Transferable Time-Series Forecasting under Causal Conditional Shift

Zijian Li, *Member, IEEE,* Ruichu Cai*, *Member, IEEE,* Tom Z. J. Fu and Kun Zhang

Abstract—This paper focuses on the problem of semi-supervised domain adaptation for time-series forecasting, which is an easily neglected but challenging problem due to the changeable and complex conditional dependencies. In fact, these domain-specific conditional dependencies are mainly led by the data offset, the time lags, and the variant data distribution. In order to cope with this problem, we analyze the variational conditional dependencies in time-series data and consider that the causal structures are stable among different domains, and further raise the causal conditional shift assumption. Enlightened by this assumption, we consider the causal generation process for time-series data and devise an end-to-end model for transferable time-series forecasting. The proposed method can not only discover the cross-domain *Granger Causality* but also address the cross-domain time-series forecasting problem. It can even provide the interpretability of the predicted results to some extent. We further theoretically analyze the superiority of the proposed methods, where the generalization error on the target domain is not only bounded by the empirical risks on the source and target domains but also by the similarity between the causal structures from different domains. Experimental results on both synthetic and real data demonstrate the effectiveness of the proposed method for transferable time-series forecasting.

Index Terms—Time Series, Granger Causality, Transfer Learning

1 Introduction

Science is all about generalizations. Lots of scientific applications of multivariate time-series also need to transfer the experiment conclusions from the lab or the virtual environments to different real environments. For example, the physiological mechanism in the human body among "Blood Glucose", " Glucagon" and "Insulin" should be held among people of different ages, genders, and even races. Semi-supervised domain adaptation, which leverages the labeled source domain data and a few labeled target domain data to make a prediction in the unlabeled target domain, essentially aims to address the notorious *domain shift* phenomenon and obtain a robust forecasting model.

Various methods have been proposed for domain adaptation [1], [2], [3], [4], [5], [6], [7], [8], with one of the most ordinary assumptions made. It is the *covariate shift assumption* that the marginal distribution $P(X)$ varies with domains while the conditional distribution $P(Y|X)$ is stable across domains. Based on this assumption, many methods based on MDD [9], [10] or adversarial training [1], [11], [12], which aim to extract the domain-invariant representation, have been proposed and achieving great successes in non-time series data. Considering that the covariate shift scenario may not be satisfied, Kun et.al [13] leverage the thoughts of causality and study three different possible scenarios: the target shift, the conditional shift, and the generalized target shift. Recently, Kun et.al [14] describe the changes of the data distribution by leveraging the Bayesian treatment and propose a causality-based method to address the domain adaptation problem.

Since time-series data is one of the most familiar data types, domain adaptation for time-series data is receiving more and more attentions. Many researchers expand these ideas to the field of time-series data [15], [16] and straightforwardly replace the feature extractors with the architectures like recurrent neural networks [17]. Though making a brave push on this field, these methods implicitly reuse the covariate shift assumption for non-time series data to time-series data. However, because of the complex dependencies of time series, (i.e., even the trivial 1st-order Markov dependence results in the dependence between any two-time steps), the conditional distribution of time-series data $P(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots \boldsymbol{z}_t)$ varies sharply between different domains, making the conventional covariate shift assumption hard to be satisfied in time-series data. Therefore, it is a challenging task to learn the domain-invariant representation in the time-series data.

The covariate shift assumption that is used by the existing time-series domain adaptation methods simply assumes $P^S(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \cdots, \boldsymbol{z}_1) = P^T(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t, \cdots, \boldsymbol{z}_1)$ and essentially considers all the relationships among variables, which is shown in Figure 1 (a). However, according to the covariate distribution shift assumption, a tiny change of $\boldsymbol{z}_1$ leads to a huge change of $\boldsymbol{z}_{t+1}$. Give a easily comprehensive example $z_t^1 = 1.5 \times z_{t-1}^1; z_t^2 = 3 \times z_{t-1}^2 + 2 \times (z_{t-1}^1)^3$.

- *Zijian Li is with the School of Computing, Guangdong University of Technology, Guangzhou China, 510006.E-mail: leizigin@gmail.com*
- *Ruichu Cai is with the School of Computing, Guangdong University of Technology and Guangdong Provincial Key Laboratory of Public Finance and Taxation with Big Data Application, Guangzhou China, 510006. Email: cairuichu@gmail.com*
- *Tom Z. J. Fu is with the School of Computing, Guangdong University of Technology, Guangzhou China, 510006., Email: fuzhengjia@gmail.com*
- *Kun Zhang is with the Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213 USA. E-mail: kunz1@cmu.edu*

*Manuscript received XX; revised XX; accepted XX. Date of publication XX XX, 2019; date of current version XX XX, 2019. Ruichu Cai and Zijian Li was supported in part by Natural Science Foundation of China (61876043, 61976052), Science and Technology Planning Project of Guangzhou (201902010058) and Guangdong Provincial Science and Technology Innovation Strategy Fund (2019B121203012). Tom Fu was partly supported by NSFC under Grant 61702113. (*Ruichu Cai is the Corresponding author.)*

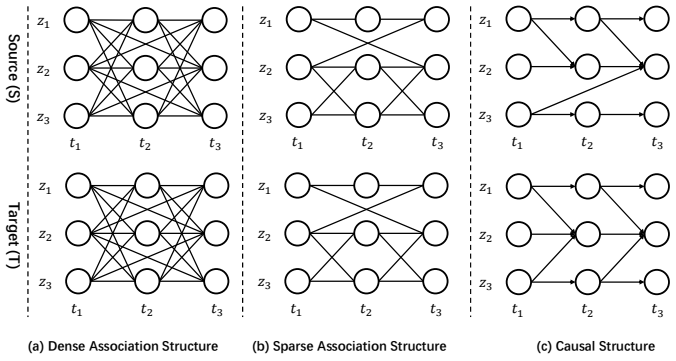


Figure 1. The illustration of various structures of three dimensions time-series data. (a) The existing methods, which reuse the covariate shift assumption, take all the relationships into account and may result in negative transfer or overfitting. (b) Other methods that leverage the sparse associative structures still can not avoid the domain-specific conditional distribution shift. (c) Considering the stable causal generalization process of time-series data, the proposed method addresses the transferable time-series forecasting problem by extracting the causal structures and modeling the domain-specific conditional distribution.

We further let $z_0^1 = 1, z_0^2 = 1$, so $z_5^2 = 208403$. If we change $z_0^1 = 1.1$, then $z_2^2 \approx 277303$. This example shows that even a small modification of z_0^1 may result in the hug change of z_2^2. Therefore, the covariate shift assumption is hard to be satisfied in the time-series data. Other methods like [2] consider that most of the previous time-series domain adaptation methods take all the relationships among variables into account and extract the stable and sparse associative structures among variables, which is shown in Figure 1 (b). However, they still can not avoid the curse of covariate shift in time-series data. What is more, they do not consider the domain-specific strengths among the sparse associative structures. As shown in Figure 1(c), we find that the causal structures between the source and the target domains are stable and compact. Since the generation process of time-series data usually follow the physical rules, which actually denote the causality. Therefore, the causal structures among the source and the target domains are also similar and stable. For example, in the physiological mechanism, the causal structures among 'Blood Glucose", "Glucagon" and "Insulin" are very stable and adequate for everyone. Inspired by the aforementioned observation, it is natural and intuitive to break away the conventional covariate shift assumption and tackle the semi-supervised time-series forecasting domain adaptation problem with the help of causality.

Following the aforementioned intuition, we proposed a solution to semi-supervised transferable time-series forecasting by assuming that the causal structures are similar and stable across domains. We first analyze the drawbacks in the existing time-series domain adaptation methods and raise the **Causal Conditional Shift Assumption**. Under this assumption, we reconsider the time-series forecasting domain adaptation problem by borrowing the idea of the causal generation process of time-series data and model this process with the help of stochastic variational inference. Based on the this insight, we devise the **G**ranger **C**ausality **A**lignment model (**GCA** in short). Technically, we devise the Recurrent Granger Causality Autoencoder to respectively and iteratively discover the Granger Causality with different lags from the source and the target domains. Considering that the strengths of causal structures from different domains are different, we further devise the domain-sensitive Granger Causality inference module. We further propose the Granger Causality similarity regularization, which is beneficial to reconstruct the target causal structures and restrict the similarity of the causal structures from different domains. Moreover, by discovering and aligning the causal structures, the proposed method can not only provide the model interpretability but also benefit the usage of expert knowledge. Our proposed method outperforms the state-of-the-art domain adaptation methods for time-series data on the synthetic and real-world data. Meanwhile, we not only provide theoretically analysis of the advantages of the intrinsic mechanisms, we also provide the visualization and interpretability results.

2 Related Works

In this section, we mainly focus on the existing techniques on domain adaptation on non-time series data, time-series domain adaptation and the Granger Causality.

2.1 Domain Adaptation on Non-Time Series Data

Domain Adaptation [7], [8], [9], [11], [12], [13], which leverages the labeled source domain data and limited labeled or unlabeled target domain data to make predictions in the target domain, have applications in various fields [3], [4], [5], [6]. Most of the approaches of domain adaptation follow the covariate shift assumption and aim to extract the domain-invariant representation. Technically, these methods can be grouped into the Maximum Mean Discrepancy-based methods and the adversarial training methods.

In view of the causal generation process, some researchers find that the conditional distribution $P(Y|X)$ usually changes and the covariate shift assumption does not hold, so they address such limitations from the view of causality. Kun et.al [13], [18] address this problem by assuming only $P(Y)$ or $P(X|Y)$ change and raising target shift and conditional shift. Cai et.al [1], [3] are motivated by the causal generation process and raise disentangled semantic representation for unsupervised domain adaptation. Recently, Kun et al. [14] consider domain adaptation as the problem of graphical model inference and directly model the changes of the conditional distribution. In this paper, we focus on semi-supervised domain adaptation on time-series forecasting, which is more challenging because the complicated conditional dependencies among time stamps are hard to be modeled.

2.2 Domain Adaptation on Time-Series Data

Time-series data is another type of common dataset and there are several pieces of research works about it. Recently, increasing attention is paid to the domain adaptation on time-series data. Da Costa et al. [15] straightforwardly extend the idea of domain adaptation for non-time series data and leverage the RNN as the feature extractor to extract the domain-invariant representation. Purushotham et al. [16] further improve it by using the variational recurrent

neural network [17]. However, these methods are not able to well extract the domain-invariant information because of the complicated dependency between timestamps. Recently, Cai et al. [2] consider that the sparse associate structure of variables are stable across domains, so they propose the sparse associative structure alignment methods for adaptative time series classification and regression tasks. Despite of achieving better performance, the sparse associate structure alignment method not only misses the domain-specific substructures (the domain-invariant substructures are caused by the time-lags.) but also is swamped with the variant strength of association. In order to address these problems, we assume that the causal structures are stable across domains and model how the conditional distribution changes.

2.3 Granger Causality

Granger Causality [19], [20], [21], [22], [23], [24], which is a set of directed dependencies among multivariate time-series data, is widely used to determine which of the past portion of a univariate time-series data aids in predicting the future evolution of another univariate time-series. Inferring Granger Causality is a traditional research problem and has been applied in several fields [25], [26], [27]. One of the most classical methods for inferring Granger Causality is the vector autoregressive (**VAR**) model [28], [29], which utilizes the linear time-lag effects as well as some sparsity regularizer like the Lasso [30] or the group lasso [31]. With the quick growth of the computation power, more Granger Causality inference methods that borrow the expressive power of neural networks have been proposed. Tank et al. [23] devise a neural network based autoregressive model and apply the sparsity penalties on the weights of the neural networks. [22] are motivated by the interpretability of self-explaining neural networks and propose the generalized vector autoregression model to detect the signs of Granger Causality. In this paper, we are motivated by the generation process of time-series data and take the Granger-causal structure as the latent variables. We combine the variational inference framework into the vector autoregression and further used it for semi-supervised domain adaptation of time-series forecasting.

3 Semi-Supervised Time-Series Domain Adaptation Model Under Causal Conditional Shift Assumption

In this section, we first give a definition of the time-series generation process based on Granger Causality. Based on this process, we further define the problem of domain adaptation of time-series forecasting. Finally, we will provide insights and the main idea of the proposed method.

3.1 Time-Series Generation Process under Granger Causality

In order to distinctly illustrate the time-series causal generation process, we first consider the discrete multivariate time-series data with the length of Υ as $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_t, \cdots \boldsymbol{z}_\Upsilon)$, in which $\boldsymbol{z}_t \in \mathbb{R}^D$ are variables with D dimensions. With the abuse of notation, we further let $\boldsymbol{z}^d$ be the univariate time-series of d-th variables and z_t^d be the value of d-th variables at t-th timestamp. Intuitively, according to the real-world observation, we usually observe that the future values of univariate time-series data are not only influenced by its past values but also decided by the other time-series.

In fact, the data generation processes are usually controlled by a stable causal mechanism among variables. In order to describe the data generation process in the causal view, the nonlinear Granger Causality [19], [20], [21], [22], [23], [24], [32] is applied for modelling the dependencies between the child nodes and their parents.

