Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models

Biwei Huang[1] Kun Zhang[1] Mingming Gong[1 2] Clark Glymour[1]

Abstract

In many scientific fields, such as economics and neuroscience, we are often faced with nonstationary time series, and concerned with both finding causal relations and forecasting the values of variables of interest, both of which are particularly challenging in such nonstationary environments. In this paper, we study causal discovery and forecasting for nonstationary time series. By exploiting a particular type of state-space model to represent the processes, we show that nonstationarity helps to identify causal structure and that forecasting naturally benefits from learned causal knowledge. Specifically, we allow changes in both causal strengths and noise variances in the nonlinear state-space models, which, interestingly, renders both the causal structure and model parameters identifiable. Given the causal model, we treat forecasting as a problem in Bayesian inference in the causal model, which exploits the time-varying property of the data and adapts to new observations in a principled manner. Experimental results on synthetic and real-world data sets demonstrate the efficacy of the proposed methods.

1. Introduction

One of the fundamental problems in empirical sciences is to make prediction for passively observed data (a task that machine learning is often concerned with) or to make prediction under interventions. In order to make prediction under interventions, one has to find and make use of causal relations. Discovering causal relationships from observational data, known as causal discovery, has recently attracted much attention. In many scientific fields, we are often faced with nonstationary time series, and concerned with both finding causal relations and forecasting the values of variables of interest, both of which are particularly challenging in such nonstationary environments.

Traditional methods of causal discovery usually focus on independent and identically distributed (i.i.d.) data or stationary processes, and assume that the underlying causal model is fixed. Such methods include constraint-based methods (Spirtes et al., 1993), score-based methods (Chickering, 2003; Heckerman et al., 1995; Huang et al., 2018), and functional causal model-based approaches (Shimizu et al., 2006; Zhang & Chan, 2006; Hoyer et al., 2009; Zhang & Hyvärinen, 2009). Specifically, constraint-based methods and score-based methods recover the causal graph up to the Markov equivalence class, within which some causal directions may not be identifiable. Presuming certain constraints on the class of causal mechanisms, functional causal model-based approaches exploit asymmetries between causal and anti-causal directions.

Those traditional methods may not be practical in a number of situations. The assumption of a fixed causal model may not hold in practice, especially for time series, where the underlying data generating processes may change over time. For example, neural connectivity in the brain may change over time or across different states. The influences between macroeconomic variables may be affected by latent common factors, e.g., economic policies, which may change across different time periods and contribute to nonstationarity of observed macroeconomic variables. If we directly apply causal discovery methods which are designed for a fixed causal model, they may give us misleading results, e.g., spurious edges and wrong causal directions; see e.g., Zhang et al. (2017). A second issue is that, with functional causal model-based approaches, there are cases where causal directions are not identifiable, such as the linear-Gaussian case and the case with a general functional class. Hence, this criterion for direction identification is not generally applicable (Zhang et al., 2015a). Therefore, it is beneficial to investigate other asymmetric criteria for the purpose of causal discovery.

Interestingly, several research papers have shown that nonstationarity contains useful information for causal discovery

[1] Department of Philosophy, Carnegie Mellon University, Pittsburgh [2] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh. Correspondence to: Biwei Huang <biweih@andrew.cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

(Hoover, 1990; Tian & Pearl, 2001; Huang et al., 2015; Zhang et al., 2017; Huang et al., 2017; Peters et al., 2016; Ghassami et al., 2018). Nonstationarity may result from a change in the underlying mechanisms, which is related to soft intervention (Korb et al., 2004) in the sense that both result in probability distribution changes, while nonstationarity can be seen as a consequence of soft interventions done by nature. Furthermore, from a causal view, it has been postulated that if there is no confounder, the marginal distribution P(cause) and the conditional distribution P(effect|cause) represent independent mechanisms of nature (Pearl, 2000; Janzing & Schölkopf, 2010), which is related to the exogeneity notion (Engle et al., 1983; Zhang et al., 2015b). How to characterize such an independence or exogeneity condition is an issue. Thanks to nonstationarity, the independence between probability distributions can be characterized statistically; in the causal direction, the causal modules P(cause) and P(effect|cause) change statistically independently, while P(effect) and P(cause|effect) change dependently generically.

On the other hand, forecasting from nonstationary data is usually hard. In this paper, we argue that forecasting can benefit from causal knowledge for nonstationary processes. First, from the causal view, the distribution shift in nonstationary data is usually constrained–it might be due to the changes in data generating processes of only a few variables. By detecting these key variables, we only need to update the distributions corresponding to these variables. In complex models, the savings can be enormous; a reduction in the number of modeling variables can translate into substantial reduction in the sample complexity. Second, by making use of the information from causal structure, each causal module changes independently and thus can be considered separately. The changes in the causal modules are usually simpler (or more natural) than those in conditional distributions that do not represent causal mechanisms, which also reduces the difficulty of prediction. Third, the causal knowledge makes the forecasts more interpretable. We can gain insight into which factors affect the target variable and how to manipulate the system properly.

In this paper, we study causal discovery and forecasting for nonstationary time series. We provide a principled investigation of how causal discovery benefits from nonstationarity and how the learned causal knowledge facilitates forecasting. Particularly, we formalize causal discovery and forecasting under the framework of nonlinear state-space models. Our main contributions are as follows:

- In Section 3, we formalize a time-varying causal model to represent the underlying causal process in nonstationary time series. We allow changes in both causal strengths and noise variances, as well as changes of causal structure in the sense that some causal influences may vanish or appear over some periods of time.
- In Section 4, we show the identifiability of the proposed causal model under mild conditions; both the causal structure and model parameters are identifiable.
- In Section 5, we give a way to estimate the proposed causal model. It can be transformed to the task of standard estimation of nonlinear state-space models.
- In Section 6, we show that causal models benefit forecasting. Given the causal model, we treat forecasting as a Bayesian inference problem in the causal model, which exploits the time-varying property of the data and adapts to new observations in a principled manner.

