UNIFIED FRAMEWORK FOR CAUSAL DISCOVERY AND LONG-TERM FORECASTING IN NON-STATIONARY ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-stationary data is prevalent in various real-world domains such as climate science, economics, and neuroscience, presenting significant challenges for tasks like forecasting and causal discovery from observational data. Existing approaches often operate under the assumption that the data is stationary. In this work, we introduce a unified framework that combines long-term forecasting and causal discovery with non-linear relations in a non-stationary setting. Specifically, we assume that the nonlinear causal relations in the observed space can be transformed into linear relations in the latent space via projections. In addition, we model the non-stationarity in the system as arising from time-varying causal relations. The proposed model demonstrates that adopting a causal perspective for long-term forecasting not only addresses the limitations of each individual task but also makes the causal process identifiable, enhances interpretability, and provides more reliable predictions. Moreover, our approach reformulates causal discovery into a scalable, non-parametric deep learning problem. Through experiments on both synthetic and real-world datasets, we show that our framework outperforms baseline methods in both forecasting and causal discovery, underscoring the benefits of this integrated approach.

1 INTRODUCTION

Causal discovery over observational time series data is a crucial task, with far reaching applications across various real-world domains such as climate science, economics, and neuroscience. Causal discovery methods aim to uncover contemporaneous and/or time-lagged dependencies, from the observed data. These dependencies not only provide useful insights into the underlying processes governing the system but also can be extremely beneficial in tasks such as making informed predictions, identifying key drivers of change, and designing effective interventions.

Most widely-used causal discovery methods are constraint based approaches. They aim to recover a directed acyclic graph (DAG) representing the causal structure among the observed variables by using conditional independence tests. Under certain assumptions and and given enough data points, constraint based methods, provide a theoretical guarantee to be able to recover the true causal graph. Notable examples include approaches like PCMCI (Runge (2022))and tsFCI (Entner & Hoyer (2010)) which extend the classic PC and the FCI algorithms respectively(Spirtes et al. (1993), Spirtes (2001)) for time series data. Additionally, many approaches also utilize traditional statistical methods such as Granger causality and vector auto-regression (VAR). A variable x_i Granger causes x_j if past values of x_i provide unique and statistically significant information about future values of x_j. However, all these methods share a critical limitation: they assume stationarity, meaning the causal relationships between variables remain constant over time. This assumption often fails in real-world scenarios where relationships can evolve due to changing system dynamics. Furthermore, these approaches tend to struggle with non-linearities and could face scalability issues when applied to large datasets.

Long-term forecasting for non-stationary time series is equally challenging. In recent years, this problem has been widely studied, and numerous deep learning-based approaches have been proposed. Transformer based models (Wu et al. (2022), Zhou et al. (2022), Zhou et al. (2021)) have gained popularity due to their exceptional performance in forecasting based tasks. However, recent works

(Zeng et al. (2022) Das et al. (2024))have questioned the necessity of such complex architectures, advocating for simpler MLP-based models that can achieve competitive performanc. Despite their simplicity, linear models are inherently limited in their ability to capture the non-linearities commonly found in real-world data. Transformer-based models, on the other hand can be sensitive to distributional shifts over time. They may struggle when faced with abrupt changes, evolving trends, or time-varying dependencies typical of non-stationary systems. It is also important to note that the majority of datasets benchmark the performance of these models are often stationary, with clear seasonal and trend dynamics, simplifying the forecasting task Wang et al. (2023).

Given a non-stationary time series, with non-linear causal relations, we present a unified framework that combines the tasks of causal discovery and long-term forecasting over the observed data. Specifically, we aim to find a projection of the observables to a latent space such that the non-linear causal relations in the observed space are transformed into linear relations in the latent space. Further, we assume the non-stationarity in the data arises from time varying causal weights. As a result, the causal model in the latent space can be represented as a time-varying coefficients regression model. This formulation guarantees identifiability of the parameters of the causal model in the latent space, and by extension, in the observation space. In our current approach we only assume time-lagged causal relations.

We demonstrate that incorporating a causal perspective into long-term forecasting effectively addresses the limitations of both tasks. Discovery of the changing causal relations enables us to capture and leverage the uncertainty in the system, to enhance interpretability and provide more reliable predictions over a longer time periods. Additionally, our approach reformulates causal discovery into a scalable, non-parametric, deep-learning based task. Through experiments on both synthetic and real-world datasets, we show that our proposed method outperforms baseline approaches in both causal discovery and long-term forecasting.

To summarize, the proposed approach has the following features:

- It accommodates non-linear causal relations among the observed variables by learning a projection to a latent space where these relations become linear.
- It models non-stationary time series by allowing causal mechanisms to change over time.
- It imposes minimal constraints on the underlying causal structure while ensuring identifiability in both the latent and observation spaces.
- We provide experimental results on both synthetic and real-world datasets, demonstrating the efficacy of our model.

2 Related Work

Causal discovery for time series Causal discovery in time series, particularly in non-stationary environments, has been explored through various modeling approaches. One notable work Huang et al. (2019) uses state-space models to represent non-stationary processes, treating the problem as a nonlinear state-space model that captures changes in both causal strengths and noise variances. The underlying hypothesis is that non-stationarity, often seen as a challenge, can actually aid in identifying causal structures, and that forecasting accuracy naturally improves when informed by these learned causal relationships. While effective, this approach has limitations: it relies on linear assumptions, and focuses mainly on the next-step prediction. Building upon this work, our goal is to relax these conditions by accommodating non-linear relationships between observed causal variables and dynamically changing causal weights, while extending the framework to support long-term prediction on both synthetic and real-world datasets. PCMCI$-\Omega$ (Gao et al. (2024b)) builds upon the PCMCI (Runge (2022)) framework to present a constraint-based, non-parametric algorithm designed for semi-stationary environments. In these settings, the data is characterized by a finite number of causal mechanisms that occur periodically. Another prominent method is CDNOD (Gao et al. (2024a)), a non-parametric, constraint-based causal discovery technique tailored for heterogeneous and nonstationary data. CDNOD focuses on detecting variables with changing causal mechanisms and models these changes using a surrogate variable. It orients the causal edges from the surrogate variable to the mechanisms that exhibit variability, effectively leveraging data heterogeneity and distributional shifts to identify both causal structures and their directionalities.

DYNOTEARS (Pamfil et al. (2020)) represents another significant contribution, adapting the well-known NOTEARS framework to time series data. DYNOTEARS formulates causal discovery as a continuous optimization problem, where the graph search is expressed through a differentiable objective function that quantifies the "DAG-ness" (Directed Acyclic Graph property) of the causal graph. A comprehensive overview of these methods, among others, is provided in the survey paper Causal Discovery in Temporal Data (Gong et al. (2023)), which summarizes and categorizes various techniques used in causal discovery for temporal settings.