Given the observed time-series $\boldsymbol{z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t, \cdots, \boldsymbol{z}_\Upsilon)$, each value of z_t^d can be represented in terms of some functions F_d and its parents $\boldsymbol{pa}(z_t^d)$, and we let the nonlinear Granger Causality follow the structural equation model shown as Equation (1).

$$z_t^d = F_d(\boldsymbol{pa}(z_t^d), \epsilon_d), \tag{1}$$

where ϵ_d are the independent noise terms and $F_d(\cdot)$ denotes any flexible types of nonlinear function. Actually, Equation (1) specifies that how the future variables $\boldsymbol{z}^i$ rely on the past values of $\boldsymbol{z}$, and $\boldsymbol{pa}(\cdot)$ essentially coincides the nonlinear Granger Causality.

Based on the aforementioned Granger-causal generation process for time-series data, it is natural to find that the value of z_t^d actually relies on all the previous observation $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{t-1})$. Therefore, small changes on $\boldsymbol{z}_1$ (e.g., time lags, value ranges) will result in violent changes on z_t^d even when the trivial 1st-order Markov dependence exists in any two timestamps.

3.2 Semi-Supervised Domain Adaptation for Time-Series Forecasting

Due to the notorious phenomenon named domain shift, the observed time-series data violate the independent and identically distributed (**I.I.D**) assumption. In order to mitigate this problem, domain adaptation on time-series data is proposed. We first respectively let the past observation and the future ground truth be $\boldsymbol{x} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t)$ and $\boldsymbol{y} = (\boldsymbol{z}_{t+1}, \cdots, \boldsymbol{z}_{t+\tau})$. It is noted that $\boldsymbol{y}$ can be the multivariate time-series or univariate time-series, we let $\boldsymbol{y}$ be multivariate time-series for generalization. We further assume that $\{(\boldsymbol{x}_i^S, \boldsymbol{y}_i^S)\}_{i=0}^N$ and $\{(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)\}_{i=0}^M$ are respectively independent and identically distributed (I.I.D.) drawn from the source domain $P^S(\boldsymbol{x}, \boldsymbol{y})$ and the target domain $P^T(\boldsymbol{x}, \boldsymbol{y})$, where M and N are the numbers of random samples drawn from the source and the target domains respectively, and we have $N \gg M$. Semi-supervised Domain adaptation for time-series forecasting aims to find the model that performs well on the dataset drawn from the target distribution by leveraging the labeled source domain data and the limited target domain data.

Existing methods on domain adaptive time-series forecasting straightforwardly follow the idea of the conventional domain adaption methods for non-time series data and assume that the conditional distribution $P(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t)$ remain fixed. However, according to the aforementioned Granger-causal generation process for time-series data, since $\boldsymbol{z}_{t+1}$ depends on $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t)$, a little

change of $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t)$, the little value change of $\boldsymbol{z}_1$ may magnify or minify the influence to $\boldsymbol{z}_{t+1}$ by the stable causal relationships, which makes the conditional distribution from different domains do not fix. Therefore, the covariate shift assumption, which is shown in Equation (2), is very difficult to be satisfied in time-series data.

$$\begin{aligned} P^S(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t) &= P^T(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t) \\ P^S(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t) &\neq P^T(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t). \end{aligned} \tag{2}$$

In order to bypass the aforementioned difficulties, one straightforward solution is to replace the covariate shift assumption with the "parent shift assumption", which is shown in Equation (3). This assumption can mitigate the accumulated changes from $\boldsymbol{z}_1$ and essentially hints that the future values only depend on their direct parents.

$$\begin{aligned} &P^S(\boldsymbol{z}_{t+1}|\boldsymbol{pa}(\boldsymbol{z}_1), \cdots, \boldsymbol{pa}(\boldsymbol{z}_t)) = \\ &P^T(\boldsymbol{z}_{t+1}|\boldsymbol{pa}(\boldsymbol{z}_1), \cdots, \boldsymbol{pa}(\boldsymbol{z}_t)) \\ P^S(\boldsymbol{pa}(\boldsymbol{z}_1), \cdots, \boldsymbol{pa}(\boldsymbol{z}_t)) &\neq P^T(\boldsymbol{pa}(\boldsymbol{z}_1), \cdots, \boldsymbol{pa}(\boldsymbol{z}_t)). \end{aligned} \tag{3}$$

Though mitigating the butterfly effect to some extent, the non-negligible dependencies still exist in time-series data with the aforementioned assumption, few changes of the parents $\boldsymbol{pa}(\boldsymbol{z}_t)$ like the different time-lag and value ranges still results in palpable variation of the future values, which further leads to the unsatisfactory of Equation (3).

Fortunately, given any domains, the causal mechanism among variables are stable, while the conditional distributions are easily influenced by the different value ranges, time lags, and offsets. An intuitive example is that the physiological mechanism in each person is stable. Therefore, with the inspiration of the domain-invariant causal mechanism between the source and the target domain, we propose the following **causal conditional shift assumption**.

Assumption 1. ***causal conditional shift assumption:*** *Given causal structures A, $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t$, and $\boldsymbol{z}_{t+1}, \cdots, \boldsymbol{z}_{t+\tau}$, we assume that the conditional distribution $P(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t)$ varies with different domains while the causal structures remain fixed. Formally, it can be shown as Equation (4).*

$$\begin{aligned} P^S(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t) \neq & P^T(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t), \\ \boldsymbol{A}^S = & \boldsymbol{A}^T \end{aligned} \tag{4}$$

in which $\boldsymbol{A}^S$ and $\boldsymbol{A}^T$ are the causal structures from the source and the target domain.

Based on the aforementioned **causal conditional shift assumption**, we can find that the key of semi-supervised domain adaptation from time-series forecasting is to leverage the domain-invariant causal structures to model the domain-specific conditional distribution. In summary, given the labeled source domain data and the limited labeled target domain data, our goal is to (1) discover the causal structures of time-series data from the source and the target domains; (2) forecast the future time-series data over the target distribution with the help of Granger causality inference.

However, the causal structures are usually changed by other domain-specific along with different domains, which further leads to the disobedience of the **causal conditional shift assumption**. For example, the responding time between "Blood Glucose" and "Insulin" of the elder is longer than that of the young, which leads to the different causal structures. In order to address this issue, we aim to extract the common causal structures between the source and the target domains.

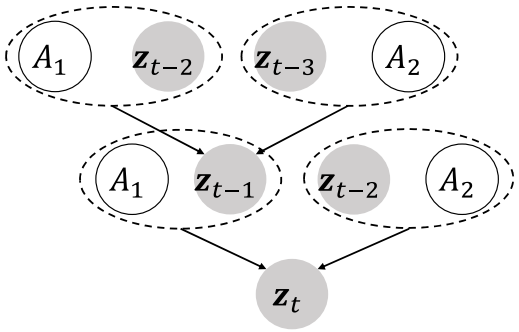


Figure 2. The illustration of how time-series data is iteratively generated via Granger-causal structures with a maximum lag of two. (The parent of $\boldsymbol{z}_{t-2}$ are omissive due to the limited space.) A_1, A_2 are the substructures of Granger-causal structure with the lag of one and two. The Granger-causal structures are stable in all the timestamps. $\boldsymbol{z}$ is the observed data of different timestamps.

Therefore, the objective function of the semi-supervised domain adaptation for time-series forecasting can be summarized as follow:

$$\begin{aligned} \min \mathbb{E}_{\boldsymbol{A}^S} P\left(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t, \boldsymbol{A}^S\right) \\ + \mathbb{E}_{\boldsymbol{A}^T} P\left(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t, \boldsymbol{A}^T\right) + Disc(\boldsymbol{A}^S, \boldsymbol{A}^T), \end{aligned} \tag{5}$$

where $Disc(\boldsymbol{A}^S, \boldsymbol{A}^T)$ is the similarity constraint of the Granger-causal structures between the source and the target domains and used to extract the domain-invariant causal structures. Note that $\mathbb{E}_{\boldsymbol{A}^*} P\left(\boldsymbol{z}_{t+1}|\boldsymbol{z}_1, \cdots, \boldsymbol{z}_t, \boldsymbol{A}^*\right))$ can be considered as an autoregression of time-series data.

4 Algorithm and Implement

According to the aforementioned objective function, we aim to simultaneously discover the Granger-causal structures from the different domains and predict the future values with the help of causal inference. Therefore, we follow the philosophy of Granger Causality and intend to address the semi-supervised domain adaptation problem for time-series forecasting by discovering and inferring the Granger Causality.

In order to extract the Granger causality from the observed time-series data, we start with the graphical model of the causal generation process for time series data, which is shown in Figure. 2. First, we let $\boldsymbol{A} = (A_1, \cdots, A_j, \cdots, A_k)$ be the causal structures with the maximum lag of k, in which the italic symbol A_j denotes the local substructures with the lag of j. Given the observed data $\boldsymbol{z}_t$, it is respectively generated from the $\boldsymbol{z}_{t-1}$ and $\boldsymbol{z}_{t-2}$ under the substructure of Granger-causal graph whose lag is one and two (We assume that the maximum lag is two for convenient illustration.). Furthermore, $\boldsymbol{z}_{t-1}$ and $\boldsymbol{z}_{t-2}$ are iteratively generated from the previously observed data under the same substructures of Granger Causality.

In the light of the power of variational autoencoder (VAE) [33], [34] in reconstructuring the latent variables, we

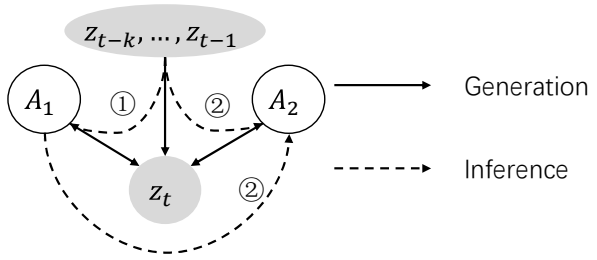


Figure 3. The illustration of the proposed RGC-VAE model with the maximum lag of 2. The solid lines denote the generation process and the dashed lines denote the inference process. The inference process iteratively reconstructs the Granger-Causal structures. In this toy example, the inference process consists of two steps. Step ① is used to infer A_1 and step ② is used to infer A_2 based on A_1. The generative process forecasts the future data with the help of the observed data and the reconstructed causal structures.

consider the Granger-causal structures as the latent variables and propose an extension of VAE, namely, Recurrent Granger Causality Variational Autoencoder (RGC-VAE), to iteratively reconstruct the Granger-causal structures with different lags, which is shown in Figure 3. As the generation process shown in Figure 3, since each component in causal structures is independent, we can find that $(A_1, \cdots, A_k)$ are independent of each other. As for the inference process in Figure 3, two steps are used to respectively reconstruct A_1 and A_2. In step ①, $\boldsymbol{z}_t, A_1$ and $(\boldsymbol{z}_{t-k}, \cdots, \boldsymbol{z}_{t-1})$ consist of a "V-structure", so we can reconstruct A_1 given $\boldsymbol{z}_t$ and $(\boldsymbol{z}_{t-k}, \cdots, \boldsymbol{z}_{t-1})$. We also find that A_1, A_2 and $(\boldsymbol{z}_{t-k}, \cdots, \boldsymbol{z}_{t-1})$ are not independent give $\boldsymbol{z}_t$. Therefore, we can infer A_2 given $A_1, \boldsymbol{z}_t$ and $(\boldsymbol{z}_{t-k}, \cdots, \boldsymbol{z}_{t-1})$ in the ② step.

Based on the data generation process shown in Figure 2, we derive the evidence lower bound (ELBO). Assuming the maximum lag is k, the logarithm of joint likelihood $lnP(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, \cdots, \boldsymbol{z}_{t-k})$ can be written as follows:

$$
\begin{aligned}
lnP(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, \cdots, \boldsymbol{z}_{t-k}) &= \mathcal{L}_{ELBO} + \sum_{j=1}^{k} D_{KL}(Q_j||P_j) \\
Q_j &= Q(A_j|A_1, \cdots, A_{j-1}, \boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}) \\
P_j &= P(A_j|A_1, \cdots, A_{j-1}, \boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}),
\end{aligned}
\tag{6}
$$

in which the second term is the summation of the Kullback–Leibler divergence between the approximate distribution and the true posterior. In practice, inferring A_j given $(A_1, \cdots, A_{j-1})$ benefits to obtain the sparse and precise causal structures. If we estimate the causal structures in the form like $p(A_j|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k})$, some overlapped and redundant edges might be brought in.

Based on Equation. (6) we can derive the variational lower bound $\mathcal{L}_{ELBO}$, which is shown as follows:

$$
\begin{aligned}
\mathcal{L}_{ELBO} &= \mathbb{E}_{Q_1, \cdots, \mathbb{E}_{Q_k}} \left[\ln P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}, A_1, \cdots, A_k)\right] \\
&- D_{KL}(Q_1||P_1) - \sum_{j=2}^{k} \mathbb{E}_{Q_1} \cdots \mathbb{E}_{Q_{j-1}} D_{KL}(Q_j||P_j),
\end{aligned}
\tag{7}
$$

in which $Q_j = Q(A_j|A_1, \cdots, A_{j-1}, \boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k})$ is the inference model and used for reconstructing the Granger-causal substructure with the lag j and assumed to follow the Bernoulli distribution. Similar to variational inference for categorical latent variables [34], we use Gumbel-Softmax trick to estimate the Granger-causal structures. (More details are shown in Appendix A.)