2. Motivation and Related Work

Identification of causal relationships from observational data is attractive for the reason that traditional randomized experiments may be hard or even impossible to do. Over the past decades, prominent progress has been made in this area. Constraint-based methods use statistical tests (conditional independence tests) to find causal skeleton and determine orientations up to the Markov equivalence class; widely-used methods include PC and FCI (Spirtes et al., 1993). Score-based methods define a score function that measures how well an equivalence class fits the observed data and search through possible equivalence classes to find the best scored one (Heckerman et al., 1995; Chickering, 2003; Huang et al., 2018). It was later shown that with functional causal model-based approaches, it is possible to recover the whole causal graph with certain constraints on the functional class of causal mechanisms, by making use of asymmetries between causal and anti-causal directions. For example, in the case of linear causal relationships, the non-Gaussianity of noise terms helps to identify the causal direction; in the causal direction, the noise term is independent of hypothetical causes, while independence does not hold in the anti-causal direction. For instance, the linear non-Gaussian acyclic model (LiNGAM) (Shimizu et al., 2006) uses this property for causal discovery.

Granger causality (Granger, 1969) is widely applied in time series analysis, especially in economics. It concerns time-lagged relationships and assumes that the underlying causal strengths and noise variances are fixed. A more recent method based on structural vector-autoregressive models further incorporates contemporaneous causal relationships (Hyvärinen et al., 2010). However, these methods are only appropriate for stationary time series, while in real-world problems, it is commonplace to encounter nonstationary data. If we directly apply the above approaches to nonstationary data, it may lead to spurious edges or wrong causal directions; see e.g., Zhang et al. (2017).

More recently, causal discovery methods for nonstationary data have been proposed (Tian & Pearl, 2001; Peters et al., 2016; Zhang et al., 2017). In particular, in Zhang et al. (2017), it adds a surrogate variable, e.g., time or domain index, to the causal system to account for changing causal relations and to determine causal directions by exploiting the independent change between P(cause) and P(effect|cause). Particularly, it uses kernel distribution embeddings to describe shifting probabilistic distributions in a non-parametric way. Despite its general applicability in theory, in practice, it may be limited in several aspects. With kernels, the computational complexity is $O(N^3)$, where N is the sample size, which is expensive and makes it intractable in large data sets. Moreover, in practice, it is not easy to choose an appropriate kernel width, and the kernel width can heavily affect the results.

Another set of studies have tried to model time-varying relationships - such relationships are either not necessarily causal, or causal relationships in which the causal direction is already known in advance, e.g., one can assume that past causes future without contemporaneous causal relationships. Hence, they do not have the phase of discovering causal structure from observational data. For the former case, representative work includes the estimation of time-varying precision matrix by minimizing the temporally smoothed L_1 penalized regression (Kolar & Xing, 2012). For the latter, it includes research studies in dynamic Bayesian networks (Dagum et al., 1992; Song et al., 2009). However, in practice, it is often the case that some causal interactions occur in the same time period, and thus it is important to consider contemporaneous causal relations, especially in time series with low temporal resolutions (Hyvärinen et al., 2010; Gong* et al., 2015), in aggregated data (Gong et al., 2017), or in equilibrium data.

For forecasting with nonstationary data, basically two types of methods are usually used: active approaches and passive approaches (Alippi & Roveri, 2008; Elwell & Polikar, 2011). Specifically, the active approach updates the model only when a change is detected, which limits its applicability for time series with gradual changes. The passive approach, such as the dynamic linear model, does not actively detect the drift in environments, but performs a continuous adaptation of the model every time new data arrive.

To the best of our knowledge, the present paper is the first work on simultaneous causal discovery (covering both contemporaneous and time-lagged causal relationships) and forecasting in nonstationary environments, where forecasting directly benefits from causal modeling in a natural way.

3. Time-Varying Linear Causal Models

Suppose that we have m observed time series $X_t = \big(x_{1,t}, \cdots, x_{m,t}\big)^{\mathrm{T}}$, satisfying the following generating process:

$$x_{i,t} = \sum_{x_j \in \mathbb{PA}_i} b_{ij,t} x_{j,t} + e_{i,t}, \tag{1}$$

where $\mathbb{PA}_i$ is the set of direct instantaneous causes of x_i, $x_j \in \mathbb{PA}_i$ is the jth direct cause of x_i, $b_{ij,t}$ is the causal coefficient from $x_{j,t}$ to $x_{i,t}$, and $e_{i,t}$ is the Gaussian noise term with $e_{i,t} \sim \mathcal{N}(0, \sigma_{i,t}^2)$, which indicates influences from unmeasured factors. The noise distribution does not have to be Gaussian; here we make this assumption mainly for the purpose of showing that even when causal relationships are linear, and the noise terms are Gaussian, the causal model is still identifiable by using nonstationarity. In real-world problems, other appropriate noise distributions can be applied. We will see in Section 4 that the identifiability of the time-varying causal model does not require the Gaussian assumption.

The causal process is assumed to have the following properties.

- Let B_t be the $m \times m$ causal adjacency matrix with entries $b_{ij,t}$, and denote by G_t the corresponding binary matrix (quantitative causal adjacency matrix), with $G_t(j,i) = 1$ if and only if $b_{ij,t} \neq 0$ and zero otherwise. We assume that the graph union $G = G_1 \cup \cdots \cup G_T$ is acyclic.
- We allow each causal coefficient $b_{ij,t}$ and noise variance $\sigma_{i,t}^2$ to change over time and model the changes by the following autoregressive models:

$$\begin{aligned} b_{ij,t} &= \alpha_{ij,0} + \sum_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t}, \\ h_{i,t} &= \beta_{i,0} + \sum_{q=1}^{q_l} \beta_{i,q} h_{i,t-q} + \eta_{i,t}, \end{aligned} \tag{2}$$

  respectively, where $\epsilon_{ij,t} \sim \mathcal{N}(0, w_{ij})$, $\eta_{i,t} \sim \mathcal{N}(0, v_i)$, and $h_{i,t} = \log(\sigma_{i,t}^2)$ models the volatility of the observed time series. Each causal coefficient and log-transformed noise variance changes independently. Again here the distributions of $\epsilon_{ij,t}$ and $\eta_{i,t}$ are not necessarily Gaussian; for example, we can easily extend it to mixture of Gaussian distributions. Note that this formulation includes the case where only causal coefficients change with time, while noise distributions stay constant, i.e., $e_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$, $\forall i \in \mathbf{N}^+$.
- We allow changes in causal structure that some causal edges may vanish or appear over some periods of time.

