Identification of Time-Dependent Causal Model: A Gaussian Process Treatment

Biwei Huang[1] **Kun Zhang**[1,2] **Bernhard Schölkopf**[1]
[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Information Sciences Institute, University of Southern California
{biwei.huang, kun.zhang, bernhard.schoelkopf}@tuebingen.mpg.de

Abstract

Most approaches to causal discovery assume a fixed (or time-invariant) causal model; however, in practical situations, especially in neuroscience and economics, causal relations might be time-dependent for various reasons. This paper aims to identify the time-dependent causal relations from observational data. We consider general formulations for time-varying causal modeling on stochastic processes, which can also capture the causal influence from a certain type of unobserved confounders. We focus on two issues: one is whether such a causal model, including the causal direction, is identifiable from observational data; the other is how to estimate such a model in a principled way. We show that under appropriate assumptions, the causal structure is identifiable according to our formulated model. We then propose a principled way for its estimation by extending Gaussian Process regression, which enables an automatic way to learn how the causal model changes over time. Experimental results on both artificial and real data demonstrate the practical usefulness of time-dependent causal modeling and the effectiveness of the proposed approach for estimation.

Introduction

In this paper we are concerned with the problem of causal discovery, i.e., how to discover causal relations from purely observational data. Traditionally, it has been noted that under appropriate assumptions, one could recover an equivalence class of the underlying causal structure based on conditional independence relationships of the variables [Pearl, 2000; Spirtes *et al.*, 2000]. In contrast, functional causal models provide a useful tool to model causal relationships [Pearl, 2000]; recently, it has been shown that with appropriately restricted functional causal models [Shimizu *et al.*, 2006; Hoyer *et al.*, 2009; Zhang and Hyvärinen, 2009b; Peters *et al.*, 2013], it is possible to identify the causal structure from purely observational data. These restricted causal models either assume that the analyzed data is in equilibrium states or assume that the causal model is time-invariant.

However, in many practical situations, especially in neuroscience, economics and climate analysis, the causal relations may change over time; if one still uses a fixed causal model, the discovered causal relations might be misleading. Let us consider the following situations.

- The causal sufficiency assumption, which states that there is no unobservable common cause of any two observed variables, holds, but causal effects (e.g., causal strength or involved parameters) change over time.
- The causal sufficiency assumption is not satisfied – there is an unobserved confounder whose influence to the observed processes changes over time. Ignoring this effect will make the causal relations between the observed processes appear to change over time.

The above situations may even happen at the same time. A conceivable example is the causal interactions between different brain areas that change greatly during different tasks and states [Shafi *et al.*, 2012; Zhao *et al.*, 2013]; thus, we see that the inside causal influences change over time. Here we are interested in how they vary along with time, and how the outside environment, which is not directly measurable, influences the inside activities, if it does so. In this paper, we focus on the case where the causal influences between observed processes change smoothly over time and the influence from unobserved confounders, if it exists, can be approximated as a smooth function of time.

To account for the issues above, we present a novel approach to modeling such time-dependent causal influences. We show that by introducing time information as a common cause for the observed processes, we can model the time-varying causal influences between the observed processes, as well as the influence from a certain type of unobserved confounders. We propose a linear time-dependent causal model and a nonlinear one with additive noise. Two questions then naturally arise: 1) Is such a causal model, including the causal direction, identifiable? 2) If it is, how can we estimate it accurately from data? Regarding the former question, we show that under mild assumptions, the causal model is identifiable.

In order to identify the time-dependent causal model, one needs to fit candidate models and then do model checking. Existing methods for estimating the time-varying models mainly use adaptive filters or sliding windows [Karlsson *et al.*, 2013; Barrett *et al.*, 2012]. These methods might lead

to large estimation errors, especially when the causal influence varies quickly over time. On the other hand, Gaussian Process (GP) regression provides a promising non-parametric Bayesian approach to regression problems [Rasmussen and Williams, 2006]. It not only enables the distributions of various quantities to be calculated explicitly, but also brings a convenient way to infer model hyperparameters such as those that control the kernel shape and noise level [Chu and Ghahramani, 2005]. We will show that GP can be used as a prior to automatically capture the smoothness level of the time-varying causal influence; as a consequence, the time-dependent causal model can be estimated in a non-parametric manner. After this step, we evaluate whether a certain causal model is valid or not by testing for the independence between the estimated noise terms and the corresponding hypothetical causes by Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2007].

Our main contribution in this paper is two-fold. *1*. We formulate time-dependent causal models to account for the time-varying causal influences between the observed processes, and/or a specific type of unobservable confounders whose influence on the observed processes can be approximated as a function smooth in time. Based on the proposed models, on the theoretical side, we discuss the identifiability of the causal structure. *2*. On the practical side, we propose a non-parametric way to estimate the time-dependent causal influences, with the model complexity automatically inferred from data.

Model Definition

By including the time information T as a special variable (a common cause), we extend the functional causal model to solve causal discovery problems in more general cases; see Figure 1. In particular, it can represent time-varying causal influences between the observed processes and the influence from unobserved confounders that is approximately a smooth function of time. Below we formulate both a linear time-dependent function causal model, and a nonlinear one with additive noise.

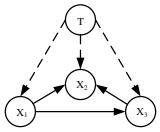

Figure 1: Causal graph $\mathcal{G}(V, E)$ ($V = \{x_1, x_2, x_3, T\}$) of the time-dependent causal model, where time information T can be considered as a common cause to the other observed variables. Dashed lines represent influences from T, and we are mostly interested in the identifiability of the causal structure between x_1, x_2 and x_3 represented by solid lines.

