Constructing a Constant-Scaling Density Estimator

Sarah You[1] and Friends[1*]

[1]Department of Statistics and Actuarial Science, University of Western Ontario

Abstract

To be completed eventually. Tbh I'm not sure what I can write on here. I'm not even sure if this is ever going to get published. I don't think this is getting anywhere and I'm just wasting my own and my professor's time. I feel stupid and worthless and I will never be able to produce any research that is worthwhile. I am never getting into graduate school. I should commit toaster bath.

Introduction

The current most popular estimation methods of unknown underlying probability density function from an observed sample of data is to apply kernel density estimation. The construction is as follows: Let $(x_1, x_2..., x_n)$ be independent and identically distributed samples drawn from some univariate distribution with unknown density f at at any given point x_i. To estimate the shape of function f, we take a moving average estimate, called a *kernel* , shown below,

$$\hat{f}_\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(x - x_i) = \frac{1}{n\lambda}\sum_{i=1}^{n} K(\frac{x - x_i}{\lambda}) \quad (1)$$

where K is the smoothing kernel and λ is a smoothing parameter called the bandwidth (cite Faraway edition 2 page 299). The kernel function K must obey several properties. It must be a smooth function where $\int K(x)dx = 1$, $\int xK(x)dx = 0$, and $\sigma_K^2 = \int x^2K(x)dx > 0$ (cite CMU page 6). This technique provides a smooth estimate of the pdf and uses all sample points, which is not possible to do by looking at the histogram alone, the once popular method of estimating a sample's underlying distribution (cite Weglarczyk).

However, there are several disadvantages to performing Kernel Density Estimation. Firstly, kernel density estimation produces a constant bias, particulary near the boundaries of datasets with a bounded support. Furthermore, if the underlying distribution has a longer tail/tails, the "main" component of the distribution risks over-smoothing (cite Zambom). Considering the Beta distribution both has a a bounded support over $(0, 1)$, and has heavy-tailed distributions depending on the magnitude of their shape and scale parameters, Kernel Density Estimation is wholly not suitable. This article attempts to find a method to use constant scaling to estimate if an underlying distribution is a beta distribution, which could significantly simply the estimation process both in time and computational resources.

*Corresponding author: aryan@email.com

Methodology

Simulation Procedure for Trend Estimation

Data for the following analysis were simulated using the following procedure.

1. Let $X_1, X_2, ...X_n \sim Beta(\alpha, \beta)$, $n \in \mathbb{N}$ and is sufficiently large, and $\alpha,\ \beta > 0$.
2. Let $Y_1, Y_2, ...Y_n$ be the upper p percent of $X_1, X_2, ...X_n$. Define
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$
and
$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{n-1}$$
.
3. Construct a constant c such that $cs + Y_{(n)} = 1$, where $Y_{(n)}$ denotes the sample maximum. Then
$$c = \frac{1 - Y_{(n)}}{s}$$
.
4.

Graphical Output - Base Scenario

Issues in Research Process

Results

To be filled when there are results

Discussion