Equation (1) can be represented in the matrix form, with

$$X_t = (I_m - B_t)^{-1} E_t, \tag{3}$$

where I_m is an $m \times m$ identity matrix, and $E_t = \big(e_{1,t}, \cdots, e_{m,t}\big)^{\mathrm{T}}$. Thus, by combining causal process (3) and autoregressive functions (2), we have the following causal model:

$$\begin{cases} X_t &= (I_m - B_t)^{-1}E_t, \\ b_{ij,t} &= \alpha_{ij,0} + \sum\limits_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t}, \\ h_{i,t} &= \beta_{i,0} + \sum\limits_{q=1}^{q_l} \beta_{iq} h_{i,t-q} + \eta_{i,t}, \end{cases} \tag{4}$$

with $\epsilon_{ij,t} \sim \mathcal{N}(0, w_{ij})$ and $\eta_{i,t} \sim \mathcal{N}(0, v_i)$, for $t = \max(p_l, q_l), \cdots, T$. Figure 1 gives the graphical representation of generating processes of the time-varying causal network.

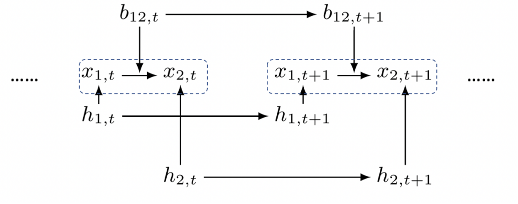


Figure 1. A graphical representation of generating processes of the time-varying causal network.

In real-world problems, there may also exist time-delayed causal relations. To consider both time-delayed and instantaneous causal relations, we modify equation (1) to

$$x_{i,t} = \sum_{x_j \in \mathbb{PA}_i} b_{ij,t} x_{j,t} + \sum_{s=1}^{s_l} \sum_{x_k \in \mathbb{PL}_i} c_{ik,t}^{(s)} x_{k,t-s} + e_{i,t}, \tag{5}$$

where $\mathbb{PL}_i$ is the set of lagged causes of x_i, and $c_{ik}^{(s)}$ represents the s-lagged causal strength from x_k to x_i. Similarly, we model the time-varying lagged causal strength with an autoregressive model,

$$c_{ij,t}^{(s)} = \gamma_{ij,0}^{(s)} + \sum_{r=1}^{r_l} \gamma_{ij,r}^{(s)} c_{ij,t-r}^{(s)} + \nu_{ij,t}^{(s)}, \tag{6}$$

with $\nu_{ij,t}^{(s)} \sim \mathcal{N}(0, u_{ij}^{(s)})$.

In the next section, we are mainly concerned with the identifiability of instantaneous causal relations, while the results are also extended to handle the above delayed causal relations.

It is worth noting that although the model (1) is linear in the processes, it is actually nonlinear in the latent processes b_{ij} and h_i. Therefore, the time-varying linear causal model is actually a specific type of nonlinear state-space model with respect to hidden variables b_{ij} and h_i. In fact, in Section 5, we will estimate the proposed model by extending methods for estimating nonlinear state-space models.

4. Model Identifiability

We show in Theorem 1 that the proposed causal model, including causal structure and model parameters, is identifiable under the following conditions:

- The underlying instantaneous causal structure is acyclic.
- Each causal coefficient varies with time and follows an autoregressive model, and distributions of $e_{i,t}$ are fixed.

Note that for identifiability, we do not require the additive noise terms to be Gaussian. Furthermore, we do not assume faithfulness (Spirtes et al., 1993), which is commonly assumed in traditional constraint-based causal discovery.

Theorem 1. *Suppose the observed time series, $X_t = \big(x_{1,t}, \cdots, x_{m,t}\big)^T$, were generated by*

$$\begin{cases} x_{i,t} = \sum_{x_j \in \mathbb{PA}_i} b_{ij,t} x_{j,t} + e_{i,t}, \\ b_{ij,t} = \alpha_{ij,0} + \alpha_{ij,1} b_{ij,t-1} + \epsilon_{ij,t}, \end{cases} \tag{7}$$

where $x_{j,t}$ is the cause of $x_{i,t}$, and $b_{ij,t}$ is the corresponding causal coefficient from $x_{j,t}$ to $x_{i,t}$, which satisfies a first-order autoregressive model with $\alpha_{ij,0}, \alpha_{ij,1} \in (-1, 1)$. The additive error, $e_{i,t}$, represents a stationary zero-mean white noise process, i.e., $E[e_{i,t}] = 0$, $E[e_{i,t}e_{i,t'}] = \sigma_i^2 \delta_{tt'}$, and $E[e_{i',t}e_{i,t}] = \sigma_i^2 \delta_{ii'}$, where $\sigma_i^2 < \infty$ and $\delta_{tt'}$ is the delta function. Similarly, the error in the autoregressive model of $b_{ij,t}$ satisfies $E[\epsilon_{ij,t}] = 0$ and $E[\epsilon_{ij,t}\epsilon_{ij,t}] = w_{ij}$. In addition, the underlying instantaneous causal structure over X_t is assumed to be acyclic.

Then the model in (7) is identifiable, including the causal order between x_i's and model parameters, when time series are long enough.

Here we give a sketch of the proof. For complete proofs of the theoretical results reported in the paper, please refer to the supplementary material.

Proof sketch.

1. First identify the root cause.

   Let
   $$S(t, t+p)_i := E[x_{i,t}^2 x_{i,t+p}^2].$$

   Let r_0 be the index of the root cause, and $\mathbf{V}_s = \mathbf{V} \backslash r_0$ denote the indices of the remaining processes, with $\mathbf{V} = \{1, \cdots, m\}$. Then we will have

   $$\begin{array}{l} S(t, t+p)_{r_0} - S(t, t+p-1)_{r_0} = 0; \\ S(t, t+p)_{r_s} - S(t, t+p-1)_{r_s} < 0, \quad \forall r_s \in \mathbf{V}_s. \end{array}$$

   The reason is that the root cause does not receive changing influences from other processes. For the root cause, $S(t, t+p)_{r_0} = \sigma_{r_0}^4$, where $\sigma_{r_0}^2$ is the noise variance in the causal model of x_{r_0}, so we can also identify the noise variance of x_{r_0}.