Preliminary

Before introducing our time-dependent functional causal model, we first briefly review ordinary functional causal models [Pearl, 2000].

In its general form, a functional causal model consists of a set of equations of the form

$$x_i = f_i(\boldsymbol{pa}_i, u_i), \qquad i = 1, \cdots, N \tag{1}$$

where $\{x_i\}$ is the set of variables in a Directed Acyclic Graph, $\boldsymbol{pa}_i$ is the set of direct causes of variable x_i, u_i represents the disturbance term due to omitted factors, and the disturbance terms are independent of each other. Each of the functions f_i represents a causal mechanism that determines the value of x_i from the causes and noise terms on the right side.

In practice, we restrict the function form of f_i to make the causal structure identifiable [Shimizu *et al.*, 2006; Hoyer *et al.*, 2009; Zhang and Hyvärinen, 2009b].

- Linear causal model:
$$x_i = \sum_j a_{i,j} \cdot \boldsymbol{pa}_i^j + u_i,$$
where $\boldsymbol{pa}_i^j$ is the jth variable of $\boldsymbol{pa}_i$, and $a_{i,j}$ is the linear causal coefficients. If $(\boldsymbol{pa}_i, u_i)$ is not jointly Gaussian distributed, the linear causal model is identifiable.
- Nonlinear additive noise model:
$$x_i = f_i(\boldsymbol{pa}_i) + u_i.$$
It has been shown to be identifiable except when f_i is linear and $(\boldsymbol{pa}_i, u_i)$ is jointly Gaussian, as well as a few "non-generic" cases.
- Post-nonlinear causal model:
$$x_i = g_i\big(f_i(\boldsymbol{pa}_i) + u_i\big),$$
where g_i denotes an invertible post-nonlinear distortion. It has been shown to be identifiable except when both g_i and f_i are linear and $(\boldsymbol{pa}_i, u_i)$ is jointly Gaussian and a few "non-generic" cases. All the non-identifiable cases have been listed in [Zhang and Hyvärinen, 2009b].

The restricted functional causal model has also been applied to the time series by taking into account the temporal constraints that the effect cannot precede the cause [Zhang and Hyvärinen, 2009a]. In particular, [Peters *et al.*, 2013] considered the nonlinear causal relations,

$$x_i(t) = f_i\big(\boldsymbol{pa}_i^P(t-P), \cdots, \boldsymbol{pa}_i^1(t-1), \boldsymbol{pa}_i^0(t)\big) + u_i,$$

where $\boldsymbol{pa}_i^p(t-p)$ denote the p time-lagged direct causes of variable x_i. Note that here it is assumed that all causal relations are fixed.

Linear Time-Dependent Causal Model

In order to capture time-varying causal relations and confounder influences which are smooth in T explicitly, we first formulate a linear time-dependent functional causal model. We assume that a multivariate time series $\mathbf{x}(t) = \big(x_1(t), x_2(t), \cdots, x_N(t)\big)^\intercal$ with finite dimensionality N has the following data generating process,

$$x_i(t) = \underbrace{\sum_{j=1}^{N}\sum_{p=1}^{P} a_{i,j,p}(t)x_j(t-p)}_{\text{lagged terms}} + \underbrace{\sum_{k\neq i} b_{i,k}(t)x_k(t)}_{\text{instantaneous terms}} + \underbrace{g_i(t)}_{\text{confounder term}} + \varepsilon_i(t), \tag{2}$$

where $\varepsilon_i(t)$ is i.i.d. (independent and identically distributed) noise and independent of the causes $x_j(t-p)$, $x_k(t)$ and t, $a_{i,j,p}(t)$ represent the time-varying lagged causal coefficients, $b_{i,k}(t)$ give the instantaneous causal coefficients which are essential especially for low time resolution data, and $g_i(t)$ represent the causal influences from unobserved confounders that are assumed to be functions smooth in time.

This formulation includes lagged influences, instantaneous effects, and influences from a certain type of unobserved confounders. In practice it might suffice to consider its special cases, e.g., drop the instantaneous terms if one believes that they do not exist.

Nonlinear Time-Dependent Causal Model

In practice the relations among the observed processes are usually nonlinear. For a nice trade-off between the identifiability and generality of the causal model, we define a time-dependent nonlinear model with additive noise:

$$x_i(t) = f_i(t, \{x_j(t-p)\}_j, \{x_k(t)\}_{k\neq i}) + \varepsilon_i(t), \qquad (3)$$

where $\varepsilon_i(t)$ is i.i.d. noise, and is independent of the causes $x_j(t-p)$, $x_k(t)$ and t. The argument t inside the nonlinear function f_i explains both time-varying causal relations and the confounder influence.

Discussion on the Identifiability of Causal Structures

In this part, we discuss the identifiability of causal structure implied by the time-dependent functional causal models. Identifiability implies that the causal model is asymmetric in causes and effects and is capable of distinguishing between them. More specifically, for the correct causal direction, the noise is independent of the hypothetical causes, as assumed in the model, but not for the backward direction.

From equation 2, we see that if the time-varying coefficients and the influences from unobserved confounders can be represented as functions of time, the time information can be seen as a cause to variable x_i, and the same in equation (3). Therefore, the variable T can be viewed as an additional argument of the causal model, and it is represented as a common cause in the causal graph (Figure 1).

