Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – August, 2021
DOI: 10.11159/icsta21.003

On the Analytic Power of Divide & Recombine (D&R)

William S. Cleveland
Purdue University, USA
wsc@purdue.edu

Abstract

In D&R (aka Split & Conquer), the data are divided into subsets. The division serves as a base for analysis of big data and for data visualization. Different analytic processes are applied to the subsets that constitute a recombination of the information in the data. For big data there are three scenarios. (1) The division is based on the subject matter, e.g., financial data for 100 banks; the division is by bank, and the 100 outputs of analytic methods are further analyzed. (2) An analytic method is applied to each subset, and the outputs are recombined with a recombination method applied to get one result for all of the data. This can provide, for all if the data, estimates of parameters or more complex information such as a likelihood function. D&R research consists of finding division and recombination methods that maximize statistical accuracy. Parallel distributed environments like Hadoop and Spark provide high computational performance for (1) and (2). (3) Similarly, an analytic method is applied to all subsets, but an iterative MM algorithm is used for optimization, e.g., maximum likelihood, that among other nice properties can avoid very large matrix inversion, turn a non-differentiable problem into a smooth problem, etc. For visualization, subsets are created by conditioning on one more variables of the analysis to create subsets of the other variables in the analysis. The subsets are displayed using the Trellis Display framework of multi-panel display. This provides a very powerful mechanism for exploratory study of multi-dimensional datasets, modeling the data, and understanding the results of analysis.