Applying Causal Discovery to Intensive Longitudinal Data

Brittany L. Stevenson, PhD STEVE473@UMN.EDU
University of Minnesota, Dept of Psychiatry

Erich Kummerfeld, PhD ERICHK@UMN.EDU
University of Minnesota, Dept of Health Informatics

Jennifer E. Merrill, PhD JENNIFER_MERRILL@BROWN.EDU
Brown University, Dept of Behavioral and Social Sciences

Editor: Sisi Ma, Erich Kummerfeld

Abstract

Intensive longitudinal data (ILD) could be a solution for two problems in psychology: 1) In traditional experiments and survey studies, findings are not necessarily representative of the real-life constructs and relationships studied, and 2) Group-level analyses commonly mischaracterize or obscure relationships for individuals. Popular analytic methods within psychology are currently not well-equipped to use ILD for causal discovery and causal inference, however. We have performed the first causal discovery analysis on ILD, encountered some challenges, and developed some solutions to these challenges. This paper describes our application of causal discovery to an example ILD dataset, and addresses two particular challenges that arose: 1) How should one address variables measured on different timelines, and 2) What number of observations is needed for individual-level analysis.

Keywords: causal discovery, intensive longitudinal data, ecological momentary assessment, precision medicine, alcohol use, mood

1. Introduction

For the last several decades, there has been growing interest among psychological scientists in using intensive longitudinal data (ILD) as a solution for two problems in psychology: 1) In traditional experiments and survey studies, findings are not necessarily ecologically valid, or representative of the real-life constructs and relationships studied (Reis and Gable, 2000) and 2) Group-level analyses commonly mischaracterize or obscure relationships between variables at the individual level (Kievit et al., 2013; Molenaar, 2004). Accordingly, ILD are increasingly used to identify ecologically valid targets for interventions (i.e., variables that, when manipulated, should cause a change in clinical outcome) and map out causal relationships between variables within people. However, popular analytic methods within psychology are currently not well-equipped for causal discovery. Because causal discovery methods are not commonly used for ILD, a number of challenges may arise for which there are no widely known solutions. The purpose of this paper is to describe the application of causal discovery algorithms to an example ILD dataset and address two challenges that arose in the process: how to address variables measured on different timelines and the number of observations needed for individual-level analysis. We first begin by describing common analytic techniques used for analyzing ILD in psychology, then move to the application of causal discovery to ILD.

2. Intensive Longitudinal Data (ILD)

In ILD, researchers sample constructs of interest in real time, often via self-report surveys delivered to a participant via mobile device, or by passively collecting information from a wearable or portable device, such as heart rate or GPS location (Bolger and Laurenceau, 2013). ILD is distinguished from non-intensive longitudinal data by the higher volume of assessments per person, usually with a daily or multiple-per-day assessment schedule, as compared to assessments being separated by weeks, months, or years in traditional longitudinal studies (Bolger and Laurenceau, 2013). Several terms are used to describe ILD and similar data types. Though the terms are not always applied uniformly, in general, 'experience sampling methodology', 'ambulatory assessment', and 'ecological momentary assessment (EMA)' often refer to similar types of ILD. The important common factor is that the method involves sampling the constructs of interest in real time (i.e., momentary assessments). In this paper, we use the term ILD to refer to methods that include repeated momentary assessments captured within participants.

ILD could also be referred to as 'time series data' in that repeated measurements are collected over time. In psychology, the term 'time series' invokes connotations of the methods used to analyze time series data, such as controlling for autocorrelation and predicting patterns, rather than explaining them, as is more common in psychology (Jebb et al., 2015). Further, time series data are typically sampled at regular intervals, but in many psychological ILD studies, surveys are administered at irregular time intervals. As a result of these differences, the term 'time series data' is not as commonly used to describe datasets in psychological literature.

Daily diary datasets are also considered a subtype of ILD, but daily diary protocols entail a once-daily survey that typically measures constructs over the past day rather than in real time, a methodological distinction which must be taken into account when analyzing and interpreting the causal structure between variables.

