REU in Random Walk

Xiaoyu Liu

December 8, 2022

Contents

Note 6: Adapted K-S Test for discrete sampling

Sun 20 Nov 2022 18:32

1 Adapted K-S Test for discrete sampling

Motivation

Simulation of continuous random processes are discrete in nature. Two factors primarily contribute to the discretization:

1. Numerical precision limit.
2. The approximation of continuous process via discrete process.

Discretization gives rise to the **oversampling effect:** When we generate a sample, the sample we get is from a discretized version of the underlying distribution. Hence, if we increasing the size of the sample, the empirical cdf cannot converge uniformly to the true underlying distribution. This brings huge problem to goodness-of-fit tests, since their consistency relies on Glivenko-Cantelli Theorem, the fact that empirical cdf must uniformly converge to true distribution.

Therefore, we need to develop techniques to perform goodness-of-fit tests on discrete simulations and avoid the oversampling effect.

Problem Background

Partition. We define a partition P of $\mathbb{R}$ as a countable set of distinct real numbers. We say P is finer or equal to P' if $P' \subset P$, and write $P > P'$.

Discretization. Let F be a continuous real-valued distribution. Let P be a partition of $\mathbb{R}$. We say $F^{(P)}$ is a discretization of F with respect to partition P if $\tilde{F}(x) = F(y)$, where $y \in P$ minimizes $|x - y|$. We may verify that $F^{(P)}$ is indeed discrete, since it takes values in $F(P)$ which has at most countably many values.

Contribution

We have developed and implemented techniques to perform k-s test to samples from a discretized distribution. The techniques are applicable to the two following cases:

1. (One sample test) The underlying distribution F_0 is continuous real-valued. We have a sample $\mathbf{X} = X_1, \ldots, X_n$ drawn independently from some distribution $F_X^{(P)}$ that is discretization of continuous F_X with respect to P. We want to test if $F_0 \equiv F_X$.
2. (Two sample test) We have samples $\mathbf{X}$ and $\mathbf{Y}$ from discretizations $F_X^{(P)}$ and $F_Y^{(P')}$ with respect to partitions P and P', where $P \leq P'$. We want to test if $F_X \equiv F_Y$.

Development

Our null hypothesis in the one-sample case is:

$$H_0 : F_0 \equiv F_X.$$

Under the null hypothesis, we must have

$$H_0' : F_0^{(P)} \equiv F_X^{(P)}.$$

That is, the discretizations of F_0 and F_X are equal.

We see that $H_0 \Rightarrow H_0'$, so whenever we reject H_0' we must also reject H_0. Since our sample is from $F_X^{(P)}$, we have just enough information to test H_0'. Hence, the decision criterion is set to reject H_0 if we reject H_0'.

On one hand, we can control type I error if we assume we can control type I error in testing H_0', since

$$\begin{aligned}
&\Pr\left[H_0 | \text{reject } H_0\right] \\
&= \Pr\left[H_0 | \text{reject } H_0'\right] \\
&= \Pr\left[H_0 | H_0' \wedge \text{reject } H_0'\right] \Pr\left[H_0' | \text{reject } H_0'\right] \\
&\quad + \Pr\left[H_0 | \neg H_0' \wedge \text{reject } H_0'\right] \Pr\left[\neg H_0' | \text{reject } H_0'\right] \\
&= \Pr\left[H_0 | H_0' \wedge \text{reject } H_0'\right] \Pr\left[H_0' | \text{reject } H_0'\right] \\
&\leq \Pr\left[H_0' | \text{reject } H_0'\right].
\end{aligned}$$

On the other hand, we cannot control type II error, and the method above yields upper bound 1. Intuitively, even if H_0' is true, we can say nothing about probability that H_0 is true, unless we have a notion of measure within the space of all continuous distributions. But we are treating F_0 as an *arbitrary* distribution, so we cannot obtain meaningful probability bound.

What remains is to find a way to test H_0'. K-S test is not directly applicable to discretizations, since it only works for continuous distributions. Nevertheless, we can use **linear interpolation** to transform $F_0^{(P)}$ and $F_X^{(P)}$ to continuous piecewise-linear distributions. Then the K-S test would be applicable. It is easy to see that

Claim. There is a 1-1 correspondence between discrete distributions wrt P and piecewise linear functions wrt P. Moreover, $F_0^{(P)} \equiv F_X^{(P)}$ iff $\tilde{F}_0^{(P)} \equiv \tilde{F}_X^{(P)}$.

