HiBaNG: Hierarchical Bayesian Nonparametric Granger Causal Discovery in Low-Data Regimes

Anonymous authors
Paper under double-blind review

Abstract

We present a principled probabilistic framework for discovering Granger causal relationships from multivariate time-series data in low-data regimes, where short sequences limit the applicability of modern deep learning approaches. While deep neural vector autoregressive (VAR) models perform well in high-data settings, they often struggle to generalize with limited samples and provide little insight into model uncertainty. To address these challenges we introduce HiBaNG, a hierarchical Bayesian nonparametric framework for Granger causal discovery. HiBaNG places a hierarchical factorized prior over binary Granger causal graphs that encodes structured sparsity and enables interpretable, uncertainty-aware inference. We develop a tractable Gibbs sampling algorithm that exploits conjugacy and augmentation for scalable posterior estimation. Extensive experiments on synthetic, semi-synthetic, and real-world climate datasets demonstrate that HiBaNG consistently outperforms both classical and deep VAR baselines, achieving improved accuracy and calibrated uncertainty.

1 Introduction

Multivariate time-series (MTS) data consist of observations of multiple variables recorded at multiple timestamps and are fundamental to a wide range of applications in economics, healthcare, climatology, and neuroscience. In these domains, uncovering causal relationships among time-series is often essential for understanding system dynamics and supporting decision-making.

In this paper, we focus on discovering such relationships from observational MTS data using Granger Causality (GC) (Granger, 1969; Lütkepohl, 2005; Shojaie & Fox, 2022), which posits that one variable is causal for another if its past values provide statistically significant information for predicting the future of the latter, beyond what is contained in its own past. While GC does not necessarily imply a structural or interventional causal relationship in the sense of, e.g., Pearl's do-calculus (Pearl, 2009) or Rubin's potential outcomes (Imbens & Rubin, 2015), it remains widely used for time-series causality[1]. The most common implementation of GC is via Vector Autoregressive (VAR) models (Lütkepohl, 2005), which assume that each variable is a function of the lagged values of other variables.

Bayesian VARs (Woźniak, 2016; Miranda-Agrippino & Ricco, 2019) extend classical VARs by incorporating prior beliefs, enabling uncertainty quantification and improved estimation under limited data. More recently, deep learning-based VAR models (Montalto et al., 2015; Tank et al., 2018; Wang et al., 2018; Nauta et al., 2019; Khanna & Tan, 2020; Wu et al., 2020; Marcinkevičs & Vogt, 2021; Gong et al., 2022; Fan et al., 2023) have gained popularity for their ability to capture complex, nonlinear dynamics, provided ample data are available. In practice, however, many real-world applications do not have the luxury of abundant data, especially during the early stages of data collection. In these low-data regimes, where the number of samples is small relative to the number of variables, deep VARs often underperform. This is mainly due to: **1)** High variance and overfitting risks when learning complex models with limited data, leading to unreliable predictions and poor uncertainty estimates (Geffner et al., 2022; Annadani et al., 2023; Deleu et al., 2023); **2)** The difficulty of model selection without access to sufficient validation data or ground-truth causal graphs,

[1]Throughout this paper, causality refers specifically to Granger causality.

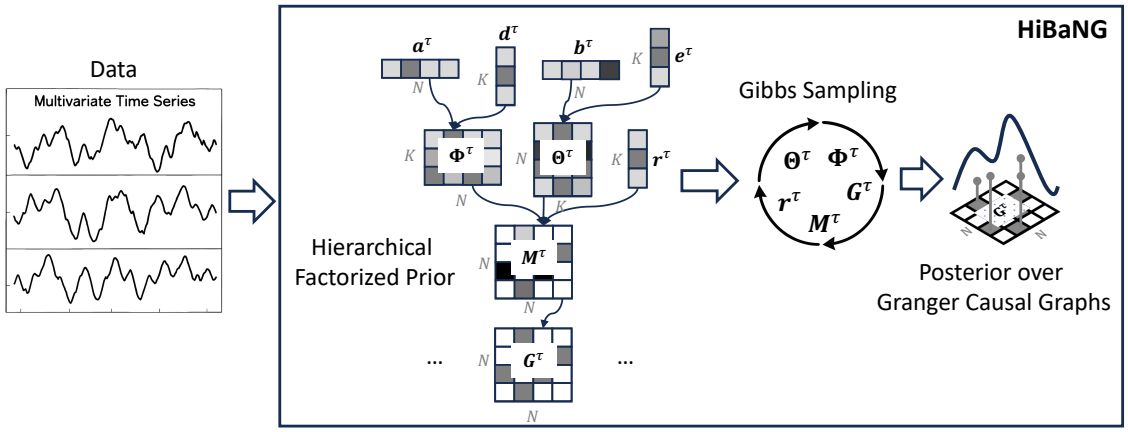


Figure 1: HiBaNG models the given MTS data via a set of binary Granger causal graphs $\boldsymbol{G}^\tau$ as latent variables with a hierarchical factorized prior. It learns to draw samples from the posterior distributions over the graphs via an efficient Gibbs sampling algorithm that exploits the nonparametric Bayesian nature of the model and conjugacy, improving stability and significantly reducing the number of hyperparameters to tune. Because of this and because of the information shared across related parameters through a structured prior, HiBaNG is particularly suited to low-data settings, allowing the model to remain expressive while mitigating overfitting and enhancing generalization.

making hyperparameter tuning unreliable and potentially biased and, ultimately, compromising robustness and interpretability.

To address these challenges, we propose a new hierarchical Bayesian VAR framework for Granger causal discovery, specifically designed for low-data regimes. Our method addresses the aforementioned issues by:

1. Introducing a hierarchical factorized prior over binary Granger causal graphs, which encodes structured sparsity and incorporates inductive bias in the absence of abundant data;
2. Decomposing Granger causality into discrete (graph structure) and continuous (causal strength) components, allowing the binary graph to constrain parameter estimation and reduce overfitting;
3. Leveraging Bayesian nonparametric techniques to integrate out latent factors and reduce the number of tunable hyperparameters.

Overall contribution: We develop **HiBaNG**, a **hi**erarchical **Ba**yesian **n**onparametric framework for **G**ranger causal discovery in data-scarce settings, which integrates interpretable priors, principled uncertainty quantification, and enables tractable posterior inference via Gibbs sampling. Our extensive experiments on synthetic, semi-synthetic, and real-world climate datasets show that HiBaNG attains improved or competitive performance relative to both classical and deep VAR baselines.

2 Preliminaries

Consider a collection of MTS data of N time series or variables in T timestamps, stored in the matrix of $\boldsymbol{X} \in \mathbb{R}^{N\times T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ where $\boldsymbol{x}_t \in \mathbb{R}^N$ consists of the samples/values of the N variables at timestamp $t \in \{1, \ldots, T\}$. A VAR model for Granger causality (Lütkepohl, 2005; Hyvärinen et al., 2010) assumes that $\boldsymbol{x}_t$ can be predicted from the $\tau_{\max}$ time lags $\{\boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-\tau_{\max}}\}$ by learning a coefficient matrix $\boldsymbol{A}^\tau \in \mathbb{R}^{N\times N}$ for each lag $\tau \in \{1, \ldots, \tau_{\max}\}$:

$$\boldsymbol{x}_t = \sum_{\tau=1}^{\tau_{\max}} \boldsymbol{A}^\tau \boldsymbol{x}_{t-\tau} + \epsilon_t, \tag{1}$$

where ϵ_t is an independent noise variable. Conventionally, variable j (the parent) does not Granger-cause variable i (the child) ($i, j \in \{1, \ldots, N\}$) if and only if for all τ, $A^\tau_{ij} = 0$. For deterministic VARs, learning

can be done by minimizing a regression error: $\min_{\{\boldsymbol{A}^\tau\}_\tau^{\tau_{\max}}} \|\boldsymbol{x}_t - \sum_{\tau=1}^{\tau_{\max}} \boldsymbol{A}^\tau \boldsymbol{x}_{t-\tau}\|_2^2 + \lambda \operatorname{reg}(\{\boldsymbol{A}^\tau\}_\tau^{\tau_{\max}})$ where $\operatorname{reg}(\{\boldsymbol{A}^\tau\}_\tau^{\tau_{\max}})$ is a sparsity-inducing penalty e.g., a group lasso penalty (Yuan & Lin, 2006; Lozano et al., 2009): $\sum_{i=1,j=1}^{N} \|A_{i,j}^\tau\|_2$. Other alternative penalties can be found in Nicholson et al. (2017).

Bayesian VARs (BVARs) (Litterman, 1986) are another important line of research especially in econometrics and finance. A standard method may model $\boldsymbol{x}_t$ with multivariate normal distributions: $\boldsymbol{x}_t \sim \mathcal{MN}\left(\sum_{\tau=1}^{\tau_{\max}} \boldsymbol{A}^\tau \boldsymbol{x}_{t-\tau}, \Sigma\right)$, where various priors can be imposed on $\{\boldsymbol{A}^\tau\}_{\tau=1}^{\tau_{\max}}$ (e.g., sparsity-inducing priors) and Σ (e.g., Inverse-Wishart priors). Learning BVARs involves inferring the posterior of $\{\boldsymbol{A}^\tau\}_{\tau=1}^{\tau_{\max}}$ and Σ. In standard BVARs, one may need to "convert" $\{\boldsymbol{A}^\tau\}_{\tau=1}^{\tau_{\max}}$ into GC graphs.

Deep VARs have recently become popular, as they use deep neural networks to model nonlinear dynamics between timestamps, which essentially generalize Eq. (1) using: $x_{it} = f_i\left(\sum_{\tau=1}^{\tau_{\max}} \boldsymbol{A}^\tau \boldsymbol{x}_{t-\tau}\right) + \epsilon_t$, where f_i is typically implemented with nonlinear neural networks.

3 Method

In this section, we first present our BVAR model in a general form, which separates the coefficients into binary GC graphs and weight matrices; then we propose a new link function that helps build a hierarchical model on binary GC graphs; subsequently, we provide details of the full model and its Bayesian inference algorithm. At the end of the section, we discuss some properties and design choices in our model that allow us to tackle the challenges mentioned in the introduction.