ILD are now collected widely in psychology to measure constructs of interest in natural environments and at the temporal specificity required to capture meaningful variation in things like mood (Bolger & Laurenceau, 2013).

3. Standard Analytic Approach to ILD

The average ILD protocol includes 3-4 scheduled assessments per day over the course of a month with a 75% completion rate for scheduled assessments, resulting in an average of 65-90 completed assessments per person (Jones et al., 2019; Wen et al., 2017). In addition to scheduled assessments, datasets may also include passive data collection methods (e.g., step count via fitness watch) or event-based surveys that are initiated by the participant when an event of interest (e.g., social conflict, cigarette or alcohol use) occurs. Although this data structure lends itself well to performing idiographic (i.e., individual-level) analyses because of the relatively high volume of observations collected per person, group-level analyses are much more common.

The predominant modeling method used to analyze ILD in psychology is referred to as multilevel modeling, or mixed effects modeling (Hoffman, 2015). This approach enables researchers to model fixed effects (effects that are the same for all individuals) and random

effects (which vary by person; Hoffman (2015)). In this paper, we use an ILD dataset measuring alcohol use and mood in young adults multiple times per day. Using the field-standard analysis method, multilevel modeling, we can look at the between-person association between mood and alcohol use (e.g., people who have higher positive mood also tend to drink more) and the within-person associations, though these within-person relationships are typically still summarized and interpreted at the group level (e.g., across all subjects, higher positive mood on a given day is associated with higher likelihood of drinking that day). In the full dataset, there are N = 100 subjects, but in order to facilitate comparisons of the two analysis methods, we have used just the same eight individuals analyzed in causal discovery. Using these eight individuals, a multilevel logistic model was used to predict the odds of drinking from the individual mood variables, including the lagged effects (each variable at the most immediately preceding time point). The multilevel model included separate terms for within-subject and between-subject effects. Results showed that, among these eight people as a group, there were several significant correlates of drinking at a given time point. These were higher than usual within-subject relaxation ($OR = 1.44$, $p = .012$), within-subject energy ($OR = 2.26$, p ¡ .001), and lower than usual within-subject stress ($OR = 0.73$, $p = .032$). Between-subject drinking and mood were not related to odds of drinking. Among lagged variables, endorsing drinking at the previous assessment predicted endorsement of drinking at the current assessment ($OR = 8.56$, p ¡ .001), and within-subject sadness ($OR = 0.66$, $p = .009$) and relaxation ($OR = 0.64$, $p = .003$) predicted lower odds of drinking. With enough observations, it is also possible to explore random effects of these relationships–for example, if there were a random effect for sadness, this would indicate that the relationship between drinking and sadness varies by individual. Exploring this effect could reveal that the relationship is very negative for some people and neutral for others. However, exploring random effects still does not characterize any given individual within the dataset, and is costly in degrees of freedom. Hence, using multilevel modeling, we can discover associations between variables in groups, but not individuals.

In addition, a primary disadvantage of multilevel modeling for ILD is that this method is not capable of discovering the causal structures that underlie the data (e.g., energetic mood causes drinking), limiting the usefulness of many researchers' data to inform interventions, a primary aim of many ILD studies.

4. Application of Causal Discovery to ILD

Applying causal discovery to ILD offers the important advantage of uncovering the potential causal structures that underlie momentary variables, which is critical information for the development of interventions. Because this method is novel, there are a number of challenges that need to be addressed when applying causal discovery to ILD. Below we discuss two primary issues: variables measured over different time periods and how many observations are needed to conduct individual-level analyses for ILD.

Addressing variables that were measured over variable time periods. It is important to understand the assessment schedule for each ILD study before analysis, because it is exceedingly common for variables to be measured on different time spans. For example, many researchers measure mood as a momentary variable (i.e., "How [happy] are you feeling right now?"), but many also measure it over a certain time period (e.g., "Over

the last hour, how [happy] have you been on average?"). Further, it is common for ILD studies to include a mix of momentary (i.e., real-time) and retrospective variables, often measuring the same constructs in both ways (e.g., current alcohol use and use since the last assessment).

