Computers & Geosciences 99 (2017) 50–60

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Causal discovery in the geosciences—Using synthetic data to learn how to interpret results

Imme Ebert-Uphoff[a,*], Yi Deng[b]

[a] Department of Electrical & Computer Engineering, Colorado State University, Fort Collins, CO, USA
[b] School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Probabilistic graphical model
Structure learning
Causal discovery
Information flow
Geoscience
Climate
Atmospheric science

ABSTRACT

Causal discovery algorithms based on probabilistic graphical models have recently emerged in geoscience applications for the identification and visualization of dynamical processes. The key idea is to learn the structure of a graphical model from observed spatio-temporal data, thus finding pathways of interactions in the observed physical system. Studying those pathways allows geoscientists to learn subtle details about the underlying dynamical mechanisms governing our planet. Initial studies using this approach on real-world atmospheric data have shown great potential for scientific discovery. However, in these initial studies no ground truth was available, so that the resulting graphs have been evaluated only by whether a domain expert thinks they seemed physically plausible. The lack of ground truth is a typical problem when using causal discovery in the geosciences. Furthermore, while most of the connections found by this method match domain knowledge, we encountered one type of connection for which no explanation was found. To address both of these issues we developed a simulation framework that generates synthetic data of typical atmospheric processes (advection and diffusion). Applying the causal discovery algorithm to the *synthetic* data allowed us (1) to develop a better understanding of how these physical processes appear in the resulting connectivity graphs, and thus how to better interpret such connectivity graphs when obtained from real-world data; (2) to solve the mystery of the previously unexplained connections.

1. Introduction

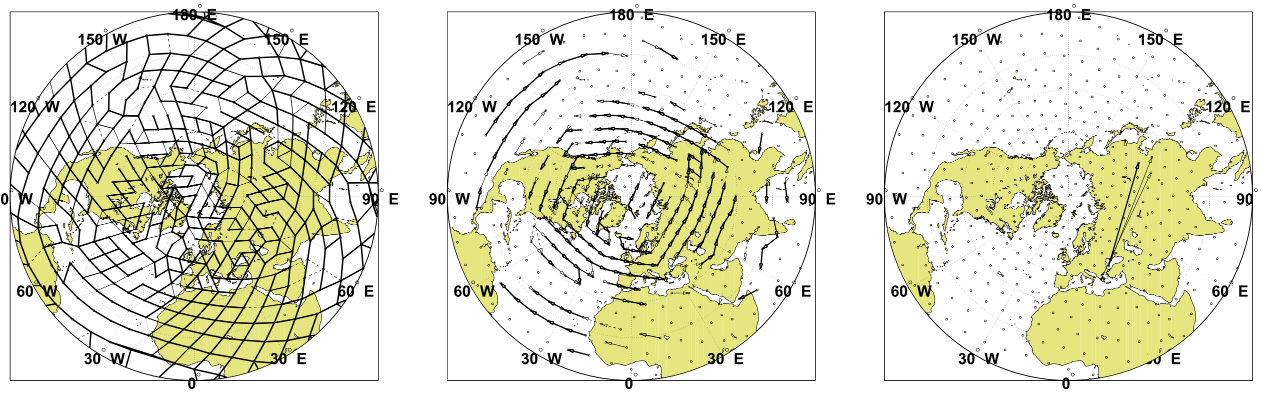
Recent research has shown great potential for causal discovery algorithms to track information flow from observed data for geoscience applications. The key idea for tracking information flow in geoscience is to interpret large-scale dynamical processes as information flow and to identify the pathways of this information flow by learning graphical models from observational data. Since probabilistic graphical models are based on information-theoretical measures, they provide an ideal tool to track such information flow. We have obtained very promising results by applying constraint-based structure learning of probabilistic graphical models to real-world atmospheric data. For example, we compared information flow in two case studies, (1) boreal winter vs. summer (Ebert-Uphoff and Deng, 2012) and (2) current climate vs. projected climate in 100 years under global warming (Deng and Ebert-Uphoff, 2014), that provided new insights into the change of large-scale dynamics for these cases. (Obviously, the latter comparison is based on data generated by climate models, in addition to observed data).

One challenge of using causal discovery in climate science (and many other geoscience applications) is that there is never any exact ground truth available in climate data,[1] i.e. the only way to evaluate the results we obtained was to have the domain expert (second author of this paper) visually inspect the resulting graphs of information flow and consider whether they *seemed physically plausible* given the current knowledge in climate science about interactions in the atmosphere. While this evaluation confirmed the potential of this new methodology, it leaves much to be desired. In particular, we did not have the tools to evaluate the accuracy of the method or to know how *exactly* to interpret the resulting networks. The lack of ground truth is a typical problem when using causal discovery in the geosciences, simply because the earth is too complex a system and not all connections are known—which is precisely the reason why we want to apply causal discovery in the first place, but it is also a major challenge when evaluating and interpreting the results, as illustrated in the following

* Corresponding author.
E-mail addresses: iebert@engr.colostate.edu (I. Ebert-Uphoff), yi.deng@eas.gatech.edu (Y. Deng).

