Uncovering Latent Causal Structures from Spatiotemporal Data

Kun Wang[1]* **Sumanth Varambally**[2]* **Duncan Watson-Parris**[2,3] **Yi-An Ma**[2,4] **Rose Yu**[4,2]
[1]Dept. of Mathematics [3]Scripps Institution of Oceanography
[2]Halıcıoğlu Data Science Institute [4]Dept. of Computer Science and Engineering
UC San Diego
{kuw010, svarambally, dwatsonparris, yianma, roseyu}@ucsd.edu

Abstract

Many important phenomenon in scientific fields such as climate, neuroscience and epidemiology are naturally represented as spatiotemporal gridded data with complex interactions. For example, in climate science, researchers aim to uncover how large-scale events, such as the North Atlantic Oscillation (NAO) and the Antarctic Oscillation (AAO), influence other global processes. Inferring causal relationships from these data is a challenging problem compounded by the high dimensionality of such data and the correlations between spatially proximate points. We present SPACY (SPAtiotemporal Causal discoverY), a novel framework based on variational inference, designed to explicitly model latent time-series and their causal relationships from spatially confined modes in the data. Our method uses an end-to-end training process that maximizes an evidence-lower bound (ELBO) for the data likelihood. Theoretically, we show that, under some conditions, the latent variables are identifiable up to transformation by an invertible matrix. Empirically, we show that SPACY outperforms state-of-the-art baselines on synthetic data, remains scalable for large grids, and identifies key known phenomena from real-world climate data.

1 Introduction

In various scientific domains such as climate science, neurology, and epidemiology, high-dimensional observational data are naturally represented as gridded time series with complex spatiotemporal interactions. Unveiling causal relationships within these datasets is crucial for predicting future states, intervening in harmful trends, and gaining insights into underlying mechanisms. In climate science, understanding teleconnections—the interactions between regions across thousands of kilometers—is essential for analyzing how climate events, such as the North Atlantic Oscillation (NAO) [Hurrell, 1995, Hurrell et al., 2003, Chen and den Dool, 2003] and the Antarctic Oscillation (AAO) [Mo, 2000, Thompson and Solomon, 2002], in one area affect weather patterns elsewhere [Liu et al., 2023].

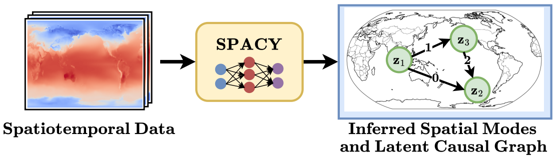


Figure 1: SPACY jointly infers latent time series and the underlying causal graph from gridded time-series data by identifying spatial modes of variability.

While several methods have been developed for causal structure learning from time-series data [Granger, 1969, Hyvärinen et al., 2010, Runge, 2020a, Tank et al., 2021, Gong Wenbo and Nick, 2022, Cheng et al., 2023], applying them to spatiotemporal data presents significant challenges.

[*] equal contribution.

Causal Representation Learning Workshop (NeurIPS 2024).

The high dimensionality of large gridded data makes it difficult for many of these techniques, especially those relying on conditional independence tests, to scale effectively [Glymour et al., 2019]. Additionally, spatially nearby time series are often highly correlated. Conditioning on nearby correlated points can obscure true causal relationships between distant locations, reducing statistical power and leading to inaccurate results [Tibau, 2022].

Recent advances in spatiotemporal causal discovery have sought to address these challenges. One common approach is a two-stage process: first, dimensionality reduction is applied to extract a small number of latent time series from the original grid of time series; then, causal discovery is performed on these reduced-dimensional representations. Examples of this approach include Tibau [2022] and Falasca et al. [2024]. However, these methods perform dimensionality reduction independent of the causal structure, potentially leading to low-dimensional representations that obscure the relationships among causally relevant entities. Another important line of research is causal representation learning from time series data [Schölkopf et al., 2021]. While approaches like those in Yao et al. [2022a,b], Chen et al. [2024] model latent time series from high-dimensional data, they do not incorporate spatial priors, making them less suitable for spatiotemporal causal discovery. Causal Discovery with Single-parent Decoding (CDSD) [Brouillard et al., 2024, Boussard et al., 2023] learns a mapping from the observational time series to latent variables to infer the latent time series. However, it assumes that each observed variable is influenced by only one latent variable.