Long-term forecasting Transformer-based models have seen significant advancements in time series forecasting, with various extensions enhancing their capability to capture complex temporal patterns. Among these, Autoformer (Wu et al. (2022)) is notable for its series decomposition approach, where the input is divided into trend and seasonal components using a specialized series decomposition block. A key component in Autoformer is the use of an auto-correlation mechanism, which identifies period-based dependencies by measuring the similarity between the input and lagged inputs in the Fourier domain. FEDFormer (Zhou et al. (2022)) takes a different approach with its Frequency Enhanced Transformer architecture, which operates in the frequency domain to extract relevant features. Spacetimeformer (Grigsby et al. (2023)) introduces a novel modification in temporal encoding by flattening the feature vector, to encode both spatial and temporal dimensions.

Linear-based models have gained an increasing popularity as an alternative to more complex architectures, particularly challenging the effectiveness of attention mechanisms in time series forecasting. DLinear (Zeng et al. (2022)) argues that attention mechanisms, being permutation invariant, often fail to capture the intricate dependencies inherent in time series data and primarily serve to reduce computational complexity rather than enhance predictive power. Instead, DLinear proposes a straightforward linear model augmented with a seasonal decomposition block, demonstrating that such a simple approach can achieve performance comparable to, or even surpass, more sophisticated models that rely heavily on attention. Time Series Dense Encoder (TiDE Das et al. (2024)) further emphasizes the potential of simpler architectures, proposing a method based solely on multi-layer perceptrons (MLPs). State space models represent another category of alternatives, with S4 (Gu et al. (2022))standing out for its capability to model long sequences effectively. S4 addresses the traditional bottlenecks associated with state-space models and is able to capture long-range dependencies with remarkable efficiency.

3 Problem Formulation

3.1 Hypothesis

Time-varying coefficient models provide a powerful framework to address non-stationarity by allowing the coefficients that govern the relationships between variables, to change over time. Instead of assuming fixed relationships, time-varying coefficient models adapt to the dynamic nature of the data, enabling the modeling of evolving dependencies and interactions. We can describe the model in terms of a regression relation between the observables and a state equation describing the evolution of the coefficients over time.

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{x}_t \boldsymbol{\beta}_t + \mathbf{e_t} \\ \boldsymbol{\beta}_{t+1} = \boldsymbol{A} \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \end{cases} \tag{1}$$

$\mathbf{x}$ and $\mathbf{y}$ represent the observables, $\boldsymbol{\beta}$ are the time varying coefficients, $\mathbf{A}$ is the matrix governing the transitions of $\boldsymbol{\beta}$ and $\mathbf{e_t}$, $\boldsymbol{\epsilon}_t$ are independent zero mean noise terms. The identification of the system is quite challenging as the identification of $\boldsymbol{\beta}_t$ depends on identification of the transition matrix $\mathbf{A}$, the covariance matrix (Q) of the noise term ϵ_t and the initial conditions of the process β_0.

Wall (1987), provides the identifiability conditions for such a system. Further, Huang et al. (2019) reformulates the problem as a causal discovery task where β constitutes the time varying causal weights. The work provides a theoretical baseline for causal discovery and forecasting in a non-stationary time series characterized by a time varying linear causal model. The core result can be summarized as follows - Given a state space model describing the linear causal relations between observed variables, the causal weights and their associated noise variance are allowed to change in an auto-regressive manner. This system is identifiable under both contemporaneous and time-lagged causal relations.

Our hypothesis builds upon this proof. Given a non-stationary time series data, with observed casual variables, we assume that the causal relations are nonlinear in the observation space. Non-stationarity in the system arises due to smoothly changing casual relations between the observed variables (as a function of time). We aim to find an alternative mapping or projection of the observations in a latent space such that the causal relations between the projections in the latent space are linear. The causal relations in the latent space (similar to the observation space) also change as a function of time. We assume that the causal weights change smoothly over time and at least one of the causal weight matrices is full rank.

Given no other constraints, and following the proof from Huang et al. (2019), we claim that in this projection space the system is identifiable. The projection is a learnt one-to-one mapping, therefore the identifiability claim can be extended to the causal mechanisms in the observed space.

3.2 Non-linear time varying causal model

We observe an n-dimensional multivariate time-series at discrete time steps, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_T}\}, \mathbf{x}_t \in \mathbb{R}^n$. The observations are assumed to have time-lagged non-linear causal relations with a maximum lag of L time steps. The time series is assumed to be non-stationary such that these causal relations change over time.

For the observation space,

$$x_{i,t} = G_{it}(\{x_{j,t-\tau} | x_{j,t-\tau} \in \text{Pa}(x_{i,t})\}, \eta_{i,t}) \tag{2}$$

where $G_{i,t}$ represents some non-linear function, $\text{Pa}(x_{i,t})$ is the set of all possible parents of $x_{i,t}$, $\tau \in \{1, 2 \cdots, L\}$ is the time lag, and η_t is a zero mean noise variable. The corresponding latent variables $\mathbf{Z} = \{\mathbf{z_1}, \mathbf{z_2}, \cdots, \mathbf{z_T}\}, \mathbf{z}_t \in \mathbb{R}^d$, are assumed to exist in a d-dimensional space where the causal relations among $\mathbf{Z}$ are linear.

$$\mathbf{z_t} = \Phi(\mathbf{x_t}) \qquad \mathbf{x_t} = \Psi(\mathbf{z_t}) \tag{3}$$

$$z_{i,t} = \sum_{\tau=1}^{L} \sum_{z_{j,t-\tau} \in \text{Pa}_i} \beta_{i,j,t}^{(\tau)} z_{j,t-\tau} + e_{i,t} \tag{4}$$

Since, we assume a feature-wise transformation, the encoder Φ produces a feature-wise map of the observations. Thus, $d = n$. $\beta_{it} \in \mathbb{R}^{d \times L}$ represents the adjacency matrix for $z_{i,t}$ such that $\beta_t[i, j] \neq 0$, implies that there is a causal connection between $z_{i,t}$ and $z_{i,t-j}$ and by extension between x_t and x_{t-j} as well. The feature-wise formulation ensures that the causal adjacency matrix remains the same for the observed as well as the latent space.

