The Relationship Between Empirical Process and Gaussian Process: An Example in Kolmogrov-Smirov Test

Stochostic Process: Final Project

Chi-Ning,Chou
PROFESSOR RAOUL NORMAND
July 24, 2015

Abstract

Kolmogrov-Smirov test is a famous non-parametric goodness of fitting test. The Kolmogrov statistics: $D_n = \sup_{x \in \mathcal{R}} |\hat{F}_n(x) - F(x)|$ is the central idea in this statistical test. D_n is a *distribution-free* statistics. The convergence of D_n provides us a way to see that whether a source is sampled from the guessing distribution. Moreover, since the probability distribution of D_n will converge to that of a Brownian Bridge, the confidence interval can be calculated.

A distribution-free statistics, the Kolmogrov statistics, of empirical distribution converging to the Brownian Bridge is so amazing that we further dig into the relationship between empirical process and Gaussian process. Looking forward to find some interesting behaviour among them.

Keywords: Kolmogrov-Smirov test, Empirical Process, Brownian Bridge, Gaussian Process

Contents

1 Kolmogrov-Smirov Test

1.1 Empirical Distribution

As observers, all we can see from a random experiment is the sampling results from an underlying distribution (if there exists one). In almost every case, we don't know the true probability distribution behind it. What we want to do is to make inferences about the underlying distribution.

As long as we only have the samples, it's intuitively to make a histogram and observe the structure. Furthermore, we can consider the *empirical distribution function*

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_i \leq x\}}$$

where $X_1, X_2, ..., X_n$ are i.i.d. sample from a cumulative distribution function F. And $\mathbf{I}$ is the indicator function.

The intuition is that we record the number of samples from the small to large and draw a cumulative function.

The empirical distribution function has some nice properties such as point-wise convergence to the underlying distribution: $\hat{F}_n(z) \xrightarrow{P} F(z)$. The result is followed from the observation that the distribution of $n\hat{F}_n(z)$ for some $z \in \mathcal{R}$ is the same as $binomial(n, F(z))$. This observation can be easily seen in figure 1.

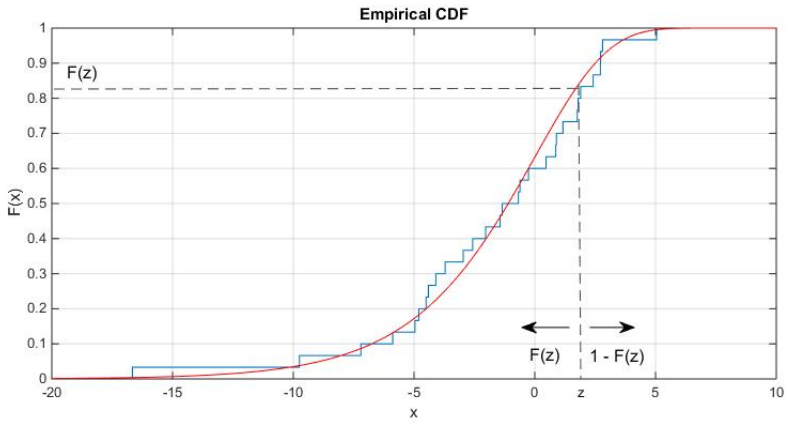

Figure 1: Empirical distribution and its point-wise convergence property.

Now we have the point-wise convergence of empirical distribution and the corresponding asymptotic rate. Based on this, we can construct confidence interval for point-wise estimation. However, what if we want to estimate the behaviour of two points? Or, the behaviour in an interval?

1.2 Kolmogrov Statistics

The Kolmogrov statistics is defined on an empirical distribution function $\hat{F}_n$ and a cumulative objective function F as follow:

$$D_n := \sup_{x \in \mathcal{R}} |\hat{F}_n(x) - F(x)|$$

where n is the number of samples.

We can see that the Kolmogrov statistics D_n is the supremum point-wise distance between the empirical distribution and the target function. The smaller the D_n is we can some how think of that the closer the two distribution are.

As long as we consider the Kolmogrov statistics between the empirical distribution and its underlying distribution, there are some nice convergence behaviours.

Theorem 1 (Glivenko-Cantelli)**.** *The Kolmogrov statistics will converge to zero as the number of samples grows to infinity. That is,*

$$D_n \xrightarrow{P} 0$$

, as $n \to \infty$

With this theorem, we have the uniform convergence of empirical distribution. Namely, for any $\epsilon > 0$ there exists a N such that for all $n > N$, the underlying distribution will lies in the ϵ-neighborhood of the empirical distribution.

In addition, the Kolmogrov statistics has a very important property: *distribution-free*. It means that no matter what underlying property is, the behaviour of the Kolmogrov statistics will be the same! Concretely, the distribution will related to the uniform distribution.

Theorem 2 (Distribution-Free Property)**.** *The distribution of the Kolmogrov statistics* D_n *is the same for all continuous underlying cumulative distribution.*

1.3 Empirical Process Theory

2 Brownian Bridge

2.1 Gaussian Process

2.2 Brownian Bridge

References

1. *Crypto Corner*, http://crypto.interactive-maths.com

2. *The Hunger Games*, Suzanne Collins. 2008. Scholastic. U.S. https://sites.google.com/site/the74thhungergamesbyced/download-the-hunger-games-trilogy-e-book-txt-file

Appendix

The code of this project can be found on Github: https://github.com/jerrychou82/MCMC_Break_St
It's welcome to discuss the code with me!