For two-sample case, the null hypothesis is:

$$H_0 : F_X \equiv F_Y.$$

We weaken the null hypothesis to the partition P (the coarser partition).

$$H_0' : F_X^{(P)} = F_Y^{(P)}.$$

Then the same reasoning holds.

Linear Interpolation

Calculation of discretization of a continuous distribution is straightforward from the definition. What we need to develop here is a method to perturb the sample in the way that performs linear interpolation on its underlying distribution.

Let $X_1, X_2, \ldots, X_m$ be samples from $F_X^{(P)}$. We define the noisy sample $\tilde{X}_i$ as follows: if $X_i = x_i \in P$, then let $\tilde{X}_i$ be a random sample from the distribution with pdf

$$\tilde{f}_i(x) = \begin{cases} 0 & x < \frac{x_{i-1}+x_i}{2} \\ \frac{1}{x_i - x_{i-1}} & x < x_i \\ \frac{1}{x_{i+1}-x_i} & x < \frac{x_i+x_{i+1}}{2} \\ 0 & x \geq \frac{x_i+x_{i+1}}{2} \end{cases}.$$

Then $\tilde{X}_1, \ldots, \tilde{X}_m$ is a sample from $\tilde{F}_X^{(P)}$, the linear interpolation of $F_X^{(P)}$.

In the special case when $P = \{q\Delta x + r : q \in \mathbb{Z}\}$ for some fixed $\Delta x > 0$ and $0 \leq r < \Delta x$, we have

$$\tilde{X}_i \equiv_d X_i + \text{Unif}(-\Delta x/2, \Delta x/2).$$

Application to SSRW simulation

For SSRW simulation, we take F_0 to be the cdf of standard normal distribution. Because the discreteness is due to independent steps of same length, the partition P falls into the special case. It remains to find Δx and r.

Suppose we simulate n steps of a SSRW. Then the range of scaled distribution is

$$\frac{n[-1,1]}{\sqrt{n}} = [-\sqrt{n}, \sqrt{n}].$$

When n is even, due to parity of steps, there are $n+1$ possible final positions after n steps. So the range is evenly divided into n intervals. Hence the resolution, Δx, is

$$\Delta x = \frac{2\sqrt{n}}{n} = \frac{2}{\sqrt{n}}.$$

Since 0 gets mapped to 0 during the rescaling, $r = 0$.

Using these knowledge of P, we can calculate the discretized-interpolated $\tilde{F}_0^{(P)}$ and noised samples.

Result

See notebook for implementation and experiment result.

Some considerations regarding different partitions

- Justify that in the two-sample case where $\mathbf{X}$ and $\mathbf{Y}$ are drawn from $F_X^{(P)}$ and $F_Y^{(P')}$, if $P < P'$, then the decision power does not depend on how fine P' is.
- In the two-sample case, if $P \cap P' = \emptyset$, i.e. P and P' does not share common points, can we obtain any information about goodness of fit? What extra assumptions do we need about F_X and F_Y? May we conjecture some F_0 and test it on both $\mathbf{X}$ and $\mathbf{Y}$? Impose $c-$Lipchitz continuity on F_0?

Further steps

- When studying edge reinforced random walks, where discrete simulation is applied extensively, the adapted k-s test can give reliable goodness-of-fit test.

Note 7: Test for Limiting Distribution

Wed 30 Nov 2022 02:14

2 Testing Limiting Distribution

Motivation

The oversampling effect. We need to numerically verify the limiting distribution of a random walk. We cannot sample from the limiting distribution directly, and instead can only sample from a location distribution after finitely many steps. Therefore, if we run a goodness-of-fit test directly between the empirical distribution and the hypothesized limiting distribution, the test will reject the null hypothesis H_0 after taking enough samples, even when H_0 is true. For experiment outputs, see `221119_ssrw_adapted_ks`. This phenomenon introduces difficulty making precise statements of the limiting distribution.

Therefore, we want to develop a method for testing limiting distributions. Our method should account for the difference between location distribution at step m and the limiting distribution. Our method should consistently fail to reject H_0 when H_0 is true. Meanwhile, it should not make too much beta error, i.e. it should reject H_0 with high probability when H_0 is not true.

Background

Let A_m be the location distribution of a random walk after m steps.

Let $X_m^{(k)}, k = 1, \ldots, n$ be i.i.d samples from A_m.

Let $S_{m,n}$ be the empirical cdf of the scaled scamples:

$$S_{m,n}(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\left(\frac{X_m^{(k)}}{\sigma\sqrt{m}} < x \right).$$

And let F_m be cdf of $\frac{A_m}{\sigma\sqrt{m}}$, the scaled distribution at step m.