The time-dependent nonlinear model (3) can be seen as a nonlinear additive noise model (ANM) on the variable set $\{x_i\} \cup \{T\}$. The linear one, (2), is actually a constrained version of ANM where variable T is also included. Assuming the ANM, it has been shown that the causal structure, including the causal direction, is identifiable in the bivariate case under mild assumptions on the nonlinear functions and data distributions [Hoyer *et al.*, 2009; Zhang and Hyvärinen, 2009b]; in the sense that for the backward direction, the noise term is not independent from the hypothetical cause. This identifiability result has been further extended to the multivariate case [Peters *et al.*, 2011].

This directly implies that based on the formulated time-dependent nonlinear causal model (3), the causal structure is identifiable under mild assumptions (for details of the assumptions, see [Hoyer *et al.*, 2009; Zhang and Hyvärinen, 2009b; Peters *et al.*, 2011]). Furthermore, the functional class of the linear time-dependent causal model (2) is contained by that of the nonlinear one (3); under those assumptions, given any function in the latter class, the noise is not independent from the hypothetical causes for the backward direction, implying that this is also the case given any function in the former class. In other words, generally speaking, the causal structure is identifiable under the linear time-dependent causal model (2).

Model Estimation

In this section, we propose a non-parametric method to estimate the time-dependent causal models, where we use GP [Rasmussen and Williams, 2006] as a prior to capture the smoothness level of the time-varying causal influence. With certain tricks, the estimation procedure for both the linear time-dependent causal model and the nonlinear one can be formulated as specific GP regression problems. Below we mainly focus on the estimation of the linear model, and only briefly mention the estimation procedure for the nonlinear model in the end, since they are similar.

In matrix form, the linear model (2) can be written as

$$\begin{aligned}
&\big(I - B(t)\big)\mathbf{x}(t) = \sum_{p=1}^{P} A_p(t)\mathbf{x}(t-p) + G(t) + \varepsilon(t), \text{ or,}\\
&\mathbf{x}(t) = \underbrace{\sum_{p=1}^{P} \big(I - B(t)\big)^{-1} A_p(t)}_{A'_p(t)}\,\mathbf{x}(t-p) + \\
&\qquad \underbrace{\big(I - B(t)\big)^{-1} G(t)}_{G'(t)} + \underbrace{\big(I - B(t)\big)^{-1}\varepsilon(t)}_{\varepsilon'(t)}, \qquad (4)
\end{aligned}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_N(t))^\intercal$, I the $N \times N$ identity matrix, $B(t)$ the matrix with entries $b_{i,k}(t)$, which could be permuted to be strict lower triangularity to imply instantaneous causal relations, $A_p(t)$ the matrix with entries $a_{i,j,p}(t)$, $G(t)$ the vector of $g_i(t)$, and $\varepsilon(t)$ the vector of $\varepsilon_i(t)$.

This inspires a computationally efficient two-step procedure to estimate $a_{i,j,p}(t)$, $b_{i,k}(t)$, $g_i(t)$, and the noise term $\varepsilon_i(t)$, by extending the procedure in [Hyvärinen *et al.*, 2010]: we first estimate $A'_p(t)$, $G'(t)$, and $\varepsilon'(t)$ in the model (4), which does not have instantaneous effects; after that, we estimate $B(t)$ and then $A_p(t)$, $G(t)$, and $\varepsilon(t)$ by making use of the relationship between $\varepsilon'(t)$ and $\varepsilon(t)$.

Step 1: We first only consider the lagged terms and the confounder terms, but no instantaneous terms. For convenience, we consider each row in equation (4), i.e.,

$$x_i(t) = \underbrace{\sum_{j=1}^{N}\sum_{p=1}^{P} a'_{i,j,p}(t)x_j(t-p)}_{\text{lagged terms}} + \underbrace{g'_i(t)}_{\text{confounder term}} + \varepsilon'_i(t), \qquad (5)$$

separately. To use the GP prior, we collect all data points and represent equation (5) in matrix notation:

$$\mathbf{y} = \mathbf{D}_{\mathcal{X}} \cdot \mathbf{f} + \boldsymbol{\varepsilon}', \qquad (6)$$

where

$$\mathbf{y} = (\mathbf{x}^\intercal(P+1), \cdots, \mathbf{x}^\intercal(T))^\intercal,$$

$$\mathbf{D}_{\mathcal{X}} = \begin{pmatrix} \tilde{\mathcal{X}}_1 & 0 & \dots & 0 \\ 0 & \tilde{\mathcal{X}}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{\mathcal{X}}_{T-P} \end{pmatrix},$$

$$\mathcal{X} = \begin{pmatrix} \mathbf{x}^\intercal(1) & \dots & \mathbf{x}^\intercal(P) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}^\intercal(T-P) & \dots & \mathbf{x}^\intercal(T-1) & 1 \end{pmatrix},$$

$$\tilde{\mathcal{X}}_i = (\mathcal{X})_{i\cdot} \otimes I,$$

$$\mathbf{f} = [\{a'_{i,j,p}(t), g'_i(t)\}_{i,j,p,t}]^\intercal,$$

$$\boldsymbol{\varepsilon}' = [\{\varepsilon'_i(t)\}_{i,t}]^\intercal.$$

Here $\otimes$ denotes kronecker product, and the entries in vectors $\mathbf{f}$ and $\boldsymbol{\varepsilon}$ have been aligned according to $\mathbf{D}_{\mathcal{X}}$.

