distribution, life is much easier and parametric tests can be used. However, most of the time we cannot fit a parametric model and a nonparametric test is necessary. (See for example Fan (1998) and Li (1996)). The same duality appears in estimation problems. Dias (2000) proposed a nonparametric estimator for densities based on a proxy of the symmetrized Kullback-Leibler distance which is consistent (Section 2). Based on this estimator, in Section 3 we propose a test statistic (henceforth called SKL test) which is asymptotically normal. Simulation results show that for mixture of normal distributions SKL test is more powerful than Kolmogorov-Smirnov (K-S) test (Section 4). Also, the normal approximation is achieved even for small samples when the underlying distribution is normal.

As an application, in Section 5, we present an example that comes from linguistic and deals with clustering the natural languages into rhythmic classes. In the linguistic literature it has been conjectured that natural languages are divided into rhythmic classes (cf. Abercrombie (1967), Pike 1945 among others). During half a century no reliable phonetic evidence was presented to support this claim. Recently Ramus, Nespor and Mehler (1999), gave evidence that simple statistical properties of the speech signal could discriminate between different rhythmic classes. They analyzed the acoustic signal of 20 sentences of each of the following languages: English, Polish, Dutch, Catalan, Spanish, Italian, French and Japanese. They computed for each sentence the standard deviation of the consonantal intervals (ΔC) and the proportion of time spent in vocalic intervals ($\%V$) and found that based on these statistics the languages appear to cluster into three groups which correspond precisely to the intuitive notion of rhythmic classes: English, Polish and Dutch represent the accentual class, French, Spanish, Catalan and Italian represent the syllabic class and Japanese represents the moraic class. In their work there is no study for Portuguese. In Section 5, we apply the proposed nonparametric test to some of these languages and find that there is no significant evidence of difference between European and Brazilian Portuguese, English and Dutch and English and European Portuguese, while there is significant difference between Brazilian Portuguese and Catalan and English and Japanese.

2 Previous results

Suppose we have two independent random samples $\mathbf{X} = (X_1, X_2, \ldots, X_{n_1})$ with distribution F and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{n_2})$ with distribution G and we would like to test whether $F = G$. First assume that both F and G are absolutely continuous cumulative distribution functions with $F \ll \mu$ and $G \ll \mu$ for a Lebesgue dominant measure μ. Moreover, assume that $f = \frac{dF}{d\mu}$ and $g = \frac{dG}{d\mu}$, the respective densities of F and G, have compact support $\mathcal{X}$. Define $\mathcal{F}_\mu$ be the class of density functions such that,

$$\mathcal{F}_\mu = \{h : \mathbb{R} \to [0, \infty) : h(x) = \frac{e^{S(x)}}{\int_{\mathcal{X}} e^{S(x)} d\mu(x)} \text{ and } \int_{\mathcal{X}} e^{S(x)} d\mu(x) < \infty\},$$

where the function S is of the class $C^2(\mathbb{R})$. It is easy to see that the elements in $\mathcal{F}_\mu$ are not identifiable since for any function S_1 such that $S_1 = S + c$, we have $e^{S_1}/(\int e^{S_1}) = e^S/(\int e^S)$. We are going to require, as Dias (1998), that $\int_{\mathcal{X}} S = 0$, to ensure uniqueness of the elements in $\mathcal{F}_\mu$.

Assume further that for any density $h \in \mathcal{F}_\mu$, there exists K and a vector $\theta = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^k$ such that

$$S_h = \langle \theta, M \rangle_K = \sum_{j=1}^{K} \theta_j M_j,$$

where M_j, $j = 1, \ldots, K$ are normalized basis functions such that $\int M_j = 1$. In order to enforce one-to-one correspondence we restrict $\int S_h = 0$ and then $\sum_{j=1}^{n} \theta_j = 0$, since $\int M_j = 1$. For any $K > 0$, let $\Theta_0 = \{\theta \in \mathbb{R}^K : \sum_j^K \theta_j = 0\}$.

Assuming that the densities f and g belong to $\mathcal{F}_\mu$, we have that there exist K_1, K_2 and vectors $\theta = (\theta_1, \ldots, \theta_{K_1})$, $\psi = (\psi_1, \ldots, \psi_{K_2})$ such that the log-likelihood of $\mathbf{X}$ and $\mathbf{Y}$ are given by

$$L_{K_1}(\theta|\mathbf{X}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \langle \theta, M(X_i) \rangle_{K_1} - \log \int e^{\langle \theta, M \rangle_{K_1}} \tag{2.1}$$

and

$$L_{K_2}(\psi|\mathbf{Y}) = \frac{1}{n_2} \sum_{i=1}^{n_2} \langle \psi, M(Y_i) \rangle_{K_2} - \log \int e^{\langle \psi, M \rangle_{K_2}} . \tag{2.2}$$

The next results (Lemma 2.3, Theorem 2.4, Lemma 2.10 and Proposition 2.12) were proved by Dias (2000) in the case that functions M are the normalized B-splines. We are going to enunciate the results for the (2.1) but obviously the results are also valid for (2.2).