2. Next, iteratively identify the remaining causal graph. Suppose that we have identified n processes that are the earliest according to the causal order. We then identify the next variable according to the causal order. Let $\mathbf{V}_n$ represent variable indices of the first n processes

and let $\mathbf{V}_{\tilde{n}} = \mathbf{V} \backslash \mathbf{V}_n$. For any $r_s \in \mathbf{V}_{\tilde{n}}$, we show that if and only if x_{r_s} is the next according to the order, $S(t, t+p)_{r_s}$ is a linear combination of cross-statistics of different orders of $x_{\mathbf{V}_n,t}$. In this way, we can identify the causal graph of the first $n+1$ processes. In addition, the corresponding parameters are also identifiable, according to the identifiability of the varying coefficient regression models (Wall, 1987).

Repeating this procedure until we go through all processes, we have the identifiability of the whole causal model.

□

It is easy to extend the above identifiability result to the case when there are both time-lagged and instantaneous causal relations, which is given in Corollary 1, since for lagged causal relations, their causal directions are fixed (from past to future), and thus, it reduces to a parameter identification problem.

Corollary 1. *Suppose that the m observed time series, $X_t = \big(x_{1,t}, \cdots, x_{m,t}\big)^T$, satisfy the following generating process:*

$$\begin{cases} x_{i,t} = \sum\limits_{x_j \in \mathbb{PA}_i} b_{ij,t} x_{j,t} + \sum\limits_{s=1}^{s_l} \sum\limits_{x_k \in \mathbb{PL}_i} c_{ik,t}^{(s)} x_{k,t-s} + e_{i,t}, \\ b_{ij,t} = \alpha_{ij,0} + \alpha_{ij,1} b_{ij,t-1} + \epsilon_{ij,t}, \\ c_{ij,t}^{(s)} = \gamma_{ij,0}^{(s)} + \sum\limits_{r=1}^{r_l} \gamma_{ij,r}^{(s)} c_{ij,t-r}^{(s)} + \nu_{ij,t}^{(s)}, \end{cases} \tag{8}$$

where $c_{ij,t}^{(s)}$ represents the s-lagged causal coefficient, which satisfies an autoregressive model with $\gamma_{ij,0}^{(s)}, \gamma_{ij,r}^{(s)} \in (-1, 1)$. The additive error, $\nu_{ij,t}^{(s)}$, represents a stationary zero-mean white noise process,with $E[\nu_{ij,t}^{(s)}] = 0$ and $E[\nu_{ij,t}^{(s)} \nu_{ij,t}^{(s)}] = u_{ij}^{(s)}$. Other notations $b_{ij,t}$, $e_{i,t}$, $\alpha_{ij,0}$, $\alpha_{ij,1}$, and $\epsilon_{ij,t}$ are the same as in Theorem 1. In addition, the underlying instantaneous causal structure over X_t is assumed to be acyclic.

Then the model in (8) is identifiable, including the causal order between x_i's and model parameters, when time series are long enough.

The above results do not take into account the changeability of σ_i^2. For the general case where σ_i^2 futher changes, our empirical results strongly suggest that the causal model is also identifiable, although currently there is no straightforward, concise proof for it.

5. Model Identification

The model defined in equation (4) can be regarded as a nonlinear state-space model, with causal coefficients and log-transformed noise variances being latent variables $Z = \big\{\{b_{ij}\}_{i,j=1}^m, \{h_i\}_{i=1}^m\big\}$, and model parameters $\theta = \big\{\{\alpha_{ij,p}\}, \{\beta_{i,q}\}, \{w_{ij}\}, \{v_i\}\big\}$. Therefore, it can be transformed to a nonlinear state-space model estimation problem. In particular, we exploit an efficient stochastic approximation expectation maximization (SAEM) algorithm (Delyon et al., 1999), combined with conditional particle filters with ancestor sampling (CPF-AS) in the E step (Lindsten et al., 2012; Lindsten, 2013), for model estimation.

5.1. SAEM Algorithm

For a traditional EM algorithm, the procedure is initialized at some $\theta_0 \in \Theta$ and then iterates between two steps, expectation (E) and maximization (M):

(E) Compute $p_{\theta^{k-1}}(Z|X)$ and the lower bound of the log-likelihood, $\mathcal{Q}(\theta, \theta^{k-1})$, with

$$\mathcal{Q}(\theta, \theta^{k-1}) = \int p_{\theta^{k-1}}(Z|X) \log p_\theta(Z, X)\, dZ.$$

(M) Compute $\theta^k = \arg\max_{\theta \in \Theta} \mathcal{Q}(\theta, \theta^{k-1})$.

In the E step, we need to compute the expectation under the posterior $p_{\theta^{k-1}}(Z|X)$, which is intractable in our case, since $p(X, Z)$ is not Gaussian. To address this issue, SAEM computes the E step by Monte Carlo integration and uses a stochastic approximation update of the quantity $\mathcal{Q}$:

$$\tilde{\mathcal{Q}}_k(\theta) = (1-\lambda_k)\tilde{\mathcal{Q}}_{k-1}(\theta) + \lambda_k \sum_{j=1}^{M} \frac{\omega_T^{(k,j)}}{\sum_l \omega_T^{(k,l)}} \log p_\theta(X_{1:T}, \mathring{Z}_{1:T}^{(k,j)}), \tag{9}$$

where $\mathring{Z}$ indicates sampled particles of Z, $\omega_T^{(k,j)}$ the weight of jth particle at kth iteration, M the generated number of particels, $X_{1:T} = \{X_t\}_{t=1}^T$, $\mathring{Z}_{1:T}^{(k,j)} = \{\mathring{Z}_t^{(k,j)}\}_{t=1}^T$, and $\{\lambda_k\}_{k \geq 1}$ is a decreasing sequence of positive step size, with $\sum_k \lambda_k = \infty$ and $\sum_k \lambda_k^2 < \infty$. The E-step is thus replaced by the following:

(E′) At each iteration, generate M particles of $\mathring{Z}_{1:T}^{(k,j)}$ from $p_{\theta^{k-1}}(Z|X)$ and compute $\tilde{\mathcal{Q}}_k(\theta)$ according to (9). (A method for sampling from $p_{\theta^{k-1}}(Z|X)$ is introduced in the next section.)

Under appropriate assumptions, SAEM is shown to converge for fixed M, as $k \to \infty$ (Delyon et al., 1999). The model parameters in the M step are updated by setting $\frac{\partial \tilde{\mathcal{Q}}_k(\theta)}{\partial \theta} = 0$. The detailed derivations are given in Section $S3$ in supplementary materials. The computational complexity in each iteration is $O(m^3 \times M \times T)$, where m is the number of variables, M the number of sampled particles (we used $M = 15$), and T the length of time series.