We put the GP prior on each time-varying coefficient and confounder term to describe their uncertainty:

$$\begin{aligned} a'_{i,j,p}(\boldsymbol{t}) &\sim GP(\mu_{i,j,p}(\boldsymbol{t}), K_{i,j,p}(\boldsymbol{t},\boldsymbol{t})), \\ g'_i(\boldsymbol{t}) &\sim GP(\mu_i(\boldsymbol{t}), K_i(\boldsymbol{t},\boldsymbol{t})), \end{aligned} \tag{7}$$

where $\mu_\cdot$ and $K_\cdot$ (with appropriate subscripts) denote the corresponding mean and covariance in GP (we use a zero mean and squared exponential covariance function), and $\boldsymbol{t}$ is the vector of collected time points. We have assumed that the priors for $a'_{i,j,p}$ and g'_i are independent of each other for different i, j, p. Then we represent $\mathbf{f}$ as

$$\mathbf{f}(\boldsymbol{t}) \sim GP(\boldsymbol{\mu}(\boldsymbol{t}), \mathbf{K}(\boldsymbol{t},\boldsymbol{t})), \tag{8}$$

where $\boldsymbol{\mu}(\mathbf{t}) = \{\mu_{i,j,p}(\mathbf{t}), \mu_i(\mathbf{t})\}_{i,j,p}$, and $\mathbf{K}(\boldsymbol{t},\boldsymbol{t})$ is a block-diagonal matrix, with $\{K_{i,j,p}(\boldsymbol{t},\boldsymbol{t})\}_{i,j,p}$ and $\{K_i(\boldsymbol{t},\boldsymbol{t})\}_i$ on its diagonal and aligned according to $\mathbf{f}(\boldsymbol{t})$.

To simplify the estimation procedure, we assume that the noise is i.i.d. Gaussian random variable, $\varepsilon_i(t) \sim N(0, \sigma^2)$, and consequently we can derive various quantities explicitly. With the GP priors and assumed Gaussian noise, the marginal likelihood of the observations can be represented as a multivariate Gaussian distribution $\mathcal{N}(\mathbf{y}; \mathbf{m}, \boldsymbol{\Sigma})$, with $\mathbf{m} = \mathbf{D}_{\mathcal{X}}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{D}_{\mathcal{X}}\mathbf{K}\mathbf{D}_{\mathcal{X}}{}^{\mathrm{T}} + \sigma^2 I$.

We maximize the marginal likelihood to learn the hyperparameters in the mean functions, covariance functions of GP, and the variance σ^2 of the noise. Then by Bayes' Theorem, we can derive the posterior distribution of $\mathbf{f}$, which also follows a Gaussian distribution with mean

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + (\mathbf{D}_{\mathcal{X}}\mathbf{K})^{\mathrm{T}}[\mathbf{D}_{\mathcal{X}}\mathbf{K}\mathbf{D}_{\mathcal{X}}{}^{\mathrm{T}} + \sigma^2 I]^{-1}(\mathbf{y} - \mathbf{m}). \tag{9}$$

The posterior mean here gives the estimated lagged coefficients $\{\hat{a}'_{i,j,p}(t)\}_{i,j,p,t}$ and confounder terms $\{\hat{g}'_i(t)\}_{i,t}$.

In practice, specific knowledge of the physical system may imply that some coefficients vary at a similar level of smoothness, that is, with the same hyperparameters on those coefficients. This will greatly decrease the number of hyperparameters to be learned.

Under the assumption of *a priori* independence among $a'_{i,j,p}(\boldsymbol{t})$ and $g'_i(\boldsymbol{t})$ for different i, j, and p and the assumption that the coefficients share the same hyperparameters (e.g., the kernel width), we can make the calculation computationally more efficient. We can do matrix operations of product $\mathbf{D}_{\mathcal{X}}\mathbf{K}\mathbf{D}_{\mathcal{X}}^{\mathrm{T}}$ and inversion $[\mathbf{D}_{\mathcal{X}}\mathbf{K}\mathbf{D}_{\mathcal{X}}^{\mathrm{T}} + \sigma^2 I]^{-1}$ in equation (9) in terms of a $T \times T$ matrix, and then go back to the original space by making use of kronecker product, instead of operating on a $N(NP+1)T \times N(NP+1)T$ kernel matrix. This greatly improves the efficiency of calculation.

Step 2: We then consider the instantaneous terms if necessary. For low time resolution data, it is important to consider the instantaneous causal influence, and [Hyvärinen *et al.*, 2010] demonstrated that neglecting it might lead to misleading interpretations of causal relations.

Suppose that we are given a candidate causal ordering for the instantaneous causal effects, denoted by $\mathcal{O}$; we aim to estimate the causal model following this instantaneous causal ordering and test whether it is plausible. Denote by $\hat{\varepsilon}'(t) = \big(\hat{\varepsilon}'_1(t), \hat{\varepsilon}'_2(t), \cdots, \hat{\varepsilon}'_N(t)\big)^\intercal$ the estimated noise from the first step; we then do the following.