In order to analyze datasets with variable time spans, it is preferable to select a subset of variables that were measured on the same time span, though this will likely reduce the number of usable observations. If this is not done, it is possible that variables will be from overlapping or variable time periods, obscuring interpretation of causal relationships. To help ameliorate this problem, sometimes ILD researchers include extra assessment questions that can be used to fill in missing data from another time point. For example, in our example ILD dataset, alcohol use was measured in real time via participant-initiated surveys, but if the participant forgot to initiate a survey, they could report alcohol use the next morning, as well as the precise time range they were drinking (for full study design details, see Carpenter and Merrill (2021)). This allows the analyst to impute missing real-time information for alcohol use the previous day, improving the number of observations that can be used in an analysis of only real-time variables. Using logic from retrospective questions in our dataset, we were able to improve the number of complete real-time drink reports from 73.96% to 99.18%, substantially increasing the number of observations that could be used for each participant. If it is known at the time of study design that idiographic causal discovery analyses are intended to be used, it will greatly facilitate data analysis to design each assessment to measure the variables of interest on comparable time spans, and to include the same variables in every assessment.

Example ILD Dataset. Causal relationships between mood and alcohol use were modeled in a sample of eight individuals who completed an ILD protocol consisting of multiple daily assessments of mood and alcohol use. Several other variables were measured at various time points in this study, but they were not measured at every assessment, producing a very low number of observations per person. Mood and alcohol use were measured multiple times per day and at almost every assessment, therefore producing the highest number of observations per person. Mood states included were happiness, stress, energy, irritability, relaxation, and sadness. Alcohol use was measured as the number of standard drinks (defined as 12 oz beer, 5 oz wine, or 1-1.5 oz liquor) consumed since the last assessment. Lagged variables (i.e., mood and alcohol use at the last assessment) were also added to the model for all variables. We analyzed the eight individual graphs with the most observations in the dataset in greater detail. In a clinical application of causal discovery, these graphs could be used to identify targets for a personalized intervention for alcohol use by examining pathways that lead to alcohol use. For example, subject 1 (see Figure 1) appears to follow a cycle where drinking leads to happiness, which leads to energy, which leads to more drinking. In addition, drinking for this individual predicts more drinking at the next hourly assessment. For this individual, a successful intervention would include alternative strategies for achieving happiness and energy, and strategies for preventing a drinking episode from starting, suggesting that perhaps abstinence would be easier for this individual to maintain than moderation.

How many observations are needed? An average ILD study produces 65-90 assessments per person and recruits an average of 150 participants (Jones et al., 2019). When conducting group-level analyses on ILD, these sample sizes are sufficient to detect even