Our overall model can be seen as a Bayesian VAR of the form:

$$\boldsymbol{x}_t \sim \mathcal{MN}\left(\sum_{\tau=1}^{\tau_{\max}} \left(\boldsymbol{A}^\tau \odot \boldsymbol{G}^\tau\right) \boldsymbol{x}_{t-\tau}, \Sigma\right), \tag{2}$$

where $\boldsymbol{G}^\tau \in \{0,1\}^{N\times N}$ is the adjacency matrix for the binary GC graph of lag τ and $\odot$ denotes the Hadamard product. We further impose the following conjugate prior distributions (Miranda-Agrippino & Ricco, 2019) on $\boldsymbol{A}^\tau$: $\psi_{i,j}^\tau \sim \operatorname{Gamma}(1,1), A_{i,j}^\tau \sim \mathcal{N}\left(0, (\psi_{i,j}^\tau)^{-1}\right)$ and on Σ: $\lambda_i \sim \operatorname{Gamma}(1,1), \Sigma = \operatorname{diag}(\lambda_1,\ldots,\lambda_N)^{-1}$ where $\operatorname{diag}(\lambda_1,\ldots,\lambda_N)$ returns a matrix with its diagonal elements as $\lambda_1,\ldots,\lambda_N$. In our model, the impact of variable j on i is modelled by two components: $G_{i,j}^\tau \in \{0,1\}$ indicating whether there is a link between i and j and $A_{i,j}^\tau \in \mathbb{R}$ indicating the weight of the link. If $G_{i,j}^\tau = 0$, j does not impact i in lag τ regardless of the value of $A_{i,j}^\tau$ while if $G_{i,j}^\tau = 1$, $A_{i,j}^\tau$ captures the influence from j to i. We also note that variable j does not Granger-cause variable i ($i,j \in \{1,\ldots,N\}$) if and only if for all τ, $G_{i,j}^\tau = 0$.

3.1 Generalized Bernoulli Poisson Link

Before describing the model in full detail, we propose a new link function named *generalized Bernoulli Poisson link* (GBPL) that thresholds a random Poisson variable m at $V \in \{1,2,\ldots\}$ to obtain a binary variable b, as one of the key building blocks in our approach.

Definition 3.1. (Generalized Bernoulli Poisson Link)

$$m \sim \operatorname{Poisson}(\gamma), b = \mathbf{1}(m \geq V),$$

where $\mathbf{1}(\cdot)$ is function returning one if the condition is true otherwise zero.

Property 3.2. Given γ and V, one can marginalize m out to get: $b \sim \operatorname{Bernoulli}\left(1 - \sum_{v=0}^{V-1} \frac{e^{-\gamma}\gamma^v}{v!}\right)$.

Remark. As $b = 0$ if and only if $m < V$, $p(b=0) = \sum_{v=0}^{V-1} p(m=v)$, thus, $p(b=1) = 1 - \sum_{v=0}^{V-1} p(m=v)$. Moreover, as $\mathbb{E}(b) = p(b=1)$, larger V leads to lower expected probability of b being one, under the same γ.

Property 3.3. Given b, the conditional posterior of m is in close-form:

$$m \sim \begin{cases} \operatorname{TPoisson}_V(\gamma), & \text{if } b > 0 \\ \operatorname{Categorical}_V([\ldots, f(v,\gamma), \ldots]), & \text{otherwise} \end{cases}$$

where $\text{TPoisson}_V(\gamma)$ is the Poisson distribution with parameter γ left-truncated at V (i.e., the samples from that Poisson distribution are greater than or equal to V) and $f(v,\gamma) = \frac{\frac{e^{-\lambda}\lambda^v}{v!}}{\sum_{v'=0}^{V-1}\frac{e^{-\lambda}\lambda^{v'}}{v'!}}$ is the normalized Poisson probability mass function. To sample from a truncated Poisson distribution, common approaches are rejection sampling (Geyer) or inverse transform sampling with inverse CDF (NumPyro). Although either approach is efficient when V is large, as shown later, V in our case takes a small number, leading to a relatively efficient sampling algorithm.

Remark. If $b > 0$, $m \geq V$ almost surely (a.s.) and one can sample m from the truncated Poisson distribution at V efficiently by computing the inverse Poisson cumulative distribution function (Giles, 2016). If $b = 0$, $m \in \{0, \ldots, V-1\}$ is sampled from the categorical distribution with normalized Poisson probability masses. The close-form conditional posterior contributes to the development of an efficient algorithm of our model.

Property 3.4. When V is set to 1, GBPL reduces to the link function proposed in Zhou (2015).

3.2 Poisson Factorized Granger-Causal Graph

Now we introduce our Bayesian construction on binary GC graphs with GBPL. To assist clarity, we discuss our method with only one lag, i.e., $\tau_{\max} = 1$, temporally omitting the notation of lag τ, and introduce the extension to multiple lags later.

The general idea is that we assume a binary GC graph $\boldsymbol{G} \in \{0,1\}^{N\times N}$ is a sample of a probabilistic factorization model with K latent factors: $\boldsymbol{G} \sim p(\Theta\Phi^{\mathrm{T}})$ where $\Theta \in \mathbb{R}_+^{N\times K}$ each entry of which $\theta_{i,k}$ indicates the weight of the k^{th} factor for variable i of being a child in a GC relation and $\Phi \in \mathbb{R}_+^{N\times K}$ each entry of which $\phi_{j,k}$ indicates the weight of the k^{th} factor for variable j of being a parent in a GC relation. In this way, whether j Granger-causes i depends on their interactions with all the K factors: $G_{i,j} \sim p\left(\sum_{k=1}^K \theta_{i,k}\phi_{j,k}\right)$. Conditioned on Θ and Φ, we have: $p(\boldsymbol{G}|\Theta,\Phi) = \prod_{i=1}^N\prod_{j=1}^N p(G_{i,j}|\Theta,\Phi)$, meaning that the links in $\boldsymbol{G}$ can be generated independently.

With the help of GBPL, we propose to impose the following hierarchical Bayesian prior $p(\boldsymbol{G})$:

$$r_k \sim \text{Gamma}\left(1/K, 1/c\right), \quad \theta_{i,k} \sim \text{Gamma}\left(a_i, 1/d_k\right), \quad \phi_{j,k} \sim \text{Gamma}\left(b_j, 1/e_k\right),$$
$$M_{i,j} \sim \text{Poisson}\left(\sum_{k=1}^K r_k\theta_{i,k}\phi_{j,k}\right), \quad G_{i,j} = \mathbf{1}(M_{i,j} \geq V), \tag{3}$$

where noninformative gamma priors $\text{Gamma}\,(1,1)$ are used for a_i, b_j, d_k, e_k, and c.

In the above model, variable r_k is introduced to capture the global popularity of the k^{th} factor (Yang & Leskovec, 2012; 2014; Zhou, 2015). Intuitively, we can understand the vector of $\theta_{i,:}$ and $\phi_{i,:}$ as the embeddings of variable i and j respectively. Whether i is a child of j is determined by the inner product between their embeddings. Moreover, each dimension of the embeddings weights differently and r_k indicates the weight of the k^{th} dimension.

Mathematically, the construction on Θ, Φ, and $\boldsymbol{r}$ can be viewed as the truncated version of a gamma process (Ferguson, 1973; Wolpert et al., 2011; Zhou, 2015) on a product space $\mathbb{R}_+ \times \Omega$: $\mathfrak{G} \sim \Gamma\text{P}(\mathfrak{G}_{a,b}, 1/c)$, where Ω is a complete separable metric space, c is the concentration parameter, $\mathfrak{G}_{a,b}$ is a finite and continuous base measure over Ω. The corresponding Lévy measure is $\nu(drd\theta d\phi) = r^{-1}e^{-cr}dr\mathfrak{G}_{a,b}(d\theta d\phi)$. In our case, a draw from the $\mathfrak{G}_{a,b}$ is a pair of $\theta_{:,k}$ and $\phi_{:,k}$ where $\theta_{:,k} = [\theta_{1,k}, \ldots, \theta_{N,k}]$ $\theta_{:,k} = [\theta_{1,k}, \ldots, \theta_{N,k}]$ and $\phi_{:,k} = [\phi_{1,k}, \ldots, \phi_{N,k}]$. A draw from the gamma process is a discrete distribution with countably infinite atoms from the base measure: $\mathfrak{G} = \sum_{k=1}^\infty r_k\delta_{\theta_{:,k},\phi_{:,k}}$ and r_k is the weight of the k^{th} atom. Although there are infinite atoms, the number of atoms with r_k greater than $\rho \in \mathbb{R}_+$ follows $\text{Poisson}(\int_\rho^\infty r^{-1}e^{-cr}\text{d}r)$ and the expectation of Poisson decreases when ρ increases. In other words, the number of atoms that have relatively large weights will be finite and small, thus, a gamma process based model has an inherent shrinkage mechanism. In our case, if we set the maximum number of latent factors K (i.e., the truncation level) large enough, the model will automatically learn the number of active factors.

The Bayesian nonparametric construction of our model guards against overfitting in a principled way. Intuitively, Bayesian nonparametrics provides a principled way to build models whose complexity adapts

to the data, rather than being fixed ahead of time. Even if K is set to a large value, the model will not use all of the latent factors, the prior acts as a complexity penalty, helping prevent overfitting by preferring simpler models unless there is strong evidence to the contrary. In our model, K is quite different from hyperparameters in parametric models such as the number of layers of a neural network. K is a truncation level that tells the model the number of maximum latent factors it can potentially use while the actual number it will use is determined by the data. In addition, the model will self-regularize if it uses more latent factors than necessary.

Finally, we refer to the model in Equations (2) and (3) as Poisson Factorized Granger-Causal Graph (PFGCG) and denote $\boldsymbol{G} \sim \text{PFGCG}(V)$. In the case of multiple lags, we summarize the model as:

$$\boldsymbol{G}^\tau \sim \text{PFGCG}^\tau(V), \quad \boldsymbol{x}_t \sim \mathcal{MN}\left(\sum_{\tau=1}^{\tau_{\max}} \left(\boldsymbol{A}^\tau \odot \boldsymbol{G}^\tau\right) \boldsymbol{x}_{t-\tau}, \Sigma\right), \tag{4}$$

where we have a separate generative process for the GC graph at each lag τ.

3.3 Inference via Gibbs sampling

Here we introduce how to estimate the posterior over the parameters of the above model using Gibbs sampling (Casella & George, 1992), which adheres to the detailed balance condition (Gilks et al., 1995), a fundamental property of Markov Chain Monte Carlo (MCMC) methods that guarantees the Markov chain converges to the desired posterior distribution as its stationary distribution. To further enhance efficiency, our method employs a hierarchical prior structure with conjugacy properties with the help of several augmentation techniques between Poisson and gamma distributions (Zhou et al., 2012; Zhou, 2015). These conjugate priors ensure that all conditional distributions are analytically well-defined and computationally tractable, simplifying the sampling process. The conjugate structure not only improves computational efficiency but also contributes to the stability of Gibbs sampling, enabling the algorithm to effectively explore the posterior distribution even in challenging settings. Here we highlight the sampling of $\boldsymbol{G}^\tau$ and leave the other details in Section A.