Lemma 2.3 *For a fixed K_1, $L_{K_1}(\theta|\mathbf{X})$ is concave in θ. Moreover, $L_{K_1}(\theta|\mathbf{X})$ is strictly concave for $\theta \in \Theta_0$. Hence there exists at most one maximizer on Θ_0.*

It is not difficult to show that $L_{K_1}(\theta|\mathbf{X})$ is continuous and at least twice differentiable in θ for a fixed K. Thus, restrict to Θ_0 one may guarantee a unique density estimate.

The next theorem shows the relationship between the maximizers $\hat{\theta}$ in Θ and θ^* in Θ_0.

Theorem 2.4 *If the vector $\hat{\theta}$ maximizes $L_{K_1}(\theta|\mathbf{X})$ then $\theta^* = \hat{\theta} - \frac{1}{K}\sum_{j=1}^{K_1}\hat{\theta}_j$ maximizes $L_{K_1}(\theta|\mathbf{X})$ subject to $\sum_{j=1}^{K_1}\theta_j = 0$. Moreover, θ^* is unique.*

For fixed K_1, let $\hat{\theta}_{n_1}^{(K_1)}$ be defined as

$$\hat{\theta}_{n_1}^{(K_1)} = \arg\max_{\theta\in\Theta_0} L_{K_1}(\theta|\mathbf{X}). \tag{2.5}$$

Notice that, in fact,

$$L_{K_1}(\theta|\mathbf{X}) = \langle\theta, \bar{M}\rangle_{K_1} - \log\int e^{\langle\theta, M\rangle_{K_1}},$$

then $\hat{\theta}_{n_1}^{(K_1)}$ is the unique solution of the equation

$$h(\theta, \bar{M}(\mathbf{X})) = 0, \tag{2.6}$$

where $\bar{M}(\mathbf{X})$ is a K-dimensional vector with j-th components given by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} M_j(X_i) = \bar{M}_j, \quad j \in \{1, \ldots, K_1\}. \tag{2.7}$$

Since $L_{K_1}(\theta|\mathbf{X})$ is at least twice differentiable we have $\hat{\theta}_{n_1}^{(K_1)}$ as the unique solution of the equation,

$$\frac{\partial L_{K_1}(\theta|\mathbf{X})}{\partial\theta} := h(\theta, M^*(\mathbf{X})) = 0, \tag{2.8}$$

where, $M^* = (1/K)\sum_{j=1}^K \bar{M}_j$ and $h : \Theta_0 \times [0,\infty)^K \longrightarrow \mathbb{R}^K$ with j-th entry,

$$h_j(\theta, \mathbf{u}) = u_j - \frac{\int \exp(\langle\theta, M(z)\rangle_{K_1}) M_j(z)dz}{\int \exp(\langle\theta, M(z)\rangle_{K_1})dz}, \tag{2.9}$$

for $j \in \{1,\dots,K_1\}$. Therefore, $\hat{\theta}_{n_1}^{(K_1)}$ is an M-estimator and since $\theta \mapsto h_\theta$ is continuous we have the following result.

Lemma 2.10 *Let θ_0 be the unique solution of*

$$h(\theta, \int f(x)M(x)d\mu(x)) = 0$$

in Θ_0, then for fixed K_1, $\hat{\theta}_{n_1}^{(K_1)} \longrightarrow \theta_0$ almost surely as $n_1 \longrightarrow \infty$.

Thus, the density estimate is, for fixed K_1

$$\hat{f}_{K_1} = e^{\hat{S} - \log\int e^{\hat{S}}},$$

where $\hat{S} = \langle\hat{\theta}, M\rangle_{K_1}$ with $\hat{\theta} = \hat{\theta}_{n_1}^{(K_1)}$.

One may notice the density estimate $\hat{f}_{K_1}$ strongly depends on the number of basis functions K_1 which regularizes the optimization problem (2.1). In order to provide an appropriate K_1, one may want to compute the Kullback-Leibler distance between the true f and the random function $\hat{f}_{K_1}$.

$$d(f, \hat{f}_{K_1}) = \int (\log f - \log \hat{f}_{K_1}) f \tag{2.11}$$

Of course, we cannot compute $d(f, \hat{f}_{K_1})$ from the data, since it requires the knowledge of f. But theoretically we can investigate this distance for the choice of an optimal K_1 in the sense of minimizing $d(f, \hat{f}_{K_1})$. Then, one may define the best K_1 as

$$\hat{K_1} = \arg \min_{K\in\{1,\dots,K_{max}\}} d(f, \hat{f}_K),$$

for $K_{max} < n$. Observe that, in order to obtain $\hat{K_1}$, it is sufficient to minimize

$$D_n(K) = \int f \log \hat{f}_K.$$

Notice that $D_n(K)$ is a random function of K and also can be approximate by

$$Z_n(K) = \frac{1}{n}\sum_{i=1}^n \log \hat{f}_K(X_i).$$

Proposition 2.12 *For any fixed K,*

$$D_n(K) - Z_n(K) = \sum_{j=1}^{K} \hat{\theta}_{nj}^{(K)} \Big(\int f(x) M_j(x) d\mu(x) - \frac{1}{n} \sum_{i=1}^{n} M_j(X_i) \Big) \longrightarrow 0 \qquad (2.13)$$

$n \longrightarrow \infty$ almost surely.

