A COMPUTATIONAL MODEL FOR CAUSAL INFERENCE VIA SUBSET SELECTION

Allocation: Illinois/50.0 Knh
PI: Wendy K. Tam Cho[1]
Co-PI: Yan Y. Liu[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

Researchers in all disciplines desire to identify causal relationships. Randomized experimental designs isolate the treatment effect and thus permit causal inferences. However, experiments are often prohibitive because resources may be unavailable or the research question may not lend itself to an experimental design. In these cases, a researcher is relegated to analyzing observational data. To make causal inferences from observational data, one must adjust the data so that they resemble data that might have emerged from an experiment. The data adjustment can proceed through a subset selection procedure to identify treatment and control groups that are statistically indistinguishable. Identifying optimal subsets is a computationally complex and challenging problem but a powerful tool for discovering scientific insights in a wide variety of fields.

INTRODUCTION

The aim of the project is to design efficient computational and statistical algorithms for making causal inferences. The research builds on previously developed statistical models [1]. The proposed research advances this concept from both theoretical and applied perspectives—it furthers discrete optimization models that capture the features and subtleties of the causal inference problem and provides quantitative tools and models that address the causal structures of interest in important substantive problems in a diverse set of scholarly domains, respectively.

Enhancing the ability to make causal inferences from observational data will stimulate research in a wide variety of fields and enhance our understanding of a broad array of phenomena. In medicine or health, causal inference studies have included applications to criminality rates related to gene patterns [7], the effect of generic substitution of presumptively chemically equivalent drugs [6], in utero exposure to phenobarbital on intelligence deficits [5], and the effect of maternal smoking on birthweight [2], to name but a few. In social science, causal inference models have been applied to studies on the impact of different voting technologies [3] and the effect of affirmative action [4]. The array of potential research questions is important, diverse, and unconstrained by field of study. Additionally, the possible applications of causal inference models are limitless given a proper research design and available data.

METHODS & RESULTS

Our **novel** approach to the data matching problem proffers a paradigm change from statistical matching methods. The statistics literature focuses on identifying individual data matches with models that are dependent on a set of underlying assumptions. While we retain the same goal of post-processing observational data to resemble experimental data, we **redefine** the process to obtain the goal of covariate balance directly, bypassing the assumptions of the statistical models entirely by replacing them with a computationally challenging problem. By consolidating steps to determine the optimal subsets of the controls and treatments based on some set of balance measures, we show that important insights into many problems that have been traditionally analyzed via statistical models can be obtained by re-formulating and evaluating within a large-scale optimization framework.

Our analysis speaks to statistical frameworks with astronomical (10^{5000}) solution spaces. Here, the identified solutions need to be independent of one another. This independence requirement adds a significant challenge for standard optimization methodologies. In this vein, we designed a hybrid metaheuristic with specialized intensification and diversification protocols in the base search algorithm. We experimentally demonstrate that our diversification protocol cuts the time required to find independent solutions in half while our intensification protocol enables the identification of solutions that are difficult to find with non-collaborative processors. We extend our algorithm to the high-performance-computing realm by implementing methods for utilizing multiple processors to collaboratively hill climb, broadcast messages to one another about the landscape characteristics, diversify across the landscape, and request aid in climbing particularly difficult peaks.

For scalability, our code eliminates the costly global synchronization operation which was originally conducted for incoming message checking. With the global synchronization, the message passing time ranged from 42% on 240 cores to 72% on 960 cores. The message passing cost declined considerably with our asynchronous communication protocol (0.007% on 240 cores and 0.01% on 960 cores). Figure 1 shows that the numerical performance scales well in our weak scaling experiments. We continue to explore scalability with larger numbers of cores as the problem instances for most practical problems are vast and require a substantially greater number of processing cores to obtain satisfactory results.

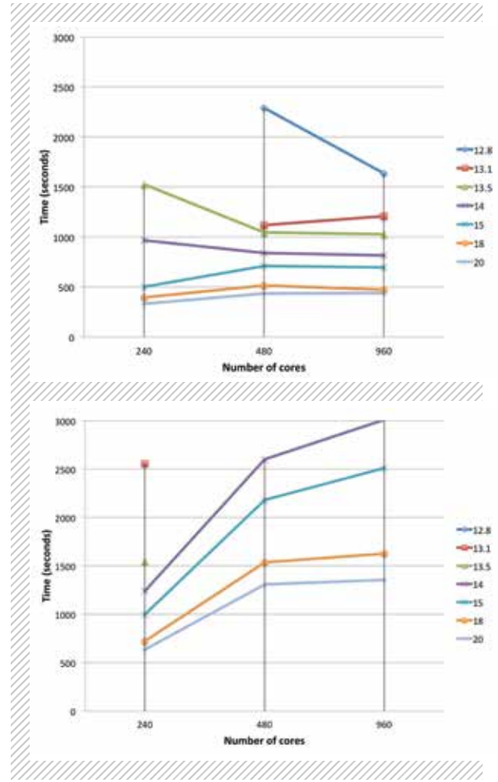

FIGURE 1: Results from Weak Scaling Experiments. Asynchronous (top) versus Synchronous (bottom) Communication.

WHY BLUE WATERS

Balance Optimization Subset Selection (BOSS)'s shift from individual matching to subset selection highlights an interesting combinatorial aspect of both the matching methodologies as well as the subset selection methodology. In particular, for even moderately sized data sets, the set of possible "solutions" is extremely large. For instance, if our control pool has 100 members, and we wish to choose a subset of size 20, there are $\binom{100}{20} = 5.359834 \times 10^{20}$ possibilities. This would be a small instance of an actual substantive problem. In a classic data set for this problem, the control pool has 16,000 members from which we choose a subset of size 185—an astronomically large problem. Given the sheer size and the non-rugged solution landscape, finding balanced subset in this solution space proves to be computationally challenging. The problem presents an extreme-scale optimization problem for which a petascale resource like Blue Waters is necessary.

PUBLICATIONS AND DATA SETS

Cho, W.K.T., and Y.Y. Liu, A parallel evolutionary algorithm for subset selection in causal inference models. Proceedings of the 2016 Annual Conference on Extreme Science and Engineering Discovery Environment. XSEDE'16: Diversity, Big Data, & Science at Scale, Miami, FL, July 17–21, 2016.

Cho, W.K.T., and Y.Y. Liu, A high-performance approach for solution space traversal in combinatorial optimization. SC15: The International Conference for High Performance Computing, Networking, Storage and Analysis. Austin, TX, November 16–19, 2015.