We present a novel variational inference-based framework for spatial-temporal causal discovery called SPAtio-temporal Causal DiscoverY (SPACY) to address these limitations (Figure 1). Our approach jointly infers both the latent time series and the underlying causal graph in an end-to-end process. The key idea of our approach is to learn the location and scale parameters of spatial factors on the grid, which we model using Radial Basis Functions (RBFs). These spatial factors determine the grid locations corresponding to each inferred latent time series. Our main contributions can be summarized as follows: (1) We introduce SPAtio-temporal Causal discoverY (SPACY), a novel variational inference-based causal discovery framework that tackles realistic and challenging settings of spatiotemporal datasets by simultaneously inferring the latent causal representation time series and the underlying causal graph. (2) Theoretically, we show that, under some conditions, the latent factors are identifiable up to transformation by an invertible matrix from the observational data when the resolution of the grid is infinite. (3) Experimentally, we demonstrate the strong performance of our method on both synthetic and real-world datasets. SPACY can infer both lagged and instantaneous causal links from high-dimensional grids in a tractable manner.

2 SPACY: Spatial-Temporal Causal Discovery

Problem Setting. We are given N samples of L-dimensional multivariate time series with T timesteps each. These L time series are arranged in a K-dimensional grid $\mathcal{G} = [0,1]^K$. We consider a two-dimensional grid ($K = 2$). This means that the time series are arranged in a rectangular grid of shape $L_1 \times L_2$ where each grid cell corresponds to a distinct time series, and $L = L_1 L_2$. We denote the observational time series as $\left\{\mathbf{X}_{1:L}^{(1:T),n}\right\}_{n=1}^{N}$. We assume that the dynamics of the observed data are driven by interactions in a smaller number of *latent* (i.e. unobservable) time series. We denote the D latent time series for each of the N samples as $\left\{\mathbf{Z}_{1:D}^{(1:T),n}\right\}_{n=1}^{N}$, with $D << L$. The latent time series is stationary with a maximum time lag of τ, meaning the present is influenced by up to τ past timesteps. Interactions in the latent time series follow an SCM represented by a DAG $\mathbf{G}$. Our goal is to infer the latent time series $\left\{\mathbf{Z}_{1:D}^{(1:T),n}\right\}_{n=1}^{N}$ and the causal graph $\mathbf{G}$ in an unsupervised manner.

2.1 Forward Model

We formalize our assumptions about the data generation process using a probabilistic graphical model (Figure 2). We assume that the latent time series $\mathbf{Z}$ is generated by an SCM with causal graph $\mathbf{G}$. The number of latent variables D is input as a hyperparameter. The spatial correlations between nearby grid points are captured by the spatial factors $\mathbf{F} \in \mathbb{R}^{L \times D}$, parameterized by $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$. These factors map the latent time series $\mathbf{Z}_{1:D}^{(1:T)} \in \mathbb{R}^{D \times T}$ to the observed time series $\mathbf{X}_{1:L}^{(1:T)} \in \mathbb{R}^{L \times T}$.

Latent SCM. We model the latent SCM that describes the dynamics of $\mathbf{Z}^{(t)}$ as an additive noise model [Hoyer et al., 2008]:

$$\mathbf{Z}_d^{(t)} = f_d\left(\text{Pa}_{\mathbf{G}}^d(< t), \text{Pa}_{\mathbf{G}}^d(t)\right) + \eta_d^{(t)}.$$

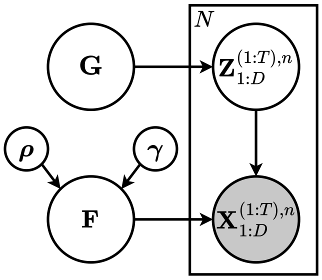


Figure 2: Probabilistic graphical model for SPACY. Shaded circles are observed and hollow circles are latent.

The causal graph $\mathbf{G}$ specifies the causal parents of each node, represented by a temporal adjacency matrix with shape $(L+1) \times D \times D$. The parent nodes from previous and current time steps are denoted by $\text{Pa}_{\mathbf{G}}^d(< t)$ and $\text{Pa}_{\mathbf{G}}^d(t)$ respectively. We assume that $\mathbf{Z}_d^t$ is influenced by at most τ preceding time steps, i.e., $\text{Pa}_{\mathbf{G}}(< t) \subseteq \{\mathbf{Z}^{t-1}, \dots, \mathbf{Z}^{t-\tau}\}$. $\mathbf{G}^{1:\tau}$ represents the lagged relationships and $\mathbf{G}^0$ represents the instantaneous edges. The time-lag τ is treated as a hyperparameter.