Lemma 2.14 *For K_1, K_2 fixed the density estimates $\hat{f}_{K_1}(\cdot) = f_{K_1}(\cdot|\hat{\theta}_{n_1})$ and $\hat{g}_{K_2}(\cdot) = g_{K_2}(\cdot|\hat{\psi}_{n_2})$ converge pointwise almost surely (a.s.) to $f_{K_1}(\cdot|\theta)$and $g_{K_2}(\cdot|\psi)$ respectively as n_1, n_2 go to infinity.*

Proof. It is enough to show one of the statements above. For fixed x, it is not difficult to check that the map $\theta \mapsto f_m(x|\theta)$ is a continuous map in $\theta \in \Theta_0$, for any $m \in \mathbb{N}$. By Lemma 2.10 we have $\hat{\theta}_n \longrightarrow \theta (\in \Theta_0)$ a.s. and so $f_m(x|\hat{\theta}) \longrightarrow f_m(x|\theta)$ almost surely as $n \longrightarrow \infty$. Notice that the null sets of the a.s. convergence do not depend on x then $f(\cdot|\hat{\theta})$ converges pointwise to $f(\cdot)$ a.s.

3 Hypothesis testing - SKL test statistic

In this section we propose a statistic to test: $H_0 : f = g$ almost surely μ versus the alternative hypothesis $H_1 : f \neq g$ over a set of positive μ-measure. Since this test statistic is based on the symmetrized Kullback-Leibler distance we will call it SKL test.

First we notice that the parameter space $\Theta_0 = \{\theta \in \mathbb{R}^K : \sum_j^K \theta_j = 0\}$ is not an open set and it is a $(K-1)$-dimensional manifold in $\mathbb{R}^K$. Therefore, in order to have any kind of asymptotic normality results we need to reparametrize the problem to $(\theta_1, \ldots, \theta_{K-1}) \in \widetilde{\Theta}_0$ such that $(\theta_1, \ldots, \theta_{K-1}, -\sum_{i=1}^{K-1} \theta_i) \in \Theta_0$. We will continue to call the parameter θ. In this case, the density will be written as

$$f(x \mid \theta) = \frac{e^{\langle \theta, \tilde{M}(x) \rangle}}{\int_{\mathcal{X}} e^{<\theta, \tilde{M}(x)>}} \in \mathcal{F}_\mu \qquad (3.1)$$

where

$$\tilde{M}_j(x) = M_j(x) - M_K(x). \qquad (3.2)$$

For fixed K_1 and K_2, we have as a consequence of Cramér's Theorem (see for example, Ferguson (1996), p.121) the asymptotic normality of the consistent estimator which solves the likelihood equation. For simplicity we will continue to denote

$$\hat{\theta}_{n_1} = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_1-1}) \tag{3.3}$$

where $\hat{\theta}_{n_1}^{(K_1)} = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_1})$ is given by (2.5). Define similarly $\hat{\psi}_{n_2}$.

Theorem 3.4 *The estimators $\hat{\theta}_{n_1}$ and $\hat{\psi}_{n_2}$ are asymptotically normal distributed. More specifically, there exists positive definite matrices Σ_1 and Σ_2 such that*

$$\sqrt{n_1}(\hat{\theta}_{n_1} - \theta) \to N_{K_1-1}(0, \Sigma_1) \tag{3.5}$$

$$\sqrt{n_2}(\hat{\psi}_{n_2} - \psi) \to N_{K_2-1}(0, \Sigma_2) \tag{3.6}$$

as $n_1, n_2 \to \infty$.

Proof. We are going to prove (3.5), (3.6) is completely analogous. We need to show that all $f \in \mathcal{F}_\mu$ satisfy the regularity conditions. First it is obvious that $\widetilde{\Theta}_0$ is an open set and the model is identifiable.

Note that $e^{\langle \theta, \tilde{M}(x) \rangle}$ is of the class $\mathcal{C}^\infty(\widetilde{\Theta}_0)$. It is easy to verify that $\int_{\mathcal{X}} e^{\langle \theta, \tilde{M}(x) \rangle} < \infty$, since $\mathcal{X}$ is a compact set and $f \in \mathcal{F}_\mu$. In addition,

$$\frac{\partial}{\partial \theta_i} e^{\langle \theta, \tilde{M}(x) \rangle} = \tilde{M}_i(x) e^{\langle \theta, \tilde{M}(x) \rangle} \tag{3.7}$$

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} e^{\langle \theta, \tilde{M}(x) \rangle} = \tilde{M}_i(x) \tilde{M}_j(x) e^{\langle \theta, \tilde{M}(x) \rangle} \tag{3.8}$$

which exist and are continuous functions for each $(\theta, x) \in \widetilde{\Theta}_0 \times \mathcal{X}$. Thus, $f(x|\theta)$ is of the class $\mathcal{C}^\infty(\widetilde{\Theta}_0)$ and the partial derivatives may be passed under the integral sign. Let

$$C(\theta) = \int_{\mathcal{X}} e^{\langle \theta, \tilde{M}(x) \rangle} \tag{3.9}$$

$$C_i(\theta) = \int_{\mathcal{X}} \tilde{M}_i(x) e^{\langle \theta, \tilde{M}(x) \rangle} \tag{3.10}$$

$$C_{ij}(\theta) = \int_{\mathcal{X}} \tilde{M}_i(x) \tilde{M}_j(x) e^{\langle \theta, \tilde{M}(x) \rangle}. \tag{3.11}$$

