Bandwidth selectors for multivariate kernel density estimation

Tarn Duong

Kernel density estimation is an important data smooting technique. It has been applied most successfully for univariate data whilst for multivariate date its development and implementation have been relatively limited. The performance of kernel density estimators depends crucially on the bandwidth selection. Bandwidth selection in the univariate case involves selecting a scalar parameter which controls the amount of smoothing. In the multivariate case, the bandwidth matrix controls both the degree and direction of smoothing so its selection is more difficult. So far most of the research effort has been expended on automatic, data-driven selectors for univariate data. There is, on the other hand, a relative paucity of multivariate counterparts. Most of these multivariate bandwidth selectors are focused on the restricted case of diagonal matrices. In this thesis practical algorithms are constructed, with supporting theoretical justifications, for unconstrained bandwidth matrices.

The two main classes of univariate bandwidth selectors are plug-in and cross validation. These unidimensional selectors are generalised to the multidimensional case. The univariate framework for theoretically analysing kernel density estimators is extended to a general multivariate version. This framework has at its core the quantification of the relative rates of convergence which provide a guide to the asymptotic behaviour of bandwidth selectors. Simulation studies and real data analysis are employed to illustrate their finite sample behaviour. It is found that unconstrained selectors possess good asymptotic and finite sample properties in a wide range of situations.

Buoyed by this success, two extensions are embarked upon. The first is variable bandwidth selection, generalising the above case where the bandwidth is fixed throughout the sample space. The variation of the bandwidths is controlled by the local properties of the data. The novel contribution is to use non-parametric clustering to summarise these local properties, along with unconstrained bandwidth matrices. The

second is in kernel discriminant analysis where unconstrained bandwidth matrices are shown to produce more accurate discrimination.

Department of Statistics
Division of Economic and Financial Studies
Macquarie University
New South Wales 2109
Australia