We implement two variants of SPACY based on the type of functional relationships being modeled.

SPACY-L. This variant models linear relationships with independent noise. f_d is defined as:

$$f_d\left(\text{Pa}_{\mathbf{G}}^d(\leq t)\right) = \sum_{k=0}^{\tau} \sum_{d'=1}^{D} (\mathbf{G} \circ W)_{d',d}^k \times \mathbf{Z}_{d'}^{t-k}, \tag{1}$$

where $\circ$ denotes the Hadamard product, and $W \in \mathbb{R}^{(\tau+1)\times D \times D}$ is a learned weight tensor. We assume that η_d^t is isotropic Gaussian noise.

SPACY-NL. This variant models non-linear relationships using Rhino [Gong Wenbo and Nick, 2022], which accounts for both instantaneous effects and history-dependent noise. We parameterize the structural equations f_d using MLPs ξ_f and λ_f shared across all nodes. We use trainable embeddings $\mathcal{E} \in \mathbb{R}^{(\tau+1)\times D \times D \times e}$ with embedding dimension e to distinguish between nodes. f_d is defined as:

$$f_d\left(\text{Pa}_{\mathbf{G}}(\leq t)\right) = \xi_f\left(\sum_{k=0}^{\tau} \sum_{j=1}^{D} \mathbf{G}_{j,d}^k \times \lambda_f\left(\left[\mathbf{Z}_j^{t-k}, \mathcal{E}_j^k\right], \mathcal{E}_0^d\right)\right). \tag{2}$$

The noise model is based on conditional spline flows [Durkan et al., 2019], with the parameters of the spline flow predicted by MLPs ξ_η and λ_η, which share a similar architecture to ξ_f and λ_f.

Spatial Factors. The low-dimensional latent time series are mapped to the high-dimensional grid by the spatial factors $\mathbf{F} \in \mathbb{R}^{L \times D}$. The d^{th} column of $\mathbf{F}$ represents the influence of the d^{th} latent variable on each grid location. To effectively capture the correlation between spatially proximate grid points under a single latent variable, we model the spatial factors using radial basis functions (RBFs). RBFs are commonly used for spatiotemporal modeling in real-life scientific fields. For example, in brain imaging [Manning et al., 2014, Farnoosh and Ostadabbas, 2021] or climate field modeling [Lozano et al., 2009, Sanford et al., 2022], data typically exhibit gradual changes across space. This continuity is effectively modeled using RBFs, which ensure locality and smoothness in the spatial factor while maintaining parameter efficiency. We assume a uniform prior over the grid $\mathcal{G}$ for the center parameter $\boldsymbol{\rho}_d$ of each kernel, and assume that the scale parameter $\boldsymbol{\gamma}_d$ comes from a standard normal distribution. Mathematically,

$$\boldsymbol{\rho}_d \sim U[0,1]^K, \boldsymbol{\gamma}_d \sim \mathcal{N}(0, I), \tag{3}$$

$$\mathbf{F}_d^\ell = \text{RBF}_d(x_\ell; \boldsymbol{\rho}_d, \boldsymbol{\gamma}_d) = \exp\left(-\frac{||x_\ell - \boldsymbol{\rho}_d||^2}{\exp(\boldsymbol{\gamma}_d)}\right), \tag{4}$$

where x_ℓ refers to the spatial coordinates of the ℓ^{th} grid point.

The observational time series is assumed to be generated by applying a grid point-wise non-linearity g_ℓ to the product of the spatial factors and latent time series, with additive Gaussian noise. We implement the nonlinearity g_ℓ as an MLP Ξ shared across all grid-points, with concatenated embeddings $\mathscr{G} \in \mathbb{R}^{L \times f}$, where f is the embedding dimension. In equations,

$$\mathbf{X}_\ell^{(t)} = g_\ell\left([\mathbf{F}\mathbf{Z}]_\ell^{(t)}\right) + \varepsilon_\ell^{(t)}, \quad \varepsilon_\ell^{(t)} \sim \mathcal{N}(0, \sigma_\ell^2 I) \tag{5}$$

$$g_\ell(x) = \Xi\left([x, \mathscr{G}_\ell]\right), \quad \mathscr{G}_\ell \in \mathbb{R}^f \tag{6}$$

