Causal Analysis of Observational Mobile Sensor Data: A Comparative Study

Yongshin Kim†, Panyu Zhang†, Gyuwon Jung, Heepyung Kim, Uichin Lee*
Korea Advanced Institute of Science and Technology
{ys.k, panyu, gonejung, heepkim}@kaist.ac.kr, uclee@kaist.edu

Abstract

As smart devices produce mobile sensor data every day, there have been attempts to discover causality among variables in observational studies. However, it requires proper methods since we cannot control the treatment in the observational study, and confounders make the causal inference more difficult. In this study, we use two well-known causal inference techniques, “matching” and “convergent cross mapping” on the mobile sensor data to show how to implement them and what are the challenges to be considered carefully.

1. Introduction

Over the past decades, smartphones and wearables have produced tons of live data every day which are collected via built-in sensors. These data are useful in monitoring the user's daily life and help better understand one's behavior. Analysis of the data could lead to designing an intervention system that suggests activities at opportune moments. Thus, previous studies have tried to explore the relationship between the variables from sensor data, especially about the "causality."

Randomized Controlled Trial (RCT), one of the experimental studies, is a useful tool for examining the causal relationship. It compares the outcome from two or more different groups, which are called “control” and “treatment”, and confirms the causality when there is a statistically significant difference in the outcome. It also randomly assigns the subjects to minimize the effect of confounders, variables that might affect both the treatment and outcome. However, RCT may not be available due to ethical issues (e.g., harmful treatment), difficulties in recruiting subjects, etc. Sometimes, when we need to run the experiment in the real world and examine the efficacy of a certain intervention, we may choose an alternative option, “observational study” [1].

In observational studies, researchers could observe the effect of treatment, but it is not well determined who will be treated or not treated. In addition, confounders among the users are not controlled so that it may be difficult to conclude whether the treatment has efficacy in changing the outcome. For instance, suppose we examine the efficacy of an intervention app for promoting physical activity. In this case, the users may be affected by other variables such as weather, emotions, or schedule, and the complex relationship among variables would make it difficult to prove that the app is effective. As there have been diverse health-related interventions such as “Digital Therapeutics”, the causal inference is getting critical to examine the therapeutic efficacy with observational data.

The purpose of this paper is to provide an overview to computer science researchers about how to analyze the causality from mobile sensor data in observational studies. Here, we focus on two well-known causal inference techniques; (1) Matching, which gives optimal pairs of subjects having similar confounders and different treatment levels, (2) Convergent Cross Mapping (CCM), which models a dynamic system without predefined treatment and outcome and determines the causality based on its prediction skill. We show a case study of conducting these techniques. We use smartphone and wearable data from one subject as it is in the N-of-1 trial, which is an important method in personal informatics and self-experimentation [8]. Also, challenges in implementing these techniques are discussed.

2. Causal Inference Techniques

There have been prior studies that apply causal inference to sensor data. Mehrotra et al. [2] developed an application that investigates the causality between emotional states and mobile phone interaction. Berkel et al. [3] suggested that the causality could be found from mobile data via CCM since human mobile interaction might be interpreted as a dynamic system.

Matching makes pairs of comparisons that are similar in confounders but different in treatment levels. Generally, the treatment is considered binary, but the matching could be extended to "non-bipartite matching" and support continuous values [6]. Matching applies distance techniques on minimizes distance measure of confounders (e.g., Mahalanobis distance, Propensity score, etc.) between pairs to minimize their influence when setting up the pairs of control/treatment groups. Causality then is estimated using the Average Treatment Effect (ATE).

CCM is a different approach that assumes data from our daily lives could be described as state variables in a dynamic system [4]. Since these variables are deterministic (i.e., future states are

---

† These authors contributed equally to this work
* Corresponding Author

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (2020R1A4A10187747).

not randomly decided), we may reconstruct the original system (denoted as M) using the sequence of observations of those variables. The core principle is that if X causes Y, X's information should be contained in the previous values of Y. Thus, we reconstruct the dynamic system M using each variable (M_X and M_Y) with time-series data; M_X and M_Y are 1:1 mapped to each other since they are both 1:1 mapped to M. Then we reconstruct M_X using its information in M_Y and vice versa (i.e., "cross-mapping"). The causality could be detected by the cross-map performance, say, if we can cross-map M_X using previous values of M_Y well, we can conclude that X causes Y. CCM works when both causer and causee are nonlinear time series. In dynamical systems, linear autocorrelation usually arises from repeated measurement errors and makes it not clear whether the reconstruction performance results from the measurement errors or its deterministic property. Since CCM is based on dynamic systems and tailored for time series data, the result would be less vulnerable to the temporal correlation between variables [5]. There are several other ways to infer causality with observational data, and they could be diverse by conditions such as pre-intervention covariates, repetitive observations, etc.