5.2. Conditional Particle Filter with Ancestor Sampling

To sample particles $\mathring{Z}$ from the posterior distribution, we use conditional particle filtering with ancestor sampling (CPF-AS) (Lindsten, 2013). The CPF-AS procedure is a sequential Monte Carlo sampler, akin to a standard particle filter but with the difference that one particle at each time step is specified as a priori. Let these prespecified particles be $\mathring{Z}'_{1:T} = \{\mathring{Z}'_t\}_{t=1}^T$. Let $\{\mathring{Z}_{1:t-1}^{(j)}, \omega_{t-1}^{(j)}\}_{j=1}^M$ be a weighted particle system targeting $p_\theta(\mathring{Z}_{1:t-1}|X_{1:t-1})$. To propagate this sample to time t, we introduce the auxiliary variable s_t^j, referred to the ancestor particle of $\mathring{Z}_t^{(j)}$. To generate a specific particle $\mathring{Z}_t^{(j)}$ at time t, we first sample the ancestor index with $P(s_t^j = i) \propto \omega_{t-1}^i$. Then $\mathring{Z}_t^{(j)}$ is sampled from $\mathring{Z}_t^{(j)} \sim f_\theta(\mathring{Z}_t|\mathring{Z}_{t-1}^{s_t^j})$, $j = 1, \cdots, M-1$. The Mth particle is sampled deterministically: $\mathring{Z}_t^{(M)} = \mathring{Z}'_t$. We sample the ancestor index s_t^M with $P(s_t^M = j) \propto \omega_{t-1}^{(j)} f_\theta(\mathring{Z}'_t|\mathring{Z}_{t-1}^j)$. Finally, all the particles are assigned importance weights, $\omega_t^{(j)} = W_{\theta,t}(\mathring{Z}_t^{(j)}, \mathring{Z}_{t-1}^{s_t^j})$, where the weight function is given by $p_\theta(X_t|\mathring{Z}_t)$. The CPF-AS is summarized in Algorithm $S1$ (Section $S4$) in supplementary materials.

5.3. Causal Graph Determination

The causal graph is determined from the sampled particles. With finite samples, there exist estimation errors; for example, even when there is no causal edge from x_j to x_i, the estimation $\hat{b}_{ij,t}$ may not be exactly zero but some small values. To determine whether there is a causal edge from x_j to x_i, we check both the mean and the variance of $\hat{b}_{ij,t}$. Specifically, if both $\bar{\hat{b}}_{ij} = \frac{1}{T}\sum_{t=1}^T \hat{b}_{ij,t} < \alpha$ and $\frac{1}{T}\sum_{t=1}^T (\hat{b}_{ij,t} - \bar{\hat{b}}_{ij})^2 < \alpha$, we determine that there is no causal edge from x_j to x_i, where α is a threshold.

In our model, we estimate the causal adjacency matrix B_t directly. Recall that LiNGAM (Shimizu et al., 2006) first estimates $A = (I - B)^{-1}$ and then recover the underlying adjacency matrix B by performing extra permutation and rescaling, since W is only identified up to permutation and scale. We directly model the causal process, represented by B_t, with the following advantages:

- It is easy to add prior knowledge of causal connections. In practice, experts may have domain knowledge about some causal edges.
- One can directly enforce sparsity constraints on the causal adjacencies; even if B_t is sparse, $(I - B_t)^{-1}$ is not necessarily sparse, so enforcing the sparsity of causal adjacency would be more difficult when working with A. Section $S5$ in the supplementary materials explains how to add sparsity constraints on causal adjacency matrix B_t and on $b_{ij,t} - b_{ij,t-1}$, which ensures smooth changes of $b_{ij,t}$ over time.
- The estimation procedure directly outputs the causal adjacency matrix, without additional steps of permutation and rescaling, which are usually expensive.

6. Forecasting with Time-Varying Causal Models

After identifying the causal model, we aim to do forecasting by taking advantage of the causal information. Suppose that we have observational data $\tilde{X}_{1:T+1}$ and $Y_{1:T}$, with $X = \{\tilde{X}, Y\}$, and we want to predict Y_{T+1}. We denote the Markov blanket of Y by $\mathcal{M}_Y = \mathcal{P}_Y \cup \mathcal{C}_Y \cup \mathcal{S}_Y$, where $\mathcal{P}_Y$ denotes the set of parents of Y, $\mathcal{C}_Y$ the set of children of Y, and $\mathcal{S}_Y$ the set of spouses of Y. Given its Markov blanket, Y is independent of remaining variables in the causal system; thus, $\mathcal{M}_Y$ contains all the information that is needed to predict Y. The posterior of Y_{T+1} given its Markov blanket at time $T+1$ can be represented as

$$\begin{aligned} & p(Y_{T+1}|\mathcal{M}_{Y,T+1}) \\ \propto\ & p(Y_{T+1}|\mathcal{P}_{Y,T+1}) \prod_{\tilde{X}_{C_i} \in \mathcal{C}_Y} p(\tilde{X}_{C_i,T+1}|\mathcal{P}_{C_i,T+1}), \end{aligned} \tag{10}$$

where $\tilde{X}_{C_i} \in \mathcal{C}_Y$ is the ith child of Y, and $\mathcal{P}_{C_i} \in \mathcal{M}_Y$ denotes the parents of $\tilde{X}_{C_i}$ in $\mathcal{M}_Y$. Let $\vec{b}_Y$ and σ_Y^2 denote the corresponding causal coefficients and noise variance in the functional causal model of Y. Let $D_T := \{\tilde{X}_{1:T}, Y_{1:T}\}$. Then we have

$$\begin{aligned} & p(Y_{T+1}|\mathcal{P}_{Y,T+1}) \\ =\ & \int\int\int\int p(Y_{T+1}|\mathcal{P}_{Y,T+1}, \vec{b}_{Y,T+1}, \sigma^2_{Y,T+1}) \\ & p(\vec{b}_{Y,T+1}|\vec{b}_{Y,T}) p(\vec{b}_{Y,T}|D_T) p(\sigma^2_{Y,T+1}|\sigma^2_{Y,T}) \\ & p(\sigma^2_{Y,T}|D_T)\, d\vec{b}_{Y,T+1}\, d\vec{b}_{Y,T}\, d\sigma^2_{Y,T+1}\, d\sigma^2_{Y,T}. \end{aligned} \tag{11}$$

Since each coefficient changes independently, $p(\vec{b}_{Y,T+1}|\vec{b}_{Y,T})$ can be written as

$$p(\vec{b}_{Y,T+1}|\vec{b}_{Y,T}) = \prod_{b_Y^j \in \vec{b}_Y} p(b^j_{Y,T+1}|b^j_{Y,T}), \tag{12}$$

where b_Y^j is the jth entry in $\vec{b}_Y$.