1. The relationship between $\varepsilon'(t)$ and $\varepsilon(t)$ in (4) gives $\big(I - B(t)\big)\varepsilon'(t) = \varepsilon(t)$, or
$$\varepsilon'_i(t) = \sum_{k \in \mathbf{pa}_i} b_{i,k}(t)\varepsilon'_k(t) + \varepsilon_i(t), \tag{10}$$
where $\mathbf{pa}_i$ denotes the set of instantaneous causes of $x_i(t)$ according to $\mathcal{O}$. To estimate $b_{i,k}(t)$ and $\varepsilon_i(t)$, we use the estimated values $\hat{\varepsilon}'(t)$ as $\varepsilon'(t)$ into the above equation. We also put a GP prior on $b_{i,k}$, i.e., $b_{i,k}(\boldsymbol{t}) \sim GP(\mu_{i,k}(\boldsymbol{t}), K_{i,k}(\boldsymbol{t},\boldsymbol{t}))$, and estimate the hyperparameters by maximizing the marginal likelihood. Here we also assume *a priori* independence among $b_{i,k}(\boldsymbol{t})$ for different i, k.
2. Then the estimated lagged causal coefficients and confounder influence from the first step can be adjusted as:
$$\hat{A}_p(t) = (I - \hat{B}(t))\hat{A}'_p(t), \quad \hat{G}(t) = (I - \hat{B}(t))\hat{G}'(t). \tag{11}$$
3. We finally evaluate whether the causal model corresponding to $\mathcal{O}$ is valid or not by testing for the independence between the estimated noise terms and corresponding hypothetical causes by HSIC; recall that in principle, the independence condition is valid for the right causal direction, but not for the wrong directions.

To find a plausible instantaneous causal ordering, one can apply Step 2 on all candidate orderings and choose the one with independent noise terms. (Sometimes one may find multiple orderings with independent noise, because of the non-identifiable cases, although very rare, and the finite sample size effect.)

Alternatively, we can estimate both the lagged terms and instantaneous terms in a single step. This, however, is computationally more demanding.

Note that given an instantaneous causal ordering, the nonlinear time-dependent model (3) is actually a nonlinear regression model. A similar procedure can be developed to estimate the model corresponding to the given causal ordering and test whether it is plausible. In the nonlinear case, we have to estimate all involved quantities in a single step.

Experimental Results

We have applied the proposed approach to time-dependent causal modeling to a variety of simulated and real data.

Simulations

Simulation 1 *(With Both Instantaneous and Lagged Causal Influence)*

We generated 1500 data points from the following equation set which includes both lagged and instantaneous causal influence, and has smoothly changing coefficients:

$$\begin{cases} x_1(t) = a_{1,1,1}(t)x_1(t-1) + a_{1,2,1}(t)x_2(t-1) + \varepsilon_1(t), \\ x_2(t) = a_{2,1,1}(t)x_1(t-1) + a_{2,2,1}(t)x_2(t-1) \\ \qquad + b_{2,1}(t)x_1(t) + \varepsilon_2(t), \end{cases}$$

where the coefficients have sinusoidal shapes: $a_{1,1,1}(t) = 0.3(\sin(c \cdot t) + 1.1)$, $a_{1,2,1}(t) = 0.2\cos(c \cdot t) + 0.05$, $a_{2,1,1}(t) = 0.2\sin(c \cdot t) + 0.1$, $a_{2,2,1}(t) = 0.5(\cos(c \cdot t) + 0.05)$, $b_{2,1}(t) = 0.2\cos(c \cdot t)$, $\varepsilon_i(t) \sim N(0, \sigma^2)$ with $\sigma = 0.1$ for $i = \{1, 2\}$.

We used the proposed linear time-dependent functional causal model to fit the data. For the order of the time lag P, we estimated it by choosing the one minimizing the cross-validated prediction error.

Figure 2 shows the estimated time-varying coefficients of our model when $c = 1$, given by our approach and the window-based method from [Karlsson *et al.*, 2013] with window lengths 50 and 100. Figure 3(A) shows the mean squared error (MSE) in the estimated coefficients along with the changing smoothness level c of the coefficients. Figure 3(B) shows the 10-step prediction error for $x_2(t)$. From them one can see that the proposed method produces much more accurate estimates of the coefficients and predictions of the time series. This verifies the usefulness of the automatically learned GP prior. For comparison, we also fitted a static structural vector auto-regressive model on the data; the causal link from $x_2(t-1)$ to $x_1(t)$ is then missing, because the estimate of the corresponding coefficient is not significant at level 0.05 according to the Wald test.

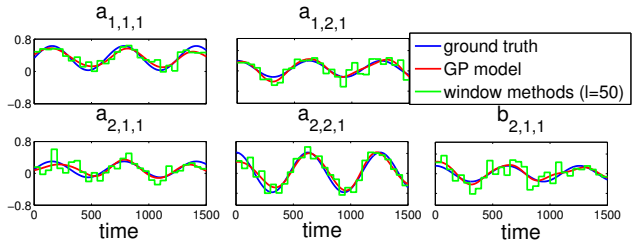

Figure 2: Estimated time-varying coefficients from Simulation 1. The mean squared error in the estimated coefficients by our GP linear time-dependent causal model, and the window methods with window length $l = 50$ and $l = 100$ are 0.0243, 0.0738, 0.1070, respectively. Here for clarity, we do not show the results for $l = 100$.