Combing the event lower bound in Equation (7) and the objective function in Equation (5), we can further derive the loss function based on the causal time-series generation process, which is shown as follow:

$$
\begin{aligned}
&\mathcal{L}_{total} = \\
&\mathbb{E}^S_{A_1 \sim Q_1} \cdots \mathbb{E}^S_{A_k \sim Q_k} \left[\ln P^S\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}, A^S_1, \cdots, A^S_k\right)\right] \\
&- D_{KL}\left(Q^S_1||P^S_1\right) - \sum_{j=2}^{k} \mathbb{E}_{Q^S_1} \cdots \mathbb{E}_{Q^S_{j-1}} D_{KL}\left(Q^S_j||P^S_j\right) \\
&+ \mathbb{E}^T_{A_1 \sim Q_1} \cdots \mathbb{E}^T_{A_k \sim Q_k} \left[\ln P^T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}, A^T_1, \cdots, A^T_k\right)\right] \\
&- D_{KL}\left(Q^T_1||P^T_1\right) - \sum_{j=2}^{k} \mathbb{E}_{Q^T_1} \cdots \mathbb{E}_{Q^T_{j-1}} D_{KL}\left(Q^T_j||P^T_j\right) \\
&+ \lambda \left(\sum_{j=1}^{k} \mathbb{E}_{A^S_j \sim Q^S_j} \mathcal{R}(A^S_j) + \sum_{j=1}^{k} \mathbb{E}_{A^T_j \sim Q^T_j} \mathcal{R}(A^T_j) \right) \\
&+ \gamma \sum_{j=1}^{k} \mathbb{E}_{Q^S_j} \mathbb{E}_{Q^T_j} Disc\left(A^S_j, A^T_j\right),
\end{aligned}
\tag{8}
$$

in which λ and γ are the hyper-parameters; Q^S_j and Q^T_j work as the inference model and approximates the prior distribution and further respectively discover the Granger-causal structures from the source and the target domains; the conditional distribution $P^S\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}, A^S_1, \cdots, A^S_k\right)$ and $P^T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \cdots, \boldsymbol{z}_{t-k}, A^T_1, \cdots, A^T_k\right)$ work as the generative model (as shown in Figure 3(c)) and are used to predict the future values with the help of Granger causality inference; and $\mathcal{R}(\cdot)$ is the sparsity-inducing penalty term. Therefore, we can simultaneously discover the Granger causality and infer the future values.

The framework of the proposed disentangled Granger Causality method is shown in Figure. 4. The proposed method is built on the causal generation process for the time-series data and we further devise it by extending extend the variational autoencoder for categorical latent variables [33], [34] to the Recurrent Granger Causality Variational Autoencoder (RGC-VAE). As shown in Figure. 4 (a), we first propose the **Recurrent Granger Causality Reconstruction** module, which is used to discover the Granger-causal structures with different lags with the observed time-series data from the source and the target domains. Moreover, we raise the **Domain-Sensitive Granger Causality Inference** module as shown in Figure. 4 (b) to model how the condition distribution changes across different domains with the help of Granger-causal structures and the historical observed data. In order to restrict the similarity of Granger Causality between the source and the target domain, we further employ the Granger Causality similarity regularization, which is shown in Figure. 4 (c). The implementation details will be introduced in the following subsections.

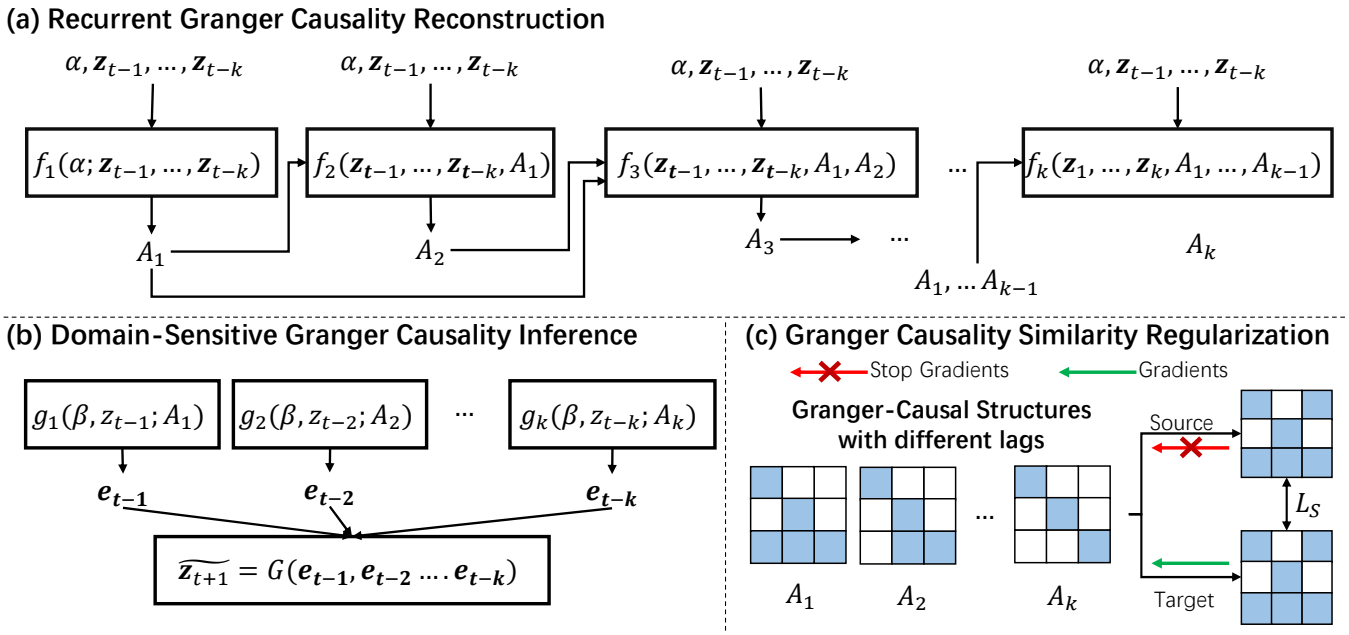


Figure 4. The framework of the proposed Granger Causality Alignment (GCA) model. (a) The recurrent Granger-causal structures of different lags reconstruction. α denotes the structure domain latent variables that are used to reconstruct the domain-specific Granger Causality. (b) The domain-sensitive Granger Causality inference process aims to model the domain-specific conditional shift under the Granger Causality. β is the conditional distribution latent variables. (c) The summary graph of Granger Causality alignment between the source and the target domain.

4.1 The Implementation of Recurrent Granger Causality Reconstruction

In this subsection, we intend to provide the details of implementation of the Recurrent Granger Causality Reconstruction. As mentioned in Equation (5), we aim to discover the Granger-causal structures from the source and the target domains by separately modeling $P^S\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1^S,\cdots,A_k^S\right)$ and $P^T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1^T,\cdots,A_k^T\right)$. It is noted that the reconstructed causal structures do not contain the strengths information and we let the strength parameters be learned in the generative model. This is because the strength parameters are usually flexible and vary with different while the causal structures are stable. If we estimate and align the causal structures with strength parameters, the strength parameters will tend to become similar, which will further lead to the suboptimal results.

According to Equation (7), it is interesting to find that the process of reconstructing the Granger-causal substructures with different lags is in a recurrent form. Therefore, we follow the iterative property and first reconstruct A_1, the Granger-causal structures whose lag is 1 then we can reconstruct A_2 of the Granger-causal structures whose lag is 2 since $Q_2 = Q(A_2|A_1,\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})$ is based on A_1, and so on so forth. By parity of reasoning, the other substructures with larger lags follow the same rule.

Technically, for $Q(A_j^*|A_1^*,\cdots,A_{j-1}^*,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k})$, we implement it the with MLPs and the categorical reparameterization trick, where we let A_j^* be the substructure with the lag of j from the source or the target domains. In details, we use Equation (9) as the universal approximator of $Q(A_j^*|A_1^*,\cdots,A_{j-1}^*,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k})$ to generate A_j^* from the previous observed data. Consequently, we implement $Q(A_j^*|A_1^*,\cdots,A_{j-1}^*,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k})$ in a functional form like Equation. (9).

$$A_j^* = f_j(A_1^*,\cdots,A_{j-1}^*,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k};\theta_j), \tag{9}$$

in which θ_j are the trainable parameters. For convenience, We further let $\Theta = (\theta_1,\cdots,\theta_j)$ be all the trainable parameters of the recurrent Granger Causality reconstruction module.

However, the causal structures may vary with different domains due to the domain-specific time lags. For example, the responding time between "Blood Glucose" and "Insulin" of the elder is longer than that of the young, which leads to the different time lags of the causal relationship between "Blood Glucose" and "Insulin". One straightforward solution is to use two different models with the same architecture for the source and the target domains. However, since the labeled target domain data is limited, it tends to result in suboptimal results on the target domain. In order to address this problem, we further introduce the trainable structural domain latent variables $\alpha^S,\alpha^T \in \mathbb{R}^{d_\alpha}$ to separately assist to model the domain-specific module of the Granger Causality, and d_α is the dimension of the trainable structural domain latent variables. Therefore, we extend Equation (9) to Equation (10) shown as follows:

$$\begin{aligned} A_j^S &= f_j(A_1^S,\cdots,A_{j-1}^S,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k},\alpha^S;\theta_j) \\ A_j^T &= f_j(A_1^T,\cdots,A_{j-1}^T,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{t-k},\alpha^T;\theta_j). \end{aligned} \tag{10}$$

4.2 The Implementation of Domain-sensitive Granger Causality Inference

In order to predict the future time-series data with the help of Granger Causality inference, we further devise the "Domain-Sensitive Granger Causality Inference" architecture, which is shown in Figure 4 (b). The Granger Causality

inference process can be separated into two steps: the intra-lag inference step and the inter-lag inference step.

In the intra-lag inference step, we aim to calculate the effectiveness of the substructures of each time lags. In details, given the j-th lag Granger-causal structure A_j^*, we first use A_j^* to mask the observed input $\boldsymbol{z}_{t-j}$, e.g., $\boldsymbol{z}_j \odot A_j^*$. Then we use MLP in the form of $g_j(\boldsymbol{z}_j \odot A_j^*, A_j^*; \phi_j)$ to calculate e_{t-j}, the effectiveness of A_j^*. And ϕ_j are the trainable parameters. In the inter-lag step, we aim to aggregate the effectiveness of all the time lags and predict the future data. In detail, we simply use another MLP-based architecture in the form of $G(e_{t-1}, e_{t-2}, \cdots, e_{t-k}; \phi_a)$ to predict the future value and ϕ_a are the trainable parameters.

However, the strengths of the causal structures might be different even the source and the target domains share the same causal structures. For example, the effectiveness of "Blood Glucose" and "Insulin" between the male and the female may be different even the age of these two population are similar. In order to cope with these difference, we further introduce another trainable domain-sensitive latent variables $\beta^S, \beta^T \in \mathbb{R}^{d_\beta}$ to model the different strengths of the causal structures, in which d_β is the dimension of the trainable domain-sensitive latent variables. Therefore, we can formulate the domain-sensitive Granger Causality inference as follows:

$$\begin{aligned}\widetilde{\boldsymbol{z}_{t+1}^*} = G\big(&g_1\left(z_1 \odot A_1, A_1; \phi_1\right), \cdots, \\ &g_j\left(z_j \odot A_j, A_j; \phi_j\right), \cdots, \\ &g_k\left(z_k \odot A_k, A_k; \phi_k\right); \beta^*\big),\end{aligned} \tag{11}$$

where $\widetilde{\boldsymbol{z}_{t+1}^*}$ is the predicted result of source or target domain. For convenience, we let $\Phi = (\phi_a, \phi_1, \cdots, \phi_k, \beta^S, \beta^T)$ be all the trainable parameters in the domain-sensitive Granger Causality inference module.

4.3 The Implementation of Granger Causality Similarity Regularization

4.3.1 A Relaxed Assumption for Domain-variant Causal Structures

According to Assumption 1, we assume that the causal structures are stable across different domains. In fact, this assumption is too strong to some extent, because the causal structures vary with different domains. For example, the corresponding time between "Blood glucose" and "Glucagon" of the young may be shorter than that of the old, which reflects that the causal structures from different domains are slightly different. If we directly forecasting the target values with the help of the source causal structures, the suboptimal performance may be achieved. Therefore, we relax the aforementioned assumption and assume that the causal structures from different domains are slightly different. Based on this relaxed assumption, a regularization term is desired, which can utilize the labeled source data to learn the domain-invariant components and exploit the limited labeled target data to learn the domain-specific components. There are many methods to achieve this goal, and we employ L1 Norm in our implementation. Formally, we have:

$$\mathcal{L}_r = \frac{1}{D \times D \times k} \sum_{j=1}^{k} ||A_j^S - A_j^T||_1, \tag{12}$$

where A_j^S and A_j^T are the structures with lag j and $||\cdot||_1$ denotes the L1 Norm.