Similarly, let $\vec{b}_{C_i}$ and $\sigma^2_{C_i}$ denote the corresponding causal coefficients and noise variance in the causal model of $\tilde{X}_{C_i}$, respectively. Then we have

$$\begin{aligned} & p(\tilde{X}_{C_i,T+1}|\mathcal{P}_{C_i,T+1}) \\ =\ & \int\int\int\int p(\tilde{X}_{C_i,T+1}|\mathcal{P}_{C_i,T+1}, \vec{b}_{C_i,T+1}, \sigma^2_{C_i,T+1}) \\ & p(\vec{b}_{C_i,T+1}|\vec{b}_{C_i,T}) p(\vec{b}_{C_i,T}|D_T) p(\sigma^2_{C_i,T+1}|\sigma^2_{C_i,T}) \\ & p(\sigma^2_{C_i,T}|D_T)\, d\vec{b}_{C_i,T+1}\, d\vec{b}_{C_i,T}\, d\sigma^2_{C_i,T+1}\, d\sigma^2_{C_i,T}, \end{aligned} \tag{13}$$

with

$$p(\vec{b}_{C_i,T+1}|\vec{b}_{C_i,T}) = \prod_{b^j_{C_i} \in \vec{b}_{C_i}} p(b^j_{C_i,T+1}|b^j_{C_i,T}), \tag{14}$$

where $b^j_{C_i}$ is the jth entry in $\vec{b}_{C_i}$.

Since $p(Y_{T+1}|\mathcal{P}_{Y,T+1})$ and $p(\tilde{X}_{C_i,T+1}|\mathcal{P}_{C_i,T+1})$ are not necessarily Gaussian, the integrations in (11) and (13) are not given in closed forms. We use Markov chain Monte

Carlo (MCMC) to do Bayesian inference; in particular, we use the Metropolis-Hastings algorithm (Robert & Casella, 2004).

$p(Y_{T+1}|\mathcal{P}_{Y,T+1})$ in equation (11) is estimated by Monte Carlo integration,

$$p(Y_{T+1}|\mathcal{P}_{Y,T+1}) = \sum_j p(Y_{T+1}|\mathcal{P}_{Y,T+1}, \vec{b}^{(j)}_{Y,T+1}, \sigma^{2(j)}_{Y,T+1}),$$

where $\vec{b}^{(j)}_{Y,T+1}$ and $\sigma^{2(j)}_{Y,T+1}$, $\forall j$, are samples from $p(\vec{b}_{Y,T+1}|\vec{b}_{Y,T})$ and $p(\sigma^2_{Y,T+1}|\sigma^2_{Y,T})$, respectively. For $\vec{b}_{Y,T}$ and $\sigma^2_{Y,T}$, we use the generated particles by CPF-AS in model estimation. Similarly, $p(\tilde{X}_{C_i,T+1}|\mathcal{P}_{C_i,T+1})$ in equation (13) is also estimated by Monte Carlo integration.

The detailed procedure of estimating $p(Y_{T+1}|\mathcal{M}_{Y,T+1})$ by Metropolis-Hastings is given in Algorithm 1, where $q(\cdot)$ is taken as a normal distribution. The first hundred samples are ignored, due to the "burn-in" period.

Algorithm 1 Forecasting of Y_{T+1} by Metropolis-Hastings

```
Initialize $Y^{(0)}$.
for $i = 1$ to $N$ do
    Propose: $Y^{\text{candi}} \sim q(Y^{(i)}|Y^{(i-1)})$.
    Acceptance probability:
        $\alpha(Y^{\text{candi}}|Y^{(i-1)}) = \min\left\{1, \frac{q(Y^{(i-1)}|Y^{\text{candi}})p(Y^{\text{candi}}|\mathcal{P}_{Y,T+1})}{q(Y^{\text{candi}}|Y^{(i-1)})p(Y^{(i-1)}|\mathcal{P}_{Y,T+1})} \cdot \frac{\prod_{\tilde{X}_{C_i} \in \mathcal{C}_Y} p(\tilde{X}_{C_i,T+1}|Y^{\text{candi}}, \mathcal{P}_{C_i,T+1}/Y^{\text{candi}})}{\prod_{\tilde{X}_{C_i} \in \mathcal{C}_Y} p(\tilde{X}_{C_i,T+1}|Y^{(i-1)}, \mathcal{P}_{C_i,T+1}/Y^{(i-1)})}\right\}$.
    Take $u \sim \text{Uniform}(0,1)$.
    if $u < \alpha$ then
        accept the propose: $Y^{(i)} = Y^{\text{candi}}$,
    else
        reject the propose: $Y^{(i)} = Y^{(i-1)}$.
    end if
end for
Output: $\hat{Y}_{T+1} = \frac{1}{N-100+1}\sum_{i=100}^{N} Y^{(i)}$.
```

7. Experimental Results

To show the efficacy of the proposed approach for simultaneous causal discovery and forecasting, we apply it to both synthetic and real-world data.

Synthetic Data We considered two types of data generating processes:

(1) Only causal strengths b_{ij} change over time according to autoregressive models, but noise variances σ_i^2 are constant over time.

(2) Both causal strengths b_{ij} and noise variances σ_i^2 change over time according to autoregressive models.

