Brazilian Journal of Probability and Statistics
2017, Vol. 31, No. 4, 697–700
https://doi.org/10.1214/17-BJPS368
Main article: https://doi.org/10.1214/17-BJPS365

Rejoinder

Steven L. Scott
Google

I thank Gelman and Vehtari (GV) and Draper and Terenin (DT) for a thoughtful discussion. Both pairs of authors have worked with big data "in the trenches" for a long time, and bring valued expertise and perspective.

Both GV and DT observe that big data is only really necessary when fitting complex models, which typically means models involving many parameters. I agree, and obliquely made the same point in the article by focusing on what happens to methods of improving on consensus Monte Carlo (CMC) as the parameter dimension p increases. Unfortunately for the methods investigated in the article, it seems the curse of dimensionality bites hard, with averaging the only method that is unaffected by dimension.

The good news on dimensionality is that in many applications from the tech industry, much of the "bigness" comes from sparse data, such as factors with vast numbers of levels. For example, one element in your model might be a dummy variable indicating a URL or web service that referred a visitor to your page. There are effectively infinitely many potential levels for such a factor, but for any particular visitor all but one are irrelevant. Modeling such a factor using random effects can turn a problem where the parameter dimension p is infinite into one where p is reasonably small. This blurs the line between two cells in DT's table, allowing sharding methods to leach into the "big n, big p" cell. Admittedly, replacing parameters with random effects is a trick that won't work in all problems, so perhaps DT were correct in applying the "model specific" label to that cell.

On the slippery question of what "big data" actually means, both GV and DT adopt the view that "big data" implies a data set which is too large for standard methods to work comfortably. I'd like to point out that this perspective, which is shared by many statisticians, can confuse engineers who view data as either "big" or not. Imagine two analysts working on the same data set. One computes an average, while the other fits a complex Bayesian model using MCMC. Does it make sense to say that one analysts has a "big data problem" while the other does not? There are obviously viewpoints from which the answer is "yes," but it is equally obvious that the difference is not the data, but the algorithms used to process it. One analyst has an algorithm that scales better than the other. Classical MCMC methods scale poorly because they assume you are able to loop over the data at will.

Received June 2017; accepted June 2017.