Non-Parametric Information-Theoretic Measures of One-Dimensional Distribution Functions from Continuous Series

PAOLO D'ALBERTO and ALI DASDAN *
Yahoo! Inc.

December 23, 2008

Abstract

We study non-parametric measures for the problem of comparing distributions, which arise during the anomaly detection of (time) series. We present an unified framework where: first, we can compare non-parametric stochastic distance-based approaches; and second, we apply and extend information-theoretic measures to continuous series.

Non-parametric measures take two distributions as input and produce two numbers as output; that is, respectively the difference between the input distributions and the statistical significance of this difference. Some of these measures, such as Kullback-Leibler measure, are defined for comparing probability distribution functions (PDFs) and some measures, such as Kolmogorov-Smirnov measure, are for cumulative distribution functions (CDFs). We first show how to adapt the PDF based measures to compare CDFs, resulting in a total of 23 CDF based measures. We then provide a unified functional form that subsumes all these measures. We present our methodology to determine the significance (of the measures) by simulations only. Finally, we evaluate these measures for the anomaly detection in continuous series.

1 Introduction

In the era of information technology, the volume of data transmitted, handled and elaborated is overwhelming. For example, Yahoo! collects multiple terabytes of data each day, and this volume is increasing. The main reason for collecting this data is to improve the software and hardware systems so as to make them understand what users want and answer user requests almost immediately. It should then be clear that there is a need for automatic detection of the system *variations* to cope with this volume and speed.

One common way to model the data is in terms of (time) series. Then, the variations can be detected by parametric and non-parametric methods. For example, Holt–Winters [18, 49] is an example of the former. In this work, we turn our attention to the latter (set of methods). In particular, we investigate the application of non-parametric measures to compute the distance between variations summarized as distributions (for an comparative study see [29] or for an overview [17]) .

Some distance measures can work with *probability distribution functions* (**PDF**s) (e.g., Kullback–Leibler [28]), and others with **cumulative distribution functions** (**CDF**s) (e.g., 1933 Kolmogorov's seminal paper [27] and expanded in [25, 8]). For reference, we will call the former as PDF measures and the latter as CDF measures. We have learned that: First, multiple measures are more useful than just relying on a single one (e.g., Gestalt effect); second, CDF measures are easier to work with because they do not suffer from the division-into-buckets problems and also they have nice properties like monotonic growth. As such, our contribution is an extension of almost all PDF measures to work with CDFs, unifying all the measures into a single functional form in the process. In addition, we determine how to provide statistical significance when working with these CDF measures.

This paper is organized as follows: in Section 2, we introduce the related work; in Section 3, we introduce our notations and the set of measures. In Section 4, we specify the measures for distribution functions and, in Section 5, we present our approach for the evaluation of the measure statistical significance; in particular, in Section 5.4, we explain how we combine measures to have an agreement-based approach and compare our approach with the state-of-the-art approaches. In Section 6, we present our experimental results and, in Section 7, our final considerations.

2 Related Work

Distribution comparison is a fundamental technique and has many applications in statistics (e.g., hypothesis testing) and other areas. Many distribution comparison measures have been proposed starting from the early 1930s. For example, a simple search at Google's scholar search engine returns thousands of results to papers and citations for each

*E-Mail the authors:{*pdalbert,dasdan*}@*yahoo-inc.com*

distribution comparison measure.

Distribution comparison measures can be applied between one discrete and one continuous distribution (i.e., model) or two discrete distributions (i.e., both empirical). We turn our attention to the latter measures because of our interests and their applications in the web search. For the discrete case, inputs to distribution comparison measures are either probability distribution functions (PDFs) or cumulative distribution functions (CDFs). In the sequel, for ease of reference, we will refer to such measures as PDF-based measures and CDF-based measures, respectively.

Due to the fundamental nature of distribution comparison, there have been many measures proposed to date. For example, Kullback–Leibler measure [28] and Kolmogorov–Smirnov measure [27] are examples of well-known PDF-based and CDF-based measures, respectively. We present a full list in Table 1 and each measure is discussed in detail in Section 4, .

Three attempts to survey the existing measures are [34] (about 64 measures), [30] surveying 7 measures for language modeling, and [38] for a broader classification. There have also been attempts to generalize and unify some of the measures. Such a work includes [1], [42]. In them, Kullback–Leibler is shown to be a special case. In this paper, we also propose a generalized functional form. To the best of our knowledge, we are unaware of any past attempts to generalize PDF-based measures so that they can also take CDFs as inputs and provide a significance (i.e., quantile).

As for anomaly detection over time series, the literature is again fairly large. A short but representative list is for example, martingale methods [44, 15, 16], Kullback–Leibler's application [6], information-theoretic approaches for material modeling [52], image formation [35] or channel denoising [46] or symbolic sequences [12]. All derive from the seminal work by [39] who introduced the concept of entropy for discrete set of probabilities for communication and the Jensen–Shannon divergence (Shannon entropy in combination with Jensen inequality [20]) has been first introduced by [37] or [50] (respectively, suggested as first reference by [31] and [9]). As the original authors said Kullback–Leibler *is* the generalization of Shannon's entropy.

Some of the anomaly detection methods require the specification of a handful of parameters or the automatic estimation by means of a minimization process. One well-known example of such parametric methods is Holt–Winters method [18, 49]. Distribution comparison falls in the category of non-parametric methods. For anomaly detection using distribution comparison measures [26], a typical approach is to specify two windows on the time series, one being the reference window and the other being a moving window whose deviation from the reference window is of interest. When each window is considered as a distribution (in the PDF or CDF form, or other using DSP techniques [51] or linked-based [10]), anomaly detection reduces to the comparison of (two or more) distributions. We also follow this last approach.

In relation to the previous work, our contributions can be summarized as follows:

1. We propose a unified functional form of almost all of the distribution comparison measures.
2. We generalize PDF-based measures to take CDFs as input.
3. We show how to provide statistical significance with these measures.
4. We give experimental results for a single application, namely, anomaly detection on time series.

3 Statistical test, unified measure, and notations

In this section, we introduce our terminology and definitions.

Statistical test. Given two arbitrary (empirical) CDFs F_R and F_W, a (non-parametric) statistic test is composed of three components:

1. The null hypothesis $H_0 : F_R \sim F_W$; that is, the distributions are the same.
2. The (distance) measure $\mathbb{D}(F_R, F_W)$, which quantifies the distance between these distributions. We use the term "measure" after [1].
3. The statistical significance, typically with significance levels at 0.05.

Unified Measure. We present a generalized measure $\mathbb{D}$ that unifies all the measures in this paper. It also helps us to abstract the components of a measure and provide a notation for them.

$$
\begin{aligned}
&\mathbb{D} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R} \\
\mathbb{D} : \mathbb{D}_{p,\varphi,\gamma_s,\psi_k}(\mathbf{r}, \mathbf{w}) =& \varphi(N) * \gamma_s\big(\|\psi_k(\mathbf{r}, \mathbf{w})\|_p\big) \\
& p \geq 0 \\
& \varphi : \mathbb{N} \to \mathbb{R} \\
& \gamma_s : \mathbb{R} \to \mathbb{R} \\
& \psi_k : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N
\end{aligned}
\tag{3.1}
$$

In Equation 3.1, the measure $\mathbb{D}$ takes in the N-element vectors $\mathbf{r}$ and $\mathbf{w}$, representing the input distributions F_R and F_W, respectively. Then, it produces the final output in four steps (described inside-out):

1. compare the elements of the input vectors in a one-to-one basis using the function ψ_k;
2. aggregate the result using a vector p-norm $\|.\|_p$;

3. scale the result using the function γ_s; and

4. normalize the result using the function φ so that the final result will be independent of N for large N.

Each of these functions in the definition of $\mathbb{D}$ takes different forms for different measures, as shown in Table 1. The only exception is the vector p-norm, which is defined as follows:

$$\|\mathbf{x}\|_p = \Big(\sum_i |x_i|^p\Big)^{\frac{1}{p}},$$
$$\|\mathbf{x}\|_\infty = \max_i |x_i|, \text{and}$$
$$\|\mathbf{x}\|_0 = \sum_i x_i.$$

Notice that the definition of $\|\mathbf{x}\|_0$ is not standardized in the literature; for example, $\|\mathbf{x}\|_0$ is $\sum sign(x_i)$ in [11], and in some other sources, $\|\mathbf{x}\|_0$ is the number of non-zero elements of x. Our vector norm helps simplifying the definition of $\mathbb{D}$.

Terminology about series. A **(time) series** S is composed of elements $s_i = (x_i, y_i)$ where i is a strictly increasing and non-negative integer, called an epoch to indicate time. The epoch helps ensure a total order of the elements of S. For reference, we identify the most recent or the last element of S by s_t.

The **reference window** R and **test window** W are the ordered set of N successive elements of S. When reduced to vectors, these windows are represented as $\mathbf{r}$ and $\mathbf{w}$. In practice, these windows typically do not overlap, and both can be made more recent (closer to "now") depending on the need for defining a new reference.

We compute the distance between the two windows R and W using $\mathbb{D}(R\|W)$, as defined in Equation 3.1. To compute the distance, we can use the input vectors for these windows to represent a PDF or CDF, both empirical.