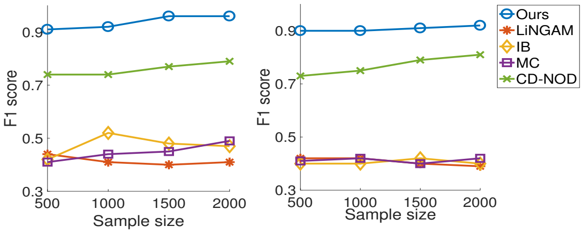
We randomly generated acyclic causal structures according to the Erdos-Renyi model (Erdős & Rényi, 1959) with parameter 0.3. Each generated graph has 5 variables. The parameters were set in the following way: the fixed noise variance $\sigma_i^2 \sim \mathcal{U}(0.1, 0.5)$, the noise variance of b_{ij}'s autoregressive model $w_{ij} \sim \mathcal{U}(0.01, 0.1)$, the noise variance of h_i's autoregressive model $v_i \sim \mathcal{U}(0.01, 0.1)$, the coefficient in b_{ij}'s autoregressive model $\alpha_{i,p} \sim \mathcal{U}(0.8, 0.998)$, and the coefficient in h_i's autoregressive model $\beta_{i,p} \sim \mathcal{U}(0.8, 0.998)$, where $\mathcal{U}(l, u)$ denotes a uniform distribution between l and u. We also considered different sample sizes $T = 500, 1000, 1500,$ and 2000. For each setting (a particular data generating process and a particular sample size), we generated 50 realizations.

For causal discovery, we identified the causal structure by the proposed method. We compared it with other well-known approaches in causal discovery, including LiNGAM, Causal Discovery from NOnstationary/heterogeneous Data (CD-NOD) (Zhang et al., 2017), the minimal change method (MB) (Ghassami et al., 2018), and the identical boundaries method (IB) (Ghassami et al., 2018). CD-NOD estimates the causal skeleton by constraint-based methods over an augmented set of variables and orients the causal direction by using the modularity property: $P(\text{cause}) \perp\!\!\!\perp P(\text{effect|cause})$. Both IB and MC are designed for multi-domain causal discovery in linear systems.

In our methods, we randomly initialized the parameters and determined the causal graph by using a threshold (we simply used 0.05) for both the mean and variance of $\hat{b}_{ij,t}$; that is, if $\bar{\hat{b}}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{b}_{ij,t} < 0.05$ and $\frac{1}{T}\sum_{t=1}^{T}(\hat{b}_{ij,t} - \bar{\hat{b}}_{ij})^2 < 0.05$, we concluded that there is no edge from x_j to x_i. For CD-NOD, the kernel width was set empirically (Zhang et al., 2017), and the significance level was 0.05. Since both IB and MC methods need data from multiple domains, we segmented the data into non-overlapping domains with sample size 100 in each domain.

In Figure 2, we reported the F1 score to measure the accuracy of learned causal graphs in both scenarios: one with only changing causal strengths (Figure 2(a)) and the other with changes in both $b_{ij,t}$ and $\sigma^2_{i,t}$ (Figure 2(b)). From the figure, one can see that our proposed method gives the best performance (the highest F1 score) in all cases, and the accuracy slightly increase along with sample size. The nonparametric method CD-NOD has the second-best performance. CD-NOD assumes that the changes are smooth, and in practice, it may be affected by inappropriately chosen kernel widths and significance level, and may need a large sample for good performance. The other three methods do not perform as well. IB and MC likely under-perform because they are designed for multi-domain systems and thus may not work well for the changes considered here. Similarly, LiNGAM is designed for fixed causal models and thus not appropriate for nonstationary data.

Then we did forecasting by making use of the estimated time-varying causal model. For each realization, we further simulated 10 values for the processes, and predicted the values of each process with one-step ahead prediction. We compared the proposed method with a collection of methods which do not consider the underlying causal model, including (vanilla) Lasso (Tibshirani, 1996), window-based Lasso, Kalman filtering (KF) (Kalman, 1960), state-space model estimated with CPF-AS (denoted by SSM(CPF)), and Gaussian process (GP) regression (Rasmussen & Williams, 2006). We used all the remaining processes as predictors for the target process. Particularly, the window size for window-based Lasso was 100.



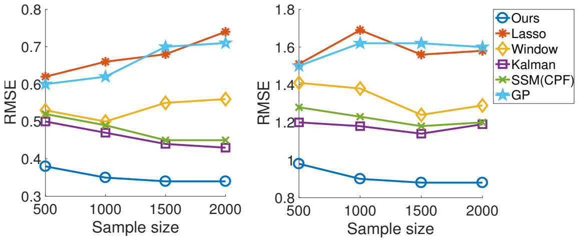
(a) Only b changes

(b) Both b and σ^2 change

Figure 2. F1 score of the estimated causal graph when (a) only $b_{ij,t}$ changes and when (b) both $b_{ij,t}$ and $\sigma^2_{i,t}$ change.

We calculated the root-mean-squared error (RMSE) of the predicted 10 values to evaluate the forecasting performance. We first did paired, one-sided Wilcoxon signed rank test between our method and each of the remaining ones (Gibbons & Chakraborti., 2011), across the two settings (with constant and changing noise variances, respectively) and across the four different sample sizes. Our methods significantly outperform all others in all cases, with the highest p-value 0.018 for the comparison with KF, 0.005 with SSM(CPF), $\times 10^{-4}$ with Lasso, $\times 10^{-3}$ with window-based Lasso, and $10^{-4}$with GP. For illustrative purposes, Figure 3 shows the mean of RMSE across different processes and parameter settings; in (a), only causal strengths $b_{ij,t}$ change, and in (b) both $b_{ij,t}$ and noise variances $\sigma^2_{i,t}$ change. We can see that the RMSE generally decreases with sample size. The Lasso and GP additionally do not consider the change of the model, and not surprisingly perform worse than others.

Real-World Economic data We investigated the causal relationships between Gross Domestic Product (GDP), inflation, economic growth, and unemployment rate, with quarterly data from 1965 to 2017 in the USA [1]. The data are normalized by subtracting the mean and dividing them by the standard deviation. We applied model (4) to estimate contemporaneous causal relations between the four macroeconomic variables. From our model, we found that inflation and economic growth affect GDP, that economic growth influences inflation, and that unemployment is directly influenced by GDP and inflation; see Figure 4. These findings seem consistent with domain knowledge [2]: for example, inflation increases the cost of products and leads to a decline in production, thus causing GDP to fall; economic growth gradually increases the price level of all goods, thereby causing inflation; inflation may increase unemployment because of the decline in competitiveness and export demand.



