Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 4, 692–696
https://doi.org/10.1214/17-BJPS365A
Main article: https://doi.org/BJPS365

Comment: Consensus Monte Carlo using expectation propagation

Andrew Gelman[a] **and Aki Vehtari**[b]

[a]*Columbia University*
[b]*Aalto University*

The article under discussion considers algorithms for performing inference on large or unwieldy datasets by partitioning the data, analyzing each piece separately, and then putting the inferences together. This is an active area of research in computational statistics (see, for example, Tresp, 2000, Ahn, Korattikara and Welling, 2012, Gershman, Hoffman and Blei, 2012, Hoffman et al., 2013, Wang and Blei, 2013, Scott et al., 2013, Wang and Dunson, 2013, and Neiswanger, Wang and Xing, 2013).

Before getting to our own ideas in this area, we would like to review the need for such divide-and-conquer algorithms.

"Big data" is sometimes defined as more than can fit into memory at once, or as any dataset that is too large for us to do what we would like with it. For example, from Wikipedia, "*Big data* is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them."

Steve Scott works at Google and sees really really big data. The problems on which we work are much smaller but the concerns he raises in the paper under discussion are relevant in our work too: for us, a survey with 100,000 respondents counts as "big data" in that the models we'd like to fit can run uncomfortably slowly. For example, it might take a couple hours to run a hierarchical regression predicting survey responses given several factors such as age, sex, ethnicity, education, and state. A few hours doesn't sound so bad, but this inhibits our ability to explore data by trying out and perturbing lots of models.

Why do we need to fit so many models to a single dataset? Because survey adjustment is complicated. As you may have heard, the 2016 U.S. presidential election polls were not far off in aggregate—Hillary Clinton had a small but consistent lead in the polls for months, and she won by three million votes—but they failed in several key states, with the key problem being differential nonresponse: some proportion of Republican voters who were not being reached in surveys (Gelman, 2016). To get the correct inferences from a survey, it is necessary to adjust for as many variables as possible so as to be able to reasonably extrapolate from sample to population. The basic idea is to model the survey response y given some large

Received June 2017; accepted June 2017.