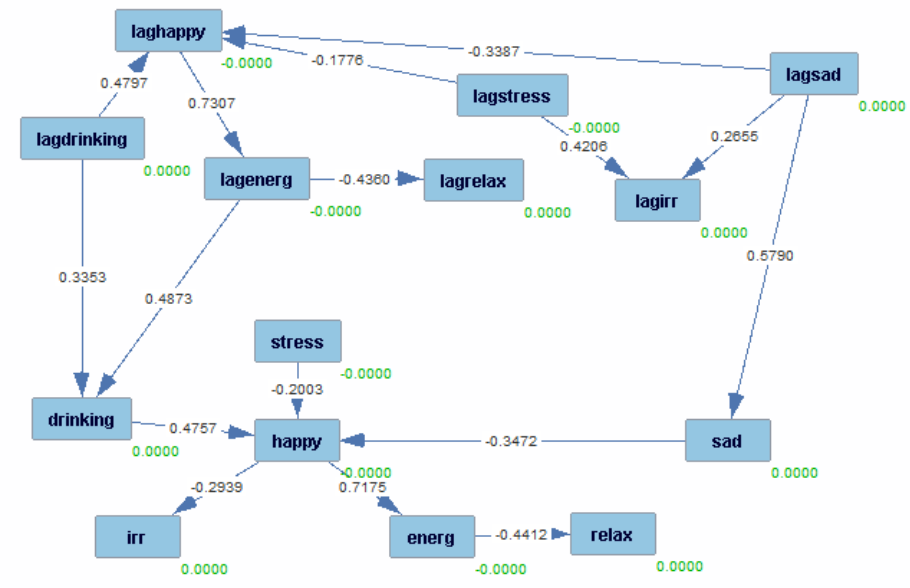

Figure 1: **Subject 1's causal network of mood and alcohol use.** Black numbers are standardized effect sizes for each edge. Green numbers are the intercept for the node (i.e., mean for the variable). Variables were centered at the individual's mean, hence all intercepts = 0.

very small effects. However, when performing idiographic analyses (analyzing each individual's data separately), the number of observations per person is key. It is well documented that in simulation studies, the precision and recall of causal discovery algorithms declines substantially as sample size decreases; for examples, see Kummerfeld and Danks (2013); Kummerfeld and Rix (2019); Biza et al.; Ogarrio et al. (2016). It is less well documented that the same is true when effect sizes change, however, inference methods typically have greater recall and precision when effect sizes are larger. We have tested this process in several individuals from our ILD dataset measuring momentary mood and alcohol, and the discovered graphs included both strong edges (i.e., connections between variables; $r \geq 0.50$) and moderate edges ($r \geq 0.30$). We analyzed eight individual subjects with a range of 30-118 observations per person ($M = 80.13$, $SD = 27.44$). We used greedy fast causal inference (GFCI) (Ogarrio et al., 2016) with default parameters in Tetrad 6.8.0 to generate causal networks of mood (6 variables) and alcohol use (1 variable) at the current assessment and at a time lag ($t - 1$), for a total of 14 variables. Background knowledge indicated that lagged variables cannot be caused by the primary (non-lagged) variables. The discovered graphs found more edges when the number of observations was higher ($r = .66$). On average, 13.5 edges were found, but the number of edges detected was higher and less variable among the four with 90-118 observations ($M = 15.75$, $SD = 1.26$) as compared to the four individuals with 30-77 observations each ($M = 11.50$, $SD = 3.11$). The number of causal edges detected was also higher and less variable for the higher-observation group ($M = 5.75$, $SD = 0.96$) than for the lower-observation group ($M = 3.75$, $SD = 3.50$), though the correlation between number of observations and causal edges detected was weak ($r = .08$).

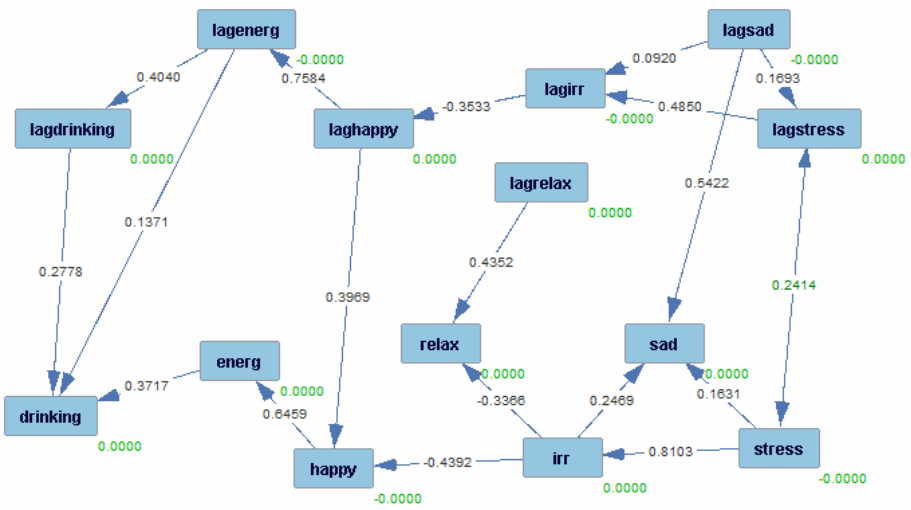

Figure 2: **Group graph of mood and alcohol use.** Black numbers are the standardized effect size for the edge. Green numbers are the intercept for the node. All variables were centered at the group mean, hence each intercept = 0.