The causal weight matrix β_t is allowed to change with time as a non-linear function of its time lagged values with some noise $\boldsymbol{\epsilon}$,

$$\beta_t = f(\{\beta_{t-r} | r \in \{1, 2, \cdots, R\}\}, t) + \epsilon_t \tag{5}$$

Each $\beta_t \in \mathbb{R}^{d \times d \times L}$ is a three-dimensional matrix. We note that, all the $d_b = d \times d \times L$ components can individually, follow a different time-varying pattern.

Objective The objective of the approach is two-fold. We aim to recover the changing causal weights $\beta_t, \ \forall t \in [T]$ for all time steps. Additionally, given an input of context length C, $\mathbf{X_{inp}} = \{\mathbf{x_{t-C+1}}, \mathbf{x_{T-C+2}}, \cdots, \mathbf{x_T}\}$., we aim to predict the next H time points of the time series $\mathbf{X_{pred}} = \{\mathbf{x_{T+1}}, \mathbf{x_{T+2}}, \cdots, \mathbf{x_{T+H}}\}$.

3.3 Identifiability Theory

The identifiability of the proposed model in the latent space is dependent only on the fact that the causal weights change over time in a smooth manner. To ensure complete identifiability of the system, we can assume that there are no interactions among the components of the causal weight matrix and each component is allowed to follow a completely independent process. Wall (1987) demonstrate why this is crucial and show that interactions among the components can result in a partially identifiable system. Further, we make no assumptions on the nature of the noise components in the model.

Theorem 1 presents the identifiability result for a time-varying causal model where the causal weights β follow a time lagged auto-regressive process.

Theorem 1 *Given a multivariate time series $Z_t = (z_{1,t}, \cdots, z_{d,t})^T$ generated by the process*

$$\begin{cases} z_{i,t} = \sum_{\tau=1}^{L} \sum_{z_{j,t-\tau} \in Pa_i} \beta_{i,j,t}^{(\tau)} z_{j,t-\tau} + e_{i,t} \\ \beta_{i,j,t}^{(\tau)} = \sum_{r=1}^{R} \alpha_{i,j,r}^{(\tau)} \beta_{i,j,t-r}^{(\tau)} + \epsilon_{i,j,t} \end{cases} \tag{6}$$

where $z_{j,t-\tau}$ is the time-lagged cause of $z_{i,t}$ with a lag of τ, and $\beta_{i,j,t}^{(\tau)}$ is the corresponding causal weight. $\beta_{i,j,t}^{(\tau)}$ follows a R-order autoregressive process with $\alpha_{i,j,r}$ being the transition coefficient for lag r. $e_{i,t}$ represents a stationary zero mean white noise process where $\mathbb{E}[e_{i,t}] = 0$, $\mathbb{E}[e_{i,t}, e_{i,t'}] = \sigma_i^2 \delta_{tt'}$ and $\mathbb{E}[e_{i,t}, e_{i',t}] = \sigma_i^2 \delta_{ii'}$ where $\sigma_i^2 < \infty$ and δ is the delta function. Similarly, for the noise in the autoregressive process $\mathbb{E}[\epsilon_{i,j,t}] = 0$, $\mathbb{E}[\epsilon_{i,j,t}, \epsilon_{i,j,t}] = w_{ij}$.

The model represented by 6 is identifiable.

The proposed Time-varying causal model can be seen as an extension of the varying coefficients regression model. The proof of theorem 1 can be extended trivially from the identifiability results for varying coefficient regression models.

Lemma 2 *Wall (1987) Given a varying coefficients regression model,*

$$\begin{cases} y_t = \sum_i \beta_{i,t} z_{i,t} + e_t \\ \beta_{i,t} = \alpha_{i,0} + \sum_{p=1}^{P} \alpha_{i,p} \beta_{i,t-p} + \epsilon_{i,t} \end{cases} \tag{7}$$

where y_t an observation variable dependent on the state variable $z_{i,t}$. $\beta_{i,t}$ follows a P-order autoregressive process with $\alpha_{i,p}$ being the transition coefficient for lag p. e_t and $\epsilon_{i,t}$ represent stationary zero mean white noise processes where $\mathbb{E}[e_t] = 0$, $\mathbb{E}[e_t, e_{t'}] = \sigma_e^2 \delta_{tt'}$ and $\mathbb{E}[\epsilon_{i,t}] = 0$, $\mathbb{E}[\epsilon_{i,t}, \epsilon_{i,t'}] = \sigma_{\epsilon_i}^2 \delta_{tt'}$. where $\sigma_e^2 < \infty$ and δ is the delta function.

Then the model parameters $\sigma_e^2, \alpha_{i,0}, \alpha_{i,p}, \sigma_{\epsilon_i}^2 \forall i, p \in \mathbb{N}^+$ are globally identifiable.

This shows that the system is identifiable in the latent space when the causal coefficients follow an R order auto-regressive process. For the general case, when the coefficients are allowed to follow any non-linear function of the lagged values described in 5, although, no formal proof exists but our empirical results strongly suggest that the system remains identifiable.

For the observation space,

$$x_{i,t} = \Phi_i(z_{i,t}) \quad \forall t \in \{1, \cdots, T\}, \forall i \in \{1, \cdots, d\} \tag{8}$$

Rearranging terms

$$\begin{aligned} \mathbf{z_t} &= \beta_t . \mathbf{z_{t-1}} + \mathbf{e_t} \\ \Psi(\mathbf{x_t}) &= \beta_t(\Psi(\mathbf{x_{t-1}})) + \mathbf{e_t} \\ \mathbf{x_t} &= \Psi^{-1}(\beta_t(\Psi(\mathbf{x_{t-1}})) + \mathbf{e_t}) \\ &= \Phi(\beta_t(\Psi(\mathbf{x_{t-1}})) + \mathbf{e_t}) \triangleq G_t(\mathbf{x_{t-1}}) \end{aligned}$$

Thus, we can extend the proof of identifiability to the causal model of $\mathbf{x_t}$.

4 Model Description

The model description can be split-up into three separate objectives - **(1)** Modeling the mapping function Φ **(2)** Modeling the β change function f **(3)** Modeling the long term prediction architecture.