Simulation 2 *(With Confounders)*

We generated 2000 data points from the following equation set:

$$\begin{cases} x_1(t) = a_{1,1,1}(t)x_1(t-1) + g_1(t) + \varepsilon_1(t), \\ x_2(t) = a_{2,2,1}(t)x_2(t-1) + g_2(t) + \varepsilon_2(t), \end{cases}$$

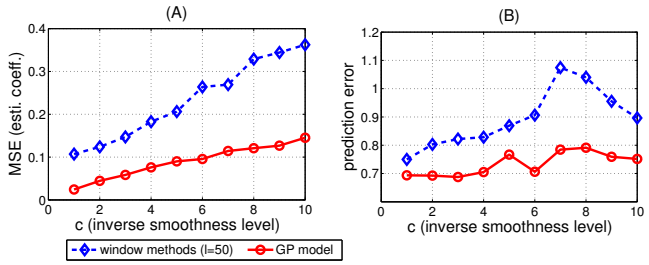

Figure 3: Mean squared error along with the changing inverse smoothness level c of the coefficients. (A) MSE of the estimated coefficients; (B) 10-step prediction error for $x_2(t)$.

where $a_{1,1,1}(t) = 0.5\sin(t)$, $a_{2,2,1}(t) = 0.35(\cos(t) + 1)$, $g_1(t) = 0.3z(t-1)$, $g_2(t) = 0.2z(t)$, $z(t) = \cos(2t + 0.5)$, $\varepsilon_i(t) \sim N(0, \sigma^2)$ with $\sigma = 0.1$ for $i = \{1, 2\}$.

Figure 4 shows the estimated causal coefficients and confounder influence under the linear time-dependent causal model. For simplification, we assumed that every coefficient and confounder term share the same hyperparameters; as a consequence, the estimated $a_{1,2,1}$ and $a_{2,1,1}$ fluctuate around zero slightly, while their true value should be zero. In order to test whether they are significant zero, we compared the cross-validated prediction error with other three cases: (1) fix $a_{1,2,1} = 0$, (2) fix $a_{2,1,1} = 0$, (3) fix $a_{1,2,1} = 0$ and $a_{2,1,1} = 0$. The results indicate that there is no time-delayed causal relation between x_1 and x_2, which is consistent with the ground truth.

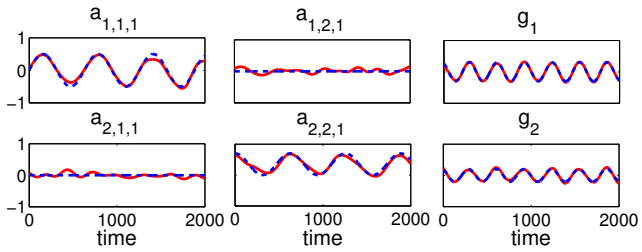

Figure 4: Estimated causal coefficients and confounder influence in Simulation 2. The blue dashed lines indicate the ground truth, and the red solid lines show our estimation.

Simulation 3 *(Linear Instantaneous Model with Various Types of Coefficients)*

We tried to estimate causal coefficients which are generated by different types of functions.

Here we assumed a simple model, which only includes instantaneous causal relations,

$$x_2(t) = b(t)x_1(t) + \varepsilon(t),$$

and we changed different types of functions for causal coefficients $b(t)$, e.g. Laplace functions, polynomial functions, step functions and noisy square-waves. Figure 5 shows the underlying coefficients, as well the estimated results with our GP linear time-dependent causal model. We can see that our method derives accurate estimations on different types of functions for causal coefficients.

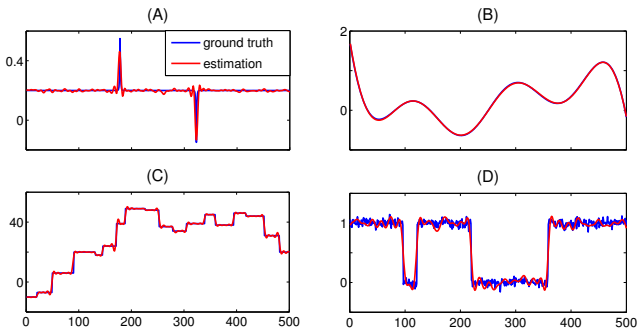

Figure 5: Different types of functions for causal coefficients, where the blue lines indicate the ground truth, and the red lines are the estimation with our GP linear time-dependent causal model. (A) Laplace function. (B) Combination of sinc function and polynomial function. (C) Step function. (D) Noisy square-waves.

Simulation 4 *(Nonlinear Model)*

Next, we generated the data according to the following time-dependent nonlinear model, which is uni-directional $(x_1(t-1) \to x_2(t))$ and does not have instantaneous effects:

$$\begin{cases} x_1(t) = x_1(t-1)(3.8 - 3.8x_1(t-1)) + \varepsilon_1(t), \\ x_2(t) = x_2(t-1)(3.5 - 3.5x_2(t-1) - A(t)x_1(t-1)) \\ \qquad\qquad + \varepsilon_2(t), \end{cases}$$

where $A(t)$ has a sigmoid shape.

We used the nonlinear time-dependent causal model to fit the data. In order to find the time-lagged causal direction, we compared the cross-validated errors in the following four cases: (1) $x_1(t-1) \to x_2(t)$, (2) $x_2(t-1) \to x_1(t)$, (3) $x_1(t-1) \to x_2(t), x_2(t-1) \to x_1(t)$, (4) no time-delayed causal relation between x_1 and x_2. Their prediction errors are 0.0073, 0.0102, 0.0076, and 0.0102, respectively. Case 1 is favored by the cross-validated prediction error. We then tested the independence between the estimated noise and $x_1(t-1), x_2(t-1)$ in the first case, and failed to reject the independence hypothesis, with the HSIC p value 0.13. This indicates the causal influence $x_1(t-1) \to x_2(t)$.