Next, we combined the four people with the highest number of observations together for a total of 405 observations and generated a group graph to compare with the individual graphs (See Figure 2). The group graph detected 19 edges, exceeding the average detected in the individual graphs, and fit was good (RMSEA = 0.0990, CFI = 0.9840). This fit was slightly superior to Subject 1's graph (RMSEA = 0.1119, CFI = 0.9690). Each edge that emerged in the group network was checked to see if it emerged in at least one of the individual networks (see Table 1). A total of 5 edges emerged in the group network that were not present in the individual networks (25% of those detected). The group graph omitted 38-50% of the total edges detected in each individual's network, including between 40-60% of all directed edges. Further, between 0 and 19% of edges with directions in the individual networks were reversed in direction in the group network. Of the 15 edges that emerged in both individual and group networks, 11 were consistent in direction across individuals. Interestingly, two edges that emerged in the group network (irritable to sad and irritable to happy) only emerged in the reverse direction in any of the individual networks. These figures raise the question of whether the group network adequately represents each individual in the dataset.

The group graph detected five edges that were not present in any of the individual graphs. To explore the possibility that the group graph was better powered to detect weaker edges, we examined edge strengths detected in individual and group graphs. The average absolute edge strength detected in the group graph was r = 0.38, in subject 1's graph was r = .42, and the average edge strength of an edge that was not detected in the individual graphs but was detected in the group graph was r = .27. At first glance, this would suggest that the group graph was better powered to find weak edges than the individual graph. However, we also tested the stability of the edges detected in the group graph as compared to Subject 1's graph with 200 bootstrapped samples.

Using 200 bootstrap resamples for the group graph and the individual with the highest number of observations (N = 118, Subject 1), we examined the stability of the edges found in each graph. There were six different edge types possible for each pair of variables (directed edge in either direction, semi-directed edge in either direction, undirected edges, and bidirected edges). If an edge emerged as one type consistently in at least 50% of bootstrap resamples, we classified this edge as moderately stable, and highly stable edges emerged 75% or more of the time. For the group graph, two edges (happy $\rightarrow$ energy and energy $\rightarrow$ drinking) emerged as highly stable, six emerged as moderately stable, and eight emerged with low stability, such that an edge emerged between the two variables more than 50% of the time, but the direction and certainty of the edge was variable. For Subject 1's network, three edges emerged as highly stable, eight emerged as moderately stable, and three had low stability. These results show that, although the individual network contains fewer edges, the edges that have emerged have higher stability than those found in a group graph with three times the observations. This suggests that combining multiple individuals' data together introduces more noise than signal. The improved ability to detect weaker edges is offset by the decreased stability in edges detected. Further, at a theoretical level, we found that many edges in the individual graphs were changed or omitted in the group graph, demonstrating that our group graph did not adequately represent any of the four individuals in the dataset.

Table 1: **Comparison of Edges in Combined 4-Person Graph vs. Individual Graphs.** The symbol → indicates a directed relationship from left to right, and – is an undirected relationship. A blank cell indicates that this edge was not detected in that individual's graph. The last three rows refer to edges that were in the individual networks but omitted or changed in the group network.