4.3.2 Causal Structure Similarity Regularization with Gradient Stopping

In the ideal case, the source structures are learned by the source labeled data while the target structures are optimized by the aforementioned regularization as well as the limited target labeled data. However, straightforwardly using L1 Norm can only make the causal structures from different domains become similar, and it can not guarantee that the model can learn correct domain-invariant components and may learn degenerated causal structures. In the worst case, the target structures are solely learned with insufficient labeled data and wrong. In the meanwhile, the source structures are totally poisoned by the regularization term and become become similar to the wrong target structures, which results in the negative transfer and finally the suboptimal results.

In order to address the aforementioned drawback, we further apply the gradient stopping operation $C(\cdot)$ [35], [36] on the source domain causal structures (as shown in Figure .4 (c)), which is devised to stop the gradients of regularization term propagating through the source structures estimation module. Therefore, it is confident to learn the correct source causal structures with the massive source labeled data and without the obstruction from the target causal structures. Furthermore, the target causal structures are guided by the similarity term and the well trained source causal structures. In this way, we let the source structures be a anchor and push the target structures to approach the source ones, so we can avoid the negative transfer and achieve the ideal case.

Therefore, we combine the gradient stopping operation with Equation (12) and the Granger Causality similarity regularization term can be formulated as follows:

$$\mathcal{L}_r = \frac{1}{D \times D \times k} \sum_{j=1}^{k} ||C(A_j^S) - A_j^T||_1, \tag{13}$$

where $C(\cdot)$ is used to stop the gradients from propagating to the source causal structure estimator.

4.4 Training and Inference

By reconstructing the Granger Causality and further predict the future value by inferring the Granger Causality, we summarize the model as follows.

During the training steps, we take the labeled source data $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X}_S, \mathcal{Y}_S)$ and the few labeled data from the target domain $(\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X}_T, \mathcal{Y}_T)$ into the proposed method and predict the future value of the next timestamps. In the testing steps, we need to predict the future τ stamps, so we use the previous predict value as the new input and obtain $\boldsymbol{z}_{t+1}, \cdots, \boldsymbol{z}_{t+\tau}$ in the autoregressive way.

If the dimension of $\boldsymbol{y}$ equals to that of $\boldsymbol{x}$ (e.g. the case of human motion prediction), the proposed **GCA** method need to predict the multivariate time-series data. If the predicted results $\boldsymbol{y}$ is the univariate time-series (e.g. PM2.5 Prediction), it is not necessary to take all the predicted time-series into the loss function. In this case, we can directly use Equation. (8) to optimize the proposed method.

If the predicted results is the univariate time-series $\boldsymbol{z}^d$ and we need to estimate the causal structure $\boldsymbol{A}$ for interpretability, we need to take all the predicted time-series into the loss function. However, the performance will easily suffer from degeneration, because the model might optimize other variables that is unrelated to $\boldsymbol{y}$. In order to address this problem, we further devise an extra strengthen loss. We formulate the extra optimization term follow:

$$\mathcal{L}_e = MSE(\widetilde{\boldsymbol{z}^d_{t+1}}, \boldsymbol{z}^d_{t+1}), \tag{14}$$

in which $\widetilde{\boldsymbol{z}^d_{t+1}}$ is the predicted time-series and $\boldsymbol{z}^d_{t+1}$ is the groud truth value; And MSE denotes the mean square error. We will investigate the effectiveness of the extra optimization term $\mathcal{L}_e$ in the ablation studies. We further use the elastic-net-style penalty term [37] for $\mathcal{R}(\cdot)$, which can be formulated as:

$$\mathcal{R}(A) = \frac{1}{2}||A||_1 + \frac{1}{2}||A||_2, \tag{15}$$

in which $||\cdot||_1$ and $||\cdot||_2$ respectively denote the L1 and L2 Norm.

Under the above objective functions, the total loss of the proposed GCA method can be summarized as the following equation:

$$\begin{aligned}
&\mathcal{L}_{total} = \\
&\mathbb{E}^S_{A_1\sim Q_1}\cdots\mathbb{E}^S_{A_k\sim Q_k}\left[\ln P^S\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A^S_1,\cdots,A^S_k\right)\right] \\
&-D_{KL}\left(Q^S_1||P^S_1\right) - \sum_{j=2}^{k}\mathbb{E}_{Q^S_1}\cdots\mathbb{E}_{Q^S_{j-1}}D_{KL}\left(Q^S_j||P^S_j\right) \\
&+\mathbb{E}^T_{A_1\sim Q_1}\cdots\mathbb{E}^T_{A_k\sim Q_k}\left[\ln P^T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A^T_1,\cdots,A^T_k\right)\right] \\
&-D_{KL}\left(Q^T_1||P^T_1\right) - \sum_{j=2}^{k}\mathbb{E}_{Q^T_1}\cdots\mathbb{E}_{Q^T_{j-1}}D_{KL}\left(Q^T_j||P^T_j\right) \\
&+\lambda\left(\sum_{j=1}^{k}\mathbb{E}_{A^S_j\sim Q^S_j}\mathcal{R}(A^S_j) + \sum_{j=1}^{k}\mathbb{E}_{A^T_j\sim Q^T_j}\mathcal{R}(A^T_j)\right) \\
&+\gamma\sum_{j=1}^{k}\mathbb{E}_{Q^S_j}\mathbb{E}_{Q^T_j}Disc\left(A^S_j,A^T_j\right) + \delta\mathcal{L}_e,
\end{aligned} \tag{16}$$

in which γ, δ and λ are the hyper-parameters.

In summary, the proposed model is trained on the labeled source and target domain data using the following procedure:

$$(\hat{\Theta}, \hat{\alpha^S}, \hat{\alpha^T}, \hat{\Phi}, \hat{\beta^S}, \hat{\beta^T}) = \underset{\Theta,\alpha^S,\alpha^T,\Phi,\beta^S,\beta^T}{\arg\min}\mathcal{L}_{total}. \tag{17}$$

5 Theoretical Insights

In this section, we discuss the theoretical insights and establish the generalization bound, with respect to the similarity of causal structures between the source and the target domain.

There are several works about the generalization theory of domain adaptation [8], [38], [39], [40], [41], but many of them require that the loss function is bounded by a positive value. In this paper, we aim to establish the generalization bound that bridges the causal structures and the error risk. We start with the relationship between the distance of the source and the target causal structures, and the error risks. As we focus on the time-series forecasting task, we let the loss function be the following form: $L : \boldsymbol{y}\times\boldsymbol{y}\boldsymbol{1} \to \mathbb{R}_+$, and it can be MSE, MAPE and so on.

We let $\mathcal{H}$ be a hypothesis space that maps $(\boldsymbol{z}, \boldsymbol{A})$ to $\mathbb{R}$, and $h \in \mathcal{H}, h : (\boldsymbol{z}, \boldsymbol{A}) \to \mathbb{R}$ denotes a function take a time-series sample $\boldsymbol{z}$ and a causal structures $\boldsymbol{A}$ as input. We further let $h^S : (\boldsymbol{z}, \boldsymbol{A}^S) \to \mathbb{R}$ and $h^T : (\boldsymbol{z}, \boldsymbol{A}^T) \to \mathbb{R}$ be the function that respectively take the source and target causal structures as input, and $h^S, h^T \in \mathcal{H}$. Therefore, we can make the following assumption:

Assumption 2. ***Causal Structures Similarity Bound Assumption****: Given any two hypotheses h^S, h^T and the causal structures $\boldsymbol{A}^S, \boldsymbol{A}^T$ from the source and target domain, a positive value K exits that make the following two inequalities hold:*

$$\begin{aligned}
\mathcal{L}^S(h^S, h^T) &= \mathbb{E}_{\boldsymbol{x}\sim P^S(\boldsymbol{x})}L(h^S, h^T) \le K||\boldsymbol{A}^S - \boldsymbol{A}^T||_1, \\
\mathcal{L}^T(h^S, h^T) &= \mathbb{E}_{\boldsymbol{x}\sim P^T(\boldsymbol{x})}L(h^S, h^T) \le K||\boldsymbol{A}^S - \boldsymbol{A}^T||_1.
\end{aligned} \tag{18}$$

This assumption implies that for any samples from any domains, the expectation of $L(h^S, h^T)$ between the h^S and h^T are bounded by the similarity between $\boldsymbol{A}^S$ and $\boldsymbol{A}^T$ over any domains. This assumption is reasonable when the loss function L is local smoothness.

Before introducing the generalization bound of the proposed method, we first give the definition of *Rademacher Complexity* and *Rademacher Complexity Regression Bounds*, which is shown as follow:

Definition 1. ***Rademacher Complexity.*** *[42] Let $\mathcal{H}$ be the set of real-value functions defined over a set X. Given a dataset $\mathcal{X}$ whose size is m, the empirical Rademacher Complexity of $\mathcal{H}$ is defined as follows:*

$$\widehat{\mathfrak{R}}_{\mathcal{X}}(\mathcal{H}) = \frac{2}{m}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}|\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)|\,\middle|\,\mathcal{X} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m)\right], \tag{19}$$

in which $\sigma_=(\sigma_1,\cdots,\sigma_m)$ are independent uniform random variables taking values in $\{-1,+1\}$. And the Rademacher Complexity of a hypothesis set $\mathcal{H}$ is defined as the expectation of $\mathfrak{R}_{\mathcal{X}}(\mathcal{H})$ over all the dataset of size m, which is shown as follows:

$$\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_\zeta\left[\widehat{\mathfrak{R}}_{\mathcal{X}}(\mathcal{H})\middle|\,|\zeta| = m\right]. \tag{20}$$

Lemma 5.1. ***(Rademacher Complexity Regression Bounds)*** *[39] Let $L : \boldsymbol{y}\times\boldsymbol{y} \to \mathbb{R}_+$ be a non-negative loss upper bounded by $M > 0$ ($L(\boldsymbol{y}, \boldsymbol{y'}) \le M$ for all $\boldsymbol{y}, \boldsymbol{y'} \in \mathcal{Y}$) and we denote by $\eta^{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$ the labeling function of domain $\mathcal{D}$. Given any fixed*

$\boldsymbol{y}' \in \mathcal{Y}, \boldsymbol{y} \to L(\boldsymbol{y}, \boldsymbol{y}')$ is $\mu-$Lipschitz for some $\mu > 0$, and the Redemacher Complexity Regression Bounds are shown as follows:

$$\begin{aligned}\mathcal{L}^{\mathcal{D}}(h^{\mathcal{D}}, \eta^{\mathcal{D}}) \leq & \frac{1}{m^{\mathcal{D}}} \sum_{i=1}^{m^{\mathcal{D}}} L(h^{\mathcal{D}}(\boldsymbol{x}_i^{\mathcal{D}}), \boldsymbol{y}_i^{\mathcal{D}}) + 2\mu \mathfrak{R}_{m^{\mathcal{D}}}(\mathcal{H}) \\ & + M\sqrt{\frac{\log \frac{2}{\sigma}}{2m^{\mathcal{D}}}},\end{aligned} \tag{21}$$

in which $m^{\mathcal{D}}$ is the dataset of domain $\mathcal{D}$ with the size of $m^{\mathcal{D}}$.

Based on the aforementioned definition and assumption, we proposed the generalization bound of the proposed **GCA** method, which is shown as follows.

Theorem 5.2. *We first let η^S and η^T be the source and target labeling function. We further let h^* be the ideal predictor that simultaneously obtains the minimal error on the source and the target domain. Formally, we have:*

$$h^* \triangleq \underset{h^S, h^T \in \mathcal{H}}{\arg\min} \{\mathcal{L}^{\mathcal{S}}(h^S, \eta^S) + \mathcal{L}^{\mathcal{T}}(h^T, \eta^T)\}. \tag{22}$$

Then we can obtain the following inequation for any $h^S, h^T \in \mathcal{H}$:

$$\begin{aligned}\mathcal{L}^T(h^T, \eta^T) \leq & \frac{1}{m^S} \sum_{i=1}^{m^S} L\left(h^S(\boldsymbol{x}_i^S), \boldsymbol{y}_i^S\right) + 2\mu \mathfrak{R}_{m^S}(\mathcal{H}) \\ & + 3M\sqrt{\frac{\log \frac{2}{\sigma}}{2m^S}} + \frac{1}{m^T} \sum_{i=1}^{m^T} L\left(h^T(\boldsymbol{x}_i^T), \boldsymbol{y}_i^T\right) \\ & + 2\mu \mathfrak{R}_{m^T}(\mathcal{H}) + 3M\sqrt{\frac{\log \frac{2}{\sigma}}{2m^T}} \\ & + 3K|\boldsymbol{A}^S - \boldsymbol{A}^T| + -\mathcal{L}^S\left(h^*, \eta^S\right),\end{aligned} \tag{23}$$

where $\mathcal{L}^S\left(\eta^S, h^*\right)$ is independent of h^T or h^S and can be considered as a constant.