An entry $G_{i,j}^\tau$ in $\boldsymbol{G}^\tau$ is involved in the generative process of data as in Eq. (2) and has a Bernoulli prior according to Eq. (3). Therefore, by denoting $p(G_{i,j}^\tau = 0|-) = s_{i,j}^{\tau,0}$ and $p(G_{i,j}^\tau = 1|-) = s_{i,j}^{\tau,1}$ ($-$ stands for all the other variables), we can derive:

$$s_{i,j}^{\tau,0} = \sum_{v=0}^{V-1} \frac{e^{-q_{i,j}^\tau} (q_{i,j}^\tau)^v}{v!}, \quad \text{and} \quad s_{i,j}^{\tau,1} = e^{-\frac{1}{2}\left((A_{i,j}^\tau)^2 \lambda_i U_j^\tau - 2A_{i,j}^\tau \lambda_i W_{i,j}^\tau\right)} \left(1 - s_{i,j}^{\tau,0}\right), \tag{5}$$

where:

$$\begin{aligned} q_{i,j}^\tau &= \sum_{k=1}^{K} \theta_{i,k}^\tau r_k^\tau \phi_{j,k}^\tau, \quad U_j^\tau = \sum_{t=1}^{T} x_{j,t-\tau}^2, \quad W_{i,j}^\tau = \sum_{t=1}^{T} x_{i,t}^{\neg\tau,\neg j} x_{j,t-\tau}, \\ x_{i,t}^{\neg\tau,\neg j} &= x_{i,t} - \sum_{j' \neq j}^{N} A_{i,j'}^\tau G_{i,j'}^\tau x_{j',t-\tau} - \sum_{\tau' \neq \tau}^{\tau_{\max}} \sum_{j'=1}^{N} A_{i,j'}^{\tau'} G_{i,j'}^{\tau'} x_{j',t-\tau'}. \end{aligned} \tag{6}$$

We can then sample $G_{i,j}^\tau \sim \text{Bernoulli}\left(s_{i,j}^{\tau,1} / \left(s_{i,j}^{\tau,0} + s_{i,j}^{\tau,1}\right)\right)$. With the above conditional posterior, one can sample the entries of $\boldsymbol{G}^\tau$ one by one using Eq. (5). After each sample, we only need to update $W_{i,j}^\tau$ and the other statistics can be updated after all the entries are sampled.

Computational complexity: With the above we have that, in one Gibbs sampling iteration, the complexity of sampling $\boldsymbol{G}^\tau$ is $\mathcal{O}(N^2)$. As can be seen in Algorithm 1 in the appendix, the whole complexity of each Gibbs sampling iteration of our model is $\mathcal{O}(N^2(V+K)\tau_{\max} + T\tau_{\max}^2)$, where N, T, V, K, $\tau_{\max}$ are the number of variables, the number of timestamps, the truncation level, the number of factors, and the maximum number of lags, respectively.

3.4 Properties, design choices and practical consequences

We refer to our method as HiBaNG (for hierarchical Bayesian nonparametric Granger causal discovery), which uses the PFGCG model and the Gibbs sampling algorithm described in Sections 3.2 and 3.3, respectively. Here we discuss some of HiBaNG's properties, its underlying design choices and their practical consequences.

Separate $\boldsymbol{A}^\tau$ and $\boldsymbol{G}^\tau$ structures modeling VAR coefficients: In our approach, $\mathbf{A} \odot \mathbf{G}$ represents coefficients, which are intrinsically sparse as $\mathbf{G}$ is binary. Thus, no sparsity-inducing penalties (Nicholson et al., 2017; Ahelegbey et al., 2016; Ghosh et al., 2018; Billio et al., 2019) or post-hoc heuristics (Nauta et al., 2019; Marcinkevičs & Vogt, 2021) are needed. Moreover, modeling binary $\mathbf{G}$ directly enables us to learn the probability of a causal link between two variables directly and one can quantify uncertainty straightforwardly.

Suitability to low-data settings: Our hierarchical Bayesian model is particularly suited to low-data settings because it allows for information sharing across related parameters through a structured prior. Specifically, our prior on $\boldsymbol{G}^\tau$ takes a factorized form, allowing $\theta_{i,k}$ or $\phi_{j,k}$ to capture the specific information for an individual variable in terms of factor k. At a deeper level in the hierarchy, $\theta_{i,k}$ is influenced by two higher-level components: a_i (capturing variable-specific traits) and d_k (capturing factor-specific traits). This setup allows the model to partially pool information to learn robust estimates for $\theta_{i,k}$ even when direct data is sparse by borrowing strength from related variables (via d_k) and related factors (via a_i). As a result, the model remains expressive while mitigating overfitting and enhancing generalization in low-data regimes.

Need for the GBPL link function: Instead of modeling binary GC graphs directly with Bernoulli distributions, we introduce GBPL as a link function that connects Bernoulli variables to underlying Poisson distributions (Zhou, 2015). This transformation enables the use of hierarchical Poisson-Gamma constructions, akin to those in Poisson matrix factorization models (Canny, 2004; Zhou et al., 2012; Gopalan et al., 2014). By leveraging the Poisson-gamma conjugacy, we gain access to a broader and more flexible set of tools for hierarchical Bayesian modeling and inference, which are difficult to apply directly to Bernoulli likelihoods.

Given the specification of the prior distributions, one can see that $\mathbb{E}\left(\sum_{k=1}^{K} r_k \theta_{i,k} \phi_{j,k}\right) = 1$ as $\mathbb{E}(\theta_{i,k}) = \mathbb{E}(\phi_{j,k}) = 1$ and $\mathbb{E}(r_k) = 1/K$. Therefore, according to Property 3.2 of GBPL, a priori, the expected sparsity of $\boldsymbol{G}$ is $N^2\left(1 - \sum_{v=0}^{V-1} \frac{e^{-1}}{v!}\right)$. Note that $V \in \{1, 2, \dots\}$ is a hyperparameter that incorporates our prior belief of the graph sparsity, i.e., larger value of V means that we encourage the model to learn sparser graphs, while the final sparsity will be determined by the model to fit the observational data. Specifically, we compute the value of $1 - \sum_{v=0}^{V-1} \frac{e^{-1}}{v!}$ with $V = \{1, 2, 3, 4\}$ as 0.2642, 0.0883, 0.0190, 0.0037, respectively. When $V = 3$, it means that less than 2% of the node pairs are expected to be connected in a graph. When $V = 4$, the sparsity is less than 0.5%, which can be overly sparse. Empirically, we observe that when $V = 4$, the sampled GC graphs from the posterior nearly have zero links, thus, we set $V \in \{1, 2, 3\}$ in practice. According to Property 3.4, the link function proposed in Zhou (2015) is a special case of GBPL (when $V = 1$). The model of Zhou (2015) cares less about the sparsity of a graph as it is given in the data. However, the expected sparsity of Zhou (2015) is 0.2642 ($V = 1$), which is too dense for many GC discovery problems. Although the expected sparsity takes finite values, the posterior graphs are not pinned to the prior grid, as data often overwhelms the prior.

Fewer hyperparameters: In low-data scenarios without ground-truths, selecting a model from a large set of hyperparameters is challenging. Our method has, by construction, a small number of hyperparameters as it integrates out the intermediate variables in its hierarchical Bayesian structure. Furthermore, it automatically learns K from data via Bayesian nonparametrics. Indeed, the main hyperparameter of our model is V that only takes a small number of discrete values, while the regularization weights in other VARs are usually continuous parameters in an infinite range.

4 Related Work

Here we focus on the following lines of related works in the literature of machine learning and refer the readers to surveys such as Shojaie & Fox (2022); Assaad et al. (2022); Gong et al. (2023) for a more comprehensive overview.

Bayesian VARs: have been widely used in econometrics and statistics (Breitung & Swanson, 2002; George et al., 2008; Fox et al., 2011; Nakajima & West, 2013; Ahelegbey et al., 2016; Ghosh et al., 2018; Billio et al., 2019; Ghosh et al., 2021). For comprehensive reviews, please see Woźniak (2016); Miranda-Agrippino & Ricco (2019). Different from most existing methods focusing on modeling real-valued coefficients of VARs, ours models the binary GC graphs directly with a novel model construction. As BVARs are across multiple disciplines, comprehensive comparisons to the latest deep VARs in the same settings and datasets have not been carefully studied before.

Deep VARs: have recently become popular in the machine learning community (Montalto et al., 2015; Tank et al., 2018; Wang et al., 2018; Nauta et al., 2019; Khanna & Tan, 2020; Wu et al., 2020; Marcinkevičs & Vogt, 2021; Gong et al., 2022; Fan et al., 2023), where different neural network architectures or learning mechanisms have been explored such as in Tank et al. (2018); Nauta et al. (2019); Khanna & Tan (2020); Marcinkevičs & Vogt (2021); Bussmann et al. (2021); Fan et al. (2023); Zhou et al. (2024). More recently, Bayesian deep VARs have been proposed, such as ACD (Löwe et al., 2022), RHINO (Gong et al., 2022), Dyn-GFN (Tong et al., 2022), and MCD (Varambally et al., 2024). Although these methods can also be considered as Bayesian approaches, many of them have different methodologies and focuses to ours. For example, ACD (Löwe et al., 2022) focuses on discovering causal relations across samples with different underlying causal graphs but shared dynamics. RHINO (Gong et al., 2022) extends VAR by modeling instantaneous causal relations (Runge, 2020; Pamfil et al., 2020) and introducing history-dependent noise, which we do not consider in this paper. Dyn-GFN (Tong et al., 2022) is a Bayesian approach based on GFlowNets (Bengio et al., 2023) focusing on discovering causal graphs varying with time. There are also recently proposed deep VAR variants focusing on causal discovery on MTS data with missing values (Cheng et al., 2023; 2024a). Wu et al. (2024) studies a different problems to ours, where their data is event sequences and each event is with a timestamp of occurrence and in a certain type.

Non-VAR methods: for causal discovery on MTS data is not the focus of our paper. To capture instantaneous causal effects that are not modeled by VARs and GC, there are functional causal models such as in Hyvärinen et al. (2010); Peters et al. (2013); Pamfil et al. (2020) and methods based on dynamic Bayesian networks (DBNs) (Dean & Kanazawa, 1989; Murphy, 2002) or structured VAR models in econometrics (Swanson & Granger, 1997; Demiralp & Hoover, 2003). For DBNs, we refer readers to surveys such as Mihajlovic & Petkovic (2001); Shiguihara et al. (2021). Moreover, there are also constraint-based approaches that extend the PC algorithm (Spirtes et al., 2000) to model time-series data (Runge, 2018; Runge et al., 2019; Runge, 2020; Huang et al., 2020).