It is easy to verify that

$$\frac{\partial^2 \log f(x \mid \theta)}{\partial\theta_i\theta_j} = \frac{C_{ij}(\theta)C(\theta) - C_i(\theta)C_j(\theta)}{C(\theta)^2} \tag{3.12}$$

Since $C(\theta)$, $C_i(\theta)$ and $C_{ij}(\theta)$ are continuous functions of θ, then in a closed neighborhood $\mathcal{N}(\theta_0)$ of the true parameter value θ_0, we have

$$\begin{aligned} C_* &:= \min_{\theta\in\mathcal{N}(\theta_0)} C(\theta) > 0 \\ \overline{C}_i &:= \max_{\theta\in\mathcal{N}(\theta_0)} C_i(\theta) < \infty \\ \overline{C}_{ij} &:= \max_{\theta\in\mathcal{N}(\theta_0)} C_{ij}(\theta) < \infty. \end{aligned}$$

Thus,

$$\left|\frac{\partial^2 \log f(x \mid \theta)}{\partial\theta_i\theta_j}\right| \le \frac{|\overline{C}_{ij}|}{C_*} + \frac{|\overline{C}_i\overline{C}_j|}{C_*^2} =: C(i,j) < \infty. \tag{3.13}$$

In a completely analogous way, the third partial derivatives can be bounded.

Let $\mathcal{I}(\theta)$ the Hessian matrix of $\log f(x|\theta)$ with entries

$$\mathcal{I}_{i,k}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial\theta_k\theta_i}\right].$$

Then $\mathcal{I}(\theta)$ is positive definite matrix. To see this, observe that

$$\mathcal{I}_{i,k}(\theta) = \mathrm{Cov}(\tilde{M}_i(X), \tilde{M}_j(X)) = \mathrm{Cov}(M_i(X) - M_K(x), M_j(X) - M_K(X)),$$

which is nonnegative definite. But M_i's form a basis for the finite dimensional approximant space (e.g., natural cubic spline space) and so the columns of $\mathcal{I}(\theta)$ are linear independent. Consequently, $\mathcal{I}(\theta)$ is a positive definite matrix. □

To measure the distance between the two distributions F and G we can use the divergent given by:

$$I_S(F,G) := \int (\log f(x) - \log g(x)) f(x) d\mu(x) + \int (\log g(y) - \log f(y)) g(y) d\mu(y). \tag{3.14}$$

Define the following estimator for $I_S(F,G)$,

$$I_S(\hat{f},\hat{g}) = \int (\log \hat{f}_{K_1}(x) - \log \hat{g}_{K_2}(x))dF_{n_1}(x) + \int (\log \hat{g}_{K_2}(y) - \log \hat{f}_{K_1})dG_{n_2}(y) \tag{3.15}$$

where F_{n_1} and G_{n_2} are the empirical distribution of F and G respectively. In fact, if we have the random samples $\mathbf{X} = (X_1, X_2, \ldots, X_{n_1})$ with distribution F and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{n_2})$ with distribution G, this estimator is of the form

$$\begin{aligned} I_S(\hat{f},\hat{g}) &= \frac{1}{n_1}\Big(\sum_{i=1}^{n_1} \log \hat{f}_{K_1}(X_i) - \sum_{i=1}^{n_1} \log \hat{g}_{K_2}(X_i)\Big) \\ &+ \frac{1}{n_2}\Big(\sum_{i=1}^{n_2} \log \hat{g}_{K_2}(Y_i) - \sum_{i=1}^{n_2} \log \hat{f}_{K_1}(Y_i)\Big) \\ &= \hat{I}_{1,n_1}(\mathbf{X}) + \hat{I}_{2,n_2}(\mathbf{Y}) \end{aligned} \tag{3.16}$$

Lemma 3.17 *For fixed K_1 and K_2, $(\hat{I}_{1,n_1}(\mathbf{X}) + \hat{I}_{2,n_2}(\mathbf{Y})) \longrightarrow (I_1 + I_2) = I_S$ almost surely as $n_1, n_2 \longrightarrow \infty$.*

Proof. It is enough to show that $\hat{I}_{1,n_1}(\mathbf{X}) \longrightarrow I_1$ almost surely as $n_1, n_2 \longrightarrow \infty$, the result for $\hat{I}_{2,n_2}(\mathbf{Y})$ is done similarly. Let

$$D^1_{n_1}(K_1) = \int f(x_1) \log \hat{f}_{K_1}(x_1) d\mu(x_1),$$

$$D^2_{n_1,n_2}(K_2) = \int f(x_1) \log \hat{g}_{K_2}(x_1) d\mu(x_1),$$

$$Z^1_{n_1}(K_1) = n_1^{-1} \sum_{i=1}^{n_1} \log \hat{f}_{K_1}(X_{1,i}),$$

and

$$Z^2_{n_1,n_2}(K_2) = n_1^{-1} \sum_{i=1}^{n_1} \log \hat{g}_{K_2}(X_{1,i}).$$

To simplify the notation we will write $\hat{I}_{1,n_1}(\mathbf{X}) = \hat{I}_1$, $\hat{I}_{2,n_2}(\mathbf{Y}) = \hat{I}_2$, $D^1_{n_1}(K_1) = D^1$, $D^2_{n_1,n_2}(K_2) = D^2$, $Z^1_{n_1}(K_1) = Z^1$, $Z^2_{n_1,n_2}(K_2) = Z^2$.