Proof.

$$\begin{aligned}& \mathcal{L}^T\left(h^T, \eta^T\right) = \mathcal{L}^T\left(h^T, \eta^T\right) + \mathcal{L}^S\left(h^S, \eta^S\right) - \mathcal{L}^S\left(h^S, \eta^S\right) \\ \leq & \mathcal{L}^S\left(h^S, \eta^S\right) + \mathcal{L}^T(h^T, h^S) + \mathcal{L}^T\left(h^S, \eta^T\right) + \mathcal{L}^S\left(h^S, h^T\right) \\ & - \mathcal{L}^S\left(h^T, \eta^S\right) \\ \leq & 2K|\boldsymbol{A}^S - \boldsymbol{A}^T| + \mathcal{L}^S\left(h^S, \eta^S\right) + \mathcal{L}^T\left(h^S, \eta^T\right) - \mathcal{L}^S\left(h^*, \eta^S\right) \\ \leq & 2K|\boldsymbol{A}^S - \boldsymbol{A}^T| + \mathcal{L}^S\left(h^S, \eta^S\right) + \mathcal{L}^T\left(h^S, h^T\right) \\ & + \mathcal{L}^T\left(h^T, \eta^T\right) - \mathcal{L}^S\left(h^*, \eta^S\right) \\ \leq & 3K|\boldsymbol{A}^S - \boldsymbol{A}^T| + \mathcal{L}^S\left(h^S, \eta^S\right) + \mathcal{L}^T\left(h^T, \eta^T\right) \\ & - \mathcal{L}^S\left(h^*, \eta^S\right) \\ \leq & \frac{1}{m^S} \sum_{i=1}^{m^S} L\left(h^S(\boldsymbol{x}_i^S), \boldsymbol{y}_i^S\right) + 2\mu \mathfrak{R}_{m^S}(\mathcal{H}) + 3M\sqrt{\frac{\log \frac{2}{\sigma}}{2m^S}} \\ & + \frac{1}{m^T} \sum_{i=1}^{m^T} L\left(h^T(\boldsymbol{x}_i^T), \boldsymbol{y}_i^T\right) + 2\mu \mathfrak{R}_{m^T}(\mathcal{H}) + 3M\sqrt{\frac{\log \frac{2}{\sigma}}{2m^T}} \\ & + 3K|\boldsymbol{A}^S - \boldsymbol{A}^T| - \mathcal{L}^S\left(h^*, \eta^S\right)\end{aligned} \tag{24}$$

□

Table 1
Different setting for the dataset of different domains.

Domain	Variance of Noise ϕ	Sample interval	c
Domain 1	1	1	0.02
Domain 2	5	2	0.04
Domain 3	10	3	0.06

According to the aforementioned generalization bound, we can find that the expected risk of h^T is not only controlled by the empirical loss on the labeled source and target data, but also by the distance between the source and target domain structures.

6 Experiment

6.1 Dataset

In this section, we give a brief introduction to the dataset that we used. For each dataset, we split it into the training set, validation set, and test set. For all the methods, we try five different random seeds and report the mean and variance. We choose the model with the best validation and evaluate the chosen model on the test set. All the code for simulation and prepossessing of the dataset will be released.

6.1.1 Simulation Dataset

In this part, we design a series of the controlled experiments on the random causal structure with given sample size, variable size. we simulate three different datasets with the following nonlinear function.

$$\begin{aligned}\boldsymbol{z}_t = & A_1 \cdot (\boldsymbol{z}_{t-1} + c \cdot Sin(\boldsymbol{z}_{t-1})) + \cdots \\ & + A_k \cdot (\boldsymbol{z}_{t-k} + c \cdot Sin(\boldsymbol{z}_{t-k})) + \epsilon,\end{aligned} \tag{25}$$

in which c is the hyper-parameters of the nonlinear term; ϵ is the variance of Gaussian distributions; A_j is the substructure of causal structure with j lag. We further employ different sample intervals for different domains. The details of the dataset are shown in Table 1. And Figure 4 shows the example of the simulated data. Since the value range of this dataset varies with different cities, we use Z-score Normalization for each domain.

Note that we can use the simulated dataset for both time-series forecasting and cross-domain Granger-causal discovery tasks. Since it is very hard to achieve the ground truth Granger-causal structures, we prove the effectiveness of the proposed GCA method and show that the importance of Granger-causal inference.

6.1.2 Air Quality Forecasting Dataset

The air quality forecast dataset [43] is collected in the Urban Air project [1] from 2014/05/01 to 2015/04/30, which contains the air quality data, the meteorological data, and the weather forecast data, etc. This dataset contains four major Chinese cities: Beijing (B), Tianjin (T), Guangzhou (G) and Shenzhen (S). We employ air quality data as well as meteorological data to predict PM2.5. The feature of this data is shown in Table 2. We choose several air quality stations that contain a few missing values and take each

1. https://www.microsoft.com/en-us/research/project/urban-air/

Table 2
Features of Air Quality Forecast Dataset.

Feature Type	Feature Name
Air quality	PM10 PM2.5 NO2 Concentration CO Concentration O3 Concentration SO2 Concentration
Meteorology	Weather Temperature Pressure Humidity Wind Speed

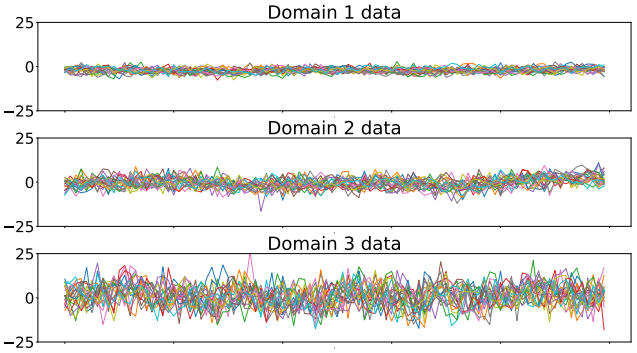


Figure 5. The Sample of three domains, we can find that the data are totally different, but they share the same summary graph.

city as a domain. We use this dataset because the air quality data is familiar and the sensors in the smart city systems usually contain complex and compact causality. Since the value range of this dataset varies with different cities, we use Z-score Normalization for each domain.

6.1.3 Human Motion Capture Dataset

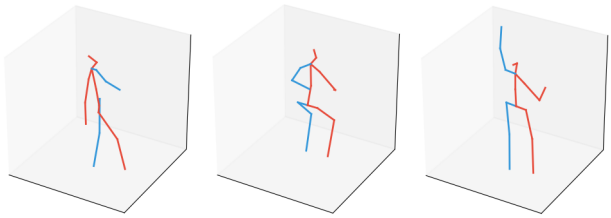
We further show the effectiveness of our method on the human motion capture dataset. Human motion has been studied in many fields, e.g. Granger Causality Inference [23], [44], the linear and the nonlinear dynamical systems [45]. Many researchers use it to study the human motion capture dataset for Granger Causality since the relationships among the joints of humans is a natural causal structure. In this paper, we use the Human3.6M dataset [46] [2], one of the most popular benchmark for human motion prediction [47], for cross-domain human motion prediction. This dataset contains 15 motions and we choose three of them as three different domains: "Walking", "Greeting" and "Eating", the example of this dataset are shown in Figure 6.1.3. We choose 9 primary sensors which record the three-dimensional coordinate and the dimension of the processed dataset is 27.

6.2 Baseline and Model Variants

Baselines of domain adaptation for time-series forecasting.

- **LSTM_S+T** only uses the labeled source domain data and the labeled target domain data to train a vanilla LSTM model and applies it to the target domain validation and test set.

2. http://vision.imar.ro/human3.6m/description.php



(a) Walking (b) Eating (c) Greeting

Figure 6. The visualized examples of a frame in the human motion capture dataset. (a), (b) and (c) respective denote "Walking", "Eating" and "Greeting".

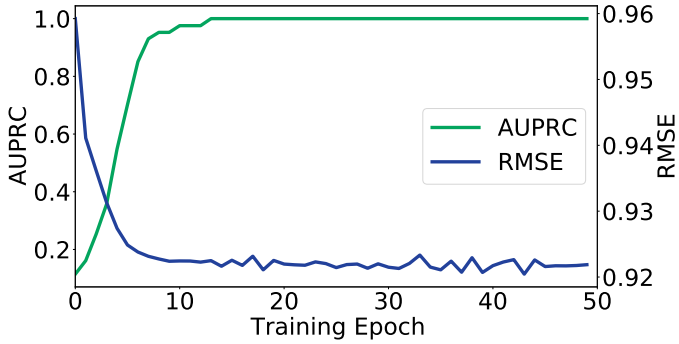
- **R-DANN** [15] uses the domain adaptation architecture proposed in [11] with Gradient Reversal Layer on LSTM, which is a straightforward solution from time-series domain adaptation.
- **RDC**: Deep domain confusion is a domain adaptation method proposed in [48] which minimizes the distance between the source and the target distribution by using Maximum Mean Discrepancy (MMD). Similar to the R-DANN method, we also employ LSTM as the feature extractor for time-series data.
- **VRADA** [16] is a time-series domain adaptation method that combines the Gradient Reversal Layer and VRNN [17].
- **SASA** [2] is one of the state-of-the-art domain adaptation approaches for time-series data, that extracts and aligns the sparse associative structures.

In order to evaluate the modules of the proposed method, we devise the following model variants.

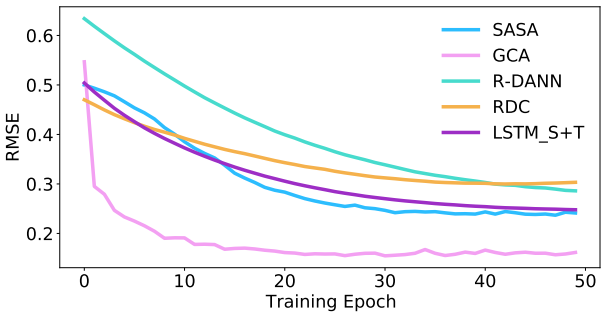
- **GCA-r:** In order to evaluate the effectiveness of the causal graph alignment regularization, we further devise a variant that removes the Granger causality similarity regularization $\mathcal{L}_r$.
- **GCA-e:** In the case that the predicted results are univariate time-series and the Granger-causal structures is desired for interpretability, we aim to evaluate the effectiveness of the extra strengthen term, so we devise a variant that removes the strengthen term $\mathcal{L}_e$.
- **GCA-s:** In the case that the predicted results are univariate time-series and the Granger-causal structures is not necessary to be obtained, we only predict and optimize the target unvariate time-series.

6.3 Result on Simulation Dataset

The Mean Squared Error(MSE) and Mean Absolute Error (MAE) in the simulated dataset are shown in Table. 3. The proposed GCA method significantly outperforms the other baselines on all the tasks with a large gap. It is worth mentioning that our Granger-causality alignment method has a remarkable MSE promotion on most of the tasks (more than 10% improvement compared with SASA). For some easy tasks like $domain1 \rightarrow domain2$ and $domain2 \rightarrow domain1$, the proposed method achieves the largest improvement among all the tasks. For the other harder tasks like $domain3 \rightarrow domain2$ and $domain2 \rightarrow domain3$, our method also achieves a very good result. As for the very challenging tasks like $domain1 \rightarrow domain3$ and



(a) AUPRC and RMSE of GCA.



(b) Convergence of GCA and baselines.

Figure 7. (a) The convergence of the GCA method. We can find that the GCA method achieves the best performance when the Granger Causality is well reconstructed. (b) The convergence of all the methods, we can find that the proposed GCA method converges faster than the other methods. *(best view in color.)*

$domain3 \rightarrow domain1$, in which the value ranges between the source and the target domain are very large, our method also obtains a comparable result.

In order to study how the inferred Granger-causality has influenced the performance of the model, we further apply the proposed method to the test dataset under different accuracy of the Granger-Causality. In detail, we evaluate the GCA on the test set every epoch and record the Area Under Precision-Recall Curve (AUPRC) of the inferred Granger-Causality. The experiment result is shown in Figure 6. According to the experiment result, we can learn the following lessons. (1) The proposed method can well discover the Granger-causality among time-series data. (2) As the increasing of the accuracy of the Granger-causality, the performance of our method increases. The proposed method achieves the best performance when the Granger-causality is exactly discovered. This phenomenon proves the rationality of GCA. (3) Compared with the other baselines, the proposed GCA method enjoys faster convergence, this is because the GCA approach can avoid the side effects of superfluous variables.