Causal discovery for non-time-series data: (e.g., I.I.D.) (Glymour et al., 2019) is another area with a different focus. Here we consider Bayesian methods (Lorch et al., 2021; Cundy et al., 2021; Geffner et al., 2022) that model binary causal graphs as loosely related works to ours among the rich literature. These methods usually leverage gradient-based Bayesian inference algorithms such as variational inference and use reparameterization techniques (Maddison et al., 2016; Jang et al., 2016) to relax the optimization over binary causal graphs to a continuous one, while ours models binary graphs directly. Moreover, they require the discovered graphs to be directed acyclic graphs. We believe that extending their methods to time-series data with multiple lags is nontrivial.

Other related works: Poisson factor analysis is a generic Bayesian framework used in various areas such as graph learning (Zhou, 2015; Zhao et al., 2017), topic modeling (Zhao et al., 2018), and dynamical data modeling (Schein et al., 2019). To the best of our knowledge, it has not been adapted for Granger causal discovery. The closest works to ours include Kalantari et al. (2018); Kalantari & Zhou (2021) which learns a set of latent factors from time-series data as well as a binary graph between them. There are several key differences with our method: **1)** Ours is tailored to Granger causal discovery which learns binary graphs of time-series variables instead of latent factors; **2)** ours considers multiple-lags while they only consider one; and **3)** ours is based on the proposed GBPL while they use the original Bernoulli Poisson Link (Zhou, 2015), which is less applicable to our problem.

Table 1: AUCROC and AUPRC results on semi-synthetic datasets. VAR (FBH) and BVAR(d) failed to learn when $T = 100$ on Lorenz 96 and $T = 200$ on Lotka–Volterra. Best and second best results are highlighted in boldface and underline texts, respectively. $\beta = N/T$ indicates the ratio of the number of variables to the number of observations.

	Lorenz 96 $T = 100$, $\beta = 0.4$		Lorenz 96 $T = 500$, $\beta = 0.08$		Lotka–Volterra $T = 200$, $\beta = 0.2$		Lotka–Volterra $T = 500$, $\beta = 0.08$		FMRI $\beta = 0.075$	
	AUCROC↑	AUPRC↑	AUCROC↑	AUPRC↑	AUCROC↑	AUPRC↑	AUCROC↑	AUPRC↑	AUCROC↑	AUPRC↑
BVAR(c)	0.47±0.02	0.09±0.01	0.73±0.02	0.43±0.02	0.50±0.01	0.50±0.011	0.78±0.04	0.52±0.05	0.66±0.08	0.42±0.12
BVAR(d)	-	-	0.73±0.03	0.43±0.02	-	-	0.67±0.01	0.23.±0.03	0.68±0.06	0.40±0.06
VAR (FBH)	-	-	0.72±0.01	0.39±0.03	-	-	0.68±0.03	0.18±0.01	0.60±0.04	0.32±0.02
PCMCI$^+$	0.62±0.02	0.17±0.01	0.82±0.02	0.59±0.02	0.72±0.03	0.44±0.02	0.78±0.01	0.47±0.01	**0.89**±0.04	**0.67**±0.07
SRU	0.53±0.01	0.12±0.01	0.82±0.03	0.57±0.04	0.55±0.02	0.54±0.01	0.61±0.02	0.32±0.02	0.66±0.02	0.32±0.03
eSRU	0.54±0.03	0.12±0.01	0.84±0.01	0.63±0.05	0.64±0.03	0.58±0.01	0.67±0.03	0.36±0.01	0.72±0.01	0.47±0.01
GVAR	0.57±0.01	0.15±0.03	0.83±0.01	0.63±0.01	0.66±0.02	0.62±0.01	0.81±0.03	0.61±0.04	0.72±0.02	0.57±0.06
JRNGC	0.58±0.07	0.14±0.04	0.79±0.01	0.55±0.03	0.68±0.02	0.32±0.03	**0.87**±0.02	0.61±0.03	0.69±0.03	0.45±0.04
HiBaNG	**0.71**±0.03	**0.35**±0.04	**0.86**±0.02	**0.68**±0.02	**0.73**±0.01	**0.73**±0.01	0.84±0.03	**0.66**±0.03	0.73±0.03	0.60±0.03

5 Experiments

Here we evaluate our method using a set of synthetic and semi-synthetic datasets as well as a real application involving the analysis of climate data. As mentioned in Section 3.4, we refer to our method as HiBaNG. We compare against various classic VAR methods such as VAR (FBH) (Benjamini & Hochberg, 1995), Bayesian approaches including BVAR with diffuse/noninformative priors, BVAR(d) (Litterman, 1986; Miranda-Agrippino & Ricco, 2019) and BVAR with conjugate priors, BVAR(c), which equivalent to an ablation of our model without $\{\boldsymbol{G}^\tau\}_\tau^{\tau_{\max}}$. We also benchmark against deep/neural approaches: SRU (Oliva et al., 2017), eSRU (Khanna & Tan, 2020), GVAR (Marcinkevičs & Vogt, 2021), JRNGC (Zhou et al., 2024), and a recent non-VAR approach, PCMCI$^+$ (Runge, 2020). As evaluation metrics we use the areas under receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves and report the calibration error (CE) to evaluate predictive uncertainties. Section B shows the full experimental setting.

We report the results on toy synthetic datasets in Section C. For more comprehensive quantitative comparisons, we conduct our experiments on three widely-used benchmark datasets, Lorenz 96 (Lorenz, 1996), Lotka-Volterra (Bacaër & Bacaër, 2011) and FMRI (Smith et al., 2011). Full details of these datasets are in Section D. For each dataset we use 5 replicates for all the methods.

We present the results for AUCROC and AUPRC in Table 1. Overall, our proposed method, HiBaNG, demonstrates superior performance across most datasets and metrics. A notable observation is the enhanced performance of HiBaNG as the parameter β increases. This trend underscores the robustness of our model in severe low-data situations, where traditional methods often struggle. For the FMRI dataset, where the severity of low-data conditions is reduced, HiBaNG ranks second in both AUCROC and AUPRC metrics. This result aligns with expectations, as richer datasets reduce the relative advantage of our Bayesian approach. Comparing HiBaNG with BVAR(c), which essentially represents HiBaNG without the integration of binary GC graphs, reveals that HiBaNG consistently outperforms BVAR(c). This comparison highlights the significance of incorporating binary GC graphs into the Bayesian framework, facilitating a clearer and more interpretable understanding of causal links while enhancing predictive accuracy.

5.1 Synthetic and semi-synthetic datasets

The results for SHD and CE are displayed in Table 2. For SHD, we include methods capable of converting their coefficients into sparse graphs, while for CE, we concentrate on the top-performing methods based on AUCROC and AUPRC scores. It is important to note that SHD is influenced by the method's approach to generating binary graphs, which may introduce bias depending on the sparsity of the ground-truth graphs. Our method's capability to directly sample binary graphs without relying on arbitrary thresholds provides a distinct advantage. Regarding CE, our method achieves the lowest error rates across almost all datasets.

Table 2: SHD and CE on semi-synthetic datasets.

	Lorenz 96		Lotka–Volterr		FMRI
	$T = 100$	$T = 500$	$T = 200$	$T = 500$	
			SHD↓		
VAR (FBH)	-	98.40±2.4	-	74.20±10.4	28.8±1.3
PCMCI$^+$	405.0±15.0	411.0±5.0	494.0±20.0	423.0±20.0	70.00±2.00
GVAR	389.6±220.6	127.4±76.8	279.0±104.1	82.8±24.1	71.6±21.8
HiBaNG	**117.7**±3.3	**71.8**±4.0	**67.0**±2.4	**45.0**±4.4	**24.2**±0.7
			CE↓		
BVAR (d)	-	0.10±0.01	-	0.11±0.01	0.11±0.02
PCMCI$^+$	0.25±0.01	0.27±0.01	0.27±0.01	0.28±0.01	0.31±0.02
GVAR	0.07±0.01	0.15±0.01	0.08±0.01	0.10±0.01	0.19±0.01
JRNGC	**0.01**±0.01	0.09±0.01	0.07±0.01	0.13±0.01	0.09±0.02
HiBaNG	0.11±0.01	**0.08**±0.01	**0.05**±0.01	**0.04**±0.01	**0.07**±0.01

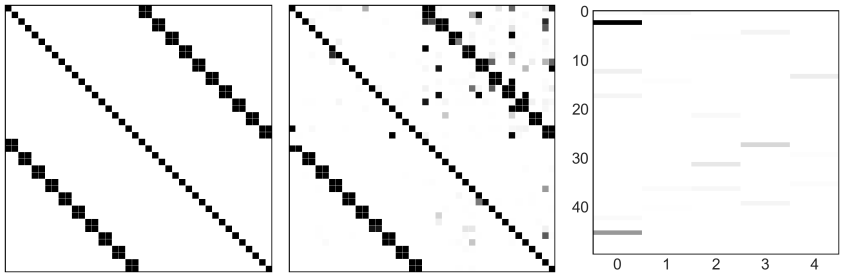


Figure 2: Qualitative analysis of HiBaNG on Lotka–Volterra. Left: Ground-truth GC graph, middle: Bernoulli posterior mean of the discovered GC graphs, right: Matrix of $\{r_k^\tau\}_{k=1,\tau}^{K,\tau_{\max}}$ showing that $K = 50$ is sufficiently large in our experiments (see Section B for more details). Each rectangle indicates a value of a matrix and brighter colors indicates larger values.

This indicates that HiBaNG's predictive confidence is highly aligned with its accuracy, a result of its inherent uncertainty-aware design. Such alignment is crucial in real-world applications, where understanding the reliability of causal predictions can inform better decision-making.

In Figure 2, we compare the Bernoulli posterior mean of HiBaNG with the ground-truth graph on Lotka–Volterra. We can see that the posterior mean discovered by our method is well aligned with the ground-truth graph, where brighter rectangles indicate higher probability of a GC link between two variables. Finally, recall that r_k^τ in Eq. (3) models the weight of latent factor k at lag τ. We plot $\{r_k^\tau\}_{k=1,\tau}^{K,\tau_{\max}}$ as a $K \times \tau_{\max}$ matrix. It can be seen that the matrix is quite sparse, where only a few entries have large values, i.e., only a few factors are active among $K = 50$. This demonstrates the shrinkage mechanism of HiBaNG on K.

In summary, our method's ability to maintain high performance, combined with enhanced interpretability and uncertainty management, makes HiBaNG a compelling choice for causal discovery on MTS data in low-data regimes.

