Scientific Research Publishing

Advances in Pure Mathematics, 2023, 13, 226-236
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Goodness-of-Fit Test for Non-Stationary and Strongly Dependent Samples

Carolina Crisci, Gonzalo Perera, Lia Sampognaro

Departamento Modelización Estadística de Datos e Inteligenica Artificial (MEDIA), CURE, Rocha, Universidad de la República, Montevideo, Uruguay
Email: gperera@cure.edu.uy

How to cite this paper: Crisci, C., Perera, G. and Sampognaro, L. (2023) Goodness-of-Fit Test for Non-Stationary and Strongly Dependent Samples. *Advances in Pure Mathematics*, **13**, 226-236.
https://doi.org/10.4236/apm.2023.135016

Received: March 31, 2023
Accepted: May 12, 2023
Published: May 15, 2023

Abstract

In this article we improve a goodness-of-fit test, of the Kolmogorov-Smirnov type, for equally distributed- but not stationary-strongly dependent data. The test is based on the asymptotic behavior of the empirical process, which is much more complex than in the classical case. Applications to simulated data and discussion of the obtained results are provided. This is, to the best of our knowledge, the first result providing a general goodness of fit test for non-weakly dependent data.

Keywords

Kolmogorov-Smirnov Test, Strongly Dependent Data, Asymptotic Behavior of Empirical Processes

1. Introduction

Kolmogorov-Smirnov (KS, for short, in the sequel) is one of the best-known goodness-of-fit tests for *iid* samples following a continuous distribution. For a small or moderate sample size, the critical values of the KS statistic, given the level of significance (or the p-value for a given KS statistic) can be computed exactly [1]. For large sample sizes, the asymptotic behavior of the empirical process, which we will recall later on, provide an approximation to the critical values. Several extensions of KS-type tests from the classical *iid* case to weakly dependent data have been developed and there are substantial recent contributions in this regard (see for instance, [2] [3]). However, the literature does not provide such a test for strongly dependent (and non-stationary) data, which is of deep interest for some applications, as we will see later on.

In previous work, we have used a model for strongly dependent and non-stationary data, that can be used in a wide series of fields, and that allows to develop

different techniques, such as High Level Exceedances [4], statistics on the mean of a random field [5], non-parametric regression [6], or asymptotic behavior of extremes [7].

The basic idea used in such a model is that data depends on two independent components, one of which is merely random noise *iid* and the other, which specifies the "state" of the system under observation, is categorical, but non-stationary and strongly dependent. For instance, if our data is the maximum wind speed in a given 10-minute period, different combinations of meteorological variables define a finite possible number of "states" of the atmosphere. These states do not determine the wind speed, but have a clear influence on the maximum speed. In general, atmosphere states are not stationary and may present strong correlations with data from many years ago, while each year corresponds to 52,560 periods of ten minutes, and therefore, a strong dependency structure must be taken into account.

As mentioned before, the KS test, and other goodness of fit tests are based on the theory of empirical processes [8]. In particular, the statistic of the KS test leads to consistency against any fixed alternative, thanks to the first theorem concerning the asymptotic behavior of the empirical process for large sample size: the well-known Glivenko-Cantelli theorem. The computation of critical values for the KS test in the case of large sample sizes relies on the second fundamental theorem of the asymptotic theory of empirical processes, namely, the Donsker invariance principle [9]. Finally, for some more intricate asymptotic computations, the so-called Hungarian embedding [10] [11] [12] [13] [14] due to Komlós-Major-Tusnády (KMT for short, in the sequel) is a powerful tool.

Let us know recall in more detail these fundamental results. Consider an *iid* sequence of real random variables $X_1,\cdots,X_n,\cdots$ such that X_1 follows the continuous distribution F, and denote F_n the empirical distribution of the sample $X_1,\cdots,X_n$ defined by

