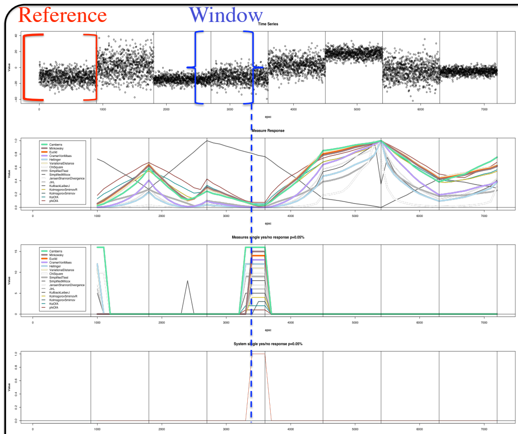
Non-Parametric Information-Theoretic Measures of One-Dimensional Distribution Functions from Continuous Time Series

Paolo D'Alberto
Ali Dasdan

Our Approach:



Series: stochastic instance

- Reference window: from series to CDF (F_r)
- Moving Window : from series to CDF (F_w)

Measures:
Apply all measures of Table 1

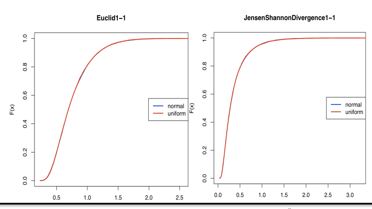
$$\mathbb{D}(\mathbf{x} = F_r, \mathbf{y} = F_w)$$

Confidence for all measures (i.e., F_r=F_w):
Given a confidence or p-value (5%) each measure flags when $F_r = F_w$ (i.e., R~W)

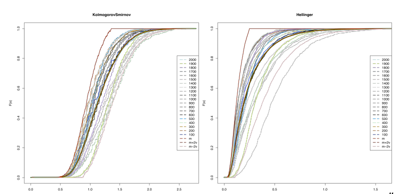
Disagreement and final vote (F_r=F_w):
We declare F_r=F_w, when no more than 1 out of 10 measures **disagrees** (disagreement 0.1)

CDF-Based measures distribution (p-value) confidence:

It is independent of the input distributions



General form for any number of samples N



Measures: (N samples)

$$\mathbb{D} : \mathbb{D}_{p,\varphi,\gamma_s,\psi_k}(\mathbf{x} = F_r, \mathbf{y} = F_w) = \varphi(N) * \gamma_s(\|\psi_k(\mathbf{x},\mathbf{y})\|_p)$$

Name	Measure	p	$\varphi(N)$	$\gamma_s(x)$	$\psi_k(x,y)$
Bhattacharyya	$\sum_i \sqrt{x_i y_i}$	0	NA	ι	$\sqrt{xy}$
Camberra	$\sum_i \frac{\lvert x_i-y_i\rvert}{x_i+y_i}$	0	$\frac{1}{\sqrt{N}}$	ι	$\frac{\lvert x-y\rvert}{x+y}$
χ^2	$\sum_i \frac{(x_i-y_i)^2}{x_i}$	0	1	ι	$\frac{(x-y)^2}{y}$
Cramer–von Mises	$\sum_i (x_i - y_i)^2$	2	1	x^2	$x-y$
Euclidean	$\sqrt{(\sum_i (x_i-y_i)^2)}$	2	1	ι	$x-y$
Hellinger	$\frac{1}{2}\sum_i (\sqrt{x_i} - \sqrt{y_i})^2$	0	1	ι	$(\sqrt{x}-\sqrt{y})^2$
Jin-K	$KLI(x, \frac{x+y}{2})$	0	$\frac{1}{\sqrt{N}}$	ι	$x\log_2(\frac{2x}{x+y})$
Jin-L	$KLI(x, \frac{x+y}{2}) + KLI(y, \frac{x+y}{2})$	0	1	ι	$x\log_2(\frac{2x}{x+y}) + y\log_2(\frac{2y}{x+y})$
Jensen–Shannon	$\frac{1}{2}(KLI(x, \frac{x+y}{2}) + KLI(y, \frac{x+y}{2}))$	0	1	ι	$0.5(x\log_2(\frac{2x}{x+y}) + y\log_2(\frac{2y}{x+y}))$
Kolmogorov–Smirnov	$\max_i \lvert x_i - y_i\rvert$	∞	$\sqrt{N}$	ι	$x-y$
Kullback–Leibler-I	$\sum_i x_i \log_2 \frac{x_i}{y_i}$	0	$\frac{1}{\sqrt{N}}$	ι	$x\log_2(\frac{x}{y})$
Kullback–Leibler-J	$\sum_i (x_i - y_i)\log_2 \frac{x_i}{y_i}$	0	1	ι	$(x-y)\log_2(\frac{x}{y})$
K_r	$\frac{1}{(r-1)}\log_2(\sum_i x_i^r y_i^{1-r})$	0	NA	$log_2(x)$	$x^r y^{1-r}$
K_s	$\frac{1}{s-1}(-1 + \sum_i x_i^s y_i^{1-s})$	0	NA	$\frac{(x-1)}{s-1}$	$x^s y^{1-s}$
K_s^2	$\frac{1}{(s-1)s}(-1 + \sum_i x_i^s y_i^{1-s})$	0	NA	$\frac{(x-1)}{s(s-1)}$	$x^s y^{1-s}$
Minkowsky	$(\sum_i \lvert x_i - y_i\rvert^r)^{\frac{1}{r}}$	$r=3$	$\log_2 N$	ι	$x-y$
ϕ	$\max_i \frac{\lvert x_i-y_i\rvert}{\sqrt{\min(\frac{x_i+y_i}{2}, 1-\frac{x_i+y_i}{2})}}$	∞	$\frac{\sqrt{N}}{\log_2 N}$	ι	$\frac{\lvert x-y\rvert}{\sqrt{\min(\frac{x+y}{2} 1-\frac{x+y}{2})}}$
Variational	$\sum_i \lvert x_i - y_i\rvert$	1	$\frac{1}{\sqrt{N}}$	ι	$\lvert x-y\rvert$
Ξ	$\max_i \frac{\lvert x_i-y_i\rvert}{\sqrt{\frac{x_i+y_i}{2} * (1-\frac{x_i+y_i}{2})}}$	∞	$\frac{\sqrt{N}}{\log_2 N}$	ι	$\frac{\lvert x-y\rvert}{\sqrt{\frac{x+y}{2}(1-\frac{x+y}{2})}}$