5.2 Qualitative analysis on climate reanalysis data

We qualitatively analyze our method's performance on climate data obtained from the Japanese reanalysis of the atmosphere (JRA55) (Kobayashi et al., 2015), detailed in Section E. We compare HiBaNG and PCMCI$^+$ with lag $\tau_{\max} = 6$ in Figure 3. We qualitatively explain physical interpretability of discovered relationships of PCMCI and ours. 1) ENSO (MEI) Autocorrelation: HiBaNG captures autocorrelations in MEI extending back to t–5, aligning with the known persistence of ENSO over approximately 6 months, as discussed in Harries &

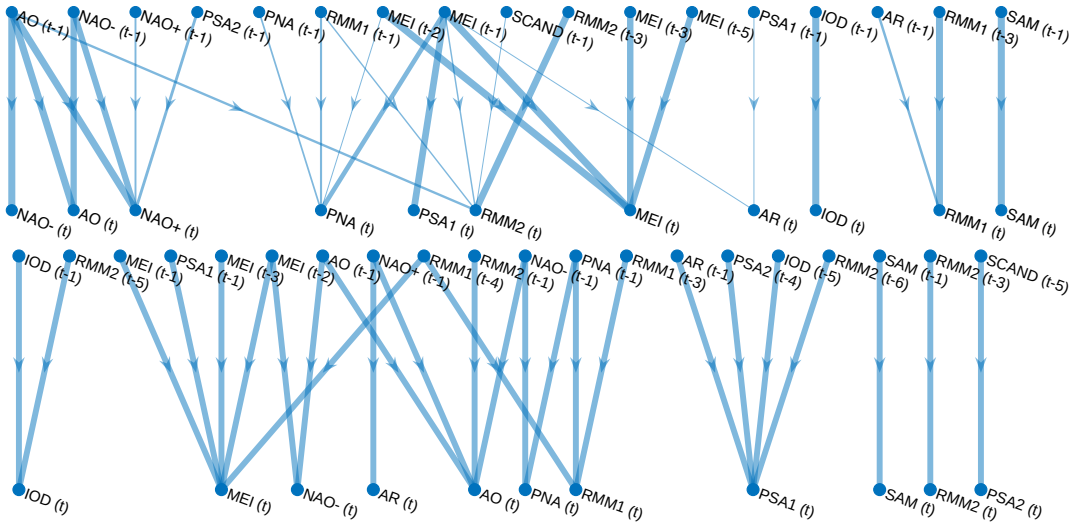


Figure 3: Results on JR55. Up: HiBaNG discovered causal links between indices on JR55 where the weights are from the Bernoulli posterior of the graph (links with weights less than 0.2 are not shown) and thicker links indicate stronger connections. Down: PCMCI$^+$ discovered causal links.

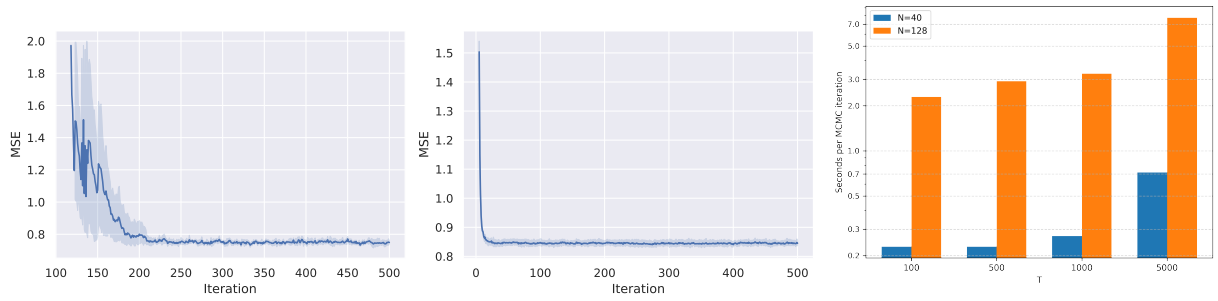
O'Kane (2021) (Figure 5). PCMCI only captures MEI autocorrelations up to t–3, suggesting a more limited temporal sensitivity. 2) MEI→PSA1 Directionality: HiBaNG correctly infers MEI(t–1)→PSA1(t), consistent with the established ENSO-to-PSA1 influence via modulation of midlatitude tropospheric flow (O'Kane & Franzke, 2025). PCMCI, in contrast, infers the reverse direction (PSA1(t–1)→MEI(t)), which contradicts known dynamics. 3) PNA Links: Both methods identify the PNA autocorrelation at one-month lag, consistent with observed behavior. However, HiBaNG additionally identifies the MEI(t–1)→PNA(t) edge, also seen in Harries & O'Kane (2021) (Figure 9b), supporting known teleconnections. PCMCI instead finds NAO–(t-1)→PNA(t), which is plausible but not as directly supported by prior work. 4) Northern Hemisphere Modes: HiBaNG provides a more complete representation of the interconnected dynamics among AO, NAO+, NAO–, and AR. For example, the HiBaNG graph includes: PSA2(t–1) → NAO+(t); PSA1(t)→AR(t); NAO+ and NAO– autocorrelations at t–1. These are consistent with Harries & O'Kane (2021) (Figures 6 and 7), and not captured by PCMCI. Overall, the HiBaNG graph shows a larger and more diverse set of edges consistent with those previously inferred from the JRA55 reanalysis via Bayesian structure learning as reported in Harries & O'Kane (2021).

5.3 Empirical computational performance

In addition to the complexity analysis given in Section 3.3, we empirically study the computational complexity by examining two aspects: running time of one Gibbs sampling iteration and the number of iterations for convergence. For all the experiments here, our method ran on an Apple laptop with an M1 Pro processor. We report the running time (seconds per Gibbs sampling iteration) of our method varying the number of variables and timestamps on Lorenz 96 in Figure 4c. It can be seen that inference in our method is efficient, scaling gracefully as a function of T. Figure 4 shows HiBaNG's MSE over the number of iterations on Lorenz 96. Although we set the maximum number of Gibbs sampling iterations to 10,000, in most cases our method converges around 200 iterations. We can also see that in low-data regimes (e.g., $T = 100$ on Lorenz 96), samples from our model may have more variance at the beginning of the inference, which is as expected as there are less timestamps. In general, our method converges within half an hour on a laptop in most cases.

6 Conclusion

We have presented a novel Bayesian VAR model tailored to Granger causal discovery on MTS data in low-data settings. Our method leverages a hierarchical Bayesian framework that separates Granger causal



(a) L96, $T = 100$ (b) L96, $T = 500$ (c) Running times

Figure 4: (a, b) HiBaNG's ($V = 2$) empirical converge on Lorenz 96. For better visualisation, we show MSE in the iterations between $[a, b]$ where a is the first iteration that MSE goes below 2.0 and $b = 500$. (c) Running times in seconds per MCMC iteration on on Lorenz 96.

relationships into binary causal graphs and real-valued weights. Through extensive experiments on synthetic, semi-synthetic, and real-world datasets, we have demonstrated that our approach can perform better in low-data regimes. For limitations, it is important to note that our method is based on the Granger causality framework, which assumes that causal relationships are reflected in time-lagged dependencies. While this assumption is appropriate for many applications, it may not hold universally, and practitioners should exercise caution when interpreting causal results.

Broader impact statement

This paper introduces a novel method for discovering causal relationships from observational data, grounded in the framework of Granger causality and a factorized representation of causal structure. By leveraging these assumptions, the method enables scalable and interpretable causal discovery, which can benefit applications in fields such as economics, neuroscience, and climate science where temporal data is abundant. The approach opens up new possibilities for causal inference in high-dimensional settings, providing a foundation for future work that can relax or adapt these assumptions to broader domains.

As the method relies on assumptions like Granger causality and factorized causal structures, there is a risk of drawing incorrect conclusions if these assumptions do not hold. Misinterpretation or misuse of inferred causal relationships could lead to flawed decisions or reinforce biases present in the data, especially in high-stakes domains.

References

Daniel Felix Ahelegbey, Monica Billio, and Roberto Casarin. Sparse graphical vector autoregression: A bayesian approach. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, (123/124): 333–361, 2016.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36, 2023.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Nicolas Bacaër and Nicolas Bacaër. Lotka, volterra and the predator–prey system (1920–1926). *A short history of mathematical population dynamics*, pp. 71–76, 2011.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Monica Billio, Roberto Casarin, and Luca Rossini. Bayesian nonparametric sparse var models. *Journal of Econometrics*, 212(1):97–115, 2019.

Jörg Breitung and Norman R Swanson. Temporal aggregation and spurious instantaneous causality in multiple time series models. *Journal of Time Series Analysis*, 23(6):651–665, 2002.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *International Conference on Discovery Science*, pp. 446–460. Springer, 2021.

John Canny. Gap: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 122–129, 2004.

George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024a.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024b.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150, 1989.

Tristan Deleu, Mizu Nishikawa-Toomey, Jithendaraa Subramanian, Nikolay Malkin, Laurent Charlin, and Yoshua Bengio. Joint bayesian inference of graphical structure and parameters with a single generative flow network. *Advances in Neural Information Processing Systems*, 36, 2023.

Selva Demiralp and Kevin D Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65:745–767, 2003.

Chenchen Fan, Yixin Wang, Yahong Zhang, and Wenli Ouyang. Interpretable multi-scale neural network for granger causality discovery. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

Emily Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on signal processing*, 59(4):1569–1585, 2011.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

Edward I George, Dongchu Sun, and Shawn Ni. Bayesian stochastic search for var model restrictions. *Journal of Econometrics*, 142(1):553–580, 2008.

Charles J. Geyer. The K-truncated Poisson distribution. URL `https://www.stat.umn.edu/geyer/aster/library/aster/doc/ktp.pdf`.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association*, 2018.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. Strong selection consistency of bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*, 49(3):1267–1299, 2021.

Michael B Giles. Algorithm 955: approximation of the inverse poisson cumulative distribution function. *ACM transactions on mathematical software (TOMS)*, 42(1):1–22, 2016.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery metuangods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, Jingping Bi, Lun Du, and Jin Wang. Causal discovery from temporal data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5803–5804, 2023.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. *Advances in neural information processing systems*, 27, 2014.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dylan Harries and Terence J O'Kane. Dynamic bayesian networks for evaluation of granger causal relationships in climate reanalyses. *Journal of Advances in Modeling Earth Systems*, 13(5):e2020MS002442, 2021.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21 (89):1–53, 2020.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Rahi Kalantari and Mingyuan Zhou. Graph gamma process linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 4060–4068. PMLR, 2021.

Rahi Kalantari, Joydeep Ghosh, and Mingyuan Zhou. Nonparametric bayesian sparse graph linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 1952–1960. PMLR, 2018.

Alireza Karimi and Mark R Paul. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4), 2010.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2020.

Shinya Kobayashi, Yukinari Ota, Yayoi Harada, Ayataka Ebita, Masami Moriya, Hirokatsu Onoda, Kazutoshi Onogi, Hirotaka Kamahori, Chiaki Kobayashi, Hirokazu Endo, et al. The jra-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93(1):5–48, 2015.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Robert B Litterman. Forecasting with bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38, 1986.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1. Reading, 1996.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. In *International Conference on Learning Representations*, 2021.