$$F_n(t)=\frac{1}{n}\sum_{i=1}^{n}1_{\{X_i\le t\}},\ \forall t\in\mathbb{R}.$$

The Glivenko-Cantelli theorem establishes that

$$\sup_{t\in\mathbb{R}}\left|F_n(t)-F(t)\right|\xrightarrow[n]{a.s.}0.$$

If we denote by b the Brownian bridge process, defined as the continuous, centered, Gaussian stochastic process of continuous parameter in $[0,1]$ characterized by the covariance structure $E(b_sb_t)=s-st$, $\forall 0\le s\le t\le 1$, then Donsker invariance principle shows that $\sqrt{n}\left(F_n(t)-F(t)\right)\xrightarrow[n]{w}b_{F(t)}$, where "$\xrightarrow{w}$" denotes weak convergence as a stochastic process (in Prohorov metric), which in turn implies that

$$\begin{aligned}&P\left(\sqrt{n}\sup_{t\in\mathbb{R}}\left|F_n(t)-F(t)\right|\ge x\right)\\&\xrightarrow[n]{}P\left(\sup_{t\in\mathbb{R}}\left|b_{F(t)}\right|\ge x\right)=P\left(\sup_{u\in[0,1]}\left|b_u\right|\ge x\right):=Q(x)\ \ \forall\ x\ge 0\end{aligned},$$

where Q is the tail of the well-known Kolmogorov-Smirnov distribution, allow-

ing to the computation of the critical values x, given a level of significance, for the KS test when n is large.

Finally, KMT provides a sequence of Brownian bridges $\left(b^m\right)_{m\in\mathbb{N}}$ and a finite, non-negative random variable C such that

$$F_n(t)=F(t)+\frac{b^m_{F(t)}}{\sqrt{n}}+R_n(t),\ \text{with}\ \ \sup_{t\in\mathbb{R}}\left|R_n(t)\right|\le C\frac{\sqrt{\log(n)}}{n}.$$

2. Main Result

We shall consider the following model: $X_1,\cdots,X_n,\cdots$ will be our data, with $X_i=f(\xi_i,Y_i)$, where $(\xi_i)_{i\in\mathbb{N}}$, $Y=(Y_i)_{i\in\mathbb{N}}$, independent among them, $(\xi_i)_{i\in\mathbb{N}}$ *iid*, and $Y=(Y_i)_{i\in\mathbb{N}}$ satisfying:

$$Y_i\in\{1,\cdots,k\}\ \ \forall i\in\mathbb{N},$$

$\forall j=1,\cdots,k$ there exists a random variable $\tau_j>0$ such that

(H1) $\tau_n(j)=\frac{1}{n}\sum_{j=1}^n 1_{\{Y_i=j\}}\xrightarrow[n]{a.s.}\tau_j(Y)$, where $\sum_{j=1}^k\tau_j=1$.

Thinking of Y_i as the state of the system at time i, even if the process Y is not stationary, assumption **(H1)** means that the observed frequencies of any state are convergent on average (which holds true under seasonal effects or monotonous trends), but since Y also exhibits strong dependence, the corresponding limits are random variables.

The empirical distribution of our sample is:

$$\begin{aligned}F_n(t)&=\frac{1}{n}\sum_{i=1}^n 1_{\{X_i\le t\}}=\frac{1}{n}\sum_{i=1}^n 1_{\{f(\xi_i,Y_i)\le t\}}\\&=\sum_{j=1}^k\left[\frac{1}{n}\sum_{i=1}^n\left(1_{\{f(\xi_i,j)\le t\}}1_{\{Y_i=j\}}\right)\right]\end{aligned}\tag{1}$$

Let us define for any $n\in\mathbb{N}$ and $j=1,\cdots,k$,

$$A_n(j)=\frac{1}{n}\sum_{j=1}^n\left(1_{\{f(\xi_i,j)\le t\}}1_{\{Y_i=j\}}\right)\tag{2}$$

Let us call S to the space of all the sequences taking values in $\{1,\cdots,k\}$. Considering **(H1)**, the subset of S defined by:

$$\Omega_Y=\left\{(y_i)_{i\in\mathbb{N}}\in S/\tau_n(j,y):=\frac{1}{n}\sum_{i=1}^n 1_{\{y_i=j\}}\xrightarrow[n\to\infty]{}\tau_j(y)\ \forall\ j=1,\cdots,k\right\}$$

fulfills that,

$$P^Y(\Omega_Y)=P(Y\in\Omega_Y)=1\tag{3}$$

therefore by conditioning $A_n(j)$ to $Y=y$, with $y=(y_i)_{i\in\mathbb{N}}$, we can assume that $y\in\Omega_Y$ and, in such a case, by the independence of $(\xi_i)_{i\in\mathbb{N}}$ and Y:

$$\left(A_n(j)\right)_{j=1,\cdots,k}/Y=y\sim\left(\frac{1}{n}\sum_{i=1}^n 1_{\{f(\xi_i,j)\le t\}}1_{\{y_i=j\}}\right)_{j=1,\cdots,k}\tag{4}$$

