Combining Experiments with Big Data to Estimate Treatment Effects

Jasjeet Sekhon, University of California, Berkeley

The rise of massive datasets that provide fine-grained information about human beings and their behavior provides unprecedented opportunities for evaluating the effectiveness of treatments across the human sciences. New methodological challenges have to be overcome to make the most of these opportunities. Among them are algorithms that do not scale, questions about how to combine experimental studies with massive observational data, and how to estimate heterogeneous treatment effects for subgroups. A problem that connects these challenges is the explosion of false positives that has come along with massive data. Progress on these challenges will require combining insights from diverse fields, particularly statistics and computer science. The core of this talk explores one issue of experimental design.

Our experiments are growing in size. Massive experiments often revolutionize their respective fields as they tend to be more representative of the population of interest and important effects, that normally are undetectable, are possible to estimate. Parallel to this development, several strides have recently been made in study of how to best design small-scale experiments. The recent literature reminds us that randomization alone does not solve all problems. In particular, most of the desirable properties that randomization brings are only guaranteed on average. In any specific sample and treatment assignment, there may be remarkable imbalances on prognostically important covariates and inferences can thereby be severely flawed. Many experiments will, when viewed unconditionally, have estimators with unnecessarily high variance; and when viewed conditionally on the observed imbalances, the estimators will be biased.

In simple small-scale experiments the standard method to avoid these problems is to block the sample before randomization. In its most stylized description, blocking is when the researcher divides the experimental sample into groups, or blocks, based on covariates and assigns treatment in fixed proportions within the groups but independently between them. If the blocks are formed so to make the units they contain as similar as possible this procedure will ensure that the treatment groups are balanced and thereby greatly improve the quality of any inferences.

Another, possibly more important, reason to block has become evident with the rise of large experiments: blocking can greatly improve inferences to other populations---be it subgroups in the experimental sample or reweighted subgroups so as to estimate treatment effects for other populations. As treatment assignment is independent between blocks, each group of units can be seen as an experiment in its own right. If the researcher is interested in the treatment effect in some subpopulation, she can simply extract the corresponding blocks and treat them as a separate experiments. Unlike most other methods for analyzing fine-grained effects, this technique guarantees unbiased estimates of the treatment effect in the subpopulations. As the blocks are predefined, it is much harder to comb through the data to find seemingly significant results. Blocking thereby improves transparency and controls the rate of false positives by design.

However, although the basic idea of blocking goes back to the infancy statistics (Fisher 1926), current blocking methods are restricted to special cases, run in exponential time, or are heuristic, providing an unsatisfactory solution in many common cases. The lack of progress in the area is explained by the fact that blocking problems are isomorphic to partitioning problems in graph theory. Net of a few special cases, these problems are known to be NP-hard. In fact, the first algorithm with proven optimality for the blocking problem with covariates of high dimensionality was introduced as late as 2004 (Greevy 2004). That algorithm exploited one of the few special cases where the underlying partitioning problem is not NP-hard. While this made the algorithm possible, it restricts the blocks to contain exactly two units and therefore severely limits its applicability.

We introduce a new blocking algorithm that can be used for multiple treatments and blocks of different sizes. We describe a constant-factor approximation algorithm with polynomial time complexity. We investigate the optimization problem where, given a minimum required group size and a distance metric, one wants to find a set of groups---a blocking---so that the maximum distance between any two units within a group is minimized. Finding this blocking is an NP-hard problem. Our algorithm produces a blocking where the maximum distance is guaranteed to be at most four times the optimal value and does so in polynomial time. Unlike previous algorithms, our algorithm works for an arbitrary group size facilitating complex experiments with several treatment arms. Simulation studies indicate that the algorithm produces solutions with a maximum distance well below the theoretically guaranteed bound, often close to the optimal solution. The algorithm can successfully be used in huge experiments; millions of observations can be blocked using a desktop computer.