We use an MLP-based autoencoder (AE) architecture to approximate Φ. The autoencoder is optimized through the reconstruction loss of the inputs. The problem formulation describes the latent space as a feature-wise transformation of the observables. Without additional constraints, this implies that the latent variables $\mathbf{Z}$ would simply correspond to an affine transformation of the inputs. To prevent this, we assume that the latent variables can have an entangled time dependent feature. To capture these temporal features, each input data point $\mathbf{x_t} \in \mathbf{X}$ is concatenated with its associated temporal information generating a learnable time-dependent embedding $\mathbf{x_{embed}}$. This modification allows the system to capture dynamic temporal features and can be defined as follows:

$$\mathbf{z} = \Phi(\mathbf{x_{embed}}) \qquad \mathbf{x_t} = \Psi^{-1}(\mathbf{z}, t), \forall t \in [T]$$

The β values at each time step are defined as a learnable parameter array. To approximate the weight change function f, we utilize a standard Transformer architecture incorporating both multi-head self-attention and cross-attention mechanisms. The previous $\beta_{[:t-L]}$ values are used to predict the subsequent β_{t+1}. To ensure the Transformer correctly attends to relevant temporal contexts, a learnable positional embedding is included in the inputs. This setup allows the Transformer to approximate the prior distribution - $p_f(\beta_t|\text{Pa}(\beta_t))$. However, since the β values are not known, this makes the prior objective unbounded. Therefore we employ a variational inference framework V to constrain and refine the learning process. Specifically we approximate the distribution $q_V(\beta_t|\text{Pa}(\beta_t), \text{Pa}(x_t))$ describing the likelihood function conditioned on the observations x_t. Reparameterization trick is then used to minimize the KL divergence between the two distributions. An MLP based model is use to approximate the distribution q_V.

Due to the enforced linear nature of the latent subspace, the next step prediction task is reduced to estimating the next H timesteps for β from the transformer and use them to estimate $\hat{\mathbf{z}}_{[\mathbf{T+1}:\mathbf{T+H}]}$ and $\hat{\mathbf{x}}_{[\mathbf{T+1}:\mathbf{T+H}]}$.

$$\hat{\mathbf{z}}_{\mathbf{i,t+1}} = \beta_{i,t+1}^T \hat{\mathbf{z}}_{\mathbf{i},[\mathbf{t-L}:\mathbf{t}]} + e_{i,t+1} \quad \hat{\mathbf{x}}_{\mathbf{t+1}} = \Psi^{-1}(\hat{\mathbf{z}}_{\mathbf{t+1}})$$

We utilize an autoregressive prediction scheme for context-based predictions and extend this to a multi-step strategy for long-term forecasting. To enhance the stability of the autoregressive process, we employ a teacher-forcing approach during training, to effectively guide the model towards more reliable predictions.

4.1 ELBO Objective

By design, $\mathbf{z}_t$ can be thought of as an alternate representation of $\mathbf{x}_t$, we can thus define the objective function solely in terms of $\mathbf{z}_t$. For simplicity, we assume the look-back window for z and β is the same.

$$p(\mathbf{z}) = \int_\beta Pr(\mathbf{z}|\beta) Pr(\beta) d\beta$$

Estimated Lower Bound (ELBO) for the variational inference framework can be written as:

$$\begin{aligned}
\log p(\mathbf{z}) &\geq \int_\beta q(\beta|\mathbf{x}) \frac{\log p(\mathbf{z}|\beta).p(\beta)}{q(\beta|\mathbf{x})}.d\beta \\
&= \mathbb{E}_{\beta \sim q_V(\beta|\mathbf{x})} \Big[\log p(\mathbf{z}|\beta) - D_{KL}\big(q_V(\beta|\mathbf{x}) || p_f(\beta)\big) \Big] \\
\log p(\mathbf{z}_t) &\geq \mathbb{E}_{\beta_t \sim q_V} \Bigg[\sum_{t=1}^{T+H} \log p\Bigg(\mathbf{z}_t | [\mathbf{z}_{t-\tau}]_{\tau=1}^{\tau=L}, [\beta_{t-\tau}]_{\tau=1}^{\tau=L} \Bigg) \\
&\quad - D_{KL}\Bigg(q_V\left(\beta_t | [\beta_{t-\tau}]_{\tau=1}^{\tau=L}, [\mathbf{x}_{t-\tau}]_{\tau=1}^{\tau=L}\right) || p_f\left(\beta_t | [\beta_{t-\tau}]_{\tau=1}^{\tau=L}\right) \Bigg) \Bigg]
\end{aligned}$$

The above expression is the ELBO for $\log p(\mathbf{z}_t)$. As, $\mathbf{x}_t = \Phi^{-1}(\mathbf{z}_t)$ we thus obtain an estimated lower bound for $p(\mathbf{x}_t)$.

Rewriting the terms -

$$\text{ELBO} \triangleq \mathcal{L}_{prior} + \mathcal{L}_{KL}$$

$$\mathcal{L}_{KL} = -D_{KL}\Big(q_V\big(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{\tau=L}, [\mathbf{x}_{t-\tau}]_{\tau=1}^{\tau=L}\big)||p_f\big(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{\tau=L}\big)\Big)$$

$$\begin{aligned}\mathcal{L}_{prior} &= \sum_{t=1}^{T+H} \log p\Big(\mathbf{z}_t|[\mathbf{z}_{t-\tau}]_{\tau=1}^{\tau=L}, [\beta_{t-\tau}]_{\tau=1}^{\tau=L}\Big) \\ &= \underbrace{\sum_{t=1}^{T} \log p\Big(\mathbf{z}_t|[\mathbf{z}_{t-\tau}]_{\tau=1}^{\tau=L}, [\beta_{t-\tau}]_{\tau=1}^{\tau=L}\Big)}_{\mathcal{L}_{present}} \quad + \underbrace{\sum_{t=T+1}^{T+H} \log p\Big(\mathbf{z}_t|[\mathbf{z}_{t-\tau}]_{\tau=1}^{\tau=L}, [\beta_{t-\tau}]_{\tau=1}^{\tau=L}\Big)}_{\mathcal{L}_{future}}\end{aligned}$$

The prediction network f and the variation network V are used to approximate p_f and q_V respectively. The KL divergence loss $\mathcal{L}_{KL}$ can be easily computed over the estimated distributions. β_t is obtained using the reparmetrization trick.

The prior term $\mathcal{L}_{prior}$ can be split into two components based on the time periods, the observed time period operating on the input context length as 'present', timesteps $t \in [1, T]$ and the prediction horizon on the prediction length as 'future', timesteps $t \in [T+1, T+H]$

In the 'present' setting, the prior of the latent space representation can be optimized by decomposing it into multiple next-step prediction tasks, effectively simplifying the prediction problem into sequential, stepwise objectives. Conversely, for the future objective, the prior of the latent space is approximated using the long-term prediction outcomes across the entire forecasting horizon. The distinction between present and future scenarios is crucial, as it highlights the differing methodologies required for approximating the likelihood priors in each context.