[1] Even when using the output from climate models, we do not have information on the large-scale dynamics, since the climate models utilize numerical equations localized in both space and time, i.e. expressing the state of the system for each location at the next time step based on that at the previous time step. These equations themselves thus do not provide explicit information on the large-scale interactions occurring in the climate system.

http://dx.doi.org/10.1016/j.cageo.2016.10.008
Received 20 May 2016; Accepted 24 October 2016
Available online 02 November 2016



(a) $T = 0$ days (b) $T = 1$ day (c) $T = 2$ days

Fig. 1. Connectivity plots of interactions in atmosphere in Northern hemisphere based on observed geopotential height data (using PC stable, D=1 d, $\alpha = 0.1$ and 800-point Fekete grid). (a)–(c) show connections found for 0, 1 and 2 days, respectively, from potential cause to potential effect.

section.

1.1. Unexplained concurrent edges in connectivity graphs

Causal discovery methods applied to observed data can never be used to *prove* causal connections between observed variables—mainly due to the potential existence of hidden common causes (aka latent variables)—, but only to *disprove* causal connections. Building on this fact, the causal discovery algorithm applied here is an elimination procedure that first assumes that any two variables can be in a cause–effect relationship, then disproves many of those relationships using conditional independence tests applied to the observed data. An important implication is that the results obtained by this approach only indicate *potential* cause–effect relationships. Thus, when applied to geoscience applications, we must always perform an evaluation step at the end of the analysis. Namely, a geoscience expert must check every link (or group of links) in the final graph. If there is a known physical mechanism that explains the link, the link is accepted as a causal interaction. Otherwise, it provides a new causal hypothesis to be studied further.

When applying this evaluation step in our analysis of connections in the atmosphere, we found that many edges were easily explained by physical means, but we often encountered one type of edge that eluded any physical interpretation, namely a spiderweb-like pattern of apparently instantaneous (or high-speed) connections between neighboring points (see Ebert-Uphoff and Deng, 2012 for the first documented occurrence). Fig. 1 provides an example of this type of unexplained edges. Fig. 1 shows interactions found in the atmosphere, based on observed daily geopotential height data, using causal discovery. The interactions shown in Figs. 1(b) and (c) are easily explained, as they represent interactions in the atmosphere due to storm tracks. However, the spiderweb-like pattern of connections in Fig. 1(a) indicates that most neighboring grid points have an instantaneous (i.e. extremely fast) interaction between them, which does not match physical observation, as no such strong and consistent motion exists, especially near the equator. Repeated simulations with similar data have shown similar patterns of unexplained connections, while all non-instantaneous connections (such as the ones in Fig. 1(b,c)) found are physically meaningful. Over the years we have increased the computational efficiency of our algorithm, thus being able to increase grid resolution, and found that with increasing resolution the number of these unexplained connections increases further. The reason for their existence—and any potential physical interpretation—remained a mystery for the past three years that we wanted to resolve.

1.2. Using synthetic data

Lack of ground truth presented a similar challenge, until recently, for a different type of network learned from climate data, namely *complex networks*. Complex networks, also known as *climate networks*, were first proposed by Tsonis and Roebber (2004) and are a much simpler concept, exclusively based on Pearson correlation. Namely, any two nodes are connected if and only if the Pearson-correlation of the corresponding data is above a chosen threshold. (Note that the purpose of complex networks in geoscience applications is to identify *similarities* between different locations, while the purpose of the causal discovery networks discussed here is to identify *interaction pathways* between different locations—a distinctly different purpose.) Complex networks have been applied to climate data for over a decade (Tsonis and Roebber, 2004; Tsonis et al., 2006, 2008; Yamasaki et al., 2008; Donges et al., 2009; Steinhaeuser et al., 2010), and many insights have been drawn from them over the years, but they had never actually been tested on simulation data until very recently. Molkenthin et al. (2014) finally filled this gap by testing complex networks on simulated data developed for that purpose and then comparing the results to the known physics of the simulation data.

Here we seek to achieve the same goal for connectivity graphs generated by causal discovery algorithms. For this purpose we developed a simulation framework, similar to the one by Molkenthin et al. (2014), that models the two most important dynamical processes in the atmosphere, diffusion and advection. These processes are also dominant in many other geoscience applications, thus allowing us to generate synthetic data sets for a great variety of different conditions and for which the exact dynamics are known. This allows us (1) to develop a better understanding of how these physical processes appear in the connectivity graphs generated by the causal discovery algorithm, and thus to better interpret connectivity graphs obtained from real-world data; (2) to resolve the mystery of the previously unexplained spiderweb connections identified from atmospheric data.

We make all of the synthetic data sets discussed here (along with results from our causal discovery approach) available to the community as benchmarks to apply other types of causal discovery algorithms.[2]

1.3. Organization of this article

The remainder of this article is organized as follows. Section 2 briefly describes the causal discovery algorithm used, sample applications, and the testbed used to generate synthetic data. Section 3

[2] See URL http://www.engr.colostate.edu/~iebert/DATA_SETS_CAUSAL_DISCOVERY/