6.4 Result on Air Quality Forecasting

In this section, we will show the experimental results on the real-world data. The experiment results on the air quality forecasting are shown in Table 6.3. According to the experiment results, we can find that the methods based on relationship modeling (GCA and SASA) outperform the other baselines. Moreover, the proposed GCA method

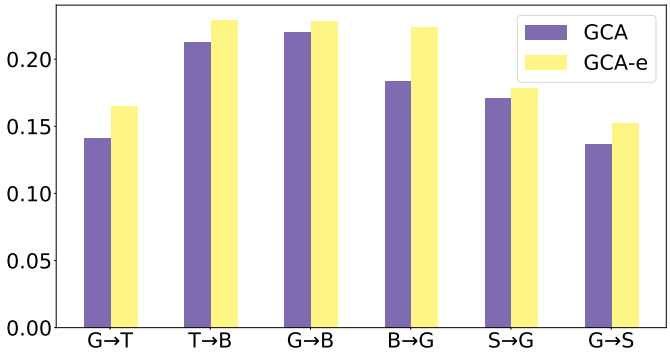


Figure 8. The Sample of three domains in the air quality forecasting dataset, we can find that the data are totally different, but they share the same summary graph. *(best view in color.)*

outperforms the SASA with a remarkable margin, which proves that the Granger Causality has more advantages than associative structures in modeling time-series data with time-lags. However, we also find that the improvements of some domain adaptation tasks such as $G \rightarrow T$, $G \rightarrow B$ and $G \rightarrow S$ are not so remarkable. This is because domain Guangzhou (G) contains many missing value data, which makes it difficult to reconstruct the real Granger-Causality.

6.5 Result on Human Motion Foresting

We also evaluate our method on human motion forecasting, another popular real-world time-series task, the experiment results are shown in Table 5. Compared with the air quality forecasting dataset, the improvement of our method is even larger. This is because (1) The human skeleton structure is a very sparse causal structure, this is why the proposed GCA method and SASA can outperform the other methods. What's more, our method can well remove the side effect of superfluous variables. (2) The human motion forecasting dataset contain more variables than the air quality forecasting dataset. (3) Different to the previous tasks that predicts only one time-series data, multivariate time-series data need to be predicted simultaneously, and it is difficult for the other baselines.

6.6 Ablation Study and Visualization

6.6.1 The study of the Effectiveness of Extra strengthen term.

Essentially, the proposed method is an autoregressive model, so all the variables are taken into consideration, which might lead to the performance degration of the target variable that needs to be predicted. In order to solve this problem, we use an extra strengthen loss. As shown in Figure. 6.6.1, the red and green bars respectively denotes the standard GCA method and the GCA-e method. We can find that (1) Totally, the performance of the proposed GCA method is better than GCA-e, which reflects the effectiveness of the extra strengthen term. (2) Furthermore, the performance of GCA-e is also comparable, which shows the stability of our method.

Table 3
The MAE and MSE on simulated datasets for the baselines and the proposed method. The value presented are averages over 5 replicated with different random seeds. Standard deviation is in the subscript.

Task	Metric	GCA	SASA	VRADA	R-DANN	RDC	LSTM_S+T
$1 \to 2$	RMSE	$\mathbf{0.9232_{\pm 0.0004}}$	$1.0001_{\pm 0.0026}$	$0.9763_{\pm 0.0009}$	$1.0148_{\pm 0.0037}$	$1.0122_{\pm 0.0033}$	$1.0078_{\pm 0.0059}$
	MAE	$\mathbf{0.7648_{\pm 0.0002}}$	$0.7961_{\pm 0.0009}$	$0.7858_{\pm 0.0003}$	$0.8017_{\pm 0.0018}$	$0.8841_{\pm 0.0015}$	$0.7990_{\pm 0.0023}$
$1 \to 3$	RMSE	$\mathbf{0.8556_{\pm 0.0010}}$	$0.8734_{\pm 0.0026}$	$0.9005_{\pm 0.0095}$	$0.8884_{\pm 0.0019}$	$0.8841_{\pm 0.0015}$	$0.8739_{\pm 0.0013}$
	MAE	$\mathbf{0.7399_{\pm 0.0005}}$	$0.7485_{\pm 0.0012}$	$0.7598_{\pm 0.0038}$	$0.7539_{\pm 0.0005}$	$0.7527_{\pm 0.0007}$	$0.7489_{\pm 0.0007}$
$2 \to 1$	RMSE	$\mathbf{0.8323_{\pm 0.0014}}$	$0.9386_{\pm 0.0012}$	$0.9277_{\pm 0.0063}$	$0.9333_{\pm 0.0065}$	$0.9397_{\pm 0.0038}$	$0.9469_{\pm 0.0066}$
	MAE	$\mathbf{0.7258_{\pm 0.0007}}$	$0.7741_{\pm 0.0004}$	$0.7682_{\pm 0.0026}$	$0.7713_{\pm 0.0029}$	$0.7740_{\pm 0.0014}$	$0.7778_{\pm 0.0030}$
$2 \to 3$	RMSE	$\mathbf{0.8570_{\pm 0.0019}}$	$0.9084_{\pm 0.0029}$	$0.9229_{\pm 0.0043}$	$0.9136_{\pm 0.0010}$	$0.9125_{\pm 0.0017}$	$0.9178_{\pm 0.0045}$
	MAE	$\mathbf{0.7493_{\pm 0.0008}}$	$0.7625_{\pm 0.0014}$	$0.7688_{\pm 0.0018}$	$0.7649_{\pm 0.0008}$	$0.7640_{\pm 0.0009}$	$0.7669_{\pm 0.0023}$
$3 \to 1$	RMSE	$\mathbf{0.8314_{\pm 0.0004}}$	$0.8596_{\pm 0.0012}$	$0.9177_{\pm 0.0008}$	$0.8632_{\pm 0.0020}$	$0.8596_{\pm 0.0032}$	$0.8631_{\pm 0.0022}$
	MAE	$\mathbf{0.7255_{\pm 0.0002}}$	$0.7379_{\pm 0.0005}$	$0.7635_{\pm 0.0003}$	$0.7393_{\pm 0.0007}$	$0.7373_{\pm 0.0012}$	$0.7395_{\pm 0.0007}$
$3 \to 2$	RMSE	$\mathbf{0.9208_{\pm 0.0009}}$	$0.9685_{\pm 0.0008}$	$0.9753_{\pm 0.0007}$	$0.9752_{\pm 0.0052}$	$0.9710_{\pm 0.0019}$	$0.9759_{\pm 0.0042}$
	MAE	$\mathbf{0.7634_{\pm 0.0004}}$	$0.7829_{\pm 0.0004}$	$0.7855_{\pm 0.0002}$	$0.7859_{\pm 0.0020}$	$0.7841_{\pm 0.0004}$	$0.7863_{\pm 0.0018}$
Average	RMSE	$\mathbf{0.8700_{\pm 0.0003}}$	$0.9249_{\pm 0.0012}$	$0.9368_{\pm 0.0021}$	$0.9314_{\pm 0.0012}$	$0.9299_{\pm 0.0018}$	$0.9309_{\pm 0.0041}$
	MAE	$\mathbf{0.7433_{\pm 0.0002}}$	$0.7670_{\pm 0.0005}$	$0.7719_{\pm 0.0021}$	$0.7695_{\pm 0.0004}$	$0.7688_{\pm 0.0006}$	$0.7693_{\pm 0.0018}$

Table 4
The MAE and MSE on Air Quality Forecasting dataset for the baselines and the proposed method. The value presented are averages over 5 replicated with different random seeds. Standard deviation is in the subscript.

Task	Metric	GCA	SASA	VRADA	R-DANN	RDC	LSTM_S+T
$B \to T$	RMSE	$\mathbf{0.1684_{\pm 0.0100}}$	$0.1952_{\pm 0.0256}$	$0.2536_{\pm 0.0145}$	$0.1868_{\pm 0.0142}$	$0.2199_{\pm 0.0650}$	$0.1910_{\pm 0.0140}$
	MAE	$\mathbf{0.2669_{\pm 0.0127}}$	$0.3037_{\pm 0.0267}$	$0.3698_{\pm 0.0253}$	$0.3036_{\pm 0.0134}$	$0.3369_{\pm 0.0514}$	$0.2975_{\pm 0.0113}$
$G \to T$	RMSE	$\mathbf{0.1497_{\pm 0.0128}}$	$0.1585_{\pm 0.0255}$	$0.2365_{\pm 0.0129}$	$0.1693_{\pm 0.0195}$	$0.1420_{\pm 0.0152}$	$0.1785_{\pm 0.0049}$
	MAE	$\mathbf{0.2618_{\pm 0.0155}}$	$0.2740_{\pm 0.0113}$	$0.3544_{\pm 0.0032}$	$0.2938_{\pm 0.0160}$	$0.2675_{\pm 0.0153}$	$0.3027_{\pm 0.0060}$
$S \to T$	RMSE	$\mathbf{0.1673_{\pm 0.0040}}$	$0.2022_{\pm 0.0159}$	$0.2662_{\pm 0.0100}$	$0.2330_{\pm 0.0111}$	$0.1877_{\pm 0.010}$	$0.2058_{\pm 0.0042}$
	MAE	$\mathbf{0.2639_{\pm 0.0040}}$	$0.3130_{\pm 0.0155}$	$0.4287_{\pm 0.0667}$	$0.3834_{\pm 0.0243}$	$0.3194_{\pm 0.0141}$	$0.3195_{\pm 0.0070}$
$T \to B$	RMSE	$\mathbf{0.2278_{\pm 0.0055}}$	$0.2366_{\pm 0.0120}$	$0.3112_{\pm 0.0267}$	$0.3222_{\pm 0.0487}$	$0.3182_{\pm 0.0732}$	$0.2652_{\pm 0.0115}$
	MAE	$\mathbf{0.2841_{\pm 0.0087}}$	$0.2962_{\pm 0.0136}$	$0.4007_{\pm 0.0284}$	$0.3522_{\pm 0.0307}$	$0.3580_{\pm 0.0393}$	$0.3243_{\pm 0.0064}$
$G \to B$	RMSE	$\mathbf{0.2169_{\pm 0.0056}}$	$0.2388_{\pm 0.0171}$	$0.2414_{\pm 0.0422}$	$0.2706_{\pm 0.0277}$	$0.2276_{\pm 0.0137}$	$0.3010_{\pm 0.0111}$
	MAE	$\mathbf{0.2844_{\pm 0.0058}}$	$0.3151_{\pm 0.0213}$	$0.3722_{\pm 0.0277}$	$0.3672_{\pm 0.0191}$	$0.3077_{\pm 0.0109}$	$0.3854_{\pm 0.0155}$
$S \to B$	RMSE	$\mathbf{0.2319_{\pm 0.0053}}$	$0.2671_{\pm 0.0244}$	$0.4222_{\pm 0.0159}$	$0.3017_{\pm 0.0184}$	$0.2744_{\pm 0.0204}$	$0.2503_{\pm 0.0108}$
	MAE	$\mathbf{0.2802_{\pm 0.0056}}$	$0.3273_{\pm 0.0266}$	$0.4621_{\pm 0.0129}$	$0.3856_{\pm 0.0216}$	$0.3499_{\pm 0.0255}$	$0.3303_{\pm 0.0139}$
$B \to G$	RMSE	$\mathbf{0.1996_{\pm 0.0046}}$	$0.2281_{\pm 0.0269}$	$0.3135_{\pm 0.0682}$	$0.2964_{\pm 0.0376}$	$0.2566_{\pm 0.0242}$	$0.2452_{\pm 0.0158}$
	MAE	$\mathbf{0.2999_{\pm 0.0114}}$	$0.3315_{\pm 0.0259}$	$0.4256_{\pm 0.0533}$	$0.4147_{\pm 0.0336}$	$0.3682_{\pm 0.0311}$	$0.3564_{\pm 0.0138}$
$T \to G$	RMSE	$\mathbf{0.1845_{\pm 0.0110}}$	$0.2225_{\pm 0.0183}$	$0.2302_{\pm 0.0316}$	$0.1941_{\pm 0.0109}$	$0.1912_{\pm 0.0132}$	$0.1959_{\pm 0.0109}$
	MAE	$\mathbf{0.2819_{\pm 0.0166}}$	$0.2892_{\pm 0.0213}$	$0.3643_{\pm 0.0299}$	$0.3259_{\pm 0.0129}$	$0.3101_{\pm 0.0122}$	$0.3069_{\pm 0.0099}$
$S \to G$	RMSE	$\mathbf{0.1769_{\pm 0.0112}}$	$0.2186_{\pm 0.0261}$	$0.4187_{\pm 0.0158}$	$0.2942_{\pm 0.0023}$	$0.2949_{\pm 0.0461}$	$0.1862_{\pm 0.0067}$
	MAE	$\mathbf{0.2887_{\pm 0.0138}}$	$0.3415_{\pm 0.0424}$	$0.5227_{\pm 0.0707}$	$0.4912_{\pm 0.0936}$	$0.4185_{\pm 0.0406}$	$0.3116_{\pm 0.0083}$
$B \to S$	RMSE	$\mathbf{0.1559_{\pm 0.0058}}$	$0.2281_{\pm 0.0269}$	$0.2240_{\pm 0.0244}$	$0.2530_{\pm 0.0312}$	$0.3350_{\pm 0.0502}$	$0.1861_{\pm 0.0140}$
	MAE	$\mathbf{0.2640_{\pm 0.0095}}$	$0.3315_{\pm 0.0259}$	$0.3698_{\pm 0.0252}$	$0.3957_{\pm 0.0358}$	$0.4521_{\pm 0.0323}$	$0.2940_{\pm 0.0111}$
$T \to S$	RMSE	$\mathbf{0.1536_{\pm 0.0172}}$	$0.1685_{\pm 0.0149}$	$0.2587_{\pm 0.0691}$	$0.2520_{\pm 0.0373}$	$0.2532_{\pm 0.0359}$	$0.1945_{\pm 0.0218}$
	MAE	$\mathbf{0.2640_{\pm 0.0025}}$	$0.3027_{\pm 0.0193}$	$0.4468_{\pm 0.0578}$	$0.3986_{\pm 0.0389}$	$0.3916_{\pm 0.0405}$	$0.2939_{\pm 0.0265}$
$G \to S$	RMSE	$\mathbf{0.1360_{\pm 0.0121}}$	$0.1552_{\pm 0.0070}$	$0.2914_{\pm 0.0297}$	$0.2017_{\pm 0.0449}$	$0.1839_{\pm 0.0172}$	$0.1529_{\pm 0.0055}$
	MAE	$\mathbf{0.2574_{\pm 0.0107}}$	$0.2844_{\pm 0.0086}$	$0.4236_{\pm 0.0262}$	$0.3418_{\pm 0.0443}$	$0.3300_{\pm 0.0192}$	$0.2940_{\pm 0.0060}$
Average	RMSE	$\mathbf{0.1810_{\pm 0.0088}}$	$0.2067_{\pm 0.0184}$	$0.2889_{\pm 0.0297}$	$0.2479_{\pm 0.0252}$	$0.2404_{\pm 0.0320}$	$0.2128_{\pm 0.0109}$
	MAE	$\mathbf{0.2750_{\pm 0.0107}}$	$0.3092_{\pm 0.0215}$	$0.4173_{\pm 0.0356}$	$0.3711_{\pm 0.0320}$	$0.3508_{\pm 0.0277}$	$0.3180_{\pm 0.0113}$