4.1.1 Prior Likelihood

For the 'present' case, we find that the prior likelihood reduces to,

$$\log p(\mathbf{z_t}|[\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) = \sum_{j=1}^{d} \log p(e_{j,t})$$

Thus, the prior can be approximated as summation over the residuals associated with the prediction of $\mathbf{z_t}$. The detailed derivation of the formulation is presented in Appendix A.1 Based on our model design, we can estimate $\mathbf{z_t}$ in two ways:

1. forecast $\hat{\mathbf{z}}_\mathbf{t}$ using β_t and
2. value obtained from the auto-encoder as $\mathbf{z_t} = \Phi(\mathbf{x_t})$ - since have access to $\mathbf{x_t}$.

The residual will be equal to the difference between the values obtained from the two prediction schemes. This is only feasible in the 'present' setting as we have access to the next time step value in the observation space. Furthermore, since β_t is required for the forecast and is calculated exclusively based on previously evaluated values, the prior estimation can be made sufficiently precise. This ensures that the results of both the prediction schemes align.

For the 'future' case,

$$\log p(\mathbf{z_t}|[\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) + \log p(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{L}]) = \sum_{j=1}^{d} \log p(e_{j,t}) + \log p(\epsilon_t) + \log \left|\frac{\partial \epsilon_t}{\partial \beta_t}\right|$$

This result arises from the fact that the transformations in β_t are completely independent of any other factor. This independence can be formally demonstrated by applying the change of variables formula exclusively to the sequences of β. Thus, we can say that the above equation can be decomposed into two independent components: one term corresponding to the latent variables and others corresponding to the causal weights β.

$$\log p(\mathbf{z_t}|[\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) = \sum_{j=1}^{d} \log p(e_{j,t})$$

and

$$\log p(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{L}]) = \log p(\epsilon_t) + \log \left| \frac{\partial \epsilon_t}{\partial \beta_t} \right|$$

Considering the first equation, direct minimization of the objective function is not feasible in the same way as in the 'present' case. However, since the objective function can be alternatively expressed in terms of the observables, the task of minimizing the objective becomes equivalent to minimizing the long term prediction loss. This allows us to effectively align the optimization process with observable data, leveraging predictive performance as a surrogate for the original objective.

The second equation describes the prediction probability of β_t conditioned on its parents. This objective is already addressed within the model by estimating a bound on the KL divergence between the prior and the variational approximation q_V, as described by the loss term $\mathcal{L}_{KL}$. This approach effectively constrains and bounds the predictive probability of β, ensuring that the inferred distributions remain consistent with the modeled dependencies.

The final loss term can be expressed as sum of reconstruction loss from the autoencoder and the ELBO terms.

$$\mathcal{L}_{total} = \mathcal{L}_{AE} + \mathcal{L}_{prior} + \mathcal{L}_{KL}$$

5 Experiments

5.1 Synthetic Data

We generate synthetic data based on the non-linear, time-varying causal model described in Section 3.2. Specifically, we explore three different settings by varying the time-varying process f governing the change of β, and the projection function Ψ applied to the latent variables.

Case 1: f is linear The causal weights are assumed to follow an AR(2) process of the form

$$\beta_t = A_1\beta_{t-1} + A_2\beta_{t-2} + \epsilon_t$$

Stability of this process can be defined in terms of the eigen values of the associated companion matrix. Define companion matrix C.

$$C = \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix}$$

If β_t has d_b features, then C is a $(2d_b \times 2d_b)$ matrix, with I being an identity matrix and 0 begin a null matrix each of dimension d_b.The underlying AR process will be stable if the largest eigen value of C is less than or equal to one. We can manually define the coefficients A_i to be such that the condition is satisfied. Alternatively, we can initialize them with random values and find the optimum values that satisfy the condition, using any gradient based optimizer. We follow the latter approach with the initial values as $A_1 = A_2 = I$ and taking ϵ_t as a zero mean Gaussian noise.

For the latent variables,

$$z_t = \beta_t^1 z_{t-1} + \beta_t^2 z_{t-2} \cdots \beta_t^L z_{t-L} + e_t$$

the companion matrix at time t can be defined as

$$C_t' = \begin{bmatrix} \beta_t^1 & \beta_t^1 & \cdots & \beta_t^{L-1} & \beta_t^L \\ I & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}$$

The β values change over time. To ensure that the eigen values of C_t' are less than one for all time steps, we can scale all the β values. The optimum scale value can be found using gradient descent. Interestingly, we observe that the scale value is approximately equal to the maximum singular value of the companion matrix of the entire time series.

Further we can generate the observables as

$$x_t = \Psi^{-1}(z_t, t) \quad \text{or equivalently} \quad x_t = \Psi^{-1}(h(z_t), s(t))$$

where h and s describe some transformations of z and t respectively. Following this scheme, we generate two datasets,

LinearZ β follows an AR(2) process and $x_t = \Psi^{-1}(h(z_t), t)$. We consider simple quadratic and exponential functions of z and use the time step as a trend component. Ψ is an MLP architecture initialized with random weights.

LinearT β follows an AR(2) process and $x_t = \Psi^{-1}(z_t, s(t))$. In this case we introduce some seasonality and trend based components in the generation of the observables.

Case 2: f is non-linear process Ta more general case where the stability of the process for β cannot be explicitly defined explicitly and depends heavily on careful initialization of parameters at $t = 0$. However, the stability conditions for the latent variable generation process, as previously mentioned, still apply. Using this scheme, we generate one dataset.

NonLinearT β_t follows a non-linear function of its past values with a lag of 2 and $x_t = \Psi^{-1}(h(z_t), s(t))$. Simple quadratic and trigonometric functions of both z and t are used to generate the observables. This dataset contains a significant seasonality component compared to the linear datasets. Both f and Ψ are characterized as MLPs initialized with random weights.

For all the three datasets, we generate 6000 samples, with 4 features and a look-back window of 5 time-steps. For long-term forecasting task, we use an input with a context length of 30 time-steps and evaluate prediction accuracy for forecast windows of 60 and 90 time-steps. The plots for the observables and latents are presented in Appendix A.2

Baselines:

- **Vector ARIMA** - VARIMA (Stock & Watson (2001)) Captures dependencies among multiple variables, accounting for autoregressive, differencing (integration), and moving average components to forecast future values.
- **N-HiTS Model** - N-HiTS (Neural Hierarchical Interpolation for Time Series) (Challu et al. (2022)) is a deep learning model which uses a hierarchical interpolation mechanism to capture temporal patterns across different resolutions.
- **Autoformer** (Wu et al. (2022))- Transformer based model that uses an auto-correlation mechanism to capture long-range dependencies, combined with a decomposition block to separate trend and seasonal components for improved long-term forecasting.
- **D-Linear** (Zeng et al. (2022))- Linear model that decomposes the time series into trend and seasonal components and uses a single linear layer.
- **TiDE** (Das et al. (2024)) - Linear model that leverages a dense encoder-decoder architecture to efficiently capture both short and long-term dependencies.