V Mihajlovic and Milan Petkovic. Dynamic bayesian networks: A state of the art. *University of Twente Document Repository*, 2001.

Silvia Miranda-Agrippino and Giovanni Ricco. Bayesian vector autoregressions: Estimation. In *Oxford Research Encyclopedia of Economics and Finance*. 2019.

Alessandro Montalto, Sebastiano Stramaglia, Luca Faes, Giovanni Tessitore, Roberto Prevete, and Daniele Marinazzo. Neural networks with non-uniform embedding and explicit validation phase to assess granger causality. *Neural networks*, 71:159–171, 2015.

Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

Jouchi Nakajima and Mike West. Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164, 2013.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

William B Nicholson, David S Matteson, and Jacob Bien. Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, 33(3):627–651, 2017.

NumPyro. Truncated and folded distributions. URL `https://num.pyro.ai/en/stable/tutorials/truncated_distributions.html`.

Junier B Oliva, Barnabás Póczos, and Jeff Schneider. The statistical recurrent unit. In *International Conference on Machine Learning*, pp. 2671–2680. PMLR, 2017.

Terence J. O'Kane and Christian L.E. Franzke. Chapter 8 - pacific–south american pattern. In Bin Guan (ed.), *Atmospheric Oscillations*, pp. 171–181. Elsevier, 2025. ISBN 978-0-443-15638-0. doi: https://doi.org/10.1016/B978-0-443-15638-0.00008-3. URL `https://www.sciencedirect.com/science/article/pii/B9780443156380000083`.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26, 2013.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Aaron Schein, Scott Linderman, Mingyuan Zhou, David Blei, and Hanna Wallach. Poisson-randomized gamma dynamical systems. *Advances in Neural Information Processing Systems*, 32, 2019.

S Seabold and PJ Statsmodels. Econometric and statistical modeling with python. In *Proceedings of the 9th Python in science conference*, pp. 57–61, 2010.

Pedro Shiguihara, Alneu De Andrade Lopes, and David Mauricio. Dynamic bayesian network modeling, learning, and inference: a survey. *IEEE Access*, 9:117639–117648, 2021.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92 (437):357–367, 1997.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality for nonlinear time series. *stat*, 1050:16, 2018.

Alexander Tong, Lazar Atanackovic, Jason Hartford, and Yoshua Bengio. Bayesian dynamic causal discovery. In *A causal view on dynamical systems, NeurIPS 2022 workshop*, 2022.

Sumanth Varambally, Yian Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. In *International Conference on Machine Learning*, 2024.

Yueming Wang, Kang Lin, Yu Qi, Qi Lian, Shaozhe Feng, Zhaohui Wu, and Gang Pan. Estimating brain connectivity with varying-length time lags using a recurrent neural network. *IEEE Transactions on Biomedical Engineering*, 65(9):1953–1963, 2018.

Robert L Wolpert, Merlise A Clyde, Chong Tu, et al. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. *The Annals of Statistics*, 39(4):1916–1962, 2011.

Tomasz Woźniak. Bayesian vector autoregressions. *Australian Economic Review*, 49(3):365–380, 2016.

Dongxia Wu, Tsuyoshi Idé, Georgios Kollias, Jiri Navratil, Aurelie Lozano, Naoki Abe, Yian Ma, and Rose Yu. Learning granger causality from instance-wise self-attentive hawkes processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 415–423. PMLR, 2024.

Tailin Wu, Thomas Breuel, Michael Skuhersky, and Jan Kautz. Discovering nonlinear relations with minimum predictive information regularization. *arXiv preprint arXiv:2001.01885*, 2020.

Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *2012 IEEE 12th international conference on data mining*, pp. 1170–1175. IEEE, 2012.

Jaewon Yang and Jure Leskovec. Structure and overlaps of ground-truth communities in networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2):1–35, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

He Zhao, Lan Du, and Wray Buntine. Leveraging node attributes for incomplete relational data. In *International conference on machine learning*, pp. 4072–4081. PMLR, 2017.

He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Inter and intra topic structure learning with word embeddings. In *International Conference on Machine Learning*, pp. 5892–5901. PMLR, 2018.

MingYuan Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, pp. 1135–1143, 2015.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2013.

Mingyuan Zhou, Lauren Hannah, David Dunson, and Lawrence Carin. Beta-negative binomial process and poisson factor analysis. In *Artificial Intelligence and Statistics*, pp. 1462–1471. PMLR, 2012.

Wanqi Zhou, Shuanghao Bai, Shujian Yu, Qibin Zhao, and Badong Chen. Jacobian regularizer-based neural granger causality. In *International Conference on Machine Learning*, 2024.

A Appendix

A Inference via Gibbs Sampling

Sampling $\boldsymbol{A}^\tau$ With the conjugacy of normal distributions, one can sample the entries of $\boldsymbol{A}^\tau$ one by one by:

$$A_{i,j}^\tau \sim \begin{cases} \mathcal{N}\left(0, (\psi_{i,j}^\tau)^{-1}\right), & \text{if } G_{i,j}^\tau = 0 \\ \mathcal{N}\left(\mu_{i,j}^\tau, \sigma_{i,j}^\tau\right), & \text{otherwise} \end{cases} \tag{7}$$

where:

$$\sigma_{i,j}^{\tau} = \left(\lambda_i G_{i,j}^{\tau} \sum_{t=1}^{T} x_{j,t-\tau}^2 + \psi_{i,j}^{\tau}\right)^{-1}, \tag{8}$$

$$\mu_{i,j}^{\tau} = \sigma_{i,j}^{\tau} G_{i,j}^{\tau} \lambda_i \left(\sum_{t=1}^{T} x_{i,t}^{\neg\tau,\neg j} x_{j,t-\tau}\right), \tag{9}$$

$$x_{i,t}^{\neg\tau,\neg j} = x_{i,t} - \sum_{j' \neq j}^{N} A_{i,j'}^{\tau} G_{i,j'}^{\tau} x_{j',t-\tau} - \sum_{\tau' \neq \tau}^{\tau_{\max}} \sum_{j'=1}^{N} A_{i,j'}^{\tau'} G_{i,j'}^{\tau'} x_{j',t-\tau'} \tag{10}$$

Sampling $\psi_{i,j}^{\tau}$ and λ_i With the conjugacy between normal and gamma distributions, one can sample $\psi_{i,j}^{\tau}$ and λ_i from their conditional gamma posteriors:

$$\psi_{i,j}^{\tau} \sim \text{Gamma}\left(1.5, 1/(A_{i,j}^{\tau}/2 + 1)\right), \tag{11}$$

$$\lambda_i \sim \text{Gamma}\left(1 + T/2, \left(1 + \sum_{t=1}^{T} (x_{i,t} - \sum_{\tau=1}^{\tau_{\max}} \sum_{j=1}^{N} A_{i,j}^{\tau} G_{i,j}^{\tau} x_{j,t-\tau})\right)^{-1}\right). \tag{12}$$

Sampling $\boldsymbol{M}^{\tau}$ With GBPL, we can sample:

$$M_{i,j}^{\tau} \sim \begin{cases} \text{Categorical}_V\left(\left[\ldots, \frac{\frac{e^{-q_{i,j}^{\tau}} (q_{i,j}^{\tau})^v}{v!}}{\sum_{v'=0}^{V} \frac{e^{-q_{i,j}^{\tau}} (q_{i,j}^{\tau})^{v'}}{v'!}}, \ldots\right]\right), & \text{if } G_{i,j}^{\tau} = 0 \\ \text{TPoisson}_V\left(q_{i,j}^{\tau}\right), & \text{otherwise} \end{cases} \tag{13}$$

where $q_{i,j}^{\tau} = \sum_{k=1}^{K} \theta_{i,k}^{\tau} r_k^{\tau} \phi_{j,k}^{\tau}$.

Sampling $M_{i,j,k}^{\tau}$ With the relationships between Poisson and multinomial distributions, we can sample:

$$[\ldots, m_{i,j,k}^{\tau}, \ldots] \sim \text{Multinomial}_K\left(m_{i,j}^{\tau}; \left[\ldots, \frac{q_{i,j,k}^{\tau}}{\sum_{k'=1}^{K} q_{i,j,k'}^{\tau}}, \ldots\right]\right), \tag{14}$$

where $q_{i,j,k}^{\tau} = \theta_{i,k}^{\tau} r_k^{\tau} \phi_{j,k}^{\tau}$.

Sampling $\theta_{i,k}^{\tau}$, $\phi_{j,k}^{\tau}$, and r_k^{τ}

$$\theta_{i,k}^{\tau} \sim \text{Gamma}\left(a_i^{\tau} + \sum_{j=1}^{N} M_{i,j,k}^{\tau}, \frac{1}{d_k^{\tau} + r_k^{\tau} \sum_{j=1}^{N} \phi_{j,k}^{\tau}}\right), \tag{15}$$

$$\phi_{j,k}^{\tau} \sim \text{Gamma}\left(b_j^{\tau} + \sum_{i=1}^{N} M_{i,j,k}^{\tau}, \frac{1}{e_k^{\tau} + r_k^{\tau} \sum_{i=1}^{N} \theta_{i,k}^{\tau}}\right), \tag{16}$$

$$r_k^{\tau} \sim \text{Gamma}\left(1/K + \sum_{i=1,j=1}^{N} M_{i,j,k}^{\tau}, \frac{1}{c^{\tau} + \sum_{i=1,j=1}^{N} \theta_{i,k}^{\tau} \phi_{j,k}^{\tau}}\right). \tag{17}$$

Table 3: Hyperparameter settings.