Remark 1

It is a key fact that the k subsamples $\left(f(\xi_i,j)\right)_{\{i:y_i=j\}}$, $j=1,\cdots,k$, are inde-

pendent and each one of size $n\tau_n(j,y)$.

If we call F^j to the distribution of $f(\xi_0, j)$ and assume that F^j is continuous we have that:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} 1_{\{f(\xi_i,j)\le t\}} 1_{\{y_i=j\}} &= \frac{n\tau_n(j,y)}{n}\frac{1}{\tau_n(j,y)}\sum_{i=1}^{n} 1_{\{f(\xi_i,j)\le t\}} 1_{\{y_i=j\}} \\ &= \tau_n(j,y) F^j_{\tau_n(j,y)n}(t)\end{aligned} \tag{5}$$

where $F^j_{\tau_n(j,y)n}$ is the empirical distribution of the subsample *j* of ***Remark 1***, with distribution F^j, which are independent among them. Then, from Glivenko-Cantelli

$$\tau_n(j,y) F^j_{\tau_n(j,y)n}(t) \xrightarrow[n]{a.s.} \tau_j(y) F^j(t) \quad \forall t \in \mathbb{R},\ j = 1,\cdots,k \tag{6}$$

and

$$\sup_{t\in\mathbb{R}} \left|\tau_n(j,y) F^j_{\tau_n(j,y)n}(t) - \tau_j(y) F^j(t)\right| \xrightarrow[n]{a.s.} 0 \tag{7}$$

Then, applying Equations (1) to (6) we have that:

$$\begin{aligned}&P\left(\sup_{t\in\mathbb{R}} \left|F_n(t) - \sum_{j=1}^{k} \tau_j F^j(t)\right| \xrightarrow[n]{} 0\right) \\ &= \int_{\Omega_Y} P\left(\sup_{t\in\mathbb{R}} \left|F_n(t) - \sum_{j=1}^{k} \tau_j F^j(t)\right| \xrightarrow[n]{} 0 / Y = y\right) \mathrm{d}P^Y(y) \\ &= \int_{\Omega_Y} P\left(\sup_{t\in\mathbb{R}} \left|\sum_{j=1}^{k} \left(\tau_n(j,y) F^j_{\tau_n(j,y)n}(t) - \tau_j F^j(t)\right)\right| \xrightarrow[n]{} 0 / Y = y\right) \mathrm{d}P^Y(y) \\ &\ge \int_{\Omega_Y} P\left(\sum_{j=1}^{k} \sup_{t\in\mathbb{R}} \left|\tau_n(j,y) F^j_{\tau_n(j,y)n}(t) - \tau_j F^j(t)\right| \xrightarrow[n]{} 0 / Y = y\right) \mathrm{d}P^Y(y) \\ &= 1\end{aligned}$$

(since each value of the last integrator equals one by (7)).

In conclusion we get:

Theorem 1

Under the previous hypotheses,

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - \sum_{j=1}^{k} \tau_j F^j(t)\right| \xrightarrow[n]{a.s.} 0$$

Remark 2

It should be noticed that

$$\sum_{j=1}^{k} \tau_j F^j(t) \tag{8}$$

is a random mixture of the distributions F^j, $j = 1,\cdots,k$.

Let us look more closely to a very simple case. Assume that $k = 2$ (therefore, $\tau_2 = 1 - \tau_1$), and that τ_1 takes values 0 or 1 with $P(\tau_1 = 0) = p$, $P(\tau_1 = 1) = 1 - p$, where $0 < p < 1$.

Then, with probability *p*, when $\tau_1 = 0$ the random mixture (8) equals $F^2(t)$, and with probability $1 - p$, when $\tau_1 = 1$ the random mixture (8) equals $F^1(t)$ and hence, (8) is just an ordinary mixture of F^1 and F^2.