Real Data Test

Stock Indices We chose the daily dividend/split adjusted closing prices from 03/16/2005 to 07/31/2014 of three stock indices: Nasdaq in US, FTSE in UK, and N300 in Japan. These three indices are among the major stock indices all over the world, so it is interesting to see how they are causally related. We analyzed the return series of the indices.

Figure 6: HSIC p values between the estimated noise and hypothetical causes.

causal direction	p value
FTSE ->Nasdaq	0.2170
Nasdaq ->FTSE	0.0210
N300 ->FTSE	0.1081
FTSE ->N300	0.0030
N300 ->Nasdaq	0.0082
Nasdaq ->N300	0

We first considered all possible pairs and tested whether the proposed approach is able to find plausible causal directions. In particular, we fitted our linear time-dependent causal model for both directions of each pair. Figure 6 shows the corresponding p values of HSIC independence test [Gretton *et al.*, 2007] between the estimated noise and hypothetical causes. We failed to reject the independence hypothesis for FTSE $\to$ Nasdaq and N300 $\to$ FTSE at significance level 0.05. As seen from the p values, the proposed approach favors N300 $\to$ FTSE $\to$ Nasdaq; this indeed matches the ordering due to time differences: the time zones corresponding to N300, FTSE, and Nasdaq are UTC+9, UTC, and UTC-5, respectively. Note that the data were aligned according to the local time.

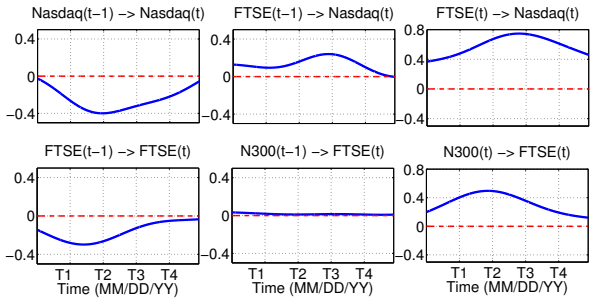

Figure 7: Part of the estimated causal coefficients on stock indices, where T1, T2, T3, and T3 stand for 12/21/2006, 12/11/2008, 12/01/2010, and 11/21/2012, respectively.

For comparison, we also fitted a structural vector autoregressive model on the data to find the instantaneous causal ordering, and found that the independence condition between the estimated noise term and the hypothetical cause does not hold in either direction for any pair. Recall that when we allow the time-dependent causal influence, we will be able to find a uni-directional causal influence.

We then fitted the model on the three time series with the above causal ordering. Figure 7 shows some of the estimated causal coefficients. The instantaneous causal effects from FTSE to Nasdaq, and from N300 to FTSE are obvious, while the one-day delayed causal effect from N300 to FTSE is quite small. These results are consistent with the influence due to the time difference. Interestingly, during the financial crisis of 2008, the causal coefficients become much larger.

Temperature in House This hourly temperature data set was recorded in six places (1 - Shed, 2 - Outside, 3 - Kitchen Boiler, 4 - Living Room, 5 - WC, 6 - Bathroom) of a house in the black forest in Germany. This house was not inhabited most of the time except for some periods including Christmas, and lacked central heating; the electric radiators in room 3, 5, and 6 started when the temperature dropped close to zero (there was no electric radiator in 4) [Peters *et al.*, 2013].

Since the temperature sensor in the shed was taken to other places occasionally, the recorded data contain a number of outliers. We only analyzed the relations between the temperature in the other five places denoted by variables 2 – 6. Here, we considered one time-lagged causal influence and the influence from possible unobserved confounders, with the prior knowledge that temperature inside does not have causal effect on the outside temperature. We found that when the temperature dropped close to zero, that is, when the electric radia-

tors started, or when there were people living there (during Christmas), the causal relations among these rooms, as well as the influences from the confounders, obviously changed. Figure 8 shows the estimated causal graph in three different states: normal state when the electric radiators were off, and there were no guests living there (state A); the period when the electric radiators started automatically for the low temperature (state B); the period when people lived in the house (e.g., during Christmas) (state C). For illustrative purposes, we only considered the causal links whose estimated coefficients are larger than 0.06. Figure 9 shows part of the estimated causal coefficients and the influences from the unobserved confounders in these three states (the three states are separated by dashed blue lines). We see that the casual relations, including causal strength and time-lagged causal structure, among these five places change across different states.

In state A, variable 2 influences 3 and 4, 5 influences the other three rooms, 4 is the sink node, and each variable is influenced by its previous value. The estimated influence from unobserved confounders g is small. In state B, the outside 2 does not have an obvious effect on the inside temperature. Since there is no electric radiator in 4, the causal directions between 3 and 4, 4 and 5 are reversed, compared to state A. The causal influence from unobserved confounders is obvious to the inside. Interestingly, the confounder seems to be consistent with the status of electric radiators. In state C, the outside 2 weakly influences the inside, the causal relations among the remaining rooms are densely connected, and the influence from unobserved confounder is obvious, which might be reasonable due to complicated human behavior. Therefore, it seems that the confounder here can be viewed as the influence from electric radiators or human behavior.