All the models are evaluated using RMSE, and results are presented in Table 1.

Methods	LinearZ		LinearT		NonLinearT	
	(30∼60)	(30∼90)	(30∼60)	(30∼90)	(30∼60)	(30∼90)
VARIMA	0.138	0.161	0.153	0.242	0.195	0.120
N-HiTS	0.292	0.321	0.249	0.225	0.221	0.321
Autoformer	0.161	0.194	0.223	0.228	0.224	0.291
D-Linear	0.109	0.157	0.162	0.229	0.073	0.095
TiDE	0.092	0.152	0.158	0.225	**0.065**	**0.078**
Ours	**0.079**	**0.139**	**0.146**	**0.212**	0.072	0.110

Table 1: RMSE scores over the synthetic datasets

5.2 Real-world data

Wang et al. (2023) demonstrates the forecastability of various popular datasets used to benchmark models for long-term forecasting objective. Following Wang et al. (2023), we choose to evaluate our model on the M4 dataset due to its highly non-stationary nature. Specifically we choose, **M4-Weekly**

and **M4-Daily** datasets. The datasets are composed of univariate time series divided into multiple categories such as finance, demographics and industry. We follow the experimental setup as described in the M4 competition.

Baselines: We compare the performance of our model against the top performing models from the original M4 competition Makridakis et al. (2020) as well as Koopman Neural forecaster because of its promising results. **Evaluation Metrics:** We evaluate our results using the sMAPE criterion, as outlined in the competition.

Table 2 compares our model performance with KNF, the top performing models (Makridakis et al. (2018)) and the baselines provided by the dataset authors. Our model performs well on the M4-Daily dataset but not on the M4-Weekly. We can also see that it is significantly better than the baselines provided, and based on the competition average, our approach scores quite highly. Our model is heavily dependent on the look-back window length and the nature of the β encoders and decoders used. There is considerable scope of improvement by exploring different hyperparameters and model architectures.

Methods	M4-Weekly (45∼13)	M4-Daily (18∼14)
Smyl S.	7.817	3.170
Montero-Manso et.al	7.625	3.097
KNF	7.254	2.990
Ours	**8.091**	**3.144**
M4-benchmark (naive)	9.161	3.045
M4-benchmark (Com)	8.944	2.980
M4-benchmark (MLP)	21.349	9.321

Table 2: sMAPE scores for M4 datasets

5.3 Causal Discovery over the observables

We compare the discovered β values with the true causal weights for the three synthetic datasets. The weights discovered by our model are sampled from the distribution estimated by the encoder and decoder, thus the weights can be scale shifted. Additionally, we threshold the weights at random values for each time point and calculate the precision recall and F1 scores. The results are promising for the linear case. For non-linear dataset, the scale disparity is much higher than the linear case, while the magnitudes of the values are low. The results are presented in Table 3

	RMSE	Precision	Recall	F1-score
LinearZ	0.15	0.68	0.63	0.62
LinearT	0.26	0.63	0.49	0.55
NonLinearT	0.07	0.43	0.38	0.40

Table 3: Causal Discovery results

6 Conclusion

In this work, we propose a unified framework for causal discovery and long-term forecasting on non-stationary time series data. By leveraging a causal perspective, our approach enhances long-term forecasting accuracy, as demonstrated by comparative results on both synthetic and real-world datasets. The framework also transforms causal discovery into a scalable, non-parametric objective, effectively addressing many limitations commonly associated with this task. Future directions include extending the framework to incorporate contemporaneous causal relations and conducting experiments with different modalities such as video.

References

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting, 2022. URL https://arxiv.org/abs/2201.12886.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder, 2024. URL https://arxiv.org/abs/2304.08424.

Doris Entner and Patrik Hoyer. On causal discovery from time series data using fci. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010*, 09 2010.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan A. Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series, 2024a. URL https://arxiv.org/abs/2407.07291.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan A. Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series, 2024b. URL https://arxiv.org/abs/2407.07291.

Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, and Jingping Bi. Causal discovery from temporal data: An overview and new perspectives, 2023. URL https://arxiv.org/abs/2303.10112.

Jake Grigsby, Zhe Wang, Nam Nguyen, and Yanjun Qi. Long-range transformers for dynamic spatiotemporal forecasting, 2023. URL https://arxiv.org/abs/2109.12218.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022. URL https://arxiv.org/abs/2111.00396.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2901–2910. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/huang19g.html.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2018.06.001. URL https://www.sciencedirect.com/science/article/pii/S0169207018300785.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169207019301128. M4 Competition.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. Dynotears: Structure learning from time-series data, 2020. URL https://arxiv.org/abs/2002.00498.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, 2022. URL https://arxiv.org/abs/2003.03685.

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

Peter Spirtes. An anytime algorithm for causal inference. In Thomas S. Richardson and Tommi S. Jaakkola (eds.), *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pp. 278–285. PMLR, 04–07 Jan 2001. URL https://proceedings.mlr.press/r3/spirtes01a.html. Reissued by PMLR on 31 March 2021.

James H. Stock and Mark W. Watson. Vector autoregressions. *Journal of Economic Perspectives*, 15 (4):101–115, December 2001. doi: 10.1257/jep.15.4.101. URL https://www.aeaweb.org/articles?id=10.1257/jep.15.4.101.

Kent D. Wall. Identification theory for varying coefficient regression models. *Journal of Time Series Analysis*, 8(3):359–371, 1987. doi: https://doi.org/10.1111/j.1467-9892.1987.tb00447.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.1987.tb00447.x.

Rui Wang, Yihe Dong, Sercan Ö. Arik, and Rose Yu. Koopman neural forecaster for time series with temporal distribution shifts, 2023. URL https://arxiv.org/abs/2210.03675.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022. URL https://arxiv.org/abs/2106.13008.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL https://arxiv.org/abs/2205.13504.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting, 2021. URL https://arxiv.org/abs/2012.07436.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022. URL https://arxiv.org/abs/2201.12740.