The preceding result shows that a KS-type test will be consistent under any given alternative in this context, but to improve the test, computing the critical value for a given significance level (or *p*-values), we need a refinement of ***Theo-***

rem 1, providing the asymptotic distribution of the test statistic. This will be obtained in ***Theorem 2***.

Given $j=1,\cdots,k$ fixed, then the sequence $\left(f\left(\xi_i,j\right)\right)_{i\in\mathbb{N}}$ is *iid* with distribution F^j and therefore, calling F_n^j to its corresponding empirical distribution, that is:

$$F_n^j\left(t\right)=\frac{1}{n}\sum\nolimits_{i=1}^{n}1_{\left\{f\left(\xi_i,j\right)\le t\right\}},$$

then, from KMT, there exists one sequence of Brownian bridges $\left(b^{m,j}\right)_{m\in\mathbb{N}}$ such as:

$$F_n^j\left(t\right)=F^j\left(t\right)+\frac{b^{n,j}_{F^j(t)}}{\sqrt{n}}+R_n^j\left(t\right),$$

where $\sup_{t\in\mathbb{R}}\left|R_n^j\left(t\right)\right|\le C^j\frac{\sqrt{\log n}}{n}$ and C^j is a finite and non-negative random variable.

Remark 3

Since the sequence of bridges $\left(b^{m,j}\right)_{m\in\mathbb{N}}$ is originated by $\left(f\left(\xi_i,j\right)\right)_{i\in\mathbb{N}}$, it depends on $\left(\xi_i\right)_{i\in\mathbb{N}}$, which is independent of $Y=\left(Y_i\right)_{i\in\mathbb{N}}$, and therefore, it must be taken into account that all the bridges $\left(b^{m,j}\right)_{m\in\mathbb{N},j=1,\cdots,k}$ are independent of *Y*.

Let us then consider $x\ge 0$ and compute:

$$\begin{aligned}&P\left(\sqrt{n}\sup\nolimits_{t\in\mathbb{R}}\left|F_n\left(t\right)-\sum_{j=1}^{k}\tau_jF^j\left(t\right)\right|\ge x\right)\\&=\int_{\Omega_Y}P\left(\sup\nolimits_{t\in\mathbb{R}}\sqrt{n}\left|F_n\left(t\right)-\sum_{j=1}^{k}\tau_jF^j\left(t\right)\right|\ge x/Y=y\right)\mathrm{d}P^Y\left(y\right)\end{aligned}\tag{9}$$

Now, given that $Y=y=\left(y_i\right)_{i\in\mathbb{N}}$,

$$\begin{aligned}F_n\left(t\right)&=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}1_{\left\{f\left(\xi_i,j\right)\le t\right\}}1_{\left\{y_i=j\right\}}\\&=\sum_{j=1}^{k}\frac{\tau_n\left(j,y\right)}{n\tau_n\left(j,y\right)}\sum_{i=1}^{n}1_{\left\{f\left(\xi_i,j\right)\le t\right\}}1_{\left\{y_i=j\right\}}\end{aligned}\tag{10}$$

But (10), as a stochastic process, has the same distribution as:

$$\sum_{j=1}^{k}\tau_n\left(j,y\right)\frac{1}{n\tau_n\left(j,y\right)}\sum_{i=1}^{i=n\tau_n\left(j,y\right)}1_{\left\{f\left(\xi_i^j,j\right)\le t\right\}}$$

where $\left(\xi_i^j\right)_{i\in\mathbb{N}}$ is *iid* with distribution equal to that of ξ_0 and such that, when *j* varies, the sequences $\left(\xi_i^j\right)_{i\in\mathbb{N}}$ are independent among them.

