Testing Group Invariance

Alex Adkins, Alexander Chatterjee, Jiajun Du, Cody Pedersen, Jesse Rinzel, Jingwen Xu, Yuankun Zou

University of Rochester, The College of New Jersey
Advisors: Alex Iosevich, Kunxu Song

2025

Outline

1. Necessary Terminology
2. Invariant Distribution
3. Construction of R Statistic
4. Choosing Epsilon
5. *R* statistic and the Choice of Epsilon
6. N=2 and 3 Cases
7. Future Directions

Necessary Terminology

Random Variable: A random variable is a measurable function

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

where:

- Ω: sample space
- $\mathcal{F}$: σ-algebra of events
- $\mathcal{B}(\mathbb{R})$: Borel σ-algebra on $\mathbb{R}$

Necessary Terminology

Expectation: The expectation of an integrable random variable

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

with respect to the probability measure P is defined by the Lebesgue integral

$$E[X] \;=\; \int_\Omega X(\omega)\, dP(\omega).$$

where:

- $\int_\Omega |X(\omega)|\, dP(\omega) < \infty$: integrability condition
- P: probability measure on $(\Omega, \mathcal{F})$

Necessary Terminology

Distribution of X: The *distribution* of a random variable X is the pushforward measure:

$$P_X(B) = P(X^{-1}(B)) \quad \text{for all } B \in \mathcal{B}(\mathbb{R})$$

That is, P_X is a probability measure on $\mathbb{R}$ induced by X.

Necessary Terminology

Isometry Group of $\mathbb{R}^n$: The set of all bijections on $\mathbb{R}^n$ that preserve the Euclidean distance.

$$\mathrm{Isom}(\mathbb{R}^n) = \left\{ f : \mathbb{R}^n \to \mathbb{R}^n \;\middle|\; \begin{array}{l} f \text{ is a bijection,} \\ \|f(x) - f(y)\| = \|x - y\| \quad \forall\, x, y \in \mathbb{R}^n \end{array} \right\}.$$

where:

- $\mathbb{R}^n$: Euclidean n-space
- $\| \cdot \|$: Euclidean norm on $\mathbb{R}^n$
- $\mathrm{Isom}(\mathbb{R}^n)$: group of all Euclidean isometries

Invariant Distribution

Consider a random variable of $\mathbb{R}^m$

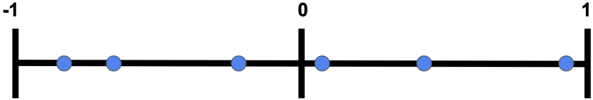
$$X : \Omega \to \mathbb{R}^m$$

and consider some isometry group of $\mathbb{R}^m$, G

GOAL: design a rule to infer if the distribution of the random variable is invariant under the group action

Invariant Distribution

Example:
Suppose we have a number line from -1 to 1 and we are sampling N points from an unknown distribution. We want to calculate a score of how symmetric the sample is, with the goal of learning if the underlying distribution is invariant under the group action.



Construction of R Statistic

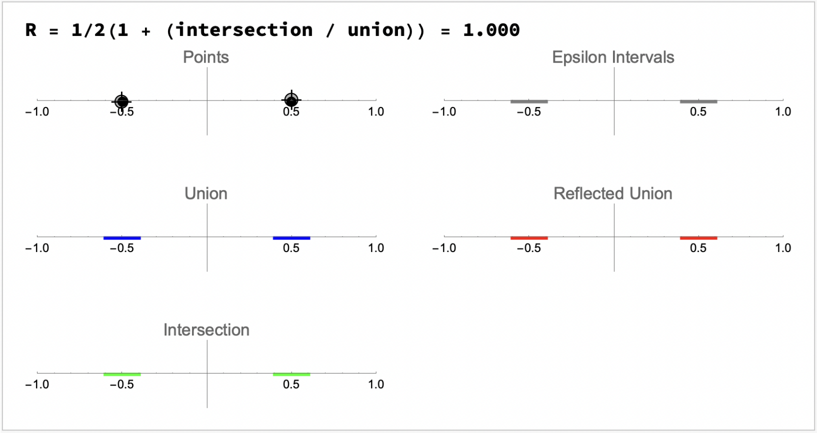
The R statistic measures the symmetry of the sample.