4 Distance Measure Specification

In practice for a finite number of samples, a measure is a quantitative comparison of the distance of two vectors. For example, the Euclidean distance of two n-dimensional vectors is a norm and a metric (i.e., $E \geq 0$, $E(\mathbf{a}, \mathbf{b}) = E(\mathbf{b}, \mathbf{a})$ and $E(\mathbf{a}, \mathbf{b}) + E(\mathbf{b}, \mathbf{c}) \geq E(\mathbf{a}, \mathbf{c})$). In this spirit, we can extend the use of the measures commonly used for vectors such as the Euclidean distance or PDFs such as information-theoretic measures and to take two CDFs as input parameters.

Considering the scenario where we want to compare two series in $\mathbb{R}$, notice that we can *always* define the intervals using CDFs, and we can *always* compare two CDFs as vectors or without any arbitrary determination of buckets or reduction to discrete values. Moreover, we notice that two series drawn from the same process will converge to the same CDF (respectively to the same vectors or PDFs) and all the measures will converge to zero. However, the measure output for *different* CDFs can be unbounded and, certainly, larger than in the case of PDFs (e.g., $[0, 2]$ [31]); however, for symmetric measure, the measure output will always be bound to the number of samples in the series.

In particular, symmetric measures are *more* suitable for our needs (such as Kullback–Leibler-J Equation 4.3, Jensen–Shannon Equation 4.5, and Variational Equation 4.9). Symmetry assures that the measure is not biased by the reference window. However, we can find application for positive measures, such as χ^2 in Equation 4.6, because these measures may give better discrimination power when applied to empirical distributions (especially with few observations or because we need to model a non-symmetric measure [30]).

We show that 17 of the measures in Table 1 have output CDFs that are independent of the input CDFs. For example, the Kolmogorv–Smirnov has a limit distribution, which is normal, independently of the input stochastic processes. We show this independence in Section 5.3 following the reasoning used for Kolmogorov–Smirnov measure and presenting experimental evidence. Unfortunately, we have found that the generalized functions K_r, K_s, and K_s^2 used for PDFs (Equation 4.11, 4.10, and 4.12), in general will not work for CDFs, because we cannot find a CDF for their output measures (see Section 5.2).

To conclude this section, we must specify that we do not use the geometric measure $\cos(\mathbf{a}, \mathbf{b})$ [45], [22]. We do not use $cos()$ because this measure compares only the direction of two vectors without considering their magnitude, which we consider important. Also, we did not investigate the **relative frequency model** proposed in [40], which was designed to overcome the drawbacks of the cosine measure but for histograms built on top of (bag of) words, thus discrete entities, and easy to put into buckets. Also, we do not consider the resistor distance [21], another symmetric version of the Kullback–Leibler measure.

4.1 Information-Theoretic measure extensions

In this Section, we present the information-theoretic measures extended, and thus applied, to CDFs, which is one of our main contributions.

Kullback–Leibler-I (KLI) [28]. It is an asymmetric measure and $F_r, F_W \neq 0$ (we assume that undefined values have no contributions). Notice that $KLI(F_R, F_W) = 0$ *iff* (i.e., if and only if) $F_R = F_W$; however if $F_R \neq F_W$, $KLI(F_R, F_W)$ can be arbitrary large.

$$KLI(F_R, F_W) = \sum_{y=s_y \in R\cup W} F_R(y) \log_2\left(\frac{F_R(y)}{F_W(y)}\right) \tag{4.2}$$

The KLI measure can be interpreted as the average of the relative information of the two windows (i.e.,

Name	Eq.	Measure	p	$\varphi(N)$	$\gamma_s(x)$	$\psi_k(x,y)$
Bhattacharyya	4.8	$\sum_i \sqrt{x_i y_i}$	0	NA	ι	$\sqrt{xy}$
Camberra	4.18	$\sum_i \frac{\lvert x_i - y_i\rvert}{x_i + y_i}$	0	$\frac{1}{\sqrt{N}}$	ι	$\frac{\lvert x-y\rvert}{x+y}$
χ^2	4.6	$\sum_i \frac{(x_i-y_i)^2}{x_i}$	0	1	ι	$\frac{(x-y)^2}{y}$
Cramer–von Mises	4.16	$\sum_i \big(x_i - y_i\big)^2$	2	1	x^2	$x-y$
Euclidean		$\sqrt{(\sum_i (x_i-y_i)^2)}$	2	1	ι	$x-y$
Hellinger	4.7	$\frac{1}{2}\sum_i \big(\sqrt{x_i} - \sqrt{y_i}\big)^2$	0	1	ι	$(\sqrt{x}-\sqrt{y})^2$
Jin-K		$KLI\big(x, \frac{x+y}{2}\big)$	0	$\frac{1}{\sqrt{N}}$	ι	$x\log_2(\frac{2x}{x+y})$
Jin-L	4.4	$KLI\big(x, \frac{x+y}{2}\big) + KLI\big(y, \frac{x+y}{2}\big)$	0	1	ι	$x\log_2(\frac{2x}{x+y}) + y\log_2(\frac{2y}{x+y})$
Jensen–Shannon	4.5	$\frac{1}{2}\big(KLI\big(x, \frac{x+y}{2}\big) + KLI\big(y, \frac{x+y}{2}\big)\big)$	0	1	ι	$0.5(x\log_2(\frac{2x}{x+y}) + y\log_2(\frac{2y}{x+y}))$
Kolmogorov–Smirnov	4.13	$\max_i \lvert x_i - y_i\rvert$	∞	$\sqrt{N}$	ι	$x-y$
Kullback–Leibler-I	4.2	$\sum_i x_i \log_2 \frac{x_i}{y_i}$	0	$\frac{1}{\sqrt{N}}$	ι	$x\log_2(\frac{x}{y})$
Kullback–Leibler-J	4.3	$\sum_i (x_i - y_i)\log_2 \frac{x_i}{y_i}$	0	1	ι	$(x-y)\log_2(\frac{x}{y})$
K_r	4.10	$\frac{1}{(r-1)}\log_2\big(\sum_i x_i^r y_i^{1-r}\big)$	0	NA	$log_2(x)$	$x^r y^{1-r}$
K_s	4.11	$\frac{1}{s-1}\big(-1 + \sum_i x_i^s y_i^{1-s}\big)$	0	NA	$\frac{(x-1)}{s-1}$	$x^s y^{1-s}$
K_s^2	4.12	$\frac{1}{(s-1)s}\big(-1 + \sum_i x_i^s y_i^{1-s}\big)$	0	NA	$\frac{(x-1)}{s(s-1)}$	$x^s y^{1-s}$
Minkowsky	4.17	$\big(\sum_i \lvert x_i - y_i\rvert^r\big)^{\frac{1}{r}}$	$r=3$	$\log_2 N$	ι	$x-y$
ϕ	4.14	$\max_i \frac{\lvert x_i - y_i\rvert}{\sqrt{\min\big(\frac{x_i+y_i}{2}, 1-\frac{x_i+y_i}{2}\big)}}$	∞	$\frac{\sqrt{N}}{\log_2 N}$	ι	$\frac{\lvert x-y\rvert}{\sqrt{\min(\frac{x+y}{2} 1-\frac{x+y}{2})}}$
Variational	4.9	$\sum_i \lvert x_i - y_i\rvert$	1	$\frac{1}{\sqrt{N}}$	ι	$\lvert x-y\rvert$
Ξ	4.15	$\max_i \frac{\lvert x_i - y_i\rvert}{\sqrt{\frac{x_i+y_i}{2} * \big(1-\frac{x_i+y_i}{2}\big)}}$	∞	$\frac{\sqrt{N}}{\log_2 N}$	ι	$\frac{\lvert x-y\rvert}{\sqrt{\frac{x+y}{2}(1-\frac{x+y}{2})}}$

Table 1: The measure $\mathbb{D}_{p,\varphi,\gamma_s,\psi_k}(F_R = \mathbf{x}, F_W = \mathbf{y})$. Note ι = identity function.

$\mathbf{E}_R[\log_2(F_r/F_w)]$).

Kullback–Leibler-J (KLJ) [28] and J stands for Jeffrey's [1]. It is a symmetric measure and $F_r, F_W \neq 0$. Notice that $KLJ(F_R, F_W) = 0$ iff $F_R = F_W$; however if $F_R \neq F_W$, $KLJ(F_R, F_W)$ can be arbitrary large.

$$KLJ(F_R, F_W) = KLI(F_R, F_W) + KLI(F_W, F_R) \tag{4.3}$$

Another example of symmetrization of the KLI is by [21].