3. Case Study: Step Count vs. Calorie Consumption

We present a simple case study on how to apply matching and CCM to sensor data. Here, we use a dataset "K-EmoPhone", which is composed of objective sensor data including motion, physiology, environment, and phone usage via an Android smartphone, Polar H10, and Microsoft Band 2 in a 1-week session per subject (n=81). Among the data, we make a common-sense scenario using steps and calorie consumption as treatment and outcome, respectively. This dataset is used to conduct a causal analysis of the following ground truth statement: "more steps will consume more calories."

We first preprocess this dataset. After iterative testing on various time windows (e.g., minutes, hours, etc.), we set them as 1 hour for matching and CCM. Moreover, we standardize the data and exclude null values by concatenating non-null consecutive time series. After the preprocessing, we implement both matching and CCM on one user's data (id: 705).

3.1. Matching

We begin the matching by identifying potential confounders. The variables in K-EmoPhone (e.g., biosignal, environment, device usage patterns, etc.) are considered as candidates, and we conduct a correlation analysis to find which of them are significantly correlated with both treatment and outcome [2]. In our case study, four variables are shown to be confounders (e.g., location, battery usage, skin temperature, and heart rate).

Next, the subjects are distributed into five ordinal groups so the first one includes subjects with the smallest steps while the last one has the largest steps. Note that we conduct a non-bipartite matching [6] since the treatment has continuous values. We then calculate Mahalanobis distance to take multiple confounders into account and pair the subjects in a way that minimizes the overall distance. In our case, we were able to reach the optimal matching (mean distance = 0.4306) with subjects having relatively high and low step counts in each pair.

Finally, we classify all the subjects into high and low treatment groups and used independent-samples t-test to check whether the confounders are well balanced for these two groups. We then calculate the ATE on the outcome (i.e., calories) for each group and conduct a Wilcoxon-signed rank test to see whether the difference is statistically significant. Our results show a significant difference between the groups ($p < .01$) with an effect size of 0.53. Therefore, by using matching methods, we could conclude steps cause calorie consumption.

3.2. Convergent Cross Mapping

In CCM, we first find the optimal number of time lags to reconstruct the original dynamic system for each variable. By doing so, we could decide how many data points in each sequence of observations are required for this method. The result in our case study shows the numbers are 4 and 3 for step counts and calorie consumption, respectively.

Next, we check the non-linearity of each variable by reconstructing a dynamic system with both linear and nonlinear models. If the reconstructed system based on the nonlinear is closer to the original one, then we conclude that the variable is nonlinear. This process is necessary for CCM, and steps and calories are tested to be nonlinear.

As we have two reconstructed dynamic systems M_{Steps} and $M_{Calories}$ solely based on each variable, now we conduct a reconstruction of M_{Steps} using the information in $M_{Calories}$ and vice versa, which is called "cross-mapping." In Fig.1, the result of the cross-mapping is shown, cross-map from steps to calories (blue line) and calories to steps (red line). Note that both lines are "converging" into certain positive values as the library size (i.e., the length of the time series) increases, which means that there exists a bidirectional causal link between steps and calories. Also, the red line converges at a higher value indicating that the steps cause calories more than calories cause steps.