6.6.2 The study of the Effectiveness of the Granger Causality Alignment.

In order to evaluate the effectiveness of the Granger Causality alignment, we remove the summary graph alignment regularization term $\mathcal{L}_r$ and devise the variant model named GCA-r. According to the experiment result shown in Figure. 9, we can find that the performance of GCA-r is better than that of SASA. This is because the GCA-r model can leverage the source domain data and the limited labeled target domain data to reconstruct and predict the future value, which shows that the inference process with Granger Causality can exclude the obstruction of redundant information. We can also find that the performance of GCA-r is slightly lower than that of GCA, this is because the small size of target domain data and the difference between the source and the target domain lead to the diversity and the inaccuracy Granger Causality of target domain, which further results in the degeneration of the experiment results. In fact, The restriction of the summary graph of Granger Causality plays a key role in transferring the domain-invariant module from the source to the target domain.

6.6.3 Comparing the Multivariate Prediction and the Univariate Prediction

In the scenario of only forecasting the object univariate time-series without interpretability, we further devise the **GCA-s** and the experiment results on the Air Quality Forecasting dataset are shown in Table 6. According to the experiment results, we can learn the following lessons: (1) Compared the experiment result between **GAC-e** and **GCA-s**, we can find that **GCA-s** performs better than **GCA-e**, which shows that simultaneously optimizing the other variables will hinder achieving the best results. (2) Compared

Table 5
The MAE and MSE on Human Motion Forecasting dataset for the baselines and the proposed method. The value presented are averaged over 5 replicated with different random seeds. Standard deviation is in the subscript.

Task	Metric	GCA	SASA	VRADA	R-DANN	RDC	LSTM_S+T
$W \to G$	RMSE	$\mathbf{0.1639_{\pm 0.0093}}$	$0.2316_{\pm 0.0397}$	$0.6770_{\pm 0.0281}$	$0.6186_{\pm 0.0699}$	$0.5160_{\pm 0.0177}$	$0.2295_{\pm 0.0048}$
	MAE	$\mathbf{0.2314_{\pm 0.0108}}$	$0.3359_{\pm 0.0355}$	$0.6204_{\pm 0.0148}$	$0.5815_{\pm 0.0260}$	$0.5301_{\pm 0.0068}$	$0.3355_{\pm 0.0038}$
$W \to E$	RMSE	$\mathbf{0.1401_{\pm 0.0284}}$	$0.2147_{\pm 0.0202}$	$0.8269_{\pm 0.0174}$	$0.5617_{\pm 0.0262}$	$0.6841_{\pm 0.0086}$	$0.2046_{\pm 0.0059}$
	MAE	$\mathbf{0.2168_{\pm 0.0337}}$	$0.3288_{\pm 0.0155}$	$0.6998_{\pm 0.0088}$	$0.5753_{\pm 0.0171}$	$0.6484_{\pm 0.0031}$	$0.3339_{\pm 0.0040}$
$G \to W$	RMSE	$\mathbf{0.1414_{\pm 0.0037}}$	$0.1585_{\pm 0.0118}$	$0.5944_{\pm 0.0785}$	$0.4044_{\pm 0.0191}$	$0.3516_{\pm 0.0104}$	$0.168_{\pm 0.0147}$
	MAE	$\mathbf{0.2163_{\pm 0.0059}}$	$0.2754_{\pm 0.0125}$	$0.5843_{\pm 0.0286}$	$0.4751_{\pm 0.0096}$	$0.4561_{\pm 0.0045}$	$0.2893_{\pm 0.0095}$
$G \to E$	RMSE	$\mathbf{0.1372_{\pm 0.0093}}$	$0.1516_{\pm 0.0151}$	$0.6872_{\pm 0.0804}$	$0.4464_{\pm 0.0446}$	$0.5655_{\pm 0.1052}$	$0.1647_{\pm 0.0066}$
	MAE	$\mathbf{0.2101_{\pm 0.0083}}$	$0.2760_{\pm 0.0141}$	$0.6265_{\pm 0.0435}$	$0.5143_{\pm 0.0279}$	$0.5971_{\pm 0.0612}$	$0.2990_{\pm 0.0063}$
$E \to W$	RMSE	$\mathbf{0.1821_{\pm 0.0117}}$	$0.2400_{\pm 0.0205}$	$0.6642_{\pm 0.0182}$	$0.4904_{\pm 0.0228}$	$0.4038_{\pm 0.0185}$	$0.1978_{\pm 0.0080}$
	MAE	$\mathbf{0.2516_{\pm 0.0193}}$	$0.3487_{\pm 0.0198}$	$0.6133_{\pm 0.0182}$	$0.5305_{\pm 0.0152}$	$0.4755_{\pm 0.0124}$	$0.3183_{\pm 0.0061}$
$E \to G$	RMSE	$\mathbf{0.1465_{\pm 0.0124}}$	$0.2225_{\pm 0.0183}$	$0.7707_{\pm 0.1184}$	$0.4896_{\pm 0.0225}$	$0.5394_{\pm 0.0867}$	$0.2411_{\pm 0.0014}$
	MAE	$\mathbf{0.2180_{\pm 0.0193}}$	$0.3268_{\pm 0.0148}$	$0.6617_{\pm 0.0428}$	$0.5257_{\pm 0.0150}$	$0.5448_{\pm 0.0866}$	$0.3454_{\pm 0.0011}$
Average	RMSE	$\mathbf{0.1541_{\pm 0.0097}}$	$0.2031_{\pm 0.0093}$	$0.7034_{\pm 0.0352}$	$0.5018_{\pm 0.0131}$	$0.5101_{\pm 0.0192}$	$0.2010_{\pm 0.0069}$
	MAE	$\mathbf{0.2262_{\pm 0.0146}}$	$0.3152_{\pm 0.0095}$	$0.6343_{\pm 0.014}$	$0.5337_{\pm 0.0070}$	$0.5420_{\pm 0.0103}$	$0.3202_{\pm 0.0051}$

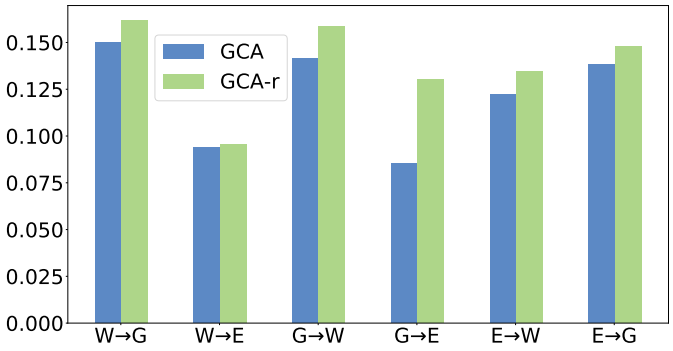
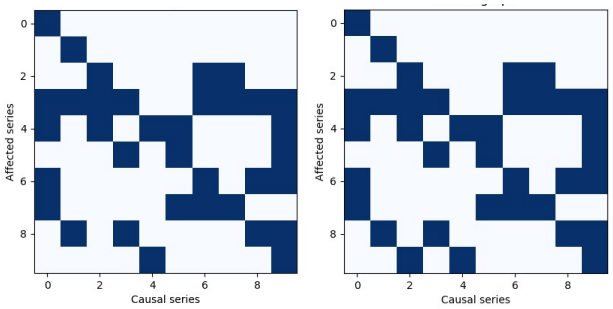


Figure 9. The Sample of three domains in the air quality forecasting dataset, we can find that the data are totally different, but they share the same summary graph. *(best view in color.)*



(a) TianJin (b) BeiJing

Figure 10. The illustration of visualization of the summary Granger-Causal structure of TianJin → BeiJing. The Blue blocks denote the Granger-causal relationship among different variables.

with the experiment results between **GCA** and **GCA-s**, we can find that the experiment results of these two methods are quite proximal, which reflects that the extra strengthen loss can mitigate the side-effect of simultaneously regressing all variables.

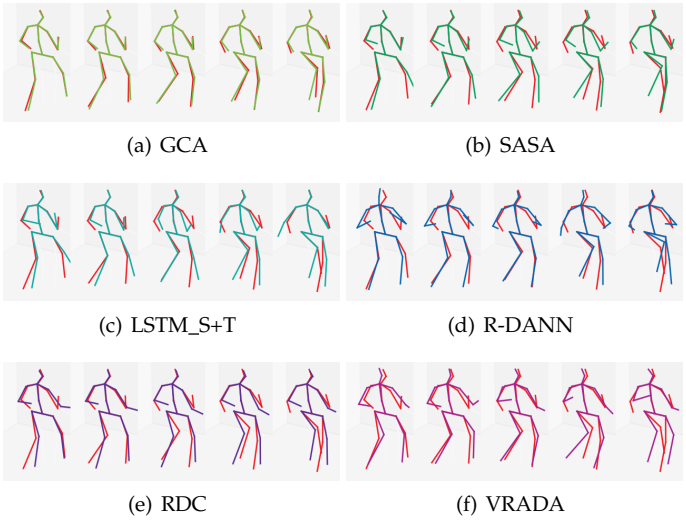
6.6.4 *Visualization and Interpretability*

Visualization of the Air Quality Forecasting dataset. To further investigate our approach, we extract the Granger-Causal structures and perform the corresponding visualization over the air quality forecasting dataset. Figure 10 shows the Granger-causal structures whose task is TianJin → BeiJing. According to the visualization, we can find that the Granger-Causal structures are very sparse, and most of the modules are similar, which means that the PM2.5 is generated by a similar causality. The aforementioned visualization also provides some interpretability to a certain extent. For example, in Figure 10 (b), we can find that the "humidity" and "PM10_Concentration" simultaneously have an influence on the generation of PM2.5, while the other variables like "pressure" and "temperature" have little influence on PM2.5. This experiment results also provide the insightful conclusion that we can mitigate the effect of PM2.5 by controlling the release of PM10.

Visualization of the Human Motion Forecasting dataset. We further conduct visualization on the human motion forecasting dataset, the experiment results are shown in Figure. 11 illustrate the walking motion with 5 steps later.

Table 6
The MAE and RMSE on simulated datasets for the different model variants. The value presented are averages over 5 replicated with different random seeds. Standard deviation is in the subscript.

