BIG-SIR a Sliced Inverse Regression Approach for Massive Data

Benoit Liquet[a]

[a]School of Mathematics and Physics
The University Of Queensland
Brisbane, Australia
b.liquet@uq.edu.au

IIn a massive data setting, we focus on a semiparametric regression model involving a real dependent variable Y and a p-dimensional covariable X. This model includes a dimension reduction of X via an index $X'\beta$. The Effective Dimension Reduction (EDR) direction β cannot be directly estimated by the Sliced Inverse Regression (SIR) method due to the large volume of the data. To deal with the main challenges of analysing massive datasets which are the storage and computational efficiency, we propose a new SIR estimator of the EDR direction by following the "divide and conquer" strategy. The data is divided into subsets. EDR directions are estimated in each subset which is a small dataset. The recombination step is based on the optimisation of a criterion which assesses the proximity between the EDR directions of each subset. Computations are run in parallel with no communication among them.

The consistency of our estimator is established and its asymptotic distribution is given. Extensions to multiple indices models, q-dimensional response variable and/or SIR_α-based methods are also discussed. A simulation study using our `edrGraphicalTools` **R** package shows that our approach enables us to reduce the computation time and conquer the memory constraint problem posed by massive datasets. A combination of `foreach` and `bigmemory` **R** packages are exploited to offer efficiency of execution in both speed and memory. Finally, results are visualised using the bin-summarise-smooth approach through the `bigvis` **R** package.

Key words: High performance computing, Effective Dimension Reduction (EDR), Parallel programming, Sliced Inverse Regression (SIR).