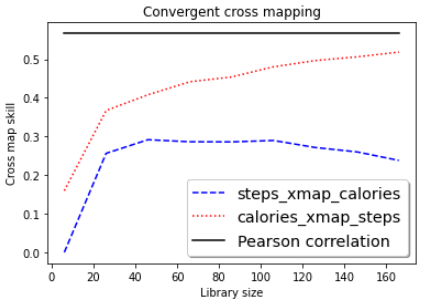

Fig 1. The result of CCM using steps and calories

Finally, we validate if the causal relationship is significant by comparing the result with that from the seasonal surrogate data [7]. The data is generated by randomly shuffling the original time series while preserving the periodicity. We create the shuffled data 100 times. Fig. 2(a) and 2(b) show the result of comparing cross-map skill of real data with that of the shuffled data (the 95% confidence level interval is shaded in light blue). If the cross map skill of original data given largest library size is within the 95 confidence level interval of cross map skills given largest library size of shuffled data, we may conclude that the causal link is not significant; otherwise, it is significant. As in Fig. 2(a) and 2(b), “steps cause calories” is significant and “calories cause steps” is not significant.

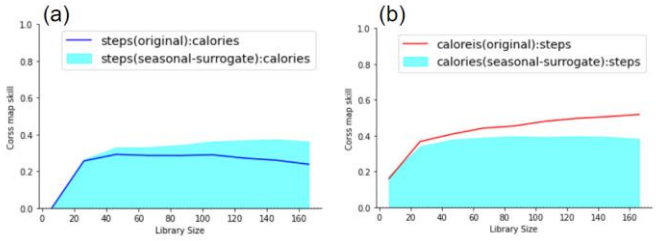

Fig 2. The causality validation using surrogate data for each direction; (a) calories cause steps and (b) steps cause calories

3.3. Challenges in causal inference techniques

There are several challenges in implementing causal inference techniques. In the data preprocessing, we find that how to set the time window may affect the result of causal inference. Thus, we should choose the time window size carefully, considering the property of data, prior domain knowledge about them, or with iterative trials. Though there may not be a gold standard for the time window, we could set it which is (1) not too large to dilute small and temporary changes of data and (2) not too small to be failed in representing the data (Note: CCM may lose nonlinearity and deterministic assumption if the time window is too small).

Next, when implementing the matching, covariate balance might not be perfectly done between the high and low treatment groups. For instance, if there is a high correlation between treatment and confounders, subjects within the same treatment group may show a similar level of confounders (e.g., subjects in the higher treatment also show high levels of covariates). Therefore, we may fail to match subjects with similar confounders coming from the different groups. Researchers should examine bias in the dataset and carefully perform confounder selection for balancing.

Lastly, causal analysis results may show the bi-directional causal links (i.e., A → B and B → A). This may happen if the treatment and outcome are interdependent on each other. In a dynamic system, CCM describes strong interdependency phenomenon as “synchrony.” This issue could be addressed partly by comparing the effect sizes of two directions in matching and comparing the original data with shuffled data in CCM.

4. Conclusion

We reviewed how to perform causal analyses on observational smartphone sensor data, by using commonly used techniques called matching and CCM. We used the K-EmoPhone dataset and showed that these techniques could be used to causality using observational data. With careful considerations of the challenges, we envision that the techniques could be leveraged to optimizing user interface design, measuring the efficacy of digital health interventions, and so on.

References

[1] Song, Jae W., and Kevin C. Chung. "Observational studies: cohort and case-control studies." Plastic and reconstructive surgery 126.6: 2234. 2010.

[2] Mehrotra, Abhinav, et al. "MyTraces: Investigating correlation and causation between users’ emotional states and mobile phone interaction." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1.3: 1-21. 2017.

[3] van Berkel, Niels, et al. "Modeling interaction as a complex system." Human–Computer Interaction 36.4: 279-305. 2021.

[4] Takens, Floris. "Detecting strange attractors in turbulence." Dynamical systems and turbulence, Warwick 1980. Springer, Berlin, Heidelberg, 366-381. 1981.

[5] Sugihara, George, et al. "Detecting causality in complex ecosystems." science 338.6106: 496-500. 2012.

[6] Lu, Bo, et al. "Optimal nonbipartite matching and its statistical applications." The American Statistician 65.1: 21-30. 2011.

[7] Chang, Chun-Wei, et al. "Empirical dynamic modeling for beginners." Ecological Research 32.6: 785-796. 2017.

[8] Li, Ian, Anind Dey, and Jodi Forlizzi. "A stage-based model of personal informatics systems." *Proceedings of the SIGCHI conference on human factors in computing systems*. p. 557-566. 2010.