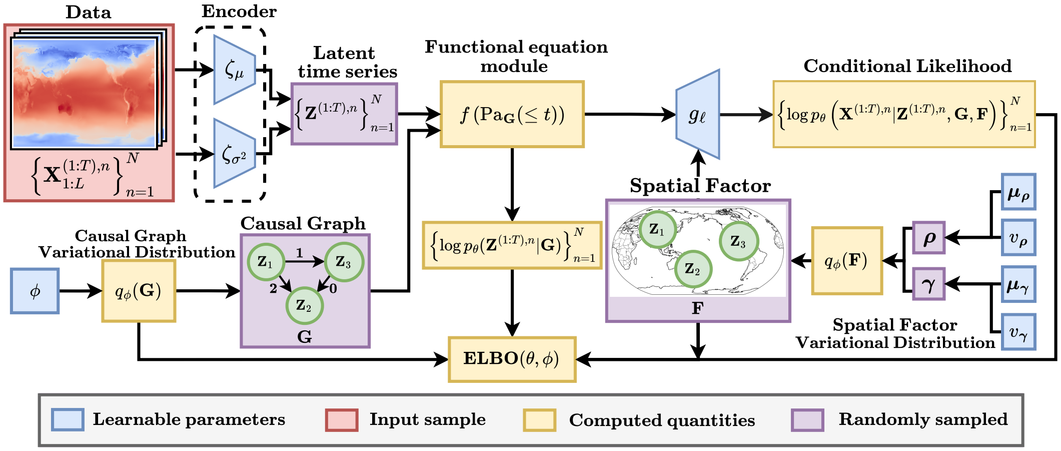

Data
Encoder
ζ_μ
ζ_{σ^2}
$\left\{\mathbf{X}_{1:L}^{(1:T),n}\right\}_{n=1}^{N}$
Latent time series
$\left\{\mathbf{Z}^{(1:T),n}\right\}_{n=1}^{N}$
Functional equation module
$f\left(\mathrm{Pa}_{\mathbf{G}}(\leq t)\right)$
g_ℓ
Conditional Likelihood
$\left\{\log p_\theta\left(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n},\mathbf{G},\mathbf{F}\right)\right\}_{n=1}^{N}$
Causal Graph Variational Distribution
ϕ
$q_\phi(\mathbf{G})$
Causal Graph
$\mathbf{Z}_1$
$\mathbf{Z}_3$
$\mathbf{Z}_2$
1
2
0
G
$\left\{\log p_\theta(\mathbf{Z}^{(1:T),n}|\mathbf{G})\right\}_{n=1}^{N}$
Spatial Factor
F
$q_\phi(\mathbf{F})$
$\boldsymbol{\rho}$
$\boldsymbol{\gamma}$
$\boldsymbol{\mu}_\rho$
v_ρ
$\boldsymbol{\mu}_\gamma$
v_γ
Spatial Factor Variational Distribution
ELBO(θ,ϕ)
Learnable parameters
Input sample
Computed quantities
Randomly sampled

Figure 3: Overview of the ELBO calculation for SPACY. The model processes spatiotemporal data $\left\{\mathbf{X}_{1:L}^{(1:T),n}\right\}_{n=1}^{N}$ to infer latent time series $\left\{\mathbf{Z}_{1:D}^{(1:T),n}\right\}_{n=1}^{N}$, where $D \ll L$. Causal relationships are modeled using a DAG $\mathbf{G}$ sampled from $q_\phi(\mathbf{G})$. Latent time series are mapped to grid locations via spatial factors $\mathbf{F}$ sampled from $q_\phi(\mathbf{F})$. Arrows in $\mathbf{G}$ are labeled with edge time-lags.

2.2 Variational Inference

Let θ denote the parameters of the forward model. Ideally, we would estimate θ using maximum likelihood estimation. However, the likelihood $p_\theta\left(\mathbf{X}\right)$ is intractable due to the presence of latent variables $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{F}$. To address this, we propose using variational inference, optimizing an evidence lower bound (ELBO) instead.