Jin-L (JinL) [31]. It is a symmetric measure and it is always defined (assuming that $0 = 0log_2(0/0)$). $JinL(F_R, F_W) = 0$ iff $F_R = F_W$; and $F_R \neq F_W$, $JinL(F_R, F_W)$ can be at most $2N$.

$$\begin{aligned} JinL(F_R, F_W) =& KLI\Big(F_R, \frac{F_R + F_W}{2}\Big) + \\ & KLI\Big(F_W, \frac{F_R + F_W}{2}\Big) \end{aligned} \tag{4.4}$$

Jensen–Shannon (JS) [20, 39]. We describe this measure using Kullback–Leibler, but historically the JS was formulated using the *entropy* (i.e., $H(F_R) = -\sum_{i=0}^{k-1} F_R(i)\log_2 F_R(i) = E[I(F_R(i))]$ expectation of the information) and Kullback–Leibler is the generalization of this entropy.

$$\begin{aligned} JS(F_R, F_W) =& \frac{1}{2}\Big[KLI\Big(F_R, \frac{F_R + F_W}{2}\Big) \\ & + KLI\Big(F_W, \frac{F_R + F_W}{2}\Big)\Big] \end{aligned} \tag{4.5}$$

χ^2 (χ^2) [23, 43, 19]. It is an asymmetric measure and it is defined for $F_R \neq 0$ (we do not count the contribution for $F_R = 0$). Notice that $\chi^2(F_R, F_W) = 0$ iff $F_R = F_W$; however, if $F_R \neq F_W$, $\chi^2(F_R, F_W)$ can be arbitrary large.

$$\chi^2(F_R, F_W) = \sum_{y=s_y \in R \cup W} \frac{\big(F_R(y) - F_W(y)\big)^2}{F_R(y)} \tag{4.6}$$

Hellinger (H) [13, 43] also Kolmogorov's [1]. It is a symmetric measure always defined. The square root operation *normalizes* the components values to make the component-wise comparison less biased (i.e., all components are between 0 and 1, components close to 0 are moved towards 1/2, and components close to 1 are moved towards 1/2). Notice that $H(F_R, F_W) = 0$ iff $F_R = F_W$.

$$H(F_R, F_W) = \frac{1}{2}\sum_{y=s_y \in R \cup W} \Big(\sqrt{F_R(y)} - \sqrt{F_W(y)}\Big)^2 \tag{4.7}$$

Bhattacharyya (B) [4, 24]. It is a symmetric measure always defined. Notice that $B(F_R, F_W) <= N$ iff $F_R = F_W$; however, if $F_R \neq F_W$, $B(F_R, F_W)$ tends to 0. Notice that if applied to $\mathbf{x}$ and $\mathbf{y}$ PDFs, Bhattacharyya and Hellinger measure are related such that $1 - B(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{y})$; however, for CDFs, Hellinger is more suitable because we can determine a significance measure. We present here Bhattacharyya for completeness.

$$B(F_R, F_W) = \sum_{y=s_y \in R \cup W} \sqrt{F_R(y)F_W(y)} \tag{4.8}$$

Variational Distance (V) [36, 1]. It is a symmetric measure always defined. Notice that $V(F_R, F_W) = 0$ iff $F_R = F_W$; however, if $F_R \neq F_W$, $V(F_R, F_W)$ is no larger than $2N$. This measure is also known as Manhattan measure or Kolmogorov's variance distance [1].

$$V(F_R, F_W) = \sum_{y=s_y \in R\cup W} |F_R(y) - F_W(y)| \tag{4.9}$$

In the following as notational device, we indicate the sum $\sum_{y=s_y\in R\cup W} F_R(y)^x F_W(y)^{1-x}$ as $\mathcal{B}_x(F_R, F_W)$ [5].

Generalized K_r (K_r) [42]. This is a generalization of measure based on the Kullback–Leibler methodology.

$$K_r(F_R, F_W) = \begin{cases} \textbf{if } r = 1 & KLI(F_R, F_W) \\ \textbf{if } r > 0 & \\ & \frac{1}{(r-1)} \log_2\big(\mathcal{B}_r(F_R, F_W)\big) \end{cases} \tag{4.10}$$

Generalized K_s (K_s) [42]. For specific values of s and for PDFs, this can generate Bhattacharyya, Hellinger and Kullback–Leibler.

$$K_s(F_R, F_W) = \begin{cases} \textbf{if } s = 1 & KLI(F_R, F_W) \\ \textbf{if } s > 0 & \\ & \frac{1}{s-1}\big(-1 + \mathcal{B}_s(F_R, F_W)\big) \end{cases} \tag{4.11}$$

Generalized K^2_s (K^2_s) [42]. For specific values of s and for PDFs, this can generate Bhattacharyya, Hellinger, Kullback–Leibler and χ^2.

$$K^2_s(F_R, F_W) = \begin{cases} \textbf{if } s = 1 & KLI(F_R, F_W) \\ \textbf{if } s = 0 & KLI(F_W, F_R) \\ \textbf{if } s > 0 & \\ & \frac{1}{(s-1)s}\big(-1 + \mathcal{B}_s(F_R, F_W)\big) \end{cases} \tag{4.12}$$

Notice the equivalence $K^2_{1/2}(\mathbf{x},\mathbf{y}){=}2K_{1/2}(\mathbf{x},\mathbf{y}) = 4(1{-}B(\mathbf{x},\mathbf{y})){=}4H(\mathbf{x},\mathbf{y})$ and $K^2_2(\mathbf{x},\mathbf{y}){=}2K_2(\mathbf{x},\mathbf{y}) = \chi^2(\mathbf{x},\mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are PDFs. We present these measures (K_s, K^2_s, and K_r) for completeness, but we could not find a significance measure (but we did for some of the instances such as χ^2).

4.2 Classic distribution-function measure We present the set of measures $\mathbb{D}(F_R, F_W)$ in the literature that are already used for CDFs.

Kolmogorov–Smirnov (KS) [27, 25, 8]. It is a symmetric measure always defined. Notice that $KS(F_R, F_W) = 0$ iff $F_R = F_W$; if $F_R \neq F_W$, $KS(F_R, F_W)$ is no larger than 1.

$$\begin{aligned} KS(F_R, f_W) &= \sup_{y\in\mathbb{R}} |F_R(y) - F_W(y)| \\ &\geq \max_{y=s_y\in R\cup W} |F_R(y) - F_W(y)| \end{aligned} \tag{4.13}$$

ϕ (ϕ) [26]. It is a symmetric measure always defined. Notice that $\phi(F_R, F_W) = 0$ iff $F_R = F_W$; if $F_R \neq F_W$, $\phi(F_R, F_W)$ is no larger than 2.

(4.14)
$$\begin{aligned} \phi(F_R, F_W) &= \sup_{y\in\mathbb{R}} \frac{|F_R(y) - F_W(y)|}{\sqrt{\min\big(\frac{F_R(y)+F_W(y)}{2}, 1 - \frac{F_R(y)+F_W(y)}{2}\big)}} \\ &\geq \max_{y=s_y\in R\cup W} \frac{|F_R(y) - F_W(y)|}{\sqrt{\min\big(\frac{F_R(y)+F_W(y)}{2}, 1 - \frac{F_R(y)+F_W(y)}{2}\big)}} \end{aligned}$$

Ξ (Ξ) [26]. It is a symmetric measure always defined. Notice that $\Xi(F_R, F_W) = 0$ iff $F_R = F_W$; if $F_R \neq F_W$, $\Xi(F_R, F_W)$ is no larger than 2.

(4.15)
$$\begin{aligned} \Xi(F_R, F_W) &= \sup_{y\in\mathbb{R}} \frac{|F_R(y) - F_W(y)|}{\sqrt{\frac{F_R(y)+F_W(y)}{2} * \big(1 - \frac{F_R(y)+F_W(y)}{2}\big)}} \\ &\geq \max_{y=s_y\in R\cup W} \frac{|F_R(y) - F_W(y)|}{\sqrt{\frac{F_R(y)+F_W(y)}{2} * \big(1 - \frac{F_R(y)+F_W(y)}{2}\big)}} \end{aligned}$$

Cramér–von Mises (W^2) [33]. It is a symmetric measure and it represents the Euclidean distance of a vector. Notice that $W^2(F_R, F_W) = 0$ iff $F_R = F_W$; if $F_R \neq F_W$, $W^2(F_R, F_W)$ is no larger than $2N$ (the number of the window samples). This definition has been recently proposed [33] and it does not follow exactly the original Andersos's definition [2].

(4.16)
$$\begin{aligned} W^2(F_R, F_W) &= \int_{-\infty}^{\infty} \big(F_R(y) - F_W(y)\big)^2 dy \\ &= \sum_{y_i=s_y\in R\cup W} (y_{i+1} - y_i)\big(F_R(y_i) - F_W(y_i)\big)^2 \\ &\sim \sum_{y=s_y\in R\cup W} \big(F_R(y) - F_W(y)\big)^2 \end{aligned}$$

Minkowsky (M_r) [3, 48]. It is a symmetric parametrized measure and the generalization of both the Euclidean ($r = 2$) and Variational ($r = 1$) distance of a vector. Notice that $M_r(F_R, F_W) = 0$ iff $F_R = F_W$. In our experiments, we set $r = 3$.

$$M_r(F_R, F_W) = \Big(\sum_{y=s_y\in R\cup W} |F_R(y) - F_W(y)|^r\Big)^{\frac{1}{r}} \tag{4.17}$$

Camberra (C) [7, 48]. It is a symmetric measure and it is a relative measure of the Euclidean distance as ϕ is a relative measure of the Kolmogorov–Smirnov distance. Notice that $C(F_R, F_W) = 0$ iff $F_R = F_W$;

$$C(F_R, F_W) = \sum_{y=s_y\in R\cup W} \frac{|F_R(y) - F_W(y)|}{F_R(y) + F_W(y)} \tag{4.18}$$

4.3 Rank function measure Wilcoxon-Mann-Whitney (Wilcox) [47, 32]. it is a symmetric test and it is based on the rank of the events happening in each series. This is a standard test and it is available in R. We also used **t-test**.