(a) Only b changes

(b) Both b and σ^2 change

Figure 3. RMSE of the forecasts when (a) only $b_{ij,t}$ change and when (b) both $b_{ij,t}$ and $\sigma^2_{i,t}$ change.

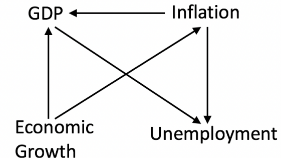


Figure 4. Identified contemporaneous causal relationships between GDP, inflation, economic growth, and unemployment.

Table 1. RMSE of the forecasts on inflation (2007 - 2017).

Methods	RMSE	Methods	RMSE
Ours	**0.32**	Lasso	0.38
Kalman filtering	0.42	Window Lasso	0.37
SSM (CPF)	0.43	GP	0.37

We then forecasted inflation from 2007 to 2017 with one-step prediction. We also included one-lagged time series as predictors. The RMSE on the normalized data is given in Table 1. Our method gives the best forecasting accuracy, as indicated by the lowest RMSE.

8. Conclusion

In this paper, we formalized causal discovery and forecasting in nonstationary environments under the framework of nonlinear state-space models. We allowed changes in causal strengths, as well as noise variances. We showed that nonstationarity helps causal model identification, and that causal knowledge improves interpretability and forecasting accuracy. The proposed methods showed promising results on macroeconomic data. As future work, we will extend our methods to cover nonlinear causal relationships, to partially observable processes, as studied in (Geiger et al., 2015), and to causal models with instantaneous cycles.

[1] Downloaded from https://www.theglobaleconomy.com/.

[2] For instance, see https://financialnerd.com/three-pillars-economy-inflation-gdp-unemployment/.

Acknowledgements

We thank Jeff Adams for helping to revise the paper. We would like to acknowledge the supported by National Institutes of Health under Contract No. NIH-1R01EB022858-01, FAINR01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN- U54HG008540, by the United States Air Force under Contract No. FA8650-17-C-7715, and by National Science Foundation EAGER Grant No. IIS-1829681. The National Institutes of Health, the U.S. Air Force, and the National Science Foundation are not responsible for the views reported in this article.

References

Alippi, C. and Roveri, M. Just-in-time adaptive classifiers-part ii: Designing the classifier. In *IEEE Trans. Neural Netw.*, volume 19(12), pp. 2053–2064, 2008.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3: 507–554, 2003.

Dagum, P., Galper, A., and Horvitz, E. Dynamic network models for forecasting. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 41–48, 1992.

Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. In *The Annals of Statistics*, volume 27(1), pp. 94–128, 1999.

Elwell, R. and Polikar, R. Incremental learning of concept drift in nonstationary environments. In *IEEE Trans. Neural Netw.*, volume 22(10), pp. 1517–1531, 2011.

Engle, R. F., Hendry, D. F., and Richard, J. F. Exogeneity. *Econometrica*, 51:277–304, 1983.

Erdős, P. and Rényi, A. On random graphs i. In *Publicationes Mathematicae*, volume 6, pp. 290–297, 1959.

Geiger, P., Zhang, K., Gong, M., Janzing, D., and Schölkopf, B. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of the 32th Conference on International Conference on Machine Learning*, 2015.

Ghassami, A. E., Kiyavash, N., Huang, B., and Zhang, K. Multi-domain causal structure learning in linear systems. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 6269–6279, 2018.

Gibbons, J. D. and Chakraborti., S. Chapman Hall/CRC Press, 2011.

Gong*, M., Zhang*, K., Tao, D., Geiger, P., and Schölkopf, B. Discovering temporal causal relations from subsampled data. In *Proceedings of the 32th Conference on International Conference on Machine Learning*, 2015.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., and Tao, D. Causal discovery from temporally aggregated time series. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. In *Econometrica*, volume 37 (3), pp. 424–438, 1969.

Heckerman, D., Geiger, D., and Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

Hoover, K. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

Hoyer, P., Janzing, D., Mooji, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Neural Information Processing Systems*, Vancouver, B.C., Canada, 2009.

Huang, B., Zhang, K., and Schölkopf, B. Identification of time-dependent causal model: A Gaussian process treatment. In *the 24th International Joint Conference on Artificial Intelligence, Machine Learning Track*, pp. 3561–3568, Buenos, Argentina, 2015.

Huang, B., Zhang, K., Zhang, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Behind distribution shift: Mining driving forces of changes and causal arrows. In *Proceedings of the Conference on IEEE International Conference on Data Mining*, pp. 913–918, 2017.

Huang, B., Zhang, K., Lin, Y., B., S., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, pp. 1551–1560, 2018.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-Gaussianity. In *Journal of Machine Learning Research*, volume 11 (5), pp. 1709–1731, 2010.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

Kalman, R. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

Kolar, M. and Xing, E. P. Estimating networks with jumps. In *Electronic journal of statistics*, volume 6, pp. 2069–2106, 2012.

Korb, K., Hope, L., Nicholson, A., and Axnick, K. *Varieties of Causal Intervention*. Springer, 2004.

Lindsten, F. An efficient stochastic approximation EM algorithm using conditional particle filters. In *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6274–6278, 2013.

Lindsten, F., Jordan, M. I., and Schön, T. B. Ancestor sampling for particle gibbs. In *Proceedings of the Conference on Neural Information Processing Systems*, 2012.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 2016.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, USA, 2006.

Robert, C. and Casella, G. *Monte Carlo Statistical Methods*. Springer, 2004.

Shimizu, S., Hoyer, P., Hyvärinen, A., and Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Song, L., Kolar, M., and Xing, E. P. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems*, pp. 1732–1740, 2009.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics, 1993.

Tian, J. and Pearl, J. Causal discovery from changes: a Bayesian approach. In *Uncertainty in Artificial Intelligence*, pp. 512–521, 2001.

Tibshirani, R. Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society*, volume 58(1), pp. 267–288, 1996.

Wall, K. D. Identification theory for varying coefficient regression models. In *Journal of Time Series Analysis*, volume 8(3), pp. 359–371, 1987.

Zhang, K. and Chan, L. Extensions of ICA for causality discovery in the hong kong stock market. In *Proc. 13th International Conference on Neural Information Processing*, 2006.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 2015a. forthcoming.

Zhang, K., Zhang, J., and Schölkopf, B. Distinguishing cause from effect based on exogeneity. In *Proc. 15th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2015)*, 2015b.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017.