Proposition 1. *The data generation model described in Figure 2 admits the following evidence lower bound (ELBO):*

$$\begin{aligned}\log p_\theta\left(\mathbf{X}^{(1:T),1:N}\right) \geq \sum_{n=1}^{N}\Bigg\{&\mathbb{E}_{q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})q_\phi(\mathbf{G})q_\phi(\mathbf{F})}\Bigg[\log p_\theta\left(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n},\mathbf{F}\right)\\ &+\left[\log p_\theta\left(\mathbf{Z}^{(1:T),n}|\mathbf{G}\right)-\log q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})\right]\Bigg]\Bigg\}+\mathbb{E}_{q_\phi(\mathbf{G})}[\log p(\mathbf{G})-\log q_\phi(\mathbf{G})]\\ &+\mathbb{E}_{q_\phi(\mathbf{F})}[\log p(\mathbf{F})-\log q_\phi(\mathbf{F})]=\mathit{ELBO}(\theta,\phi)\end{aligned} \tag{7}$$

See section A.1.1 for the derivation. We outline the computation of the ELBO in Figure 3. q_ϕ represents the variational distribution, with variational parameters ϕ. The first term $\log p_\theta(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n},\mathbf{F})$ in (7) represents the conditional likelihood of the observed data $\mathbf{X}^{(1:T),n}$ conditioned on $\mathbf{Z}^{(1:T),n}$ and $\mathbf{F}$, and represents how well the observed data is fit. The remaining terms represent the KL divergences of the variational distributions from their prior distributions. More details about the implementation of the loss terms are in Appendix A.2.

We detail the implementation of the variational distributions below:

Causal graph $q_\phi(\mathbf{G})$. The variational distribution for the adjacency matrix $q_\phi(\mathbf{G})$ is modeled as a product of independent Bernoulli distributions, indicating the presence or absence of every edge. To compute the expectation over $q_\phi(\mathbf{G})$, we sample one graph using Monte Carlo sampling, leveraging the Gumbel-Softmax trick [Jang et al., 2017].

Spatial Factor $q_\phi(\mathbf{F})$. We model the variational distributions of the center and scale parameters $\boldsymbol{\rho}_d$ and $\boldsymbol{\gamma}_d$ as normal distributions with learnable mean and log-variance parameters $(\boldsymbol{\mu}_{\rho_d}, v_{\rho_d})$, $(\boldsymbol{\mu}_{\gamma_d}, v_{\gamma_d})$. To sample from $q_\phi(\mathbf{F})$, we first sample $\boldsymbol{\rho}_d$ and $\boldsymbol{\gamma}_d$ using the reparameterization trick [Kingma and Welling, 2014], and then compute the RBF kernel using these parameters. To ensure that the coordinates of the center lie in the range $[0, 1]$, we apply the sigmoid function.

$$\boldsymbol{\rho}_d \sim \mathcal{N}\left(\boldsymbol{\mu}_{\rho_d}, \exp\left(v_{\rho_d}\right)I\right), \boldsymbol{\gamma}_d \sim \mathcal{N}\left(\boldsymbol{\mu}_{\gamma_d}, \exp\left(v_{\gamma_d}\right)I\right)$$

$$\mathbf{F}_d^\ell = \mathrm{RBF}_d(x_\ell; \boldsymbol{\rho}_d, \boldsymbol{\gamma}_d) = \exp\left(-\frac{||x_\ell - \mathrm{sigmoid}(\boldsymbol{\rho}_d)||^2}{\exp(\boldsymbol{\gamma}_d)}\right).$$

Encoder $q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})$. To obtain the latents from the observational samples, we use a neural network encoder. Specifically, the variational distribution $q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})$ is modeled as a normal distribution whose mean and log-variance are output by MLPs ζ_μ and ζ_{σ^2}. We sample $\mathbf{Z}$ from the distribution using the reparameterization trick:

$$\mathbf{Z}^{(t),n} \sim \mathcal{N}\left(\zeta_\mu(\mathbf{X}^{(t),n}), \exp\left(\zeta_{\sigma^2}(\mathbf{X}^{(t),n})\right)\right).$$

3 Identifiability Analysis

In this section, we examine the identifiability of the generative model introduced in Section 2.1. Roughly speaking, a model is said to be identifiable if the latent variables can be uniquely recovered from observational data. Several prior works have investigated the identifiability of latent parameters in various deep generative models [Khemakhem et al., 2020, Zheng et al., 2022, Yao et al., 2022a].