5 Significance or p-value of a measure

For some measure $\mathbb{D}$s, the distribution of the measure values is well studied and known, e.g., $\sqrt{N}KS(F_R, F_W)$ for CDFs extracted by windows with N points or $\chi^2(f_R, f_W)$ for PDFs with N buckets. For some others, the distribution can be determined by simulation, e.g., $\phi(F_R, F_W)$ or $\Xi(F_R, F_W)$. Thus, we aim at the determination of the measure significance by either tables or simulations, and thus we would like to avoid bootstrap, i.e., simulation on the fly. Bootstrap is a powerful approach but it will require a training set and, thus, an *a priori* knowledge of the series, giving pressure on the final user of this statistical measures.

We have found empirically that simulation suffices for most of the $\mathbb{D}(F_R, F_W)$ used in this paper and we did obtain a distribution function of the measure values. However, we could **not** find a distribution function for the following measures:

1. KLI (because it produces negative measures),
2. Bhattacharyya, generalized K_s, K_s^2, and K_r (because we cannot find a normalizing function $\varphi(N)$ see Table 1 thus a CDF).

5.1 Simulation, $\mathbb{D}$, and its CDF We can describe our simulation process as follows. We take a measure $\bar{\mathbb{D}}$, e.g., $\bar{\mathbb{D}} = KS$. We select the number of sample N, e.g., $N = 1000$. We generate (randomly) M pairs of N samples each, e.g., $M = 5000$, taken from the same stochastic process.

Thus, we collect the measure values (x) and we determine a CDF and we define it as $F_N^1(x)$. If we repeat the process k times, we may have different CDFs F_N^i with $i \in [1, k]$. In practice, what we obtain is a *cloud* of functions $\{F_N^i\}_{1 \leq i \leq k}$. When we change N, the number of samples, we are going to have *clouds*, i.e., $\{F_N^i\}_{(1 \leq i \leq k, N)}$. For any number of samples N, we want to determine the normalizing function $\varphi(N)$ that makes it possible to compare the measures with respect other sample sizes ($F_{N_0} \sim F_{N_1}$). In Figure 1, we show the simulation for $\sqrt{N} * KS()$ (i.e., $\varphi(N) = \sqrt{N}$) for different sample sizes $N \in [1, 20] * 100$ and thus the cloud of distributions.

To combine the deterministic nature of $\varphi(N)$ with the stochastic nature of the measure value, we need to estimate $\varphi(N)$ and to do so we take $\frac{1}{\varphi(N)} \sim E[\mathbb{D}_{\varphi=\iota}]$, which is the average distance for the different measures with no normalizing factor, see Table 2. Then, we plug in the value(s) in the measure $\mathbb{D}$. Before we proceed further, a bibliographic note about how to estimate the average $E[\mathbb{D}]$. This estimate boils down to the properties of a random walk and the area below its path; even though there is no clear and complete treatment for all these measures our experimental results confirm the results in the literature for the variational distance $E[V(R\|W)] = \frac{1}{4}\sqrt{\pi N}$, see [14, 41].

Thus, our simulations obtain CDF clouds as a function of N. We define a **representation of the behavior of the measure CDF** as a stochastic function

$$F_{\mathbb{D}}(x) \in \mathcal{N}(\bar{\mu}(x), \bar{\sigma}(x)), \tag{5.19}$$

where $\bar{\mu}(x)$ is an estimate of the representative CDF and $\bar{\sigma}(x)$ is a function representing our confidence about the representative function.

We assume that we have found a representative distribution when 90% of F_N are included in the intervals $\bar{\mu}(x) \pm 2\bar{\sigma}(x)$, giving empirical justification in saying the CDFs $F_N(x)$ (as functions) have a normal distribution. Moreover, the empirical $\bar{\mu}(x)$ should be a smooth function without presenting anomaly accumulations or steps because of the merging of $F_N(x)$ with different N. Thus, we may take $\bar{\mu}(x)$ as representative distribution function of the measure (i.e., for Hellinger).

In the following, we present our approach and findings: we start by showing how to determine empirically the functions $\bar{\mu}(x)$ and $\bar{\sigma}(x)$ (Section 5.2); we then show that the $\bar{\mu}(x)$ is a CDF that is independent of the inputs CDFs (Section 5.3).

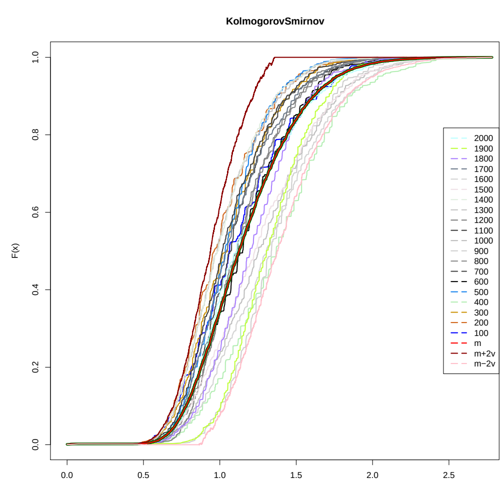

Figure 1: The Kolmogorov–Smirnov-measure CDFs

5.2 Window-Size independence, how $\varphi(N)$ comes to play In Figure 2, we present the result of the simulation for the measure Hellinger (H) and, in Figure 1, Kolmogorov–Smirnov ($\sqrt{N}KS$). For every window size between 100 and 2000 at increments of 100, we generated 1000 intervals withdrawn from the same normal distribution $\mathcal{N}(0, 1)$. Then

Table 2: Simulation of the expectation for the null hypothesis H_0 measure (i.e., $E[\mathbb{D}]$).

N	10	100	1000	10000	100000	1000000	$1/\varphi(N)$
E[ϕ]	0.70104	0.28269	0.09925	0.03391	0.01141	0.00374	$\sim log(N)/\sqrt{N}$
E[Ξ]	0.80613	0.30643	0.10466	0.03523	0.01169	0.00383	$\sim log(N)/\sqrt{N}$
E[KS]	0.34440	0.11867	0.03830	0.01227	0.00383	0.00124	$\sim 1/\sqrt{N}$ [8]
E[KLI]	1.04086	1.95563	1.00179	1.76800	-0.10939	51.66521	N/A
E[KLJ]	2.10635	2.87245	2.84168	2.94492	2.81301	3.06639	constant
E[JnK]	0.40072	0.63525	0.14532	0.51653	-0.40636	25.44923	N/A
E[JinL]	0.83340	0.74388	0.71258	0.73634	0.70326	0.76659	constant
E[JS]	0.41670	0.37194	0.35629	0.36817	0.35163	0.38329	constant
E[χ^2]	1.91610	2.11921	2.00276	2.04105	1.95051	2.12593	constant
E[V]	2.8322	8.88756	28.09942	89.08333	272.74629	915.27286	$\sim \sqrt{N}$
E[H]	0.34231	0.26507	0.24784	0.25529	0.24374	0.26568	constant
E[B]	10.15768	100.23492	1000.25215	10000.24470	100000.25625	1000000.23429	$\sim N$
E[W]	0.70672	0.67146	0.66599	0.67193	0.63212	0.71241	constant
E[E]	0.78139	0.76166	0.75571	0.75984	0.73908	0.77774	constant
E[$M_{r=3}$]	0.53319	0.35486	0.23930	0.16416	0.10910	0.07781	$\sim 1/log_3(N)$
E[C]	4.65470	16.74404	54.79298	177.49281	557.46331	1809.98318	$\sim \sqrt{N} \sim N^{\frac{540}{1000}}$

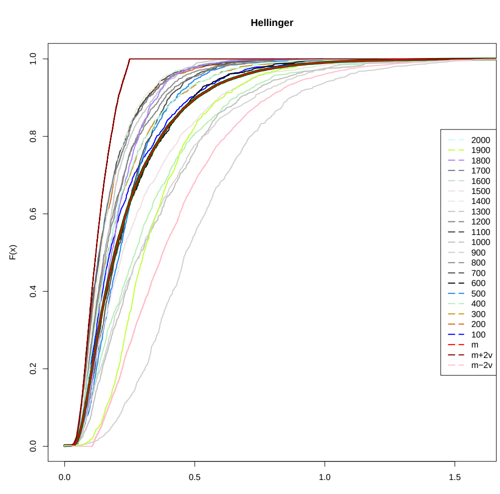

Figure 2: The Hellinger-measure CDFs

we computed the value of the measures determining the CDF supporting the similarity assumption H_0.

In Figure 1 and 2, for window sizes N_0=100 (dark blue) and N_1=2000 (azure), both measures have CDFs that are *similar*. This is possible because of $\varphi(N)$.

Average $\bar{\mu}(x)$**.** For each window size, we have a different CDF $F_N(x)$. We determine the average of the distribution as follows:

$$F_{\bar{\mu}}(x) = \frac{1}{M} \sum_N F_N(x) \tag{5.20}$$

Notice that $F_{\bar{\mu}}$ is still a distribution and it could be used as representative of the family of distributions. In fact, $F_1(x) + F_2(x)$ is not a valid distribution, however, $\frac{1}{2}(F_1(x) + F_2(x))$ is. In Figure 1, we draw the average μ in red. With our assumption about the nature of the distribution function, $F_{\bar{\mu}}(x)$ most likely should tend to $\bar{\mu}(x)$.

