Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)

Yang Wang, Department of Systems and Information Engineering

1 Introduction

In the field of behavioral, social and psychological science, the traditional practices of experiments design share a common sequences of events. First, a hypothesis is proposed. Next, an experiment is designed which would test the hypothesis. After building up the theoretical framework, participants are then recruited and those that consent are enrolled to participate in the experimental protocol. Experiment data are collected into a centralized repository. This central data repository is used to fit candidate statistical models, commonly obtaining maximum likelihood or Bayesian estimates of model parameters and fit statistics. These data are generally considered to belong to the research group that collect the observational data. Experiment results are disseminated through journal articles and/or conference talks. Finally, other research groups may replicate the experiments and generate new hypothesis on top of the previous results.

However, there are several potential problems accompanying this approach: i) There are differing views on who owns the data. ii) Protection of participants confidentiality may impose restrictions or barriers to data sharing. iii) There can be a very long interval between the generation of a hypothesis and the next opportunity for the hypothesis to be revised or the experiment replicated. iv) It is difficult, if not impossible, to perform longitudinal linking between different data repositories while maintaining participant confidentiality.

Boker et al. (2013) proposed a new paradigm wherein data remain where they were originally collected—on each participant's personal smartphone, computer, tablet, or wearable computing device—and remain private, that is to say these data are never revealed by the participant. Candidate statistical models are fit using a secure a internet connect to send a vector of model parameters to individual's device. Each device will calculate the likelihood of the data on that single device and only this likelihood value will be transmitted back to the research group. An optimizer in the research lab aggregates the likelihood values from all participants and chooses a new vector of parameters and repeat the process. The aim is to find a maximally likely set of parameter values for the model. Also, data can be collected at the same time that models are optimized. This means that when sufficient data are collected to reach a preselected statistical power, the study can automatically terminate or switch to a cross validation regime. Besides, individual level variables, repeated measurements, and time series are conveniently linked on participants' devices. Furthermore, if a participant opts into many studies simultaneously and consents to data sharing between studies, all of the studies have real time data sharing.

2 Objectives

This new paradigm would ease a variety of new study designs that involve bigger and higher quality data than previously practical. For example, it would become feasible to study patient history across hospitals even when data is legally prohibited from leaving the hospital. And without concern of personal information leakage, participants are more likely to provide true responses and reliable measurements. In general, MIDDLE permits statistical models to be fit and scientic hypotheses to be tested while maintaining the privacy of participants data. As an integrated software platform, MIDDLE will provide automatic management of opt-in and opt-out consent; facilitate longitudinal linking of within participant data and sharing of data between studies; reduce cost for the researcher and funding institute; and accelerate determination of results. Robust privacy protection will reassure participants involved in sensitive research and permit research on topics that are otherwise too sensitive or involve data that is legally required to be kept private.

3 Research methodology and plan

My proposed research is centered around the development of the MIDDLE paradigm with the aim of building a general-purpose platform prototype for data-intensive research. This platform features data privacy protection, faster data sharing and efficient experiment management.

More specifically, I will focus on distributed and peer-to-peer estimation algorithms and their software implementation. In the context of MIDDLE, distributed likelihood estimation (DLE) refers to a method for estimating maximum likelihood parameters and fitting statistics for statistical models that is very similar to the full information maximum likelihood method. It involves knowledge in sensor networks, data fusion, statistical analysis, optimization and software development. In signal processing and sensor networks, it is usually assumed that the cost of processing data in separate nodes is more efficient than transmitting the raw data back to a central repository. The literature abounds with methods and algorithms for distributed data-processing and parameter optimization. Specifically, Nowak (2003) considered a distributed expectation-maximization algorithm for density estimation. He proved that for a broad class of estimation problems, the distributed algorithms converge to the globally optimal values at a particular rate (Rabbat & Nowak, 2004). Also, Blatt and Hero (2004) proposed an algorithm for distributed maximum likelihood estimation in sensor networks. Interestingly, the aggregation of suboptimal local estimates can lead to the globally optimal solution. For MIDDLE, participants may opt in or drop out of the experiment at any time and the incoming new data may introduce new structure and alter the model. Unlike traditional distributed estimation problems, the estimation of statistical models in MIDDLE is performed on a dynamically changing sample. It requires new algorithm and convergence criteria to optimize the ever-evolving longitudinal multilevel structure.

I will cooperate with faculty members and graduate students in the psychology department on this project. There are software, privacy, and statistical challenges involved in building the MIDDLE system. A prototype of this system will allow us to explore privacy and statistical challenges in more detail than is presently possible. For example, what convergence criteria are most suitable for a distributed likelihood calculation wherein participants may choose at their leisure to enter or leave the system? What choices are available for standard errors and how do these choices influence interpretation? Can we compensate for the bias that will be introduced when participation is correlated with measured or missing data? Our work will be published with an open-source license and we will endeavor to make it convenient for other research groups to install and work with the prototype to investigate issues of interest.

In addition to statistical questions, a prototype will facilitate the investigation of privacy concerns. We will work to identify weak points in the system most vulnerable to surreptitious monitoring or malicious actors within the system. We will identify areas where cryptography may contribute to ensuring privacy. One promising refinement to enhance individual privacy is to perform peer-to-peer aggregation of objective functions prior to transmission to the MIDDLE optimizer. Peer-to-peer aggregation may help protect the privacy of the connection between participants and the likelihood of the model given their data.

A prototype MIDDLE service is too much work to build from scratch. We will endeavor to reuse existing software components wherever possible to speed us to our goal. For example, `OpenMx` is free and open source software for use with R that allows estimation of a wide variety of advanced multivariate statistical models (Boker et al., 2011). `OpenMx` consists of a library of functions and optimizers that allow workers to quickly and flexibly define a structural equation model (SEM) and estimate parameters given observed data. We will improve the modularity of `OpenMx` so it can adapt to the MIDDLE architecture wherein the optimizer is separated from the evaluation of individual data rows. Specically, we propose to build a prototype MIDDLE system with a plug-in style architecture that includes

- A distributed full information maximum likelihood optimizer as a central service
- A personal device emulation module (PDEM) with a plug-in API
- An individual record likelihood estimator that runs on the PDEM
- A peer-to-peer and peer-to-server communications module
- A personal data storage and retrieval module that runs on the PDEM

With these components, we will simulate a complete experiment on the University of Virginia Linux cluster from start to finish.

References

Blatt, D., & Hero, A. (2004). Distributed maximum likelihood estimation for sensor networks. *In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, *3*(8), iii-929.

Boker, S. M., Brick, T. R., von Oertzen, T., Estabrook, R., Pritikin, J. N., Hunter, M. D., ... Neale, M. C. (2013). *Maintained Individual Data Distributed Likelihood Estimation (MIDDLE).* Manuscript submitted for publication.

Boker, S. M., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... others (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317.

Nowak, R. D. (2003). Distributed EM algorithms for density estimation and clustering in sensor networks. *Signal Processing, IEEE Transactions on,*, *51*(8), 2245-2253.

Rabbat, M., & Nowak, R. (2004). Distributed optimization in sensor networks. *In Proceedings of the 3rd international symposium on Information processing in sensor networks (pp 20-27).*