We focus on the specific case where g_ℓ in (6) is the identity map. To analyze identifiability, we extend the notion of a gridded time series to infinite resolution. Instead of observing the time series at a finite set of grid points, we assume it can be observed at every point within the bounded K-dimensional grid $\mathcal{G} = [0,1]^K$. In this framework, $\mathbf{X}(x)$ represents a T-dimensional random variable describing the observational time series at location x on the grid.

We also generalize our assumptions about how the spatial factors are generated, and assume that they are function evaluations at the grid points of a family of linearly independent functions. Notably, the family of RBF functions are one such family of functions [Smola and Schölkopf, 1998]. To formalize this, we introduce the following definition.

Definition (Spatial Factor Process)**.** Let $\mathcal{G} = [0,1]^K$ be a $K-$dimensional grid, and let $\mathbf{Z} \in \mathbb{R}^{D\times T}$. Suppose $\mathcal{F} = \{F_{\psi_1}, ..., F_{\psi_D}\}$ is a finite linearly independent family. We define a Spatial Factor Process SFP($\mathbf{Z}, \mathcal{F}, p_\varepsilon$), denoted by $\mathbf{X} : \mathcal{G} \to \mathbb{R}^T$, as follows: for each location $x \in \mathcal{G}$ in the grid,

$$\mathbf{X}(x) = \mathbf{F}_x^\top \mathbf{Z} + \varepsilon_x, \text{ where } \mathbf{F}_x = \begin{bmatrix} F_{\psi_1}(x), & \ldots, & F_{\psi_D}(x)\end{bmatrix}^\top \tag{8}$$

and $\varepsilon_x \sim p_\varepsilon(\cdot)$ is a normally distributed noise term.

The first result demonstrates that if two SFPs are equal at all grid points, they must share the same spatial factors and latent time series, up to a permutation. This implies that the spatial factors and latent time series are identifiable from the *conditional* likelihood $\log p_\theta\left(\mathbf{X}(x)|\mathbf{Z}, \mathbf{F}_x\right)$.

Theorem 1 (Identifiability of SFPs)**.** *Given two SFPs $\mathbf{X} =$ SFP$(\mathbf{Z}, \mathcal{F}, p_\varepsilon)$ and $\mathbf{Y} =$ SFP$(\tilde{\mathbf{Z}}, \tilde{\mathcal{F}}, q_\varepsilon)$ where none of the rows of $\mathbf{Z}$ or $\tilde{\mathbf{Z}}$ are all zero, such that $p(\mathbf{X}(x)) = p(\mathbf{Y}(x))$ for every $x \in \mathcal{G}$, then $\mathbf{Z} = P\tilde{\mathbf{Z}}$ and $\mathcal{F} = \tilde{\mathcal{F}}$ for some permutation matrix P.*

We now turn to the identifiability of the latent time series from the observational distribution. The following result shows that the latent variable distribution can be recovered up to a transformation by an invertible matrix. Although not as precise as Theorem 1, it still guarantees that the latents are partially identifiable.

Theorem 2 (Identifiability of the latents)**.** *Suppose two spatial factor processes $\mathbf{X}(x)$ and $\tilde{\mathbf{X}}(x)$ with spatial factors $\mathbf{F}_x$ and $\tilde{\mathbf{F}}_x$ have the same observational distributions for all $x \in \mathcal{G}$. Then the latent variable distribution is identifiable up to transformation by an invertible matrix.*

The detailed mathematical statements and proofs for these results are provided in Appendix A.1.2.

4 Experiments

We assess SPACY's ability to capture causal relationships across various spatiotemporal contexts using both synthetic datasets with known ground truth and simulated climate datasets. Our results demonstrate that SPACY consistently uncovers accurate causal relationships while generating interpretable outputs. An implementation of SPACY is available at (`https://anonymous.4open.science/r/spacy-572B/`). The code is built with PyTorch 2.1 and tested on NVIDIA A10 GPUs.