A Appendix

A.1 Prior Likelihood derivation

For present timesteps: For derivation sake, we simplify the problem statement - Consider only two features i.e. $\mathbf{z_{t+1}} = [z_{1,t+1}, z_{2,t+1}]$ with a maximum lag of $L = 1$. For the 'present' case, β_{t+1} is obtained from the past time-lagged values $\beta_{t-\tau}$ (where $\tau \in [L]$) , thus for all purposes, β_{t+1} can be assumed independent of $\mathbf{z_t}$. Also, lets assume that some invertible function $\mathbf{f}$ exists which transforms the sequence $A := [z_{1,t}, z_{2,t}, \beta_{t+1}, z_{1,t+1}, z_{2,t+1}]$ to $B := [z_{1,t}, z_{2,t}, \beta_{t+1}, \eta_{1,t+1}, \eta_{2,t+1}]$. Therefore, by the change of variables formula:

$$\log p(A) = \log p(B) + \log |det(J_{A\to B})|$$

$$\log p[z_{1,t}, z_{2,t}, \beta_{t+1}, z_{1,t+1}, z_{2,t+1}] = \log p[z_{1,t}, z_{2,t}, \beta_{t+1}, \eta_{1,t+1}, \eta_{2,t+1}] + \log |det(J_{A\to B})|$$

Calculating the Jacobian, we observe that it is of the general form-

$$J_{A\to B} = \begin{bmatrix} \mathbb{I} & 0 \\ * & \text{diag}\Big(\frac{\partial \mathbf{f}_i^{-1}}{\partial z_{i,t+1}}\Big) \end{bmatrix}$$

for each of the i features of z. Therefore

$$\log p(\mathbf{z_t}, \beta_{t+1}, \mathbf{z_{t+1}}) = \log p(\mathbf{z_t}, \beta_{t+1}, \eta_{t+1}) + \sum_{i=j}^{n} \log \left| \frac{\partial \mathbf{f_i}^{-1}}{\partial z_{i,t+1}} \right|$$

Taking LHS:

$$\begin{aligned} \log p(\mathbf{z_t}, \beta_{t+1}, \mathbf{z_{t+1}}) &= \log p(\mathbf{z_{t+1}} | \beta_{t+1}, \mathbf{z_t}) + \log p(\beta_{t+1}, \mathbf{z_t}) \\ &= \log p(\mathbf{z_{t+1}} | \beta_{t+1}, \mathbf{z_t}) + \log p(\beta_{t+1}) + \log p(\mathbf{z_t}) \end{aligned}$$

As clearly $\beta_{t+1} \perp\!\!\!\perp \mathbf{z_t}$. For the first term on RHS, we have $\eta_{t+1} \perp\!\!\!\perp \mathbf{z_t}$ and $\beta_{t+1} \perp\!\!\!\perp \mathbf{z_t}$.

$$\log p(\mathbf{z_t}, \beta_{t+1}, \eta_{t+1}) = \log p(\mathbf{z_t}) + \log p(\beta_{t+1}) + \log p(\eta_{t+1})$$

Simplifying we get,

$$\log p(\mathbf{z_{t+1}} | \mathbf{z_t}, \beta_{t+1}) = \log p(\eta_{t+1}) + \sum_{i=j}^{n} \log \left| \frac{\partial \mathbf{f_i}^{-1}}{\partial z_{i,t+1}} \right|$$

Generalizing -

$$\log p(\mathbf{z_t} | [\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) = \sum_{j=1}^{d} \log p(\eta_{j,t}) + \sum_{i=j}^{d} \log \left| \frac{\partial \mathbf{f_i}^{-1}}{\partial z_{i,t}} \right|$$

By definition, we assume that the causal relations in the latent space are linear, therefore the Jacobian term is a constant. We can directly write the prior as :

$$\log p(\mathbf{z_t} | [\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) = \sum_{j=1}^{d} \log p(\eta_{j,t})$$

For the future case: For $t \geq T$ similar to the previous case, we can simplify the prior estimation problem into a transformation from sequence $A := [z_{1,t}, z_{2,t}, \beta_t, z_{1,t+1}, z_{2,t+1}, \beta_{t+1}]$ to $B :=$

$[z_{1,t}, z_{2,t}, \beta_t, \eta_{1,t+1}, \eta_{2,t+1}, \epsilon_{t+1}]$. This is different from the present case, since this also involves estimating β_{t+1} from β_t. (For simplicity, assuming only a lag of $L = 1$). Here ϵ_{t+1} is the noise involved in estimating β_{t+1}. Using the change of variables formula-

$$\log p(A) = \log p(B) + \log |det(J_{A\to B})|$$

$\log p[z_{1,t}, z_{2,t}, \beta_t, z_{1,t+1}, z_{2,t+1}, \beta_{t+1}] = \log p[z_{1,t}, z_{2,t}, \beta_t, \eta_{1,t+1}, \eta_{2,t+1}, \epsilon_{t+1}] + \log |det(J_{A\to B})|$
Evaluating the jacobian

$$J_{A\to B} = \begin{bmatrix} \mathbb{I} & 0 & 0 \\ * & \text{diag}\left(\frac{\partial \mathbf{f}_i^{-1}}{\partial z_{i,t+1}}\right) & 0 \\ * & * & \left(\frac{\partial \epsilon_{t+1}}{\partial \beta_{t+1}}\right) \end{bmatrix}$$

Therefore -

$$\log p[\mathbf{z_t}, \beta_t, \mathbf{z_{t+1}}, \beta_{t+1}] = \log p[\mathbf{z_t}, \beta_t, \eta_{t+1}, \epsilon_{t+1}] + \sum_{i=j}^{n} \log \left|\frac{\partial \mathbf{f_i}^{-1}}{\partial z_{i,t+1}}\right| + \log \left|\frac{\partial \epsilon_{t+1}}{\partial \beta_{t+1}}\right|$$

Taking LHS

$$\begin{aligned} \log p[\mathbf{z_t}, \beta_t, \mathbf{z_{t+1}}, \beta_{t+1}] &= \log p(\mathbf{z_{t+1}}|\beta_t, \mathbf{z_t}, \beta_{t+1}) + \log p(\beta_t, \mathbf{z_t}, \beta_{t+1}) \\ &= \log p(\mathbf{z_{t+1}}|\mathbf{z_t}, \beta_{t+1}) + \log p(\mathbf{z_t}|\beta_t, \beta_{t+1}) + \log p(\beta_t, \beta_{t+1}) \\ &= \log p(\mathbf{z_{t+1}}|\mathbf{z_t}, \beta_{t+1}) + \log p(\mathbf{z_t}|\beta_t) + \log p(\beta_{t+1}|\beta_t) + \log p(\beta_t) \\ &= \log p(\mathbf{z_{t+1}}|\mathbf{z_t}, \beta_{t+1}) + \log p(\mathbf{z_t}, \beta_t) + \log p(\beta_{t+1}|\beta_t) \end{aligned}$$