Variance $\bar{\sigma}(x)$**.** A natural definition of distribution variance is as follows:

$$F_{\bar{\sigma}}(x) = \sqrt{\frac{1}{M} \sum_N (F_N(x) - F_{\bar{\mu}}(x))^2} \tag{5.21}$$

In general $F_{\bar{\sigma}}$ is not a distribution and more precisely the application of subtraction and power reduce the result not to be a valid distribution (because the result $F_N(x) - F_{\bar{\mu}}(x)$ can be negative for some x). In Figure 1, we plot $F_{\bar{\mu}}(x) + 2F_{\bar{\sigma}}(x)$ using a dark-red color, and we plot $F_{\bar{\mu}}(x) - 2F_{\bar{\sigma}}(x)$ using a pink color.

We assume that we have found a representative distribution when 19 of the 20 CDFs, i.e., 90% of them, are included in the intervals $F_{\bar{\mu}}(x) \pm 2F_{\bar{\sigma}}(x)$ suggesting that the CDFs $F_N(x)$ (as functions) has a normal distribution $\mathcal{N}(F_{\bar{\mu}}(x), F_{\bar{\sigma}}(x))$ and as M gets larger this should converge to our assumption $\mathcal{N}(\bar{\mu}(x), \bar{\sigma}(x))$. Moreover, the empirical $F_{\bar{\mu}}(x)$ should be a smooth function without presenting anomaly accumulations or steps because of the merging of $F_N(x)$ with different N. Thus, we may take $F_{\bar{\mu}}(x)$ as representative distribution function of the measure (i.e., for Hellinger). [1]

5.3 Input-Distribution independence Take $F_R(y)$ as Y. That is, consider the output of a CDF as a stochastic variable. Assume that $G_R(Y)$ is the inverse function of F_R (with the proper definition for a finite number of samples N). The event $\{F_R(y) \leq t\}$ is identical to the event $\{y \leq G_R(t)\}$, which has probability $F_R(G_R(t)) = t$. This leads to $P[Y \leq t] = t$ with $t \in [0, 1]$ (See [8] Ch.1 Section 12). Thus, when we consider as input $Y = F_R(y)$, we actually obtain a measure for which the distribution of the input should not affect the distribution of the measure because F_R is uniformly distributed independently of R.

What we have found experimentally is that $F_{\bar{\mu}}(x)$ is independent of the distribution of the inputs (i.e., R and W) (as we show in Figure 3). Moreover, the distribution function $F_{\bar{\mu}}(x)$ could be used as representative distribution.

In practice, we repeat the previous simulation (Section 5.2) for window size 100–2000 collecting 1000, 2000, 5000, and 10000 measure samples per window size using two different stochastic processes: normal distribution ($\mathcal{N}(0, 1)$) and uniform distribution ($\mathcal{U}(0, 1)$). For each stochastic process, we obtain 4 representative distribution functions. We then compare them in a 4×4 table (by size and by input distribution and as we report in Figure 3 the last tile, 10000 sample per window size, because of space restrictions).

By visual inspection and by applying the measures here implemented (e.g., ϕ) or just already tabulated (e.g., Kolmogorov–Smirnov), we have found that the measure distributions are equivalent. This gives us strong evidence that our measures such as Jensen–Shannon applied to CDFs have the same properties of the Kolmogorov–Smirnov measure; thus, we have found that JS has a measure distribution independent of the nature of the input stochastic processes and it can be simulated just once.

We then use the larger set (the simulation using 10,000 samples) to determine different p-values and thus measures thresholds for each p-values. For example, for each measure we determine the threshold value having the equivalent p-value of 95%. This means that if we consider two interval R and W; we determine F_R and F_W; we measure $JS(F_R, F_W)$ and the measure has value larger than the threshold, we know that only 5% of intervals drawn from the same stochastic process have same or larger measure; we may decide to reject the assumption that W and R are similar because there is too little of a probability.

In the following experiments, we use the simulation distributions and we tabulate the p-value and thus the significance of the measures.

5.4 Disagreement (with Multiple Measures) A measure is designed to detect and to quantify the differences between its inputs. Different measures are keen to quantify different properties of the inputs and, thus, they are not all alike and they do not perform all the same.

In this paper, we investigate and quantify how the aggregation of different measures can affect the sensitivity of a non-parametric measuring system. Consensus is a simple approach of using M different measures and a decision is taken only when a quorum of the measures agrees. All the measures presented in this paper are designed to work *better* in verifying that two distributions are statistically equivalent (H_0 hypothesis, otherwise, they state no equivalence).

We quantify the detection power of different measures and we determine what is the minimum quorum or rate (e.g., 10% or 20% disagreement means that 1 or 2 measure in a set of 10 measures suggest that the two distributions are different) for a consensus approach. In particular, we want to show that our measure extensions as in Section 4.1 are a good addition.

In the following section, we discuss our experimental setup and results.

6 Experimental results

We organize the measures in two sets: **standard** and **extension**.

Standard. In this set we have: Wilcoxon-Mann-Whitney, t-test, Kolmogorov–Smirnov, ϕ, and Ξ (total of 5 measures).

Extension. In this set we have: Kullback–Leibler (symmetric), Jin-L, Jensen–Shannon, χ^2, Hellinger, Variational, Cramér–von Mises, Minkowsky, and Euclid (i.e., total of 9 measures), for which only Cramér–von Mises has been previously used in the literature, however none have be used for series in the continuous domain $\mathbb{R}$ (at least to our knowledge).

We show that our extension measures are competitive with the standard ones and they may be used separately or together. The overall measure will permit a better and more effective statistical test for series with real values.

6.1 Setup We apply the **standard** and **extension** measures separately and together on the following series.

For 1000 times, we repeat the following process:

[1] Notice that in practice, we could not find a CDF for Bhattacharyya measure because we could not find a smooth CDF.

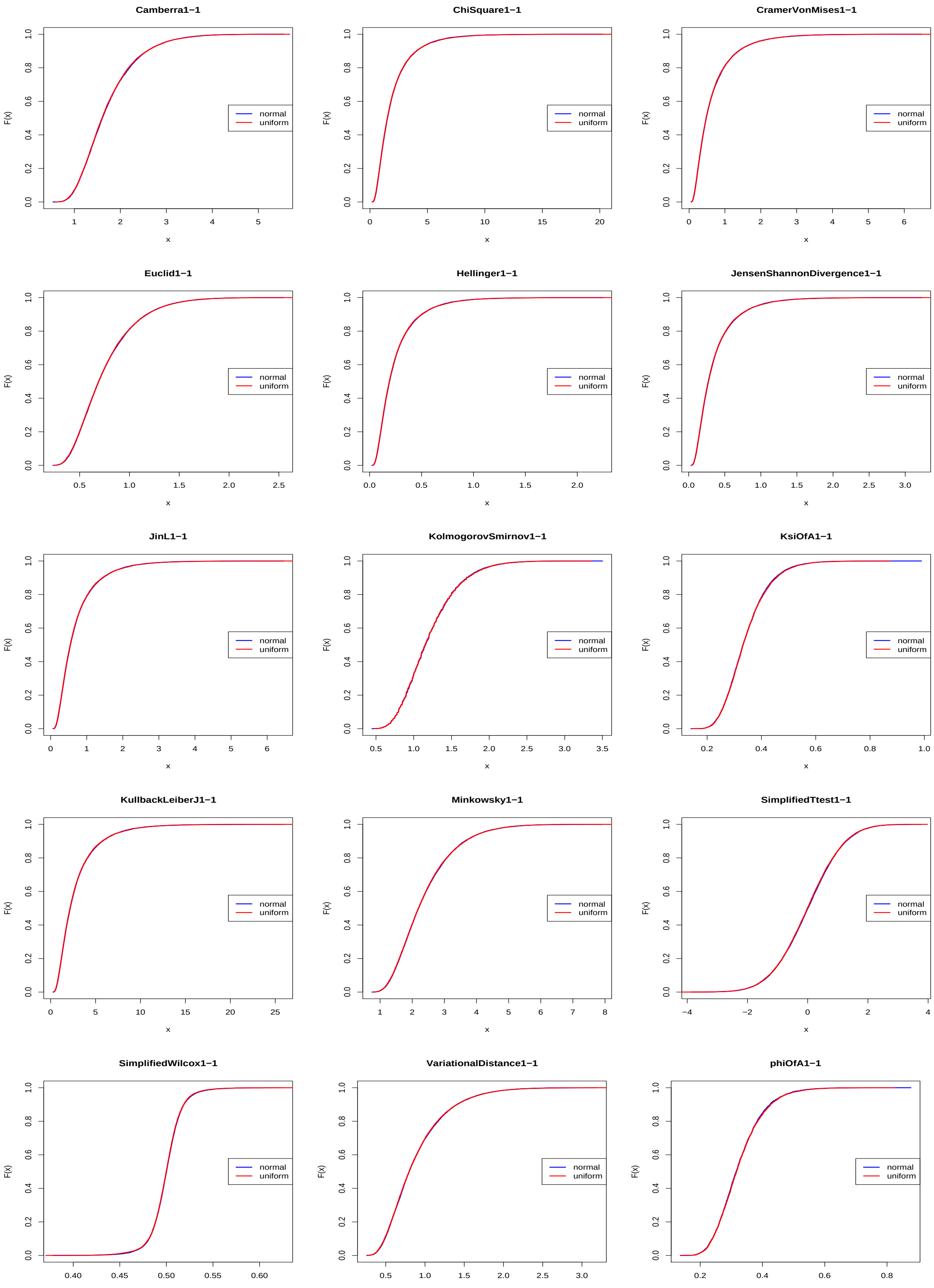

Figure 3: The 1000 measure samples for R and W drawn from the same normal $\mathcal{N}(0,1)$ and uniform $\mathcal{U}(0,1)$ stochastic processes with length 100,200, ... 2000. Notice that we plot the CDF obtained using Wilcox's and t's test, for which we know that measure CDF is independent of the input.