If we now return to *Remark 3*, building the Hungarian embedding for each $\left(\xi_i^j\right)_{i\in\mathbb{N}}$, we may assume that the KMT representation for the empirical distribution is valid with a sequence of Brownian bridges $\left(b^{m,j}\right)_{m\in\mathbb{N},\,j=1,\cdots,k}$, that are not only independent with respect to *Y* but also independent among them when *j* varies. Therefore, as keeping the distribution unchanged does not affect the probabilities, we have that (9) equals to:

$$\begin{aligned}&\int_{\Omega_Y} P\left(\sup_{t\in\mathbb{R}}\sqrt{n}\left|\sum_{j=1}^{k}\tau_n(j,y)F^j_{n\tau_n(j,y)}(t)-\tau_j F^j(t)\right|\geq x/Y=y\right)\mathrm{d}P^Y(y)\\&=\int_{\Omega_Y} P\left(\sup_{t\in\mathbb{R}}\sqrt{n}\left|\sum_{j=1}^{k}\tau_n(j,y)\left(F^j(t)+\frac{b^{n\tau_n(j,y),j}_{F^j(t)}}{\sqrt{n\tau_n(j,y)}}+R^j_{n\tau_n(j,y)}(t)\right)\right.\right.\\&\left.\left.-\tau_j F^j(t)\right|\geq x/Y=y\right)\mathrm{d}P^Y(y)\end{aligned}\tag{11}$$

Considering in (11) that the terms $R^j_{n\tau_n(j,y)}$ are negligible and that, as indicated above, the distribution as a process of the summation is not changed (and therefore neither does the probability) if instead of $\left(b^{n\tau_n(j,y)}\right)_{j=1,\cdots,k}$ we put $\left(b^j\right)_{j=1,\cdots,k}$ Brownian bridges independent among them and with respect to Y (and therefore with respect to the $\tau_n(j,.)$ and τ_j), we have that (11) equals to:

$$\begin{aligned}&\int_{\Omega_Y} P\left(\sup_{t\in\mathbb{R}}\left|\sqrt{n}\sum_{j=1}^{k}\left(\tau_n(j,y)-\tau_j\right)F^j(t)+\sum_{j=1}^{k}\frac{b^j_{F^j(t)}}{\sqrt{\tau_n(j,y)}}\right|\geq x\right)\mathrm{d}P^Y(y)\\&=P\left(\sup_{t\in\mathbb{R}}\left|\sum_{j=1}^{k}\sqrt{n}\left(\tau_n(j)-\tau_j\right)F^j(t)+\sum_{j=1}^{k}\frac{b^j_{F^j(t)}}{\sqrt{\tau_n(j)}}\right|\geq x\right)\end{aligned}\tag{12}$$

If we take the limit for n tending to infinity in (12), under the additional hypothesis.

(H2) The sequence of random vectors $\sqrt{n}\left(\tau_n(j)-\tau_j\right)_{j=1,\cdots,k}\xrightarrow[n]{w}D=(D_1,\cdots,D_k)$ where D is a random vector in $\mathbb{R}^k$, degenerated (since $\sum_{j=1}^{k}D_j=0$), but the vectors of $\mathbb{R}^{k-1}$ obtained by the suppression of one of any of the k coordinates of D are not degenerated, and where D is independent of the Brownian bridges $\left(b^j\right)_{j=1,\cdots,k}$, we finally have that (12) tends to:

$$P\left(\sup_{t\in\mathbb{R}}\left|\sum_{j=1}^{k}D_jF^j(t)+\sum_{j=1}^{k}\frac{b^j_{F^j(t)}}{\sqrt{\tau_j}}\right|\geq x\right)$$

Therefore we have:

Theorem 2

Under the previous hypotheses, $\forall x\geq 0$:

$$\begin{aligned}&P\left(\sqrt{n}\sup_{t\in\mathbb{R}}\left|F_n(t)-\sum_{j=1}^{k}\tau_jF^j(t)\right|\geq x\right)\xrightarrow[n]{}\\&P\left(\sup_{t\in\mathbb{R}}\left|\sum_{j=1}^{k}D_jF^j(t)+\sum_{j=1}^{k}\frac{b^j_{F^j(t)}}{\sqrt{\tau_j}}\right|\geq x\right):=T(x)\end{aligned}$$

Remark 4

The expression of $T(x)$ can be computed by Monte Carlo as will be seen in the next section.

Remark 5

Obviously, for practical purposes, D_j and τ_j should be often replaced by their empirical estimations.

3. A Model for Simulated Data

For our simulations, we will use the model of *Example 2* of [7], with some minor modifications.