Task	Metric	GCA	GCA-s	GCA-e
$B \to T$	RMSE	$0.1684_{\pm 0.0100}$	$\mathbf{0.1663_{\pm 0.0036}}$	$0.1685_{\pm 0.0080}$
	MAE	$0.2669_{\pm 0.0127}$	$\mathbf{0.2532_{\pm 0.0009}}$	$0.2682_{\pm 0.0082}$
$G \to T$	RMSE	$0.1497_{\pm 0.0128}$	$\mathbf{0.1452_{\pm 0.0063}}$	$0.1500_{\pm 0.0103}$
	MAE	$0.2618_{\pm 0.0155}$	$\mathbf{0.2436_{\pm 0.0047}}$	$0.2629_{\pm 0.0130}$
$S \to T$	RMSE	$\mathbf{0.1673_{\pm 0.0040}}$	$\mathbf{0.1670_{\pm 0.0043}}$	$0.1714_{\pm 0.0064}$
	MAE	$\mathbf{0.2639_{\pm 0.0040}}$	$\mathbf{0.2626_{\pm 0.0077}}$	$0.2671_{\pm 0.0088}$
$T \to B$	RMSE	$\mathbf{0.2276_{\pm 0.0055}}$	$\mathbf{0.2276_{\pm 0.0050}}$	$0.2337_{\pm 0.0084}$
	MAE	$\mathbf{0.2841_{\pm 0.0087}}$	$\mathbf{0.2817_{\pm 0.0065}}$	$0.2846_{\pm 0.0106}$
$G \to B$	RMSE	$\mathbf{0.2169_{\pm 0.0056}}$	$\mathbf{0.2150_{\pm 0.0068}}$	$0.2273_{\pm 0.0135}$
	MAE	$0.2844_{\pm 0.0058}$	$\mathbf{0.2731_{\pm 0.0087}}$	$0.2922_{\pm 0.0125}$
$S \to B$	RMSE	$\mathbf{0.2312_{\pm 0.0053}}$	$\mathbf{0.2314_{\pm 0.2801}}$	$0.2318_{\pm 0.0046}$
	MAE	$\mathbf{0.2802_{\pm 0.0056}}$	$\mathbf{0.2901_{\pm 0.0043}}$	$0.2807_{\pm 0.0050}$
$B \to G$	RMSE	$0.1996_{\pm 0.0046}$	$\mathbf{0.1969_{\pm 0.0060}}$	$0.2013_{\pm 0.0052}$
	MAE	$\mathbf{0.2999_{\pm 0.0114}}$	$\mathbf{0.2983_{\pm 0.0075}}$	$0.3024_{\pm 0.0128}$
$T \to G$	RMSE	$0.1845_{\pm 0.0110}$	$\mathbf{0.1796_{\pm 0.0135}}$	$0.1891_{\pm 0.0136}$
	MAE	$\mathbf{0.2819_{\pm 0.0166}}$	$\mathbf{0.2808_{\pm 0.0075}}$	$0.2991_{\pm 0.0195}$
$S \to G$	RMSE	$\mathbf{0.1769_{\pm 0.0112}}$	$\mathbf{0.1780_{\pm 0.0032}}$	$0.1871_{\pm 0.0119}$
	MAE	$\mathbf{0.2887_{\pm 0.0138}}$	$\mathbf{0.2880_{\pm 0.0059}}$	$0.2923_{\pm 0.0106}$
$B \to S$	RMSE	$\mathbf{0.1558_{\pm 0.0058}}$	$\mathbf{0.1568_{\pm 0.0032}}$	$0.1570_{\pm 0.0045}$
	MAE	$\mathbf{0.2640_{\pm 0.0095}}$	$\mathbf{0.2640_{\pm 0.0062}}$	$0.2652_{\pm 0.0151}$
$T \to S$	RMSE	$0.1536_{\pm 0.0172}$	$\mathbf{0.1440_{\pm 0.0065}}$	$0.1543_{\pm 0.0174}$
	MAE	$0.2640_{\pm 0.0025}$	$\mathbf{0.2532_{\pm 0.0095}}$	$0.2683_{\pm 0.0231}$
$G \to S$	RMSE	$\mathbf{0.1360_{\pm 0.0121}}$	$\mathbf{0.1356_{\pm 0.0035}}$	$0.1359_{\pm 0.0082}$
	MAE	$0.2574_{\pm 0.0107}$	$\mathbf{0.2545_{\pm 0.0085}}$	$0.2625_{\pm 0.0106}$
Average	RMSE	$\mathbf{0.1810_{\pm 0.0088}}$	$\mathbf{0.1783_{\pm 0.0055}}$	$0.1839_{\pm 0.0093}$
	MAE	$\mathbf{0.2750_{\pm 0.0107}}$	$\mathbf{0.2704_{\pm 0.0067}}$	$0.2788_{\pm 0.0124}$



(a) GCA (b) SASA (c) LSTM_S+T (d) R-DANN (e) RDC (f) VRADA

Figure 11. The illustration of visualization of the human Motion forecasting of the task Guesting→Walking. The red lines denote the ground truth motion and the others colours denote the different methods. *(best view in color.)*

We can find that the visualized 3D motions of the proposed GCA method are well fitted into the ground truth motions, while the other baselines do not fit well. For some joints like the legs and the elbows, our method markedly surpass the other methods, which shows that the Granger Causality is beneficial to capture the relationship between different joints and results in better prediction.

Note that the proposed method aims to leverage the Granger Causality to address the time-series domain adaptation problem, so we do not follow the same experiment setting of the standard human motion estimation works that forecast a long term human motion and just predict the future a few steps (e.g., 5 steps).

6.6.5 Computational Complexity Analysis

Different from the previous LSTM based method that extract the feature on multivariate level, the architecture of the proposed GCA method works on the univariate level. Given p univariate time-series data whose time lag is k, the proposed GCA method only requires p small MLP for recurrent Granger Causality reconstruction and $k \times p$ small MLP for Granger Causality inference networks. Since the hidden size for each univariate is small, we set the dimension of the hidden state to 3 in practice, the proposed method does not have very large number of parameters. However, it is incontestable that the time complexity of our method compares unfavorably with the LSTM based methods. Since we need to recurrently reconstruct Granger-causal structure. Fortunately, the time lag is small in practice (usually less than 10). Therefore, given a small time lag k and a large p, the time cost of the proposed GCA method is acceptable and is even faster than the sparse associative structure alignment model.

7 Conclusion

In this paper, we present a Granger Causality Alignment method for semi-supervised time-series forecasting. In our proposal, the proposed Granger Causality alignment method, which simultaneously reconstructs the Granger Causality and infers on the Granger-causal structure, can well transfer domain knowledge via the domain-sensitive module and the Granger Causality alignment. The proposed method not only addresses the domain adaptation problem for the time-series forecasting task but also provide insightful interpretability that may be useful in making a decision of time-series application. Though effective, the proposed method needs to take the time-lag as a prior or hyperparameters when we apply our method to the real-world dataset. Therefore, exploring how to learning the Granger Causality with adaptive time lag would be a future direction.

Acknowledgments

The authors would like to thank Jie Qiao and Jiahao Li from the Guangdong University of Technology for their help and support on this work.

References

[1] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, "Learning disentangled semantic representation for domain adaptation," in *IJCAI: proceedings of the conference*, vol. 2019. NIH Public Access, 2019, p. 2060.

[2] R. Cai, J. Chen, Z. Li, W. Chen, K. Zhang, J. Ye, Z. Li, X. Yang, and Z. Zhang, "Time series domain adaptation via sparse associative structure alignment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6859–6867, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16846

[3] R. Cai, F. Wu, Z. Li, P. Wei, L. Yi, and K. Zhang, "Graph domain adaptation: A generative view," *arXiv preprint arXiv:2106.07482*, 2021.

[4] Z. Hao, D. Lv, Z. Li, R. Cai, W. Wen, and B. Xu, "Semi-supervised disentangled framework for transferable named entity recognition," *Neural Networks*, vol. 135, pp. 127–138, 2021.

[5] C. Shui, Z. Li, J. Li, C. Gagné, C. Ling, and B. Wang, "Aggregating from multiple target-shifted sources," *arXiv preprint arXiv:2105.04051*, 2021.

[6] Z. Li, R. Cai, H. W. Ng, M. Winslett, T. Z. Fu, B. Xu, X. Yang, and Z. Zhang, "Causal mechanism transfer network for time series domain adaptation in mechanical systems," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 2, pp. 1–21, 2021.

[7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[8] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7404–7413.

[9] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[10] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010.

[11] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[12] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 5423–5432.

[13] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International Conference on Machine Learning*. PMLR, 2013, pp. 819–827.

[14] K. Zhang, M. Gong, P. Stojanov, B. Huang, Q. Liu, and C. Glymour, "Domain adaptation as a problem of inference on graphical models," *arXiv preprint arXiv:2002.03278*, 2020.

[15] P. R. d. O. da Costa, A. Akçay, Y. Zhang, and U. Kaymak, "Remaining useful lifetime prediction via deep domain adaptation," *Reliability Engineering & System Safety*, vol. 195, p. 106682, 2020.
[16] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, "Variational recurrent adversarial deep domain adaptation," 2016.
[17] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in neural information processing systems*, vol. 28, pp. 2980–2988, 2015.
[18] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
[19] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric granger causality testing," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1647–1669, 2006.
[20] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.
[21] A. Seth, "Granger causality," *Scholarpedia*, vol. 2, no. 7, p. 1667, 2007, revision #127333.
[22] R. Marcinkevičs and J. E. Vogt, "Interpretable models for granger causality using self-explaining neural networks," *arXiv preprint arXiv:2101.07600*, 2021.
[23] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, "Neural granger causality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.
[24] S. Löwe, D. Madras, R. Zemel, and M. Welling, "Amortized causal discovery: Learning to infer causal graphs from time-series data," *arXiv preprint arXiv:2006.10833*, 2020.
[25] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.
[26] S. Z. Chiou-Wei, C.-F. Chen, and Z. Zhu, "Economic growth and energy consumption revisited—evidence from linear and nonlinear granger causality," *Energy Economics*, vol. 30, no. 6, pp. 3063–3076, 2008.
[27] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.
[28] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling methods for temporal causal modeling," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 577–586. [Online]. Available: https://doi.org/10.1145/1557019.1557085
[29] W. W. Wei, "Time series analysis," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2006.
[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[31] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
[32] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.
[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[34] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
[35] G. Valvano, A. Leo, and S. A. Tsaftaris, "Self-supervised multi-scale consistency for weakly supervised segmentation learning," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, 2021, pp. 14–24.
[36] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *arXiv preprint arXiv:1611.02163*, 2016.
[37] W. B. Nicholson, D. S. Matteson, and J. Bien, "Varx-l: Structured regularization for large vector autoregressions with exogenous variables," *International Journal of Forecasting*, vol. 33, no. 3, pp. 627–651, 2017.
[38] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.
[39] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.
[40] C. Cortes and M. Mohri, "Domain adaptation in regression," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 308–323.
[41] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, p. 137, 2007.
[42] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-iid processes," 2009.
[43] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, *Forecasting Fine-Grained Air Quality Based on Big Data*. New York, NY, USA: Association for Computing Machinery, 2015, p. 2267–2276. [Online]. Available: https://doi.org/10.1145/2783258.2788573
[44] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf, "Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination," in *IJCAI: Proceedings of the Conference*, vol. 2017. NIH Public Access, 2017, p. 1347.
[45] E. B. Fox, M. C. Hughes, E. B. Sudderth, and M. I. Jordan, "Joint modeling of multiple time series via the beta process with application to motion capture segmentation," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1281–1313, 2014.
[46] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
[47] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1418–1427.
[48] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

APPENDIX A
THE VARIATIONAL LOWER BOUND

$$
\begin{aligned}
&\ln P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\boldsymbol{z}_{t-2},\cdots,\boldsymbol{z}_{t-k}) = \ln \frac{P(\boldsymbol{z}_t,A_1,A_2,\cdots,A_k|\boldsymbol{z}_{t-1},\boldsymbol{z}_{t-2},\cdots,\boldsymbol{z}_{t-k})}{P(A_1,A_2,\cdots,A_k|\boldsymbol{z}_t,\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})} \\
=&\ln P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_k) + ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{P(A_1|\boldsymbol{z}_t,\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})} \\
=&\ln P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_k) + \ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{P(A_1|\boldsymbol{z}_t,\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_t,\cdots,\boldsymbol{z}_{t-k})} \\
&+\ln\frac{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j-2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j-2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})} \\
=&\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_k|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_{k-1})}\Bigg[\underbrace{\ln P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_k)}_{\text{The reconstruction loss }\mathcal{L}_r} \\
&+\ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})} \\
&+\ln\frac{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\Bigg] \\
=&\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_k|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_{k-1})}\left[\mathcal{L}_r+\ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\right] \\
&+\sum_{j=1}^{k}D_{KL}(Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})||P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k}) \\
\geq&\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_k|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_{k-1})}\left[\mathcal{L}_r+\ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\prod_{j=2}^{k}Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\right] \\
=&\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_k|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_{k-1})}\mathcal{L}_r+\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\left[\ln\frac{P(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{QA_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\right] \\
&+\sum_{j=2}^{k}\mathbb{E}_{Q(A_j|A_1\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{\boldsymbol{t-k}})}\frac{P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}{Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})} \\
=&\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_k|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k},A_1,\cdots,A_{k-1})}\mathcal{L}_r-D_{KL}(Q(A_1|\boldsymbol{z}_{t-1},\boldsymbol{z}_{t-2},\cdots,\boldsymbol{z}_{t-k})||P(A_1|\boldsymbol{z}_{t-1},\boldsymbol{z}_{t-2},\cdots,\boldsymbol{z}_{t-k})) \\
&-\sum_{j=2}^{k}\mathbb{E}_{Q(A_1|\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\cdots\mathbb{E}_{Q(A_{j-1}|A_1,\cdots,A_{j-2},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})}\left[D_{KL}\left(Q(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\right.\right. \\
&\left.\left.||\ P(A_j|A_1,\cdots,A_{j-1},\boldsymbol{z}_{t-1},\cdots,\boldsymbol{z}_{t-k})\right)\right]
\end{aligned}
$$