FSSD Implementation

Darsh Patel

September 2021

1 Introduction

A Goodness-of-Fit test using the Finite Set Stein Discrepancy statistic and a set of paired test locations. The statistic is $n * FSSD^2$. The statistic can be negative because of the unbiased estimator.

- H_0: the sample follows p
- H_1: the sample does not follow p

p is the specified constructor in the form of an UnnormalizedDensity.

2 Pseudocode

- p: an instance of UnnormalizedDensity
- q: an unknown sample density
- k: a DifferentiableKernel object
- V: a J x dx numpy array of J locations to test the difference
- null sim: an instance of H_0Simulator for simulating the bull distribution
- α: significance level

2.1 Computing the Feature Tensor of Data

- X: an n x d instance of data
- n: 1st-dimension of X

INPUTS: X, an instance of data

OUTPUTS: X_i, the feature tensor of the data

$$\nabla p = \nabla X \quad (1)$$
$$K = Ker(k) \quad (2)$$
$$\nabla_X k_1(X, v), \nabla_X k_2(X, v), ..., \nabla_X k_n(X, v) = \nabla_X k(X, v) \quad (3)$$
$$\frac{dK}{dv} = \nabla_X k(X, v)^T \quad (4)$$
$$X_i = \frac{\nabla K + \frac{dK}{dv}}{\sqrt{d * J}} \quad (5)$$

2.2 Using the feature tensor to compute the mean and variance of the asymptotic normal distribution under H_1 of the test statistic

INPUTS: X_i, the n x d x J feature tensor of the log density of the distribution; useUnbiased, if true use the unbiased version of the mean; returnVariance, if false, avoid computing and returning the variance

OUTPUTS: mean (T) and variance (Var)

$$n, d, J = Dim(X_i) \quad (6)$$
$$\tau = X_i[dim : (n, d * J)] \quad (7)$$

if useUnbiased:

$$t_1 = \sum(\frac{\sum^n \tau}{n})^2 * \frac{n}{n-1} \quad (8)$$
$$t_2 = \frac{\sum \frac{\sum^n \tau^2}{n}}{n-1} \quad (9)$$
$$T = t_1 - t_2 \quad (10)$$

else:

$$T = \sum(\frac{\sum^n \tau}{n})^2 \quad (11)$$

if not returnVariance: return T

$$\mu = \frac{\sum^n \tau}{n} \quad (12)$$
$$Var = 4 * (\frac{\sum^n (\tau \cdot \mu)^2}{n}) - 4 * (\sum^n \mu^2)^2 \quad (13)$$

return T, Var

2.3 Compute and return the test statistic

INPUTS: X_i, the n x d x J feature tensor of the log density of the distribution; T, the mean of the feature tensor

OUTPUTS: $n * FSSD^2$, test statistic

$$M = T \tag{14}$$
$$stat = n * FSSD^2 = n * M \tag{15}$$

return stat

2.4 Approximate the p-value with the permutations

INPUTS: $n * FSSD^2$, test statistic

OUTPUTS: p-value, H_0 rejection

p-value approximated by taking the mean of the event that the simulated test statistic is greater than the calculated test statistic

Reject H_0 if p-value $< \alpha$

return α, p-value, $n * FSSD^2$, H_0 rejection