$$R(S, \epsilon) = \frac{1}{\#G} \sum_{g \in G} \frac{\mu_0\big(B(S, \epsilon) \cap B(gS, \epsilon)\big)}{\mu_0\big(B(S, \epsilon)\big)}$$

where:

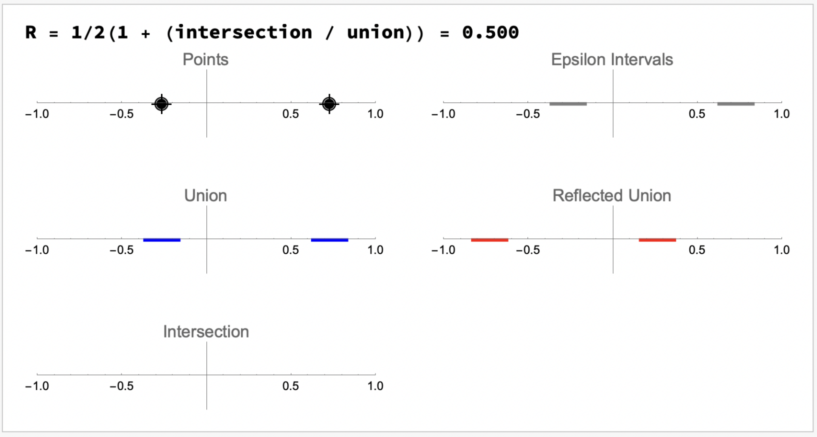
- S: Sample of N points from an unknown distribution
- $B(S, \epsilon)$: Union of epsilon neighborhoods around points in S
- μ_0: Lebesgue measure (n-dimensional area)
- G: An isometry group

Construction of R Statistic



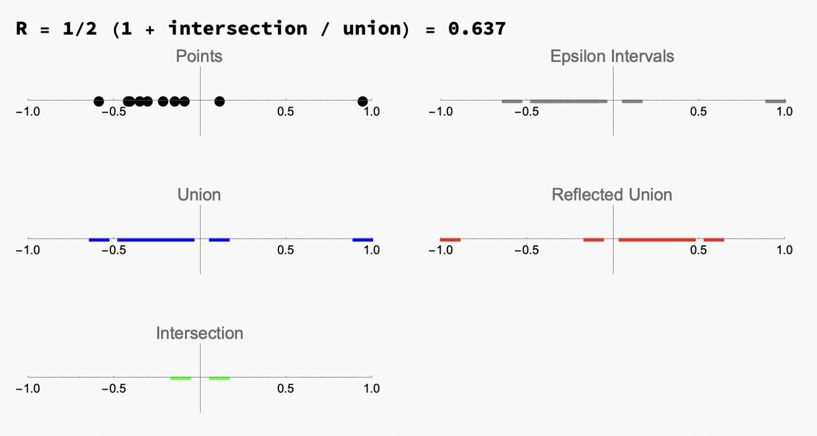
R = 1/2(1 + (intersection / union)) = 1.000

Construction of R Statistic



```
R = 1/2(1 + (intersection / union)) = 0.500
```

Construction of R Statistic



This raises the question of how small/large should epsilon be as a function of N?

Choosing Epsilon

A reasonable way to choose epsilon is to look at the general level of intersection of ϵ-neighborhoods of two random samples. More rigorously, we want to evaluate

$$\mathbb{E}[\mu_0(B(S_1, \epsilon) \cap B(S_2, \epsilon))]$$

where S_1 and S_2 are chosen from some uniform distributions (for example, the uniform distribution over a unit disk). The idea is that we will then enforce it to be some constant, and this would give us a value for epsilon.

Choose Epsilon

For example, suppose we draw a sample of N points in $\mathbb{R}^m$ and want to test if the underlying distribution is rotational invariant. We can calculate the average distance from the data points to the origin. Then, we compute the radius r for a disk on which the uniform distribution gives the same average radial distance. Then, we may consider the expected overlap area if we draw two N samples from the disk uniformly and form the ϵ-balls.