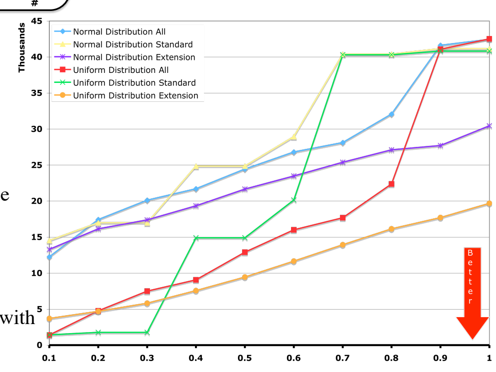
Table 1: The measure $\mathbb{D}_{p,\varphi,\gamma_s,\psi_k}(F_R = \mathbf{x}, F_W = \mathbf{y})$. Note ι = identity function.

Experimental Results:

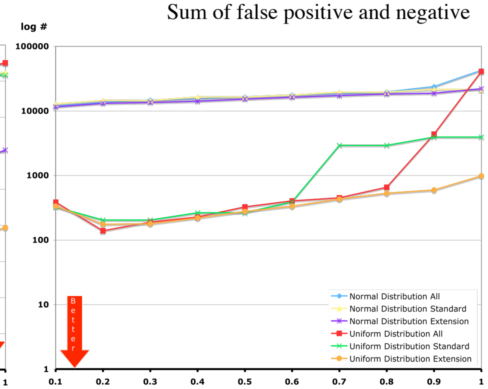
Standard: Wilcoxon-Mann-Whitney, t-test, Kolmogorov-Smirnov, ϕ, and Ξ

Extended: Kullback-Leibler, Jin-L, Jensen-Shannon, Variational, Cramer-von Mises, Minkowsky, Euclidean, and χ^2

Variation change only:
Sum of false positive and negative



Average change only:
Sum of false positive and negative



PDF is not CDF (repeat)

Summary:
An extension of non-parametric measures for comparing distributions.

Contributions:
Information-Theoretic measures are defined for probability distribution functions (PDFs), we extend them for cumulative distribution functions (CDFs).

1. We investigate a total of 23 CDF based measures Table 1
2. We provide a unified functional form
3. We test for the anomaly detection in continuous series.

Conclusions:

The new measures are less sensitive to an agreement policy delivering consistent performance

In combination with standard measures, we have a Gestalt effect and thus a sharper statistical tool

We have a (R and C) library used for: histogram comparison and series analysis with Holt-Winters, moving average and others series approaches in a single framework