Note that, $\hat{I}_1 = Z^1 - Z^2$. Notice that it is sufficient to show that,

$$(D^1 - Z^1) + (D^2 - Z^2) \longrightarrow 0,$$

almost surely as $n_1, n_2 \longrightarrow \infty$. For this, observe that,

$$(D^1 - Z^1) = \sum_{j=1}^{K_1-1} \hat{\theta}_{n_1,j}\Big(\int f(x_1)\tilde{M}_{j,1}(x)d\mu(x) - n^{-1}\sum_{i=1}^{n_1}\tilde{M}_{j,1}(X_i)\Big)$$

and

$$(D^2 - Z^2) = \sum_{j=1}^{K_1-1} \psi_{n_1,j}\Big(\int f(x)\tilde{M}_{j,2}(x)d\mu(x) \ - \ n^{-1}\sum_{i=1}^{n_1}\tilde{M}_{j,2}(X_i)\Big)$$

Following Dias (2000), we have, if $\hat{\theta}_{n_1}$ and ψ_{n_2} are the maximum likelihood estimators then $\hat{\theta}_{n_1} \longrightarrow \theta_0$ and $\psi_{n_2} \longrightarrow \psi_0$ almost surely as $n_1, n_2 \longrightarrow 0$, where θ_0 and ψ_0 are the true parameter values. Moreover, by the strong law of large numbers, $n_1^{-1}\sum_{i=1}^{n_1}\tilde{M}_{l,j}(X_i) \longrightarrow \int f(x)\tilde{M}_{j,1}(x)d\mu(x)$ almost surely as $n_1 \longrightarrow \infty$, for $j = 1, \ldots, K_1 - 1$. Thus, $\hat{I}_1 \longrightarrow I_1$ almost surely as n_1, n_2 goes to infinity. □

Theorem 3.18 *For all $f, g \in \mathcal{F}_\mu$ we have a positive constant σ_I such that*

$$\sqrt{n_1}(\hat{I}_{1,n_1}(\mathbf{X}) - I_1) + \sqrt{n_2}(\hat{I}_{2,n_2}(\mathbf{Y}) - I_2) \longrightarrow N(0, \sigma_I)$$

as $n_1, n_2 \longrightarrow \infty$. Note that under H_0 we have $I_1 = I_2 = 0$ and the result turns to be

$$\sqrt{n_1}\,\hat{I}_{1,n_1}(\mathbf{X}) + \sqrt{n_2}\,\hat{I}_{2,n_2}(\mathbf{Y}) \longrightarrow N(0, \sigma_I)$$

where $\sigma_I = \sigma_{I_1} + \sigma_{I_2}$ with

$$\begin{aligned}
\sigma_{I_1} \quad = \quad & \sum_{j=1}^{K_1-1} \theta_j^2 \mathrm{Var}\left(\tilde{M}_{j,1}(X_i)\right) + 2\sum_{j=1}^{K_1-2}\sum_{l=j}^{K_1-1} \theta_j\theta_l \,\mathrm{Cov}\left(\tilde{M}_{j,1}(X_i), \tilde{M}_{l,1}(X_i)\right) \\
& + \sum_{j=1}^{K_2-1} \psi_j^2 \mathrm{Var}\left(\tilde{M}_{j,2}(X_i)\right) + 2\sum_{j=1}^{K_2-2}\sum_{l=j}^{K_2-1} \psi_j\psi_l \,\mathrm{Cov}\left(\tilde{M}_{j,2}(X_i), \tilde{M}_{l,2}(X_i)\right) \\
& +2\sum_{j=1}^{K_1-1}\sum_{l=1}^{K_2-1} \theta_j\psi_l \,\mathrm{Cov}\left(\tilde{M}_{j,1}(X_i), \tilde{M}_{l,2}(X_i)\right) \qquad (3.19)
\end{aligned}$$

with a completely analogous expression for σ_{I_1} substituting accordingly.

Proof.

$$\begin{aligned}
\sqrt{n_1}(\hat{I}_1 - I_1) &= \sqrt{n_1}(\int f(x)\{\sum_{j=1}^{K_1-1}\theta_j\tilde{M}_{j,1}(x) - \sum_{l=1}^{K_2-1}\psi_l\tilde{M}_{l,2}(x)\}d\mu(x) \\
&- \frac{1}{n_1}\{\sum_{i=1}^{n_1}\sum_{j=1}^{K_1-1}\theta_j\tilde{M}_{j,1}(X_i) - \sum_{l=1}^{K_2-1}\psi_l\tilde{M}_{l,2}(X_i)\} \\
&+ \sqrt{n_1}(\int f(x)\{\sum_{j=1}^{K_1-1}(\hat{\theta}_j-\theta_j)\tilde{M}_{j,1}(x) - \sum_{l=1}^{K_2-1}(\hat{\psi}_l-\psi_l)\tilde{M}_{l,2}(x)\}d\mu(x) \\
&- \frac{1}{n_1}\{\sum_{i=1}^{n_1}\sum_{j=1}^{K_1-1}(\hat{\theta}_j-\theta_j)\tilde{M}_{j,1}(X_i) - \sum_{l=1}^{K_2-1}(\hat{\psi}_l-\psi_l)\tilde{M}_{l,2}(X_i)\}.
\end{aligned}$$