Model	$\tau_{\max}$	# hidden layers	# hidden units	# training epochs	Learning rate	Mini-batch size	Parameter space
VAR (FBH)	{1,3,5}	NA	NA	NA	NA	NA	NA
BVAR(d)	{1,3,5}	NA	NA	NA	NA	NA	NA
BVAR(c)	{1,3,5}	NA	NA	NA	NA	NA	NA
SRU	NA	1	10	2000	1.0e-3	50	$\mu_1 = [0.01, 0.05]$ $\mu_2 = [0.01, 0.05]$ $\mu_3 = [0.01, 1.0]$
eSRU	NA	2	10	2000	1.0e-3	50	$\mu_1 = [0.01, 0.05]$ $\mu_2 = [0.01, 0.05]$ $\mu_3 = [0.01, 1.0]$
GVAR	{1,3,5}	2	50	1,000	1.0e-4	64	$\lambda = [0.0, 3.0]$ $\gamma = [0.0, 0.1]$
HiBaNG	{1,3,5}	NA	NA	10,000	NA	NA	$V = \{1, 3, 5\}$

Sampling a_i^τ and b_j^τ By introducing auxiliary variables from the Chinese Restaurant Table (CRT) distribution (Zhou et al., 2012; Zhou & Carin, 2013), we can sample:

$$l_{i,k}^\tau \sim \text{CRT}\left(\sum_{j=1}^{N} M_{i,j,k}^\tau, a_i^\tau\right), \tag{18}$$

$$a_i^\tau \sim \text{Gamma}\left(1 + \sum_{k=1}^{K} l_{i,k}^\tau, \frac{1}{1 + \sum_{k=1}^{K} \log(1 + r_k^\tau \sum_{j=1}^{N} \phi_{j,k}^\tau / d_k^\tau)}\right), \tag{19}$$

$$o_{j,k}^\tau \sim \text{CRT}\left(\sum_{i=1}^{N} M_{i,j,k}^\tau, b_j^\tau\right), \tag{20}$$

$$b_j^\tau \sim \text{Gamma}\left(1 + \sum_{k=1}^{K} o_{j,k}^\tau, \frac{1}{1 + \sum_{k=1}^{K} \log(1 + r_k^\tau \sum_{i=1}^{N} \theta_{i,k}^\tau / e_k^\tau)}\right). \tag{21}$$

Sampling d_k^τ, e_k^τ and c^τ

$$d_k^\tau \sim \text{Gamma}\left(\sum_{i=1}^{N} a_i^\tau + 1, \frac{1}{\sum_{i=1}^{N} \theta_{i,k}^\tau + 1}\right), \tag{22}$$

$$e_k^\tau \sim \text{Gamma}\left(\sum_{j=1}^{N} b_j^\tau + 1, \frac{1}{\sum_{i=1}^{N} \phi_{j,k}^\tau + 1}\right), \tag{23}$$

$$c^\tau \sim \text{Gamma}\left(2, \frac{1}{\sum_{k=1}^{K} r_k^\tau + 1}\right). \tag{24}$$

B Details of Experimental Settings

HiBaNG We use 10,000 as the maximum Gibbs sampling iterations where the first 5,000 are burn-in iterations and we then collect the samples from the conditional posteriors of the graphs every 10 iterations[2], which are stored in $\boldsymbol{Y} \in \mathbb{R}_+^{N \times N \times \tau_{\max} \times H}$ ($H = 500$ is the number of collections). The Bernoulli conditional posterior probability of a link between i and j at lag τ in collection $h \in \{1, \ldots, H\}$ is computed by Eq. (5) as: $Y[i, j, \tau, h] = \frac{s_{i,j}^{\tau,1}}{s_{i,j}^{\tau,0} + s_{i,j}^{\tau,1}}$. Given the collections, to compare with other methods, we compute the averaged probability of the discovered GC graph by $\text{mean}(\max(\boldsymbol{Y}, \text{dim} = `\tau'), \text{dim} = `h')$ (Marcinkevičs & Vogt, 2021). As HiBaNG has an intrinsic shrinkage mechanism on K, we set $K = 50$ that is empirically large enough for our experiments. The only hyperparameter that we need to tune is V, which we vary in $\{1, 2, 3\}$.

[2]As shown in Section 5.3, our method converges in much less iterations.

Algorithm 1: Inference Algorithm for HiBaNG.

input : MTS data $\boldsymbol{X}$, number of lags $\tau_{\max}$, hyperparameter V
output : Posterior samples of $\{\boldsymbol{A}^\tau\}_\tau^{\tau_{\max}}$ and $\{\boldsymbol{G}^\tau\}_\tau^{\tau_{\max}}$
Initialize all the variables;
while *Not converged* **do**
  for $i = 1 \dots N$ **do**
    Sample λ_i;
  end
  for $\tau = 1 \dots \tau_{max}$ **do**
    for $i = 1 \dots N$, $j = 1 \dots N$ **do**
      Sample $M_{i,j}^\tau$ and $M_{i,j,k}^\tau$;
    end
    Sample c^τ;
    for $i = 1 \dots N$ **do**
      Sample a_i^τ and b_i^τ;
    end
    for $k = 1 \dots K$ **do**
      Sample d_k^τ, e_k^τ, r_k^τ;
    end
    for $i = 1 \dots N$ **do**
      for $k = 1 \dots K$ **do**
        Sample $\theta_{i,k}^\tau$ and $\phi_{i,k}^\tau$;
      end
    end
    for $i = 1 \dots N$ **do**
      for $j = 1 \dots N$ **do**
        Sample $A_{i,j}^\tau$, $\psi_{i,j}^\tau$, $G_{i,j}^\tau$;
      end
    end
  end
end

Baselines As ours is a VAR approach for GC, we mainly include baselines that are also based on the VAR framework in our comparison. **1)** We compare with the widely-used VAR with F-tests for Granger causality and the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) for controlling the false discovery rate (FDR) (at $q = 0.05$) denoted as VAR (FBH) and implemented in the statsmodels library (Seabold & Statsmodels, 2010). **2)** For Bayesian methods, we compare with two classic approaches but with different prior distributions: BVAR with diffuse/noninformative priors on the coefficients, named BVAR(d), i.e., $(\{\boldsymbol{A}^\tau\}_\tau^{\tau_{\max}}, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}}$ (Litterman, 1986; Miranda-Agrippino & Ricco, 2019), whose posterior has an analytical form. BVAR with conjugate priors on the coefficients, named BVAR(c): $\psi_{i,j}^\tau \sim \text{Gamma}\,(1,1)\,, A_{i,j}^\tau \sim \mathcal{N}\left(0, (\psi_{i,j}^\tau)^{-1}\right)$ and for Σ: $\lambda_i \sim \text{Gamma}\,(1,1)\,, \Sigma = \text{diag}\,(\lambda_1, \dots, \lambda_N)^{-1}$. This is equivalent to an ablation of our model without $\{\boldsymbol{G}^\tau\}_\tau^{\tau_{\max}}$, for which we use the Gibbs sampling with the same settings as ours. For deep/neural VARs, we compare with a method with component-wise statistical recurrent units (SRU) (Oliva et al., 2017) and its improved version (economy SRU, eSRU) (Khanna & Tan, 2020) with sample-efficient architectures. The important hyperparameters of SRU and eSRU are the strengths (μ_1, μ_2, μ_3) of three regularization terms. We also compare with GVAR (Marcinkevičs & Vogt, 2021) that uses self-explaining neural networks (Alvarez Melis & Jaakkola, 2018) and converts the weights in the neural networks into binary GC graphs with a heuristic stability-based procedure. As one of state-of-the-art methods, Jacobian Regularizer-based Neural Granger Causality (JRNGC) (Zhou et al., 2024) is also included in the comparison. For non-VAR methods, the most recent one PCMCI$^+$ (Runge, 2020) is also compared. For the baselines, we either use their original settings or follow these in Marcinkevičs & Vogt (2021), shown in Table 3. For all the compared methods, we set $\tau_{\max} = 5$ unless otherwise specified.

Evaluation Metrics Following Khanna & Tan (2020); Marcinkevičs & Vogt (2021) that aggregate graphs at multiple lags into one, we use four metrics to compare the discovered GC graph of a method on a dataset with the ground-truth graph. For all the baseline methods, we compute the score of a discovered GC graph from their learned VAR coefficients. For our method, the score of a GC graph is the mean of the Bernoulli posterior. We report the areas under receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves by comparing the score of a discovered GC graph to the ground-truth graph. Moreover, as mentioned before, VAR (FBH) and GVAR use specific post-hoc processes to convert coefficients to binary GC graphs, thus, we also report the structural Hamming distance (SHD) between the discovered binary GC graph and the ground-truth one. Note that unlike AUCROC and AUPRC, SHD is biased to the sparsity of the ground-truth graph, e.g., for a sparse ground-truth graph, a method always predicting no links achieves low SHD. To measure the predictive uncertainty of the Granger-causal graphs discovered by different approaches, we report the calibration error (CE) (Guo et al., 2017), which has been a widely used metric for model uncertainly and confidence (Liu et al., 2020). CE examines the difference between the model's probability and the true probability given the model's output, whose definition is shown in Definition 2.1 of Kumar et al. (2019). We consider the causal discovery task with N variables as a binary classification problem with N^2 samples, i.e., predicting a Granger-causal link between a pair of variables and then compute CE accordingly. For AUCROC and AUPRC, higher values indicate better performance and for SHD and CE, lower values are better.

Model Selection and Parameter Tuning To conduct model selection for each method, we split the input MTS data into a training set (first 80% timestamps) and a test set (remaining 20%) (Gong et al., 2022). We train a method on the training set and use the learned model to conduct one-step prediction on the test set. We then use the mean square error (MSE) on the test set and select the parameters of a method that give the best MSE. Our model selection is different from that of GVAR (Marcinkevičs & Vogt, 2021), where the best model is selected by comparing with the ground-truth graphs and report the best achievable performance. Importantly, we note that our task is discovering GC graphs from data without training with ground-truth, which is an unsupervised problem (e.g., akin to unsupervised clustering). As no ground-truth is given for training, we use forecasting performance (e.g., MSE) as a proxy objective for selecting model parameters for VAR-based methods, including ours. This is consistent with how Granger causality is commonly defined—based on predictive influence. We believe this constitutes a fair comparison across all VAR-based methods, as the same procedure is used consistently. PCMCI$^+$ relies on conditional independence testing to infer a causal graph. Its key hyperparameter is the p-value threshold. Unlike VAR methods, PCMCI$^+$ does not perform forecasting, and thus MSE cannot be used as a model selection criterion. Therefore, we used the standard default setting of 0.05.

Setting V As stated in Section 3.4, the expected sparsity of $\mathbf{G}$ in the proposed prior distribution is $N^2\left(1-\sum_{v=0}^{V-1}\frac{e^{-1}}{v!}\right)$, where V controls the sparsity of the graphs. We compute the value of $1-\sum_{v=0}^{V-1}\frac{e^{-1}}{v!}$ with $V=1,2,3,4$ as 0.2642, 0.0883, 0.0190, 0.0037, respectively. Thus, setting $V=3$ already induces significant sparsity, which matches the empirical needs of our experiments. We see that when $V=4$, the sampled GC graphs from the posterior nearly have zero links, thus, we set $V\in\{1,2,3\}$ in practice.

Number of latent factors K We believe that $K=50$ is sufficiently large in our experiment settings. This can be seen on the RHS of Figure 2, which shows r_k^τ for all the lags and all K factors as a $K\times\tau_{\max}$ a matrix ($K=50$ and $\tau_{\max}=5$). Recall that r_k^τ indicates the weight of latent factor k at lag τ. If we look at one column of the figure that shows the weights of the 50 factors for one lag, we can see that only a few entries have large values, meaning that only a few factors are active among $K=50$. We had similar observations in other datasets as well.