- We choose (randomly) the window size $W \in [1, 10] * 100$ (and have no bias on a single window size);
- We choose (randomly) the number of windows in the series $M \in [2, 20]$;
- We choose (randomly) a window E in the series that will be withdrawn from the same distribution as the first window (in $[2, M]$).

Then, we generate the series in three different ways as follows: changing both average and variance, average only, and variance only.

Changing Both Average and Variance. Using a normal distribution generator $\mathcal{N}(0, 10)$, we determine the reference average m_0 and variance v_0. We generate the two windows (the reference R and E) by either using a normal distribution $\mathcal{N}(m_0, v_0)$ or a uniform distribution $U(m_0 - v_0, m_0 + v0)$. Then, for every other window, we select at random m_i and v_i from $\mathcal{N}(0, 10)$. See Figure 4 for a trial with eight different windows where the first and the fourth are from the same stochastic process. The system using a 20% disagreement threshold recognizes the similarity with a sharp positive pulse and the system correctly flags all the others interval as different.

Changing Average Only. We generate the two windows (i.e., R and E) either using a normal distribution or a uniform distribution as before (m_0 and v_0). Then, for every other window we select at random $m_i = m_0 + r * \frac{m_0}{i}$ where $r = \pm 1$, switching its sign with equal probability. Thus, as M is larger, the windows become closer to the reference R. See Figure 5 for a trial with eight different windows where the first and the fourth are from the same stochastic process. The system using a 20% disagreement threshold recognizes the similarity with a sharp positive pulse and the system correctly flags all the others interval as different.

Changing Variance Only. We generate the two windows (i.e., R and E) either using a normal distribution or a uniform distribution as before (m_0 and v_0). Then, for every other window we select at random $v_i = v_0 + r * \frac{v_0}{i}$ where $r = \pm 1$, switching its sign with equal probability. Thus, as M becomes larger, the windows become closer to the reference R. See Figure 6 for a trial with eight different windows where the first and the fourth are from the same stochastic process. The system using a 20% disagreement threshold recognizes the similarity with a rather slow positive pulse; however, it recognizes other intervals as similar (false similarity or false positive).

As shown for the two experiments where we change either the average (e.g., Figure 5) or the variance (e.g., Figure 6), we generate a series that converges to the first window and, thus, the reference. Nonetheless, there is only one window in the series that has the same attributes of the reference (for a total of 1000), but for large enough M, the measures will find it harder and harder to distinguish them (e.g., Figure 6).

Window sliding. The moving window that sweeps the series use a step of 100 epochs or points.

Disagreement. We quantify the number of times the system recognizes two windows as *the same* for different level of disagreement (0.1, 0.2, 0.3, 0.4., 0.5., 0.6., 0.7, 0.8, 0.9, 1). That is, with a disagreement of 0.1, two windows are recognized as *equal* if at most 1 over 10 measures does not say so (and at least 9 do).

Matching and found. With a **match**, we identify the positive response of the system for the two windows R and E, which we know are equal. With a **found**, we identify the positive replies of the system independently of the position in the series (because a sliding window as a fixed step of 100 instead of the effective window size). For example, in Figure 5 we have 1 match and 2 found. The golden standard for matches is 1000 (the number of windows that are actually drawn from the same distribution).

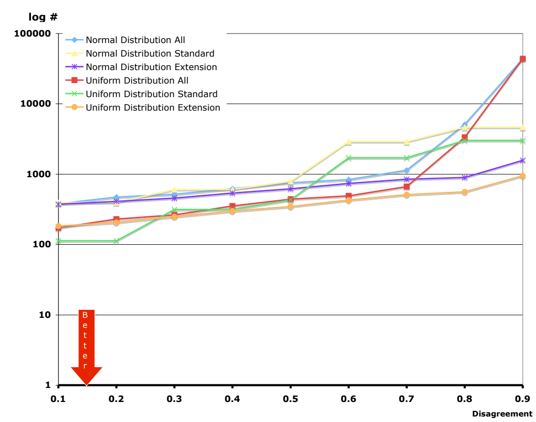

Figure 7: Average and variance change: Sum of false positive and false negative

6.2 Summary results In Figure 7, we present the number of times we commit an error ($found - 1000 + 1000 - matching$), that is, when the system says that two windows are equal minus the golden standard 1000 (false positives), plus the number of misses of the system (false negatives). This is the summary for the experiment where both average and variance change. Notice that the standard approach has a smaller error for the range 0.1 and 0.2 for uniformly distributed series. However it has the same error of our extensions for series using normal distributions. Overall, our extension work well with a steady performance as function of the disagreement factor.

In Figure 8, we present the error for average change only and variance change only. Notice that for average only,

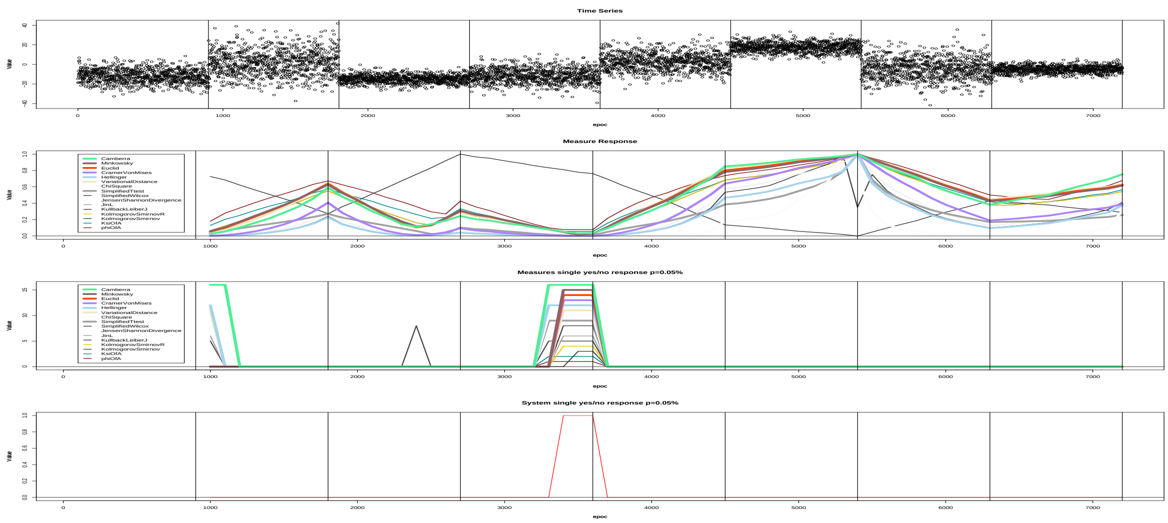

Figure 4: Average-and-variance change sample trial.

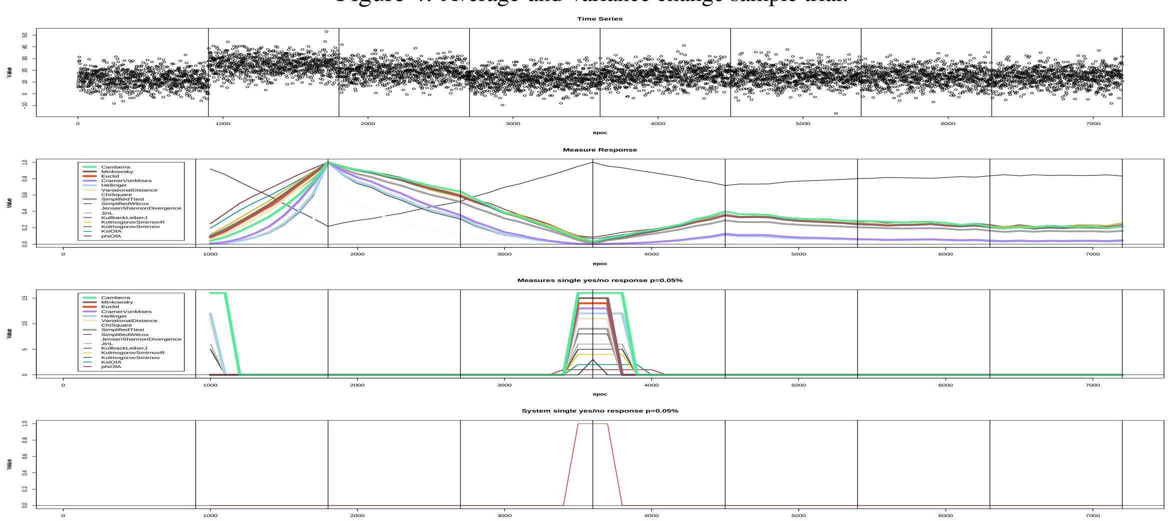

Figure 5: Average-change sample trial.