Edge in Combined Graph	Subject 1	Subject 2	Subject 3	Subject 4
(lag) Energy → (lag) Drinking		–	–	–
(lag) Happy → (lag) Energy	–	–	–	–
(lag) Irritable → (lag) Happy				
(lag) Sad → (lag) Irritable	→		–	→
(lag) Sad → (lag) Stress				
(lag) Stress → (lag) Irritable	→	–	–	→
(lag) Drinking → Drinking	→	→	→	→
(lag) Energy → Drinking	→			
(lag) Happy → Happy				→
(lag) Happy → Sad				
(lag) Relax → Relax				
(lag) Sad → Sad	→			
(lag) Stress - Stress		→	→	
Irritable → Relax			→	
Irritable → Sad			←	←
Stress → Irritable		→	→	→
Stress → Sad				
Irritable → Happy	←			
Happy → Energy	→	→	→	←
Energy → Drinking		→		←
Number of edges omitted	8 (50%)	6 (43%)	7 (41%)	6 (38%)
Number of directed edges omitted	3 (60%)	2 (40%)	3 (50%)	3 (43%)
Number of edges changed directions	1 (6%)	0 (0%)	1 (6%)	3 (19%)

To conclude the discussion of how many observations are needed to conduct an idiographic analysis of ILD, based on the results of our analyses, for edges with moderate effect sizes, it is recommended to have 90 or more observations per individual. Below 90 observations, edge detection decreased noticeably. At 118 observations, stability of edges detected was higher than for the group graph. It is worth noting that most ILD studies produce fewer than 90 observations per person, particularly when observations are restricted to those that measured the same set of constructs. Therefore, it may be difficult to apply causal discovery to ILD studies that did not consider this analytic approach in advance (i.e., during the design of the study).

5. Conclusion

The application of causal discovery to ILD is novel and may help to inform clinical interventions and be particularly useful for uncovering relationships between variables for a given individual, rather than groups. In the current study, we addressed variables that were measured on different timelines by imputing values based on questions asked in other assessments and excluding many variables that were not measured on the same time span. These practices can facilitate causal discovery analyses in ILD datasets. However, when possible, studies intending to use causal discovery analyses should be designed to measure as many variables as possible on the same time span. We also found that 90 or more observations per person is ideal for causal discovery analyses. With this number of observations, individual-level graphs detected edges of moderate strength and edges emerged with higher stability than a graph that combined four people's data together. Results suggest that causal discovery is feasible in ILD, even at the individual level, as long as the number of usable observations per person is high (e.g., 90 or more). In future work, we plan to systematically analyze the performance of causal discovery on larger ILD datasets, and explore the use of causal discovery methods that are more targeted to time-series data types, such as (Malinsky and Spirtes, 2018, 2019).

Acknowledgments

BLS was supported by funding from T32DA037183. EK was supported by funding from Grant No. NCRR 1UL1TR002494-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms.

Niall Bolger and Jean-Philippe Laurenceau. *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press, 2013.

Ryan W Carpenter and Jennifer E Merrill. How much and how fast: Alcohol consumption patterns, drinking-episode affect, and next-day consequences in the daily life of underage heavy drinkers. *Drug and alcohol dependence*, 218:108407, 2021.

Lesa Hoffman. *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge, 2015.

Andrew T Jebb, Louis Tay, Wei Wang, and Qiming Huang. Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology*, 6:727, 2015.

Andrew Jones, Danielle Remmerswaal, Ilse Verveer, Eric Robinson, Ingmar HA Franken, Cheng K Fred Wen, and Matt Field. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. *Addiction*, 114(4):609–619, 2019.

Rogier Kievit, Willem Eduard Frankenhuis, Lourens Waldorp, and Denny Borsboom. Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, 4:513, 2013.

Erich Kummerfeld and David Danks. Tracking time-varying graphical structure. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 1205–1213, 2013.

Erich Kummerfeld and Alexander Rix. Simulations evaluating resampling methods for causal discovery: ensemble performance and calibration. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2586–2593. IEEE, 2019.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47. PMLR, 2018.

Daniel Malinsky and Peter Spirtes. Learning the structure of a nonstationary vector autoregression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2986–2994. PMLR, 2019.

Peter CM Molenaar. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4):201–218, 2004.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379. PMLR, 2016.

Harry T Reis and Shelly L Gable. Event-sampling and other methods for studying everyday experience. 2000.

Cheng K Fred Wen, Stefan Schneider, Arthur A Stone, and Donna Spruijt-Metz. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. *Journal of medical Internet research*, 19(4):e132, 2017.