Choosing Epsilon

Theorem

Let $X : \Omega \to D \subseteq \mathbb{R}^m$ and $Y : \Omega \to \langle K, \mathcal{A} \rangle$ be random variables where X is uniform. Let $Q : K \to \mathcal{L}(D)$ (Lebesgue-measurable sets), and assume that $\mu_0 \circ Q$ is Lebesgue-measurable. Then,

$$\mathbb{E}_Y(\mu_0(Q(Y))) = \int_D \mathbb{P}_Y(x \in Q(Y)) d\mu_0$$

Now, let D be the disk of radius r and $K = \mathbb{R}^{2N}$ so that Y represents the two samples. We may denote the first n coordinates as a vector Y_1 and the last n coordinates as Y_2. Then, let $Q(Y) = B(Y_1, \epsilon) \cap B(Y_2, \epsilon)$.

Choosing Epsilon

Applying the previous theorem (ignoring some boundary terms that vanishes when $\frac{\epsilon}{r} \to 0$) gives

$$\mathbb{E}_Y(\mu_0(Q(Y))) \approx \pi r^2 \left(1 - \left(1 - \left(\frac{\epsilon}{r}\right)^2\right)^n\right)^2$$

Therefore, let A be the desired expected proportion of the disk covered by the overlapped area (say, 10%), we have

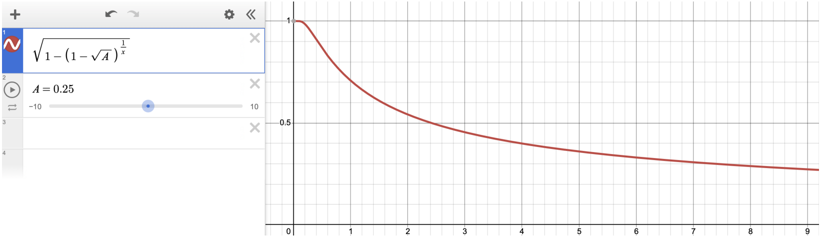
$$A\pi r^2 = \pi r^2 \left(1 - \left(1 - \left(\frac{\epsilon}{r}\right)^2\right)^n\right)^2$$

which gives

$$\epsilon = r\sqrt{1 - \sqrt[n]{1 - \sqrt{A}}}$$

Choosing Epsilon

A graph of this relation of ϵ to n is given here, with n on the x-axis and ϵ on the y-axis, and assuming $r = 1$:



$\sqrt{1-(1-\sqrt{A})^{\frac{1}{x}}}$

$A = 0.25$

This does not account for the overlap of a rotation, or any other group action, of the intervals around these random sets of points, only the overlap between the intervals around two entirely unrelated random sets of points. So, we want to understand what occurs in a case with a group action.

R statistic and the Choice of Epsilon

To this end, we consider the simplest case, where the sample is drawn uniformly from the interval $[-1, 1]$ and the group action is reflection about the origin. In this case, we have

$$R(S, \epsilon) = \frac{1}{2}\left(1 + \frac{\mu_0\big(B(S, \epsilon) \cap B(-S, \epsilon)\big)}{\mu_0\big(B(S, \epsilon)\big)}\right)$$

and since we already know that S is drawn uniformly, we can try to compute

$$\mathbb{E}_S[R(S, \epsilon)]$$

as a function of ϵ and the sample size N.

R statistic and the Choice of Epsilon

The case where $N = 1$ is fairly easy to solve using our previous theorem. Applying a twisted version of it, together with some approximations which gives vanishing errors as $\epsilon \to 0$, we are able to obtain that when $N = 2$

$$\mathbb{E}_S[R(S,\epsilon)] \approx \frac{1}{2}\left(\frac{16}{3} + \frac{5}{3}\epsilon - \frac{26}{3} \cdot \frac{1}{(2-\epsilon)} - 8\ln\frac{4-2\epsilon}{4-\epsilon}\right)$$

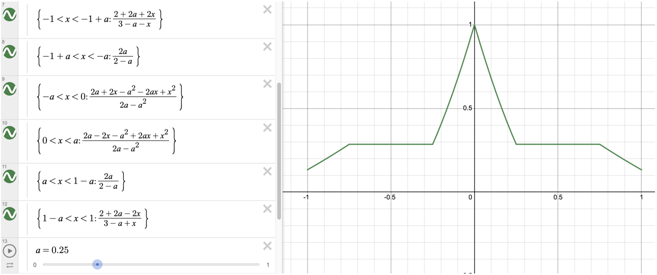
Cody's numerical simulation suggests that this estimation gives less than 5% error given that $\epsilon < 0.5$ (which is a fairly loose constrain since the interval length is just 2), and that the error is indeed converging to 0 as ϵ becomes smaller.