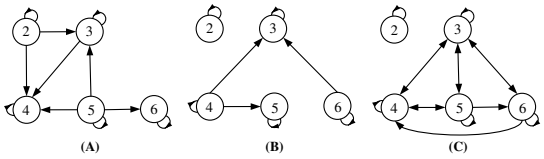

Figure 8: Causal relations among the five places in three different states. We only show causal coefficients which are larger than 0.06 for illustrative purposes. In particular, the self-loop represents the self-influence from its own previous value. (A) Normal state when the electric radiators were off and no guests lived there. (B) The period when the electric radiators were on. (C) The period when people lived in the house.

Conclusion and Discussions

By including the time information as a common cause, this paper extends ordinary functional causal models to identify the time-dependent causal influences, and proposes a nonparametric way to estimate the time-varying causal model and the influence from certain unobserved confounders which can be represented as a smooth function of time.

This work can be extended in several directions. First, we are interested in the cases when the causal structure, instead of only the causal influences, changes over time. We find that if the time information can be viewed as a common cause, the varying causal structure is still identifiable under our framework. However, it is not straightforward to find an efficient way to infer the time-varying instantaneous causal structure automatically from observational data, especially for the nonlinear model. Secondly, we can extend our formulation to incorporate heteroscedastic noise, where the noise is not i.i.d., but its distribution changes along with the cause; it can then identify the underlying causal structure in more general situations. Another line of our future work is to analyze the causal relations between brain regions from fMRI data with the proposed model and method.

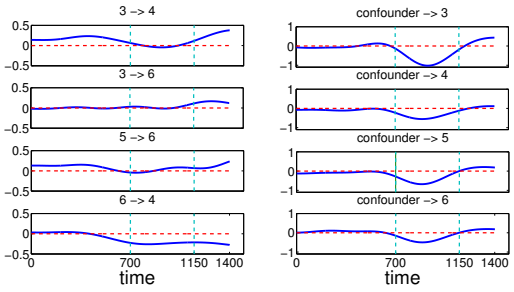

Figure 9: Part of the causal coefficients and the causal influences from the unobserved confounders in the three states (blue solid lines). The three states are separated by the dashed blue lines: state A is during $[1, 700]$; state B is during $[701, 1150]$; state C is during $[1151, 1400]$. The red dashed lines correspond to zero. In the left column, the panels from top to bottom show the causal coefficients of the following directions: $3 \to 4$, $3 \to 6$, $5 \to 6$, and $6 \to 4$, respectively. The right column, from top to bottom, shows the causal influence from the unobserved confounders to 3, 4, 5, 6, respectively.

Acknowledgements

We would like to thank Jonas Peters for helping us to revise this paper, Eric Lacosse for helpful discussions, and anonymous referees for useful suggestions. Kun Zhang was supported in part by DARPA grant No. W911NF-12-1-0034.

References

[Barrett *et al.*, 2012] A. B. Barrett, M. Murphy, M-A. Bruno, Q. Noirhomme, M. Boly, L. Steven, and K. S. Anil. Granger causality analysis of steady-state electroencephalographic signals during propofol-induced anaesthesia. *PLoS ONE*, 7(1):e29072, 2012.

[Chu and Ghahramani, 2005] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–41, 2005.

[Gretton *et al.*, 2007] A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and J. A. Smola. A kernel statistical test of independence. *In Advances in Neural Information Processing Systems*, pages 585–592, 2007.

[Hoyer *et al.*, 2009] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *In Advances in Neural Information Processing Systems*, pages 689–696, 2009.

[Hyvärinen *et al.*, 2010] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11:1709–31, 2010.

[Karlsson *et al.*, 2013] B. Karlsson, M. Hassan, and C. Marque. Windowed multivariate autoregressive model improving classification of labor vs. pregnancy contractions. *35th Annual International Conference of the IEEE EMBS*, pages 7444–7, 2013.

[Pearl, 2000] Judea Pearl. Causality: Models, reasoning and inference. *The MIT Press*, 2000.

[Peters *et al.*, 2011] J. Peters, J. Mooij, D. Janzing, and B. Schöelkopf. Identifiability of causal graphs using functional models. *In Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. AUAI Press*, pages 589–598, 2011.

[Peters *et al.*, 2013] J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using restricted structural equation models. *In Advances in Neural Information Processing Systems*, pages 154–162, 2013.

[Rasmussen and Williams, 2006] C. E. Rasmussen and K. I. C Williams. Gaussian processes for machine learning. *The MIT Press*, 2006.

[Shafi *et al.*, 2012] M. M. Shafi, M. B. Westover, M. D. Fox, and A. Pascual-Leone. Exploration and modulation of brain network interactions with noninvasive brain stimulation in combination with neuroimaging. *European Journal of Neuroscience*, 35:805–25, 2012.

[Shimizu *et al.*, 2006] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–30, 2006.

[Spirtes *et al.*, 2000] P. Spirtes, C. N. Glymour, and R. Scheines. Causation, prediction, and search. *The MIT press*, 2000.

[Zhang and Hyvärinen, 2009a] K. Zhang and A. Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. *he European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 570–585, 2009.

[Zhang and Hyvärinen, 2009b] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. AUAI Press*, pages 647–655, 2009.

[Zhao *et al.*, 2013] Y. Zhao, S. A. Billings, HL. Wei, and P. G. Sarrigiannis. A parametric method to measure time-varying linear and nonlinear causality with applications to EEG data. *IEEE Transactions on Biomedical Engineering*, 60(11):3141–8, 2013.