Consider $\sigma(1)$, $\sigma(2)$ independent, such that

$$P(\sigma(1)=1)=\delta,\ P(\sigma(1)=2)=1-\delta,\ P(\sigma(2)=1)=\eta,$$

$$P(\sigma(2)=2)=1-\eta,\ 0<\delta<1,\ 0<\eta<1, \text{ and } \delta\neq\eta.$$

Let $(\sigma_1(1),\sigma_1(2)),\cdots,(\sigma_n(1),\sigma_n(2)),\cdots$ *iid* with the same distribution as $(\sigma(1),\sigma(2))$, and consider a fixed random variable U independent of $(\sigma_1(1),\sigma_1(2)),\cdots,(\sigma_n(1),\sigma_n(2)),\cdots$, such that

$$P(U=1)=p, P(U=2)=1-p, 0<p<1$$

and define

$$Y_i := \sigma_i(U) \tag{13}$$

Then $k=2$, and:

$$\tau_n(1)=\frac{1}{n}\sum_{i=1}^{n}1_{\{Y_i=1\}}$$

Hence

$$\tau_n(1)/U=1\sim\frac{1}{n}\sum_{i=1}^{n}1_{\{\sigma_i(1)=1\}}\xrightarrow[n]{a.s.}\delta$$

(by the Strong Law of Large Numbers), and in a similar way

$$\tau_n(1)/U=2\sim\frac{1}{n}\sum_{i=1}^{n}1_{\{\sigma_i(2)=1\}}\xrightarrow[n]{a.s.}\eta, \text{ and then } \tau_n(1)\xrightarrow[n]{a.s.}\tau_1,$$

with

$$\tau_1=\begin{cases}\delta & \text{if } U=1\\ \eta & \text{if } U=2\end{cases} \tag{14}$$

On the other hand,

$$\tau_n(2)=\frac{1}{n}\sum_{i=1}^{n}1_{\{Y_i=2\}}$$

and

$$\tau_n(2)/U=1\sim\frac{1}{n}\sum_{i=1}^{n}1_{\{\sigma_i(1)=2\}}\xrightarrow[n]{a.s.}1-\delta,$$

and

$$\tau_n(2)/U=2\sim\frac{1}{n}\sum_{i=1}^{n}1_{\{\sigma_i(2)=2\}}\xrightarrow[n]{a.s.}1-\eta,$$

and then $\tau_n(2)\xrightarrow[n]{a.s.}\tau_2$, with

$$\tau_2 = \begin{cases} 1-\delta & \text{if } U=1 \\ 1-\eta & \text{if } U=2 \end{cases} \tag{15}$$

and thus, by (14) and (15), **(H1)** is satisfied.

Now consider the bivariate random vector $\sqrt{n}\left(\tau_n(j)-\tau_j\right)_{j=1,2}$.

Then

$$\sqrt{n}\left(\tau_n(1)-\tau_1, \tau_n(2)-\tau_2\right)/U=1$$
$$\sim\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\sigma_i(1)=1\}}-\delta\right), \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\sigma_i(1)=2\}}-(1-\delta)\right)\right)\xrightarrow[n]{w} Z_1$$

a bivariate Gaussian, centered, degenerated random vector with covariance matrix

$$\begin{pmatrix} \delta(1-\delta) & -\delta(1-\delta) \\ -\delta(1-\delta) & \delta(1-\delta) \end{pmatrix} \tag{16}$$

by the ordinary Central Limit Theorem, and using the fact that

$$1_{\{\sigma_i(1)=1\}}+1_{\{\sigma_i(1)=2\}}=1, \forall\, i\in\mathbb{N}.$$

On the other hand,

$$\sqrt{n}\left(\tau_n(1)-\tau_1, \tau_n(2)-\tau_2\right)/U=2$$
$$\sim\left(\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\sigma_i(2)=1\}}-\eta\right), \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} 1_{\{\sigma_i(2)=2\}}-(1-\eta)\right)\right)\xrightarrow[n]{w} Z_2$$

a bivariate Gaussian, centered, degenerated random vector with covariance matrix

$$\begin{pmatrix} \eta(1-\eta) & -\eta(1-\eta) \\ -\eta(1-\eta) & \eta(1-\eta) \end{pmatrix} \tag{17}$$

Therefore, setting:

$$D = \begin{cases} Z_1 & \text{if } U=1 \\ Z_2 & \text{if } U=2 \end{cases} \tag{18}$$

then D is a centered degenerated bivariate random vector, whose distribution is a mixture of Gaussian laws, and where the suppression of any of its two coordinates is a non-degenerated one-dimensional mixture of Gaussian distributions.