Observe that, as $n_1 \longrightarrow \infty$ we have

$$\sum_{j=1}^{K_1-1}\sqrt{n_1}(\hat{\theta}_j-\theta_j)(\int f(x)\tilde{M}_{j,1}(x)d\mu(x) - \frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{M}_{j,1}(X_i)) \longrightarrow 0,$$

in probability since

$$\sqrt{n_1}(\hat{\theta}_j-\theta_j) \longrightarrow N(0,\sigma_{\hat{\theta}})$$

and

$$(\int f(x)\tilde{M}_{j,1}(x)d\mu(x) - \frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{M}_{j,1}(X_i)) \longrightarrow 0$$

in probability. Also, by the Central Limit Theorem for i.i.d. random variables

$$\sqrt{n_1}\left(\int f(x)\{\sum_{j=1}^{K_1-1}\theta_j\tilde{M}_{j,1}(x)\}d\mu(x) - \frac{1}{n_1}\{\sum_{i=1}^{n_1}\sum_{j=1}^{K_1-1}\theta_j\tilde{M}_{j,1}(X_i)\}\right) \longrightarrow N(0,\sigma_{S_1}),$$

and

$$\sqrt{n_1}\left(\int f(x)\{\sum_{l=1}^{K_2-1}\psi_l\tilde{M}_{l,1}(x)\}d\mu(x) - \frac{1}{n_1}\{\sum_{i=1}^{n_1}\sum_{l=1}^{K_2-1}\psi_l\tilde{M}_{l,2}(X_i)\}\right) \longrightarrow N(0,\sigma_{S_2})$$

where $\sigma_{S_1} = \mathrm{Var}\left(\sum_{j=1}^{K_1-1}\theta_j\tilde{M}_{j,1}(X_i)\right)$ and $\sigma_{S_2} = \mathrm{Var}\left(\sum_{j=1}^{K_2-1}\psi_j\tilde{M}_{j,1}(X_i)\right)$. Hence,

$$\sqrt{n_1}(\hat{I}_1 - I_1) \longrightarrow N(0,\sigma_{I_1}),$$

where σ_{I_1} is given by (3.19). The result

$$\sqrt{n_2}(\hat{I}_2 - I_2) \longrightarrow N(0, \sigma_{I_2}),$$

is proved analogously and noticing that $\hat{I}_1$ and $\hat{I}_2$ are independent random variables we have the desired result. □

Extensions. This procedure can be extended to test closeness of multivariate distribution functions by using tensor product among the B-spline basis. Also, one might consider the dimension of the approximant spaces (K_1 and K_2) to be unknown and estimated from the data using either an adaptive procedure similar to H-splines (Dias (1998)) or a Bayesian approach similar to the one proposed by Dias and Gamerman (2002) for nonparametric regression.

4 Simulation results

In order to assess the range of applicability we performed some simulation for small samples using several known distributions. Figure 4.1 shows that the normal distribution for the test statistic holds even for samples of size 30 when the underlying true distribution is normal. This result was verified using 1000 bootstrap resampling of the original data and 1000 independent replications of the sampling distribution. For non-symmetric distributions such as gamma distributions we have a small skewness to the right.

Moreover, we compare SKL test with Kolmogorov-Smirnov (K-S) test which is the most used nonparametric test for comparing continuous distributions. It is well-known that K-S test presents problems in heavy-tailed distributions (see, Mason and Schuenemeyer (1983) and Mason and Schuenemeyer (1992)). Therefore, we chose to make this comparison in terms of power using using 2000 bootstrap replications of