Using the facts that $\mathbf{z_{t+1}} \perp\!\!\!\perp \beta_t|\mathbf{z_t}, \beta_{t+1}$ and $\mathbf{z_t} \perp\!\!\!\perp \beta_{t+1}$ Now for the first term in RHS -

$$\log p[\mathbf{z_t}, \beta_t, \eta_{t+1}, \epsilon_{t+1}] = \log p(\mathbf{z_t}, \beta_t) + \log p(\eta_{t+1}) + \log p(\epsilon_{t+1})$$

Using the independent noise conditions stating $\eta_{t+1} \perp\!\!\!\perp \mathbf{z_t}$, $\eta_{t+1} \perp\!\!\!\perp \beta_t$ and $\epsilon_{t+1} \perp\!\!\!\perp \mathbf{z_t}$, $\epsilon_{t+1} \perp\!\!\!\perp \beta_t$, Therefore we get, -

$$\log p(\mathbf{z_{t+1}}|\mathbf{z_t}, \beta_{t+1}) + \log p(\beta_{t+1}|\beta_t) = \log p(\eta_{t+1}) + \log p(\epsilon_{t+1}) + \sum_{i=j}^{n} \log \left|\frac{\partial \mathbf{f_i}^{-1}}{\partial z_{i,t+1}}\right| + \log \left|\frac{\partial \epsilon_{t+1}}{\partial \beta_{t+1}}\right|$$

Similar to previous case, relations in $\mathbf{z_t}$ are linear so the jacobian is constant and can be ignored.

$$\log p(\mathbf{z_{t+1}}|\mathbf{z_t}, \beta_{t+1}) + \log p(\beta_{t+1}|\beta_t) = \log p(\eta_{t+1}) + \log p(\epsilon_{t+1}) + \log \left|\frac{\partial \epsilon_{t+1}}{\partial \beta_{t+1}}\right|$$

or more generally -

$$\log p(\mathbf{z_t}|[\mathbf{z_{t-\tau}}]_{\tau=\mathbf{1}}^{\mathbf{L}}, \beta_t) + \log p(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{L}]) = \sum_{j=1}^{n} \log p(\eta_{j,t}) + \log p(\epsilon_t) + \log \left|\frac{\partial \epsilon_t}{\partial \beta_t}\right|$$

To observe independent results with β_t , we can take $A := [\beta_t, \beta_{t+1}]$ and $B := [\beta_t, \epsilon_{t+1}]$, we get -

$$\log p(\beta_t|[\beta_{t-\tau}]_{\tau=1}^{L}]) = \log p(\epsilon_t) + \log \left|\frac{\partial \epsilon_t}{\partial \beta_t}\right|$$

A.2 Synthetic Data

Figure 1 shows the plots for the generated data, the latent variables and randomly indexed β values.

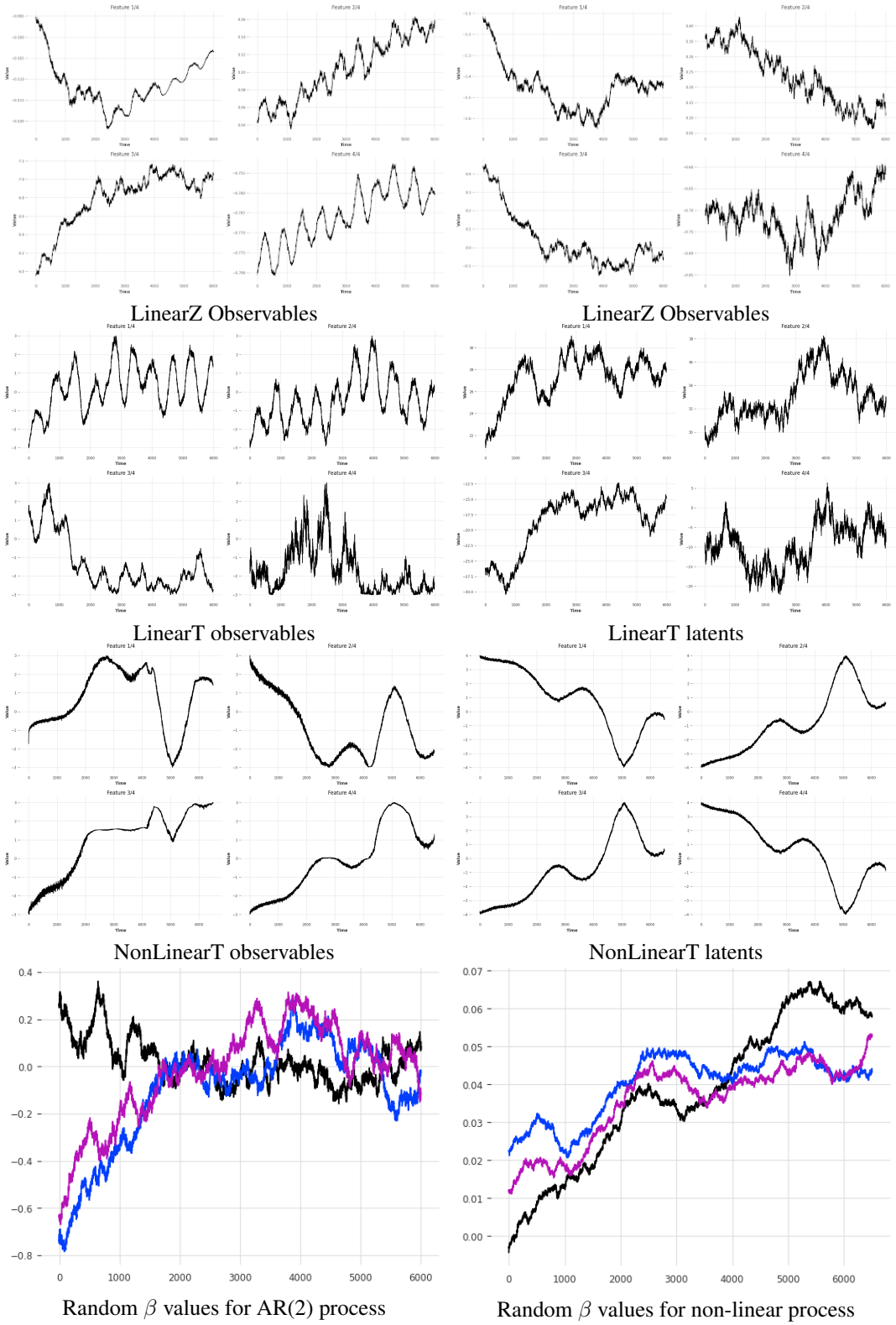

Figure 1: Synthetic data