Furthermore, if we write down $D=(D_1, D_2)$, then it is very easy to check that $D_2=-D_1$, and that D_1 may be represented as $(2-U)W_1+(U-1)W_2$, with W_1, W_2 independent among them and with respect to U, such that

$$W_1\sim N\left(0, \delta(1-\delta)\right), \quad W_2\sim N\left(0, \eta(1-\eta)\right) \tag{19}$$

and therefore, **(H2)** is satisfied.

Finally, consider F^1 and F^2, two continuous distributions such that $F^1\neq F^2$, and two independent sequences $V_1^{(1)},\cdots,V_n^{(1)},\cdots iid\sim F^1$, $V_n^{(2)},\cdots,V_n^{(2)},\cdots iid\sim F^2$ and set:

1) If $\sigma_i(U)=1$, $X_i=V_i^{(1)}$

2) If $\sigma_i(U)=2$, $X_i=V_i^{(2)}$

Then, as seen before, ***Theorem 1*** and ***Theorem 2*** apply to $(X_i)_{i\in\mathbb{N}}$ and therefore, we will simulate large samples of such type of data (for different choices of the couple F^1, F^2), improve the test of the KS type given by ***Theorem 2***, and discuss the results.

4. Application to Simulated Data

Following the model of the previous secction we simulated large samples where the KS-type test provided by ***Theorem 2*** was performed.

We have chosen the required parameters in the following way:

$p = 0.3, \delta = 0.3, \eta = 0.6$. With this choice, and assuming as the true model the corresponding mixture with F^1 a $N(0,1)$ distribution, and F^2 a $N(3,1)$ distribution, we simulated 4000 independent samples of size $n = 500$ of the true model to compute *p*-values by MonteCarlo.

We also simulated an extra independent sample, following the true model, to apply our test.

We proposed for fitting (*i.e.*, as H_0 in our test) a similar mixture model but taking as F^1 a Cauchy distribution with location parameter 0 and scale parameter 1, and as F^2 a Cauchy distribution with location parameter 3 and scale parameter 1.

In this context, the corresponding critical value for the KS statistic (maximal difference between the empirical distribution and the proposed one) was 0.3311402 and MonteCarlo computations leads to a p-value = 0.0285, what clearly implies rejection.

Figure 1 shows the comparison between the empirical distribution of the simulated sample of the true model used for testing, and the theoretical distribution of the proposed model. The difference between them is notorious and in

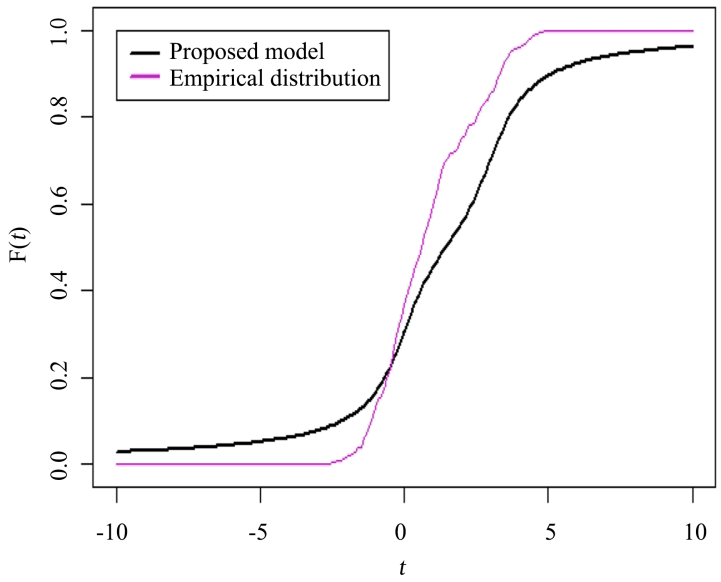

Figure 1. Proposed distribution vs empirical distribution based on simulated data of the true model.

particular, it should be noticed that the distribution of the proposed model, for larger values of the argument, is always clearly below the empirical distribution of the true model. This reflects the fact that the proposed model is much more heavy-tailed than the true model.