N=2 and 3 Cases

Using the approximation, we reduce the original problem to calculating the probability

$$\mathbb{P}_S(x \in B(S,\epsilon) \cap B(-S,\epsilon) | x \in B(S,\epsilon))$$

with a fixed $x \in (-1,1)$ and a random S with N points. The graph of this for $N = 2$ is



$\left\{-1 < x < -1 + a: \frac{2+2a+2x}{3-a-x}\right\}$

$\left\{-1 + a < x < -a: \frac{2a}{2-a}\right\}$

$\left\{-a < x < 0: \frac{2a+2x-a^2-2ax+x^2}{2a-a^2}\right\}$

$\left\{0 < x < a: \frac{2a-2x-a^2+2ax+x^2}{2a-a^2}\right\}$

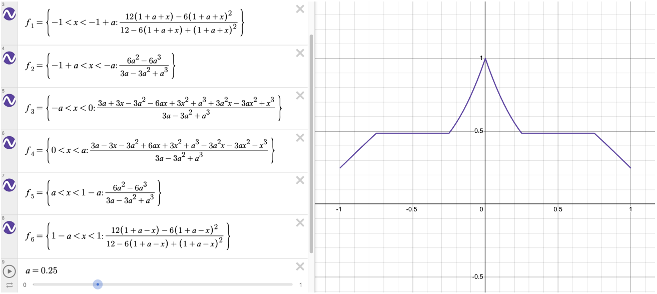
$\left\{a < x < 1 - a: \frac{2a}{2-a}\right\}$

$\left\{1 - a < x < 1: \frac{2+2a-2x}{3-a+x}\right\}$

$a = 0.25$

N=2 and 3 Cases

The graph of this for $N = 3$ is



$$f_1 = \left\{ -1 < x < -1 + a : \frac{12(1+a+x) - 6(1+a+x)^2}{12 - 6(1+a+x) + (1+a+x)^2} \right\}$$

$$f_2 = \left\{ -1 + a < x < -a : \frac{6a^2 - 6a^3}{3a - 3a^2 + a^3} \right\}$$

$$f_3 = \left\{ -a < x < 0 : \frac{3a + 3x - 3a^2 - 6ax + 3x^2 + a^3 + 3a^2x - 3ax^2 + x^3}{3a - 3a^2 + a^3} \right\}$$

$$f_4 = \left\{ 0 < x < a : \frac{3a - 3x - 3a^2 + 6ax + 3x^2 + a^3 - 3a^2x - 3ax^2 - x^3}{3a - 3a^2 + a^3} \right\}$$

$$f_5 = \left\{ a < x < 1 - a : \frac{6a^2 - 6a^3}{3a - 3a^2 + a^3} \right\}$$

$$f_6 = \left\{ 1 - a < x < 1 : \frac{12(1+a-x) - 6(1+a-x)^2}{12 - 6(1+a-x) + (1+a-x)^2} \right\}$$

$a = 0.25$

If I is the integral of this function from $(-1, 1)$ with respect to x, $\frac{1}{2}(1 + \frac{I}{\mu_0((-1,1))}) = \frac{1}{2} + \frac{I}{4}$ gives the $R(S, \epsilon)$ statistic.

Future Directions

- The first thing we will do is to continue calculating the expected value of R statistic under larger N. Although the expression gets more complicated very rapidly, we believe that our method of calculating it (using indicator functions) is quite mechanical and not hard to generalize. Therefore, we plan to develop a program that give us the analytical approximation given any N.
- We also plan to try on larger isometry groups and higher dimensional data.

Thanks

Thank you all for listening! Special thanks to Alex Iosevich[1] and all organizers of the program. Lastly, thanks to our advisor Kunxu Song!

[1] Also, please give us an A in 471 :)