We assume the scaling term is $m^{-\frac{1}{2}}$ in this note, but this can be generalized.

By the Glivenko-Cantelli Theorem, as $n \to \infty$, $S_{m,n} \to F_m$ uniformly in distribution a.s.

We hypothesize that the random walk has limiting distribution F_0, that is, $F_m \to F_0$ uniformly in distribution as $m \to \infty$.

Development of The Test

Our null hypothesis is

$$H_0 : F_m \xRightarrow{m\to\infty} F_0.$$

Our alternative hypothesis is

$$H_A : F_m \text{ does not converge or } F_m \xRightarrow{m\to\infty} F' \not\equiv F_0.$$

We define our test statistic to be

$$D_{m,n} = \sup_x \left| \tilde{S}_{m,n}(x) - F_0(x) \right|.$$

We also define

$$K_{m,n} = \sup_x \left| \tilde{S}_{m,n}(x) - \tilde{F}_m(x) \right|.$$

$$T_m = \sup_x \left| \tilde{F}_m(x) - F_0(x) \right|.$$

In summary, we have the following picture:

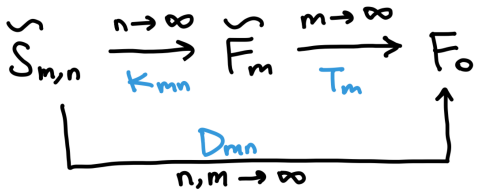

Figure 1

We bound $K_{m,n}$ using Kolmogorov Theorem:

$$\Pr\left[K_{m,n} < \frac{d}{\sqrt{n}} | H_0\right] = L(d,n).$$

We bound T_m using Berry-Esseen Theorem:

$$T_m \leq 0.4748 \frac{1}{\sqrt{m}}.$$

Now we can bound $D_{m,n}$ using triangle inequality:

$$K_{m,n} - T_m \leq D_{m,n} \leq K_{m,n} + T_m.$$

Now we try to bound the probability that $D_{m,n} < \frac{v}{\sqrt{n}}$. For lower bound,

$$\begin{aligned}
&\Pr\left[D_{m,n} < \frac{v}{\sqrt{n}} | H_0\right] \\
&\geq \Pr\left[K_{m,n} + T_m \leq \frac{v}{\sqrt{n}} | H_0\right] \\
&\geq \Pr\left[K_{m,n} < \frac{v}{\sqrt{n}} - 0.4748\frac{1}{\sqrt{m}} | H_0\right] \\
&= L\left(v - 0.4748\sqrt{\frac{n}{m}}\right).
\end{aligned}$$

Similarly, one can obtain an upper bound. Hence we can bound the cdf of $D_{m,n}$ conditioned on H_0 from above and below:

$$L\left(v - 0.4748\sqrt{\frac{n}{m}}\right) \leq \Pr\left[D_{m,n} < \frac{v}{\sqrt{n}} | H_0\right] \leq L\left(v + 0.4748\sqrt{\frac{n}{m}}\right).$$

Now the bound for p-value is given by

$$1 - L\left(v + 0.4748\sqrt{\frac{n}{m}}\right) \leq p \leq 1 - L\left(v - 0.4748\sqrt{\frac{n}{m}}\right).$$

For some fixed α, we calculate the parameter v^* using

$$\alpha = 1 - L\left(v^* + 0.4748\sqrt{\frac{n}{m}}\right)$$

The decision criterion is

$$D_{m,n} > \frac{v^*}{\sqrt{n}} = \frac{L^{-1}(1-\alpha)}{\sqrt{n}} + 0.4748\sqrt{\frac{1}{m}}$$

or equivalently,

$$1 - L\left(D_{m,n}\sqrt{n} - 0.4748\sqrt{\frac{n}{m}}\right) < \alpha.$$

Application to SSRW and results

Following are observations from applying this test to SSRW. For code and output, see `221130_adapted_ks_1samp`.

1. The oversampling effect vanishes: the test fails to reject every time, no matter how many samples we take.
2. The test is very optimistic for $m < n$: the calculated p value is almost always 1.
3. The test is still able to reject: when applied to a random walk generated using Cauchy distribution, the test consistently rejects the null hypothesis.

Discussion

Several problems still need to be addressed:

1. The lower and upper bounds are not very tight when $m \approx n$. To obtain reasonably tight bound we m bigger than n by $1-2$ orders of magnitude. This is not a big problem in general, however, since we can pick $m \gg n$ if we want tighter bounds.
2. The test is

Further Steps

1. Develop a 2-sample version of this test.