5. Conclusions & Further Work

As seen in the previous section, a KS-type test may be performed for non-stationary and strongly dependent samples of large size. Its performance, both in terms of statistical efficiency and computational complexity is satisfactory. A large variety of real data may be analyzed using this tool and other related ones.

In particular, in a forthcoming paper by the same authors, this goodness of fit test plays a key role in the determination of the number of components and relative weights of a mixture of extremal distributions. The previous paper [7] shows that these types of mixtures are suitable for extremal analysis of many environmental data where non-stationarity and strong dependence appear.

Another direction of further work is the extension of this paper to other testing tools based on the asymptotic behavior of the empirical process and related statistical procedures.

Acknowledgements

This work was partial supported by *Proyecto CSIC-VUSP* "*Análisis de eventos climáticos extremos y su incidencia sobre la producción hortifrutícola en Salto*" (*Uruguay*). The authors thank to an anonymous referee for his highly valuable suggestions.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dickinson Gibbons, J. and Chakraborti, S. (2021) Nonparametric Statistical Inference. 6th Edition, Chapman Hall, London.

[2] Tanguep, E. and Njomen, D. (2021) Kolmogorov-Smirnov APF Test for Inhomogeneous Poisson Processes with Shift Parameter. *Applied Mathematics*, **12**, 322-335. https://doi.org/10.4236/am.2021.124023

[3] Zhao, J. and Li, X. (2022) Goodness of Fit Test Based on BLUS Residuals for Error Distribution of Regression Model. *Applied Mathematics*, **13**, 672-682. https://doi.org/10.4236/am.2022.138042

[4] Bellanger, L. and Perera, G. (2003) Compound Poisson Limit Theorems for High-Level Exceedances of Some Non-Stationary Processes. *Bernoulli*, **9**, 497-515. https://doi.org/10.3150/bj/1065444815

[5] Perera, G. (2002) Irregular Sets and Central Limit Theorems. *Bernoulli*, **8**, 627-642.

[6] Aspirot, L., Bertin, K. and Perera, G. (2009) Asymptotic Normality of the Nadaraya-Watson Estimator for Nonstationary Functional Data and Applications to Telecom-

munications. *Journal of Nonparametric Statistics*, **21**, 535-551. https://doi.org/10.1080/10485250902878655

[7] Crisci, C. and Perera, G. (2022) Asymptotic Extremal Distribution for Non-Stationary, Strongly-Dependent Data. *Advances in Pure Mathematics*, **12**, 479-489. https://doi.org/10.4236/apm.2022.128036

[8] Shorack, G.R. and Wellner, J.A. (2009) Empirical Processes with Applications to Statistics. *Classics in Applied Mathematics*, xxxvi + 955. https://doi.org/10.1137/1.9780898719017

[9] Billinsgley, P. (1999) Convergence of Probability Measures. 2nd Edition, John Wiley Sons, Inc., Hoboken. https://doi.org/10.1002/9780470316962

[10] Komlós, J., Major, P. and Tusnády, G. (1975) An Approximation of Partial Sums of Independent RV'-s, and the Sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **32**, 111-131. https://doi.org/10.1007/BF00533093

[11] Komlós, J., Major, P. and Tusnády, G. (1976) An Approximation of Partial Sums of Independent RV's, and the Sample DF. II. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **34**, 33-58. https://doi.org/10.1007/BF00532688

[12] Bretagnolle, J. and Massart, P. (1989) Hungarian Constructions from the Nonasymptotic Viewpoint. *Annals of Probability*, **17**, 239-256. https://doi.org/10.1214/aop/1176991506

[13] Koning, A.J. (1994) KMT-Type Inequalities and Goodness-of-Fit Tests. *Statistica Neerlandica*, **48**, 117-132. https://doi.org/10.1111/j.1467-9574.1994.tb01437.x

[14] Van der Vaart, A.W. (2000) Asymptotic Statistics. Cambridge University Press, Cambridge.