Scalable High-Dimensional Multivariate Linear Regression for Feature-Distributed Data

Shuo-Chieh Huang[1], Ruey S. Tsay[1]

[1]*Booth School of Business, University of Chicago, U.S.A.*

Abstract

Feature-distributed data, referred to data partitioned by features and stored across multiple computing nodes, are increasingly common in applications with a large number of features. This paper proposes a two-stage relaxed greedy algorithm (TSRGA) for applying multivariate linear regression to such data. The main advantage of TSRGA is that its communication complexity does not depend on the feature dimension, making it highly scalable to very large data sets. In addition, for multivariate response variables, TSRGA can be used to yield low-rank coefficient estimates. The fast convergence of TSRGA is validated by simulation experiments. Finally, we apply the proposed TSRGA in a financial application that leverages unstructured data from the 10-K reports, demonstrating its usefulness in applications with many dense large-dimensional matrices.

Keywords: Frank-Wolfe algorithm; Reduced-rank regression; Distributed computing.