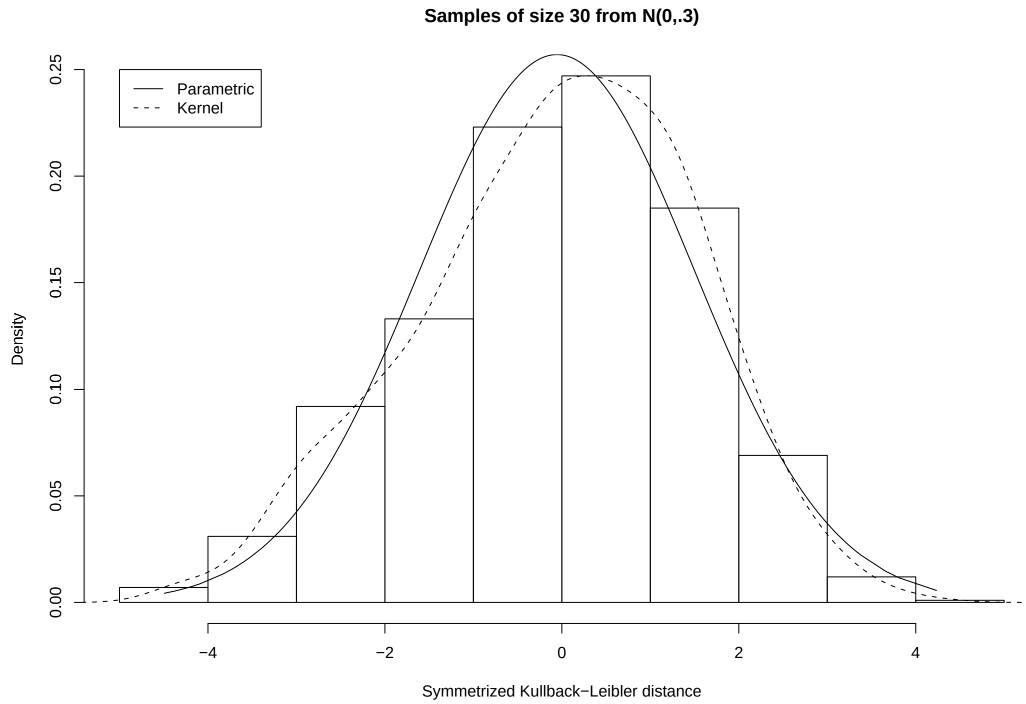

Figure 4.1: Parametric and kernel estimate for the distribution of SKL

mixture of normal distributions. The sampling distribution is given by

$$f(x) = .8\phi((x + .5)/.6) + .2\phi((x - \mu)/.6) \tag{4.1}$$

where ϕ is the standard normal density and μ is the mean of the contaminating distribution. Table 4.1 and Figure 4.2 show that SKL is consistently more powerful than K-S in this case.

μ	-.5	0	.5	.7	.9	1.0	1.1	1.2	1.5
SKL	.045	.099	.161	.256	.384	.541	.663	.782	.991
K-S	.045	.098	.124	.233	.356	.444	.500	.616	.885

Table 4.1: Power function for SKL and K-S for mixture of normal distributions

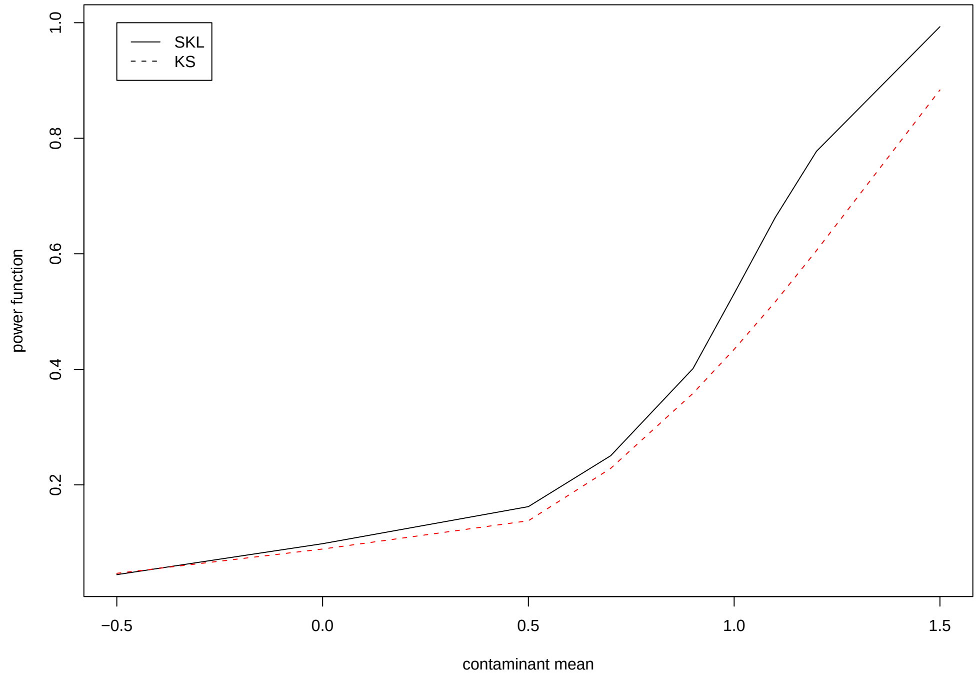

Figure 4.2: Power function for SKL and K-S for mixture of normal distributions

5 Numerical example

In this section we use data from two sources. One of them is the data from Ramus et al. (1999). The other consists of 20 sentences from Portuguese read by two speakers of Modern European Portuguese and Brazilian Portuguese (EP and BP respectively). These sentences were designed by Sonia Frota and Charlotte Galves to study several characteristics of Portuguese, not only consonantal and vocalic intervals, but also stressed syllables, secondary stressed syllables among others. These sentences were recorded at 16 kHz and 11kHz and then segmented by hand by two persons. They used both audio and visual clues to identify consonantal and vocalic intervals and used Multi Speech 3700 software to analyze the acoustic signal. The same procedure was used by Ramus et al. (1999) to record and segment the other acoustic data as well.

