Bivariate random number generating using kernel density estimator

Yaqiong Yao, Cheng Huang

10/12/2018

Motivation

The motivation of this project is to improve the process of drawing samples for bootstrap purpose. Since when doing bootstrap, we only have a set of sample data instead of the density function of the random variable. Then how to draw random numbers from the unknow distribution becomes a problem. In this project, we are going to investigate two methods. The first one is to use empirical cumulative distribution. Randomly generate numbers between 0 and 1 and then map them to certain values by the empirical cumulative distirbuiton. The second one is to use kernel density estimate, where we can obtain the smoothed CDF for this unknow distribution.

Methodology

The idea of drawing samples from kernel density estimate is inspired by the website.[1]. It gives us a naive implementation of generating random values from univariate distribution. The principle of this method is first to randomly select values from certain kernel density and then draw samples from data we have. Adding these two parts together forms what we want. It also provides a detailed proof in the website. In our project, we are going to use same idea and extend it to the case with two random variables linked by copula.

Simulate from copula

Suppose we have two random variables X and Y. Their dependency is defined by certain copula C.

$$F(x, y) = C(F_x(x), F_y(y)),$$

where $F_x(\cdot)$ and $F_y(\cdot)$ is the marginal cumulative distribution of X and Y. Based on the copula, we can randomly select (U_x, U_y), where $U_x = F_x(x)$ and $U_y = F_y(y)$. Then, by mapping the samples onto the empirical cumulative distribution function, one could easily generate sample from a specific distribution. And the second method is to construct a kernel density estimation (KDE) for the sample drawed. It could be created by first selecting kernel densities functions like Normal. We can also express the cumulative distribution of KDE explicitly

$$\hat{F}_h(x) = \sum_{i=1}^{n} \frac{1}{n} \Phi(\frac{x - x_i}{b}), \tag{1}$$

where n is the number of sample points we have, b is the bandwidth and Φ is the CDF of standard normal distribution. Base on the 1, we can also acquire a closed form solution.

[1] https://stats.stackexchange.com/questions/321542/how-can-i-draw-a-value-randomly-from-a-kernel-density-estimate/321726