Baselines. We compare SPACY with state-of-the-art baselines. We include the two-step algorithms Mapped PCMCI (Varimax-PCA + PCMCI$^+$ with Partial Correlation test) [Tibau, 2022, Runge,

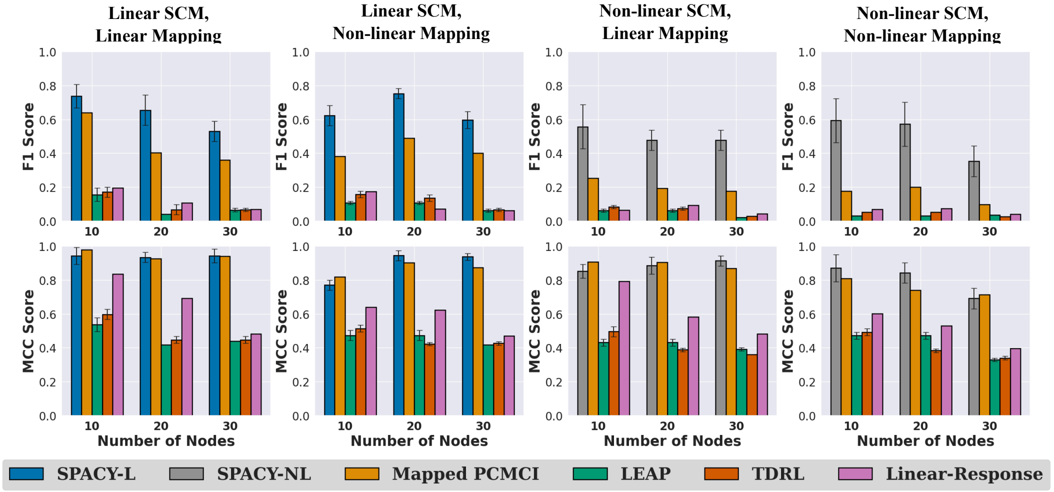


Figure 4: Results on different configurations of the synthetic datasets. We report the F1 and MCC scores for each method across different latent dimensions D. Average over 5 runs reported.

2020b] and the Linear Response method [Falasca et al., 2024]. We also evaluate against the causal representation learning approaches, LEAP [Yao et al., 2022a] and TDRL [Yao et al., 2022b].

4.1 Synthetic Data

Setup. Since real-world datasets lack ground truth causal graphs, we generate synthetic datasets with known causal relationships to benchmark SPACY's causal discovery performance. These are generated from randomly constructed ground-truth graphs, following the forward model described in Figure 2. We experiment with several configurations of synthetic data. The latent time series are generated using either (1) a linear structural causal model (SCM) with randomly initialized weights and additive Gaussian noise, or (2) a nonlinear SCM, where the structural equations are modeled by randomly initialized MLPs, combined with additive history-dependent conditional-spline noise. For more details on synthetic dataset settings and generation, refer to Appendix A.3.1.

We assess the performance of SPACY and the baselines using two metrics: the orientation F1 score of the inferred causal graph $\mathbf{G}$, and the mean correlation coefficient (MCC) between the learned and ground-truth latent representations $\mathbf{Z}$. More details on the evaluation process are in Appendix A.2.4.

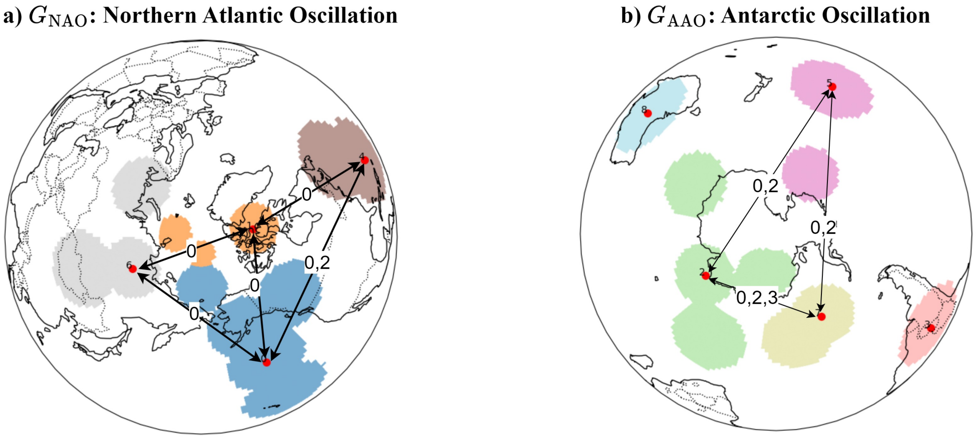
Results. The results of the synthetic experiments are shown in Figure 4. SPACY consistently outperforms all other methods across all settings of D in terms of F1 score. On the linear SCM datasets, Mapped PCMCI performs competitively, particularly when using linear spatial mapping, while LEAP, TDRL, and Linear-Response exhibit weaker performance. In the nonlinear settings, SPACY significantly outperforms the baselines, with a more pronounced performance drop observed for LEAP, TDRL, and Linear-Response, whose F1 scores decline sharply as D increases. SPACY's performance scales more effectively with increasing D, further widening the gap in performance.