For each sentence the duration of the consonantal intervals were computed, call this variable C. This variable is important in view of the work of Ramus et al. (1999) which could cluster 8 languages into into three groups which correspond precisely to the intuitive notion of rhythmic classes: English, Polish and Dutch represent the accentual class, French, Spanish, Catalan and Italian represent the syllabic class and Japanese represents the moraic class. The same variable was used by Duarte, Galves, Garcia and Maronna (2001) using a parametric approach adjusting a gamma model to fit the data from all languages. Maximum likelihood ratio tests seems to confirm Ramus et al. classification and placed European Portuguese among the accentual languages and Brazilian Portuguese among the syllabic ones. Using SKL and K-S to compare the distribution of C for some of the languages we obtained somehow different results. First of all, we cannot distinguish between Brazilian and European Portuguese (p-values: SKL=.69 and K-S=.66). Also, at a 5% significance level there is evidence of difference between Brazilian Portuguese and Catalan (p-values: SKL=.02 and K-S=.02). Figure 5.3 presents density estimates by kernel and by SKL (Dias2000) suggesting that Catalan is bimodal, maybe a mixture of two gammas and this causes the tests to reject the equality of the distributions.

As conjectured there is significant evidence for difference between English and Japanese (p-values: SKL=.01 and K-S $< 10^{-3}$) and no evidence of difference between English and Dutch (p-values: SKL=.59 and K-S=.18) and English and European Portuguese (p-values: SKL=.15 and K-S=.05).

Acknowledgments

We thank Ricardo Molina for recording the data, Janaisa Viscardi and Flaviane Fernandes for segmenting the data, Frank Ramus, Marina Nespor and Jacques Meller for making their data available. We also thank Denise Duarte, Antonio Galves and Charlotte Galves for many fruitful discussions.

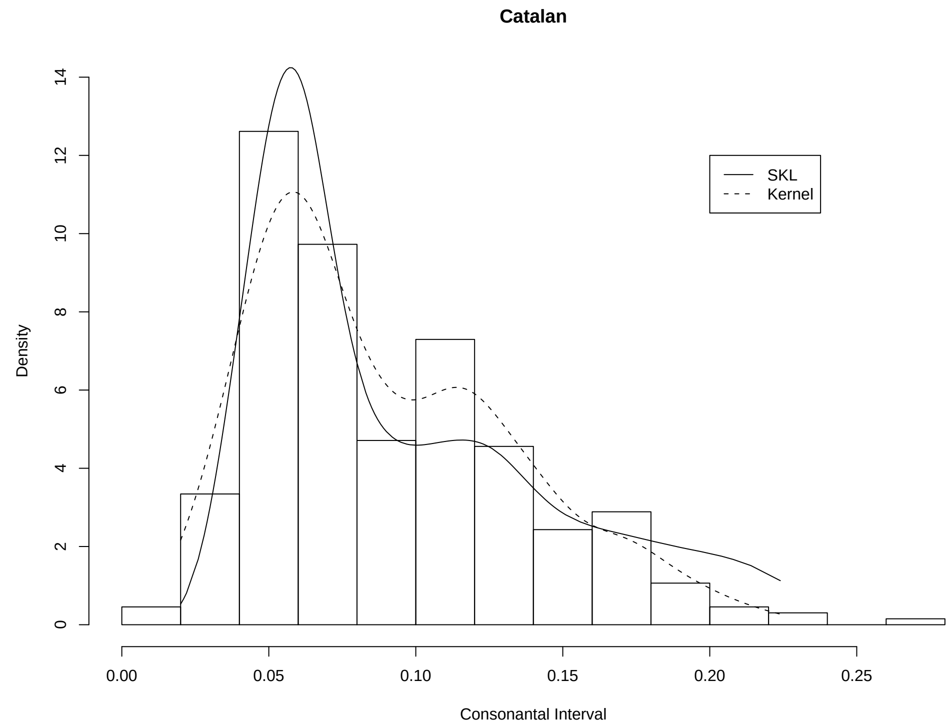

Figure 5.3: Density estimates for consonantal intervals of Catalan

References

Abercrombie, D. (1967). *Elements of general phonetics*, Aldine, Chicago.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (2000). A note on density estimation using a proxy of the Kullback-Leibler distance, *Brazilian Journal of Probability and Statistics* **13**(2): 181–192.

Dias, R. and Gamerman, D. (2002). A Bayesian approach to hybrid splines nonparametric regression, *Journal of Statistical Computation and Simulation.* **72**(4): 285–297.

Duarte, D., Galves, A., Garcia, N. and Maronna, R. (2001). The statistical analysis of acoustic correlates of speech rhythm, *Workshop on Rhythmic patterns, parame-*

ter setting and language change, ZiF, University of Bielefeld. Pode ser obtido em http://www.physik.uni-bielefeld.de/complexity/duarte.pdf p. Paper presented.

Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters, *Econometric Theory* **14**(5): 604–621.

Ferguson, T. S. (1996). *A course in large sample theory*, Texts in Statistical Science Series, Chapman & Hall, London.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions, *Econometric Rev.* **15**(3): 261–274.

Mason, D. M. and Schuenemeyer, J. H. (1983). A modified Kolmogorov-Smirnov test sensitive to tail alternatives, *Ann. Statist.* **11**(3): 933–946.

Mason, D. M. and Schuenemeyer, J. H. (1992). Correction: "A modified Kolmogorov-Smirnov test sensitive to tail alternatives" [Ann. Statist. **11** (1983), no. 3, 933–946; MR 85c:62113], *Ann. Statist.* **20**(1): 620–621.

Ramus, F., Nespor, M. and Mehler, J. (1999). Correlates of linguistic rhythm in speech, *Cognition* **73**: 265–292.