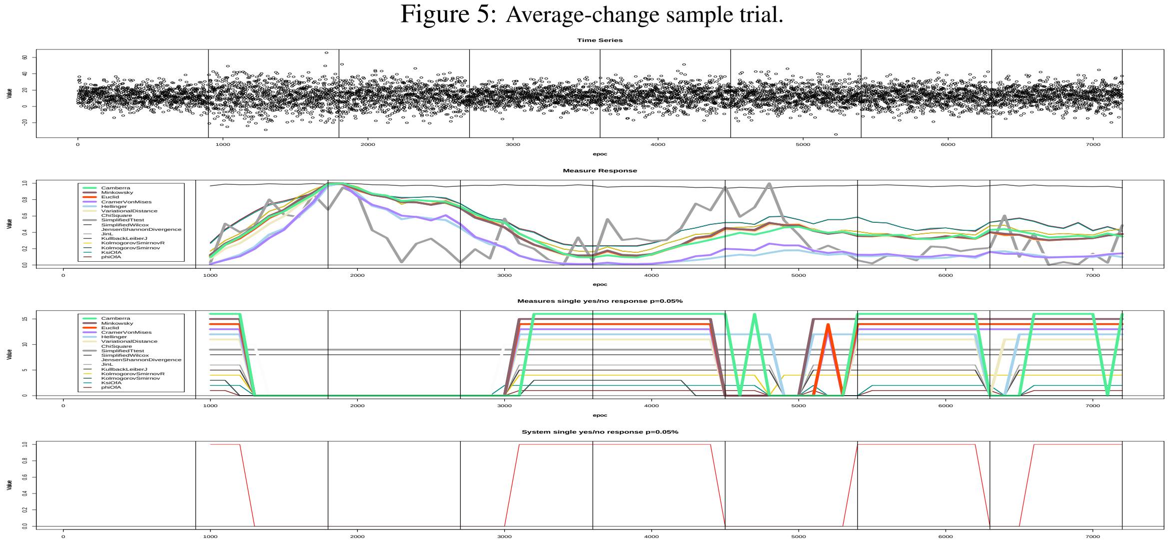

Figure 6: Variance-change sample trial: The first and the fourth windows are drawn from the same stochastic process; each windows has 900 samples and the sliding windows goes by 100 samples. We show in the first strip the series, in the second every measure normalized to 1, in the fourth the positive or no response for every measure, and, in the last, the system response with 20% disagreement. Notice that for variance only Wilcox and t-test always consider the intervals as different.

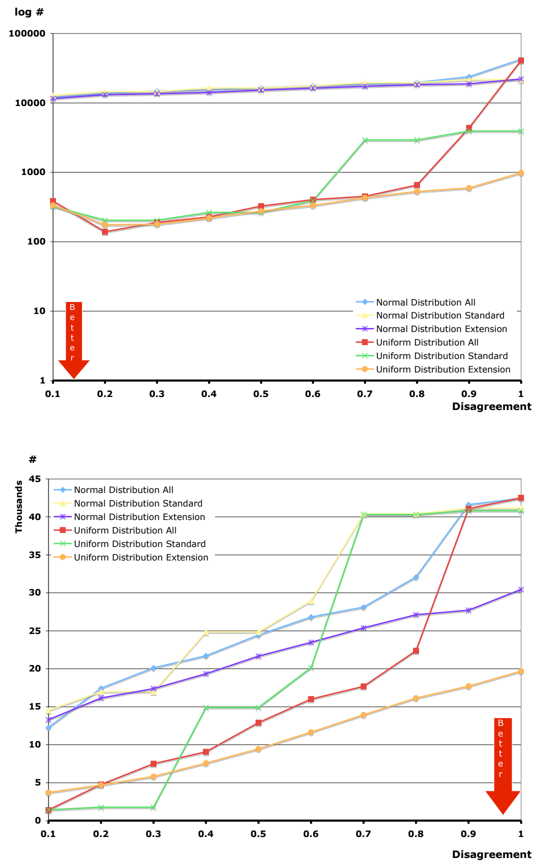

Figure 8: Sum of false positive and false negative: Average change only (top) and variance change only (bottom)

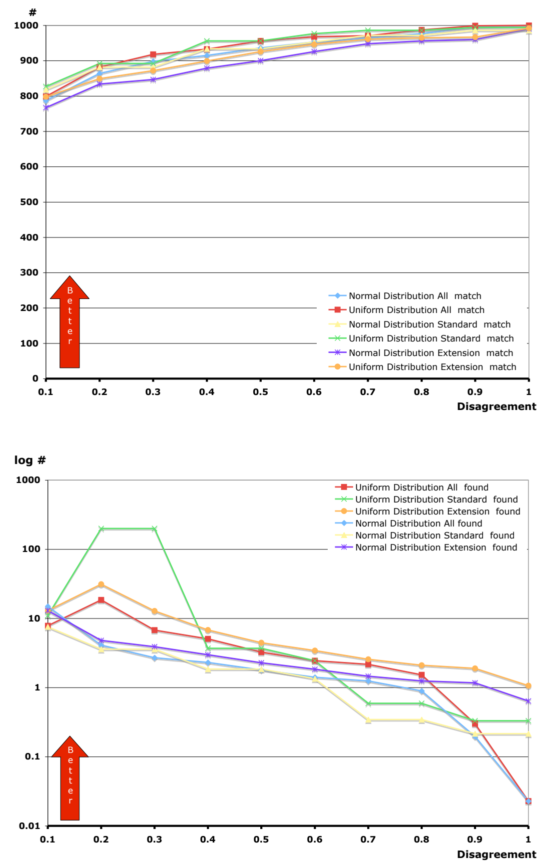

Figure 9: Average and Variance Change: Matching (top) and $\log\left(\frac{1}{|1000-Found|}\right)$ (bottom)

the combination of both standard and extension measures present the smallest error and thus a Gestalt effect.

6.3 Average and variance results In Figure 9, we present the number of perfect matching and the number of *found* matching by all approaches. The standard approach appears to be superior using the right combination of consensus/disagreement for series built using uniform distributions. However, our extensions deliver a predictable and ultimately better performance for inputs drawn from a normal stochastic process.

6.4 Average only results Our extensions seem always superior than the *standard* approach, see Figure 10. Making our approach more sensitive in detecting average variation.

6.5 Variance only results In Figure 11, we can see a similar performance as in Figure 9, where the *standard* measures offer a more sensitive tool for the true similarity (or anomaly) detection.

7 Conclusions

We presented and tested 23 measures. We compare their discriminating performance for time series in the continuous domain $\mathbb{R}$. Our main contributions are a unified treatment of all the measures and the modification of PDF-based measures to work with CDFs. We have validated our contributions with experimental results. We believe our contributions enrich the state-of-the-art tool set available to researchers for the practical evaluation of non-parametric measures.

In this work, we considered fixed the number of measures and the p-value. We are planning to investigate further the relation between the number and type of measures, the p-value factor and the nature of the series.

The authors would like to thank Deepak Agarwal, Daniel Kifer, Tony Thrall, and Ann Kalinowski for their invaluable advice and time for numerous discussions.

References

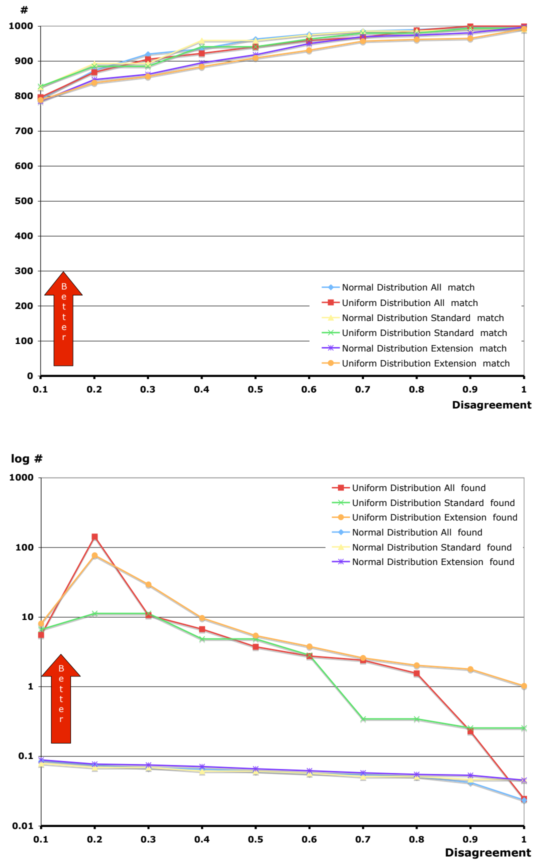

Figure 10: Average change only: Matching (top) and $\log\left(\frac{1}{|1000-Found|}\right)$ (bottom)

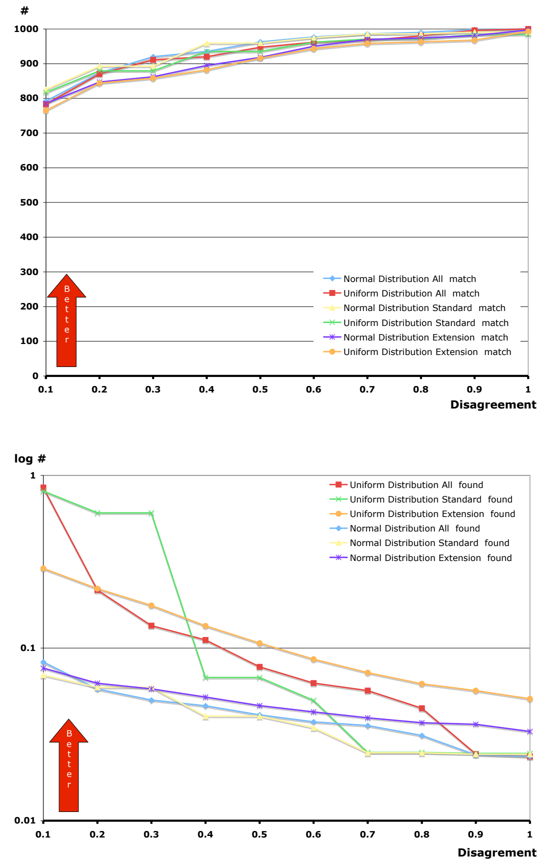

Figure 11: Variance change only: Matching (top) and $\log\left(\frac{1}{|1000-Found|}\right)$ (bottom)

[1] S. ALI AND S. SILVEY, *A general class of coefficients of divergence of one distribution from another*, Journal of the Royal Statistical Society. Series B, 28 (1966), pp. 131–142.

