Confidence interval and test of multiple correlation coefficient in high-dimensional population

Shurong Zheng

joint work with Zhidong Bai

1 abstract

Multiple correlation coefficient (MC) is an important concept and quantity in multivariate analysis, which is defined as the maximum correlation coefficient between one random variable and a linear combination of another set of random variables, usually denoted by $\overline{R}$. The MC measures the capability of a set of random variables to predict another random variable of main interest. In linear regression, $\overline{R}^2$ is used as the coefficient of determination which is defined by the ratio of sum of squares of regression to the sum of total squared variations. It measures the goodness of fit of the linear model.

The application of MC has a long history as well as a wide range. For example, in dominance analysis it establishes the predictor importance by examining the additional contribution of each predictor to the MC in a given model; In time series, MC is used to establish a portmanteau test of lack of fit (Pena and Rodriguez (2002)). In spatial process, MC is used to assess the correlation between one spatial process and several others (Dutilleul et al. (2008)).

It is known that the almost all existing results about MC are obtained under the assumptions of normality and/or fixed dimension of data. However, in practice the normality assumption does not hold and the dimension of data is often large so that the bias of the

estimation of the PMC (population MC)is not tolerable. Moreover, the large dimension usually makes SMC (sample MC) overestimating PMC. As mentioned earlier, the adjusted MC R^* may take undesirable negative values. Especially, when the dimension of data becomes larger, the adjusted R^{*2} will have a large probability to be negative. Therefore, the affect of large dimension has to be considered when limiting theory, such as the CLT, for SMC is applied. A better approach in dealing with large-dimensional data settings would be based on an asymptotic theory under nonnormality and when both n and p approach proportionally to infinity. In fact, in nowadays large dimensional data are frequently met in many modern scientific fields, such as micro-array data in biology, stock market analysis in finance, signal processing in wireless communication networks, etc. Then the motivation of this paper is to establish the limits and CLT of SMC without the normality assumption when the sample size and the dimension of the data increase proportionally.