C Synthetic Data

We conduct experiments on the synthetic data generated from a VAR model specified in Eq. (1) to test whether our method is able to discover the ground-truth graphs. Given $N=16$ and $\tau_{\max}=6$, we construct $\{\boldsymbol{A}^\tau\}_{\tau=1}^{\tau_{\max}}$ by first specifying the nonzero entries (i.e, the ground-truth causal graphs) and then for each nonzero entry we sample $A_{i,j}^\tau\sim\text{uniform}(0.1,0.2)$. We generate $T=1,000$ samples accordingly, initialize

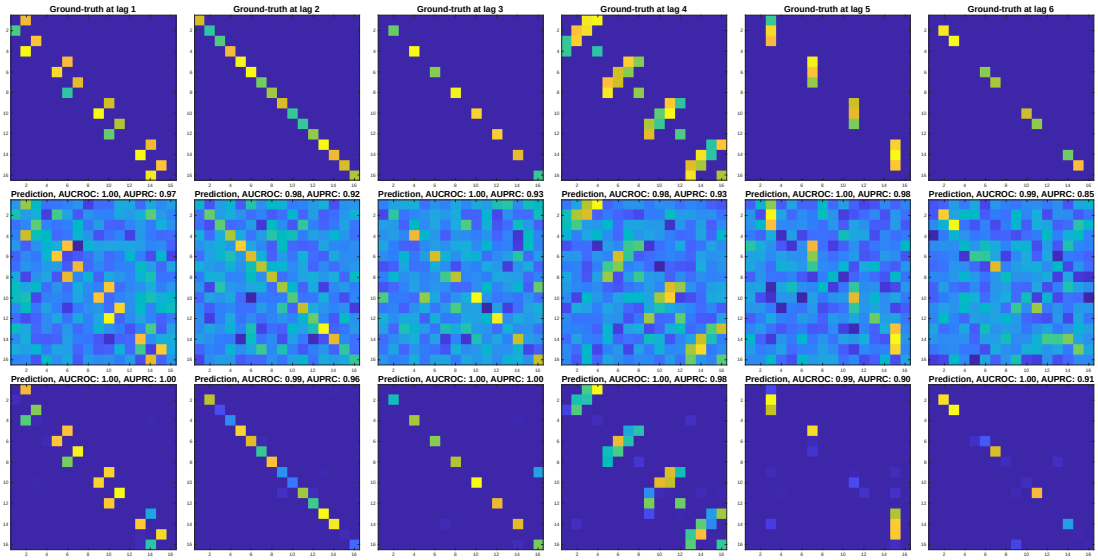
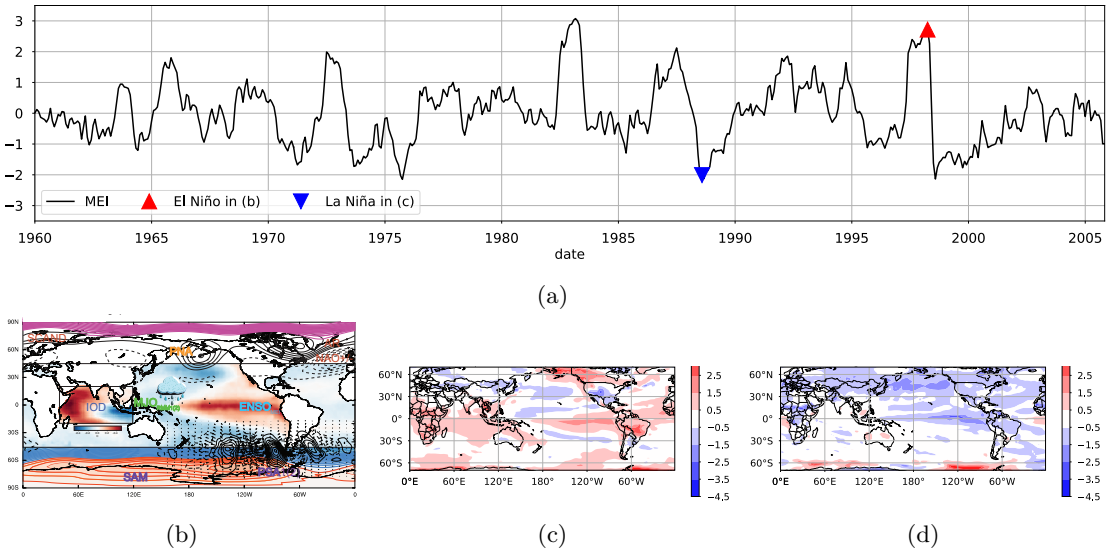


Figure 5: Synthetic dataset. First to third rows: ground-truth graphs, coefficients of VAR, and GC graphs of ours at different lags. Columns: $\tau = 1, \ldots, 6$. The AUCROC and AUPRC scores at each lag of VAR and ours are shown in the sub-captions. Mean AUCROC over all the lags is 0.99 (VAR) and 1.0 (ours); Mean AUPRC is 0.93 (VAR) and 0.96 (ours).

$\boldsymbol{x}_0$ from a standard normal distribution and sample $\epsilon_t \sim \mathcal{N}(0, 0.01)$. We show the results of our method in Figure 5, where we also fit a randomly initialized VAR (Seabold & Statsmodels, 2010) to the data as a reference. It can be seen that the ground-truth graphs at different lags have diverse patterns and our method discovers them well (also reflected by better AUCROC and AUPRC). Unlike VAR, our method directly discovers binary graphs without using thresholds or tests.

D Semi-synthetic data

1) Lorenz 96 (Lorenz, 1996) is a standard benchmark synthetic MTS dataset for GC, which is generated from the following nonlinear differential equations: $\frac{dx_{i,t}}{dt} = (x_{i+1,t} - x_{i-2,t})x_{i-1,t} - x_{i,t} + F$, for $1 \leq i \leq N$, where F is a constant that models the magnitude of the external forcing. The system dynamics become increasingly chaotic for higher values of F (Karimi & Paul, 2010). We set $N = 40$, $F = 40$, and $T = \{100, 500\}$ which mimic noisy observations with reasonably large numbers of variables but few observations. **2)** Following Marcinkevičs & Vogt (2021), we evaluate the methods on another synthetic dataset generated by the Lotka–Volterra model (Bacaër & Bacaër, 2011), where we use $N = 40$ and $T = \{200, 500\}$. For the other parameters of the Lotka–Volterra model, we use the same settings as in Marcinkevičs & Vogt (2021). **3)** We consider the FMRI dataset with realistic simulations of blood-oxygen-level dependent (BOLD) time series (Smith et al., 2011). These were generated using the dynamic causal modeling functional magnetic resonance imaging (fMRI) forward model. Following Khanna & Tan (2020); Marcinkevičs & Vogt (2021), we use 5 replicates from the simulation no. 3 of the original dataset, where $N = 15$ and $T = 200$ are pre-specified as standard settings. We introduce a straightforward metric to measure the severity of low-data of a dataset: $\beta = N/T$, meaning that larger β indicates a dataset have larger number of variables with few number of observations. We show the value of β of a dateset in Table 1. We notice that Cheng et al. (2024b) recently introduces a few new benchmark datasets with many timestamps (e.g, from 8,000 to 50,000). Our focus is on low-data regimes with less timestamps and their ground-truth graphs are undirected (i.e., the adjacency matrices are symmetric) while our method discovers directed graphs, therefore, these datasets are less applicable to our problem.



(a)

(b) (c) (d)

Figure 6: (a) MEI index. (b) The geographical locations of the indices. (c) El Niño temperature anomaly in April 1998. (d) La Niña temperature anomaly August 1988.

Table 4: SHD for VAR (FBH), GVAR, and HiBaNG with $\tau_{\max} = 1$. Means and standard derivations are computed over 5 replicates on each dataset.

	L96		LV		FMRI
	$T = 100$	$T = 500$	$T = 200$	$T = 500$	
VAR (FBH)	-	72.00±4.69	125.60±21.88	451.40±69.85	26.00±1.10
GVAR	373.20±39.77	93.60±55.23	147.80±80.36	111.00±39.75	51.00±16.79
HiBaNG	112.91±1.95	71.43±3.52	186.13±21.84	49.73±4.72	25.17±1.43

E More Introduction to the Climate Reanalysis Data

Following Harries & O'Kane (2021), we compute a set of 13 indices diagnosing the activity of the major atmospheric tropospheric and convective global climate modes at monthly resolution from 1960 to 2005, resulting in an MTS dataset where $N = 13$ and $T = 551$. The climate indices' names are shown in Table 6. Among the indices, the Multivariate El Niño Southern Oscillation Index (MEI) index is a representative example of the other climate indices in that the timeseries is associated with regionally distributed coherent responses in the atmosphere and surface ocean. In Figure 6(a), we plot the timeseries of MEI as an example, which characterizes the El Niño / La Niña cycle. Positive values of the MEI are associated with El Niño periods, negative values are associated with La Niña periods, with the magnitude of the index proportional to the strength of the event. For example, according to the MEI index, there was a large El Niño in April of 1998. Figure 6(b) illustrates how much warmer (red) and cooler (blue) the surface air temperature was in that month with respect to the average April. This map illustrates a large warm patch over the eastern Pacific ocean, which is typical of El Niño. Conversely, the MEI indicates that August 1988 was a large La Niña, with an associated surface air temperature map illustrated in Figure 6(c). This map illustrates how different the surface air temperature was with respect to an average August. As is typical of La Niña events, the eastern Pacific region of anomalously cool.

Table 5: SHD for VAR (FBH), GVAR, and HiBaNG with $\tau_{\max} = 3$. Means and standard derivations are computed over 5 replicates on each dataset.

	L96		LV		FMRI
	$T = 100$	$T = 500$	$T = 200$	$T = 500$	
VAR (FBH)	-	79.60±3.32	84.40±4.18	75.40±5.85	25.80±1.33
GVAR	577.40±47.51	133.20±107.22	213.80±117.98	128.60±68.31	50.00±18.99
HiBaNG	114.75±1.68	73.18±2.67	75.30±9.21	45.50±2.34	24.78±1.08

Table 6: Climate index names

Index name	Abbreviation
Atlantic Oscillation	AO
Indian Ocean Dipole	IOD
Multivariate El Niño Southern Oscillation Index	MEI
North Atlantic Oscillation (positive and negative phases)	NAO+/-
Atlantic Ridge patterns	AR
Scandinavian blocking patterns	SCAND
Pacific North American patterns	PNA
Pacific South American patterns	PSA1/2
Southern Annular Mode	SAM
Wheeler-Hendon Madden-Julian oscillation	RMM1/2