[2] T. W. ANDERSON, *On the distribution of the two-sample Cramer–von Mises criterion*, Annals of Mathematics Statistics, 33 (1962), pp. 1148–1159.

[3] B. BATCHELOR, *Pattern Recognition: Idea and Practice*, New York: Plenum Press, 1978.

[4] A. BHATTACHARYYA, *On a measure of divergence between two statistical populations defined by probability distributions*, Bulletin Calcutta of Mathematics Society, 35 (1943), pp. 99–109.

[5] H. CHERNOFF, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, The Annals of Mathematical Statistics, 23 (1952), pp. 493–507.

[6] T. DASU, S. KRISHNAN, S. VENKATASUBRAMANIAN, AND K. YI, *An information-theoretic approach to detecting changes in multi-dimensional data streams*, in Proceedings Symposium on the Interface of Statistics, Computing Science, and Applications (INTERFACE), Pasadena, CA, May 2006.

[7] E. DIDAY, *Recent progress in distance and similarity measures in pattern recognition*, in Second International Joint Conference on Pattern Recognition, 1974, pp. 534–539.

[8] W. FELLER, *An Introduction to Probability Theory and its Applications*, vol. 2, John Wiley & Sons, 2nd ed. ed., 1971.

[9] B. FUGLEDE AND F. TOPSOE, *Jensen-Shannon divergence and Hilbert space embedding*, in Proceedings International Symposium on Information Theory (ISIT), IEEE, Jun-Jul 2004, p. 31.

[10] A. GHOTING, M. E. OTEY, AND S. PARTHASARATHY, *Loaded: Link-based outlier and anomaly detection in evolving data sets*, in ICDM, 2004, pp. 387–390.

[11] G. GOLUB AND C. V. LOAN, *Matrix Computations*, The Johns Hopkins Univ. Press (Oct. 15, 1996), 3rd ed. ed., 1996.

[12] I. GROSSE, P. BERNAOLA-GALVAN, P. CARPENA, R. ROMAN-ROLDAN, J. OLIVER, AND H. E. STANLEY, *Analysis of symbolic sequences using the Jensen-Shannon divergence measure,*, Physical Review E, 65 (2002), pp. 1–16.

[13] H. HAHN, *Über die integrale des herrn Hellinger und die orthogonalinvarianten der quadratischen formen von unendlich vielen veränderlichen*, Journal Monatshefte für Mathematik, 23 (1912), pp. 161–224.

[14] A. HAREL, *Random walk and the area below its path*,

Mathematics of Operations Research, 18 (1993), pp. 566–577.

[15] S.-S. Ho, *A martingale framework for concept change detection in time-varying data streams*, in Proceedings International Conference on Machine Learning (ICML), Bonn, Germany, Aug 2005.

[16] S.-S. Ho and H. Wechsler, *On the detection of concept change in time-varying data streams by testing exchangeability*, in Proceedings Conference on Uncertainty in Artificial Intelligence (UAI), Edinburgh, Scotland, Jul 2005.

[17] V. Hodge and J. Austin, *A survey of outlier detection methodologies*, Artificial Intelligence Review, 22 (2004), pp. 85–126.

[18] C. C. Holt, *Forecasting seasonal and trends by exponentially weighted moving averages*, Office of Naval Research, Research Memorandum No. 52 (1957).

[19] A. A. Hope, *A simplified Monte Carlo significance test procedure*, Journal of the Royal Statistical Society. Series B (Methodological), 30 (1968), pp. 582–598.

[20] J. Jensen, *Sur les fonctions convexes et les ingalits entre les valeurs moyennes.*, Acta Mathematique, 30 (1906), pp. 175–193.

[21] D. Johnson and S. Sinanovic, *Symmetrizing the Kullback–Leibler distance*.

[22] W. Jones and G. Furnas, *Pictures of relevance: A geometric analysis of similarity measures*, Journal of American Society for Information Science, 38 (1987), pp. 420–442.

[23] A. Kagan, *Towards the theory of Fisher's amount of information*, Doklady Akademii nauk SSSR., 151 (1963), pp. 277–278. (in Russian).

[24] T. Kailath, *The divergence and Bhattacharyya distance measures in signal selection*, IEEE Transactions on Communications, 15 (1967), pp. 52–60.

[25] D. Kendall, *Andrei Nikolaevich Kolmogorov. 25 april 1903-20 october 1987*, Biographical Memoirs of Fellows of the Royal Society,, 37 (1991), pp. 301–319.

[26] D. Kifer, S. Ben-David, and J. Gehrke, *Detecting change in data streams.*, in Proceedings International Conference on Very Large Data Bases (VLDB)., Toronto, Canada, Aug 2004, Morgan Kaufmann, Elsevier, pp. 180–191.

[27] A. Kolmogorov, *Sulla determinazione empririca di una legge di distribuzione*, Giornale Istituzioni Italiane Attuari, 4 (1933).

[28] S. Kullback and R. A. Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics, 22 (1951), pp. 79–86.

[29] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, *A comparative study of anomaly detection schemes in network intrusion detection*, in In Proceedings of the Third SIAM International Conference on Data Mining, 2003.

[30] L. Lee, *Measures of distributional similarity*, in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Morristown, NJ, USA, 1999, Association for Computational Linguistics, pp. 25–32.

[31] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Transactions on Information Theory, 37 (1991), pp. 145–151.

[32] H. Mann and D. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, Annals of Mathematical Statistics, 18 (1947), pp. 50–60.

[33] M. Melucci, *On rank correlation in information retrieval evaluation*, SIGIR Forum, 41 (2007), pp. 18–33.

[34] T. Noreault, M. McGill, and M. Koll, *A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment*, in Proceedings of Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, 1981, ACM, pp. 57–76.

[35] J. O'Sullivan, R. Blahut, and D. Snyder, *Information-theoretic image formation*, IEEE Transactions on Information Theory, 44 (1998), pp. 2094–2123.

[36] M. Pinsker, *Information and information stability of random variables and processes*, Probl. Peredachi Inf., 7 (1960).

[37] C. R. Rao and T. K. Nayak, *Cross entropy, dissimilarity measures, and characterizations of quadratic entropy*, IEEE Transactions on Information Theory, 31 (1985), pp. 589–593.

[38] S. N. Rodionov, *A brief overview of the segime shift detection methods*, in Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability, V. Velikova and N. Chipev, eds., Varna, Bulgaria, Jun 2005.

[39] C. Shannon, *A mathematical theory of communication.*, The Bell System Technical Journal, 27 (1948), pp. 379–423 and 623–656.

[40] N. Shivakumar and H. García-Molina, *SCAM: A copy detection mechanism for digital documents*, in Proceedings 2nd Conference on the Theory and Practice of Digital Libraries, 1995.

[41] L. Takács, *A bernoulli excursion and its various applications*, Advances in Applied Probability, 23 (1991), pp. 557–585.

[42] I. J. Taneja and P. Kumar, *Relative information of type s, Csiszár's f-divergence, and information inequalities*, Information Sciences, 166 (2004), pp. 105–125.

[43] I. Vajda, *On the f-divergence and singularity of probability measures*, Journal Periodica Mathematica Hungarica, 2 (1972), pp. 223–234.

[44] V. Vovk, I. Nouretdinov, and A. Gammerman, *Testing exchangeability on-line*, in Proceedings International Conference on Machine Learning (ICML), Aug 2003.

[45] Z. Wang, S. Wong, and Y. Yao, *An analysis of vector space models based on computational geometry*, in Proceedings International Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, 1992, ACM, pp. 152–160.

[46] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, *Universal discrete denoising: known channel*, IEEE Transactions on Information Theory, 51 (2005), pp. 5–28.

[47] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics Bulletin, 1 (1945), pp. 80–83.

[48] D. R. Wilson and T. R. Martinez, *Improved heterogeneous distance functions*, Journal of Artificial Intelligence Research, 6 (1997), pp. 1–34.

[49] P. R. WINTERS, *Forecasting sales by exponentially weighted moving averages*, Management Science, 6 (1960), pp. 324–342.

[50] A. WONG, , AND M. YOU, *Entropy and distance of random graphs with application to structural pattern recognition*, Pattern Analysis and Machine Intelligence, 7 (1985), pp. 599–609.

[51] L. YANG, C. LIU, J. M. SCHOPF, AND I. FOSTER, *Anomaly detection and diagnosis in grid environments*, in SC '07: Proceedings of the 2007 ACM/IEEE conference on Supercomputing, New York, NY, USA, 2007, ACM, pp. 1–9.

[52] N. ZABARAS AND S. SANKARAN, *An information-theoretic approach to stochastic materials modeling*, Computing in Science and Engineering., 9 (2007), pp. 30–39.