The quality of the causal representations, measured by the MCC score, follows a similar pattern. Mapped PCMCI remains competitive with SPACY, while LEAP, TDRL, and Linear-Response consistently show lower MCC scores across all configurations. Figure 7 provides a visual illustration of the recovered spatial factors.

We also examine the runtime and performance of SPACY across spatial scales. The results are detailed in Appendix A.4. In addition, we assess SPACY's robustness to over-specifying latent variables D and the kernel function choice. Results are in Appendix A.5.

4.2 Real-World Application to Climate Science

The Global Temperature Dataset is a mixed real-simulated dataset containing monthly global temperature data from 1999 to 2001[Chen and den Dool, 2003]. Before applying SPACY, the data was deseasonalized by subtracting the monthly means. Due to the global nature of the data, we utilized the


a) G_{NAO}: Northern Atlantic Oscillation

b) G_{AAO}: Antarctic Oscillation

Figure 5: Qualitative results for Global Temperature climate dataset. The numbers on the arrow refers to the time lag of the causal links. Subgraph of G depicting learned causal relationships among regions associated with the (a) Northern Atlantic Oscillation (b) Antarctic Oscillation.

Haversine distance for calculating the RBF kernel in the spatial factors, as opposed to the Euclidean distance. Details on the dataset and preprocessing can be found in Appendix A.7.

Results. We qualitatively evaluate SPACY's inferred spatial factors and causal graph due to the lack of a ground truth causal graph. Figure 12 shows the spatial factors and causal graphs learned by SPACY from the Global Temperature Dataset, visualized using the method in Appendix A.3.3. The spatial modes identified by SPACY correspond to key regions influencing global climate patterns, such as coastlines of major land masses (e.g., East Asia, Northern Europe) and critical ocean areas (e.g., Central Pacific, South Atlantic).

Figure 5 highlights two subgraphs derived from SPACY's results: G_{NAO} and G_{AAO}, which align with the Northern Atlantic Oscillation (NAO) [Hurrell, 1995, Chen and den Dool, 2003, Hurrell et al., 2003] and the Antarctic Oscillation (AAO) [Thompson and Solomon, 2002, Mo, 2000]. These subgraphs reveal how SPACY uncovers causal connections between regions sharing similar weather characteristics, influenced by known teleconnection patterns. SPACY accurately identifies the spatial extent and connectivity of NAO-related regions (North-Eastern Canada, North Western Europe) [Chen and den Dool, 2003, Hurrell, 1995] and AAO-related regions (South-East Australia, South Atlantic, South Indian Ocean) [Thompson and Solomon, 2002]. The subgraphs reflect temperature correlations and oscillations in these areas, identifying both instantaneous and lagged links.

Additionally, SPACY infers modes that are spatially confined with distinct centers and scales and clear spatial boundaries, enhancing interpretability. In contrast, principal component analysis and Mapped PCMCI (Figure 13) produce broadly distributed components that are challenging to interpret.

5 Conclusion

In this work, we examined the problem of inferring causal relationships from spatiotemporal data. This problem has significant applications in climate, neuroscience, and biomedical science, among other fields. We proposed an end-to-end variational inference method to learn the latent causal representations and the underlying SCM, while producing an interpretable output. We discussed the structural identifiability of our model, and demonstrated the empirical efficacy of our method on both synthetic and simulated climate datasets. SPACY successfully recovers spatial patterns linked to known events like the Northern Atlantic Oscillation and Antarctic Oscillation.

As a direction for future work, our method can be extended to multivariate settings. Performing latent causal representation learning and causal discovery between multiple variables could further enhance the capability of our approach in handling complex real-world datasets. Such an extension would be particularly valuable in domains like climate science [Tibau, 2022, Brouillard et al., 2024], where interactions between multiple variables (e.g. temperature and pressure) are critical.

References

James W. Hurrell. Decadal trends in the north atlantic oscillation: Regional temperatures and precipitation. *Science*, 269(5224):676–679, 1995. doi: 10.1126/science.269.5224.676. URL `https://www.science.org/doi/abs/10.1126/science.269.5224.676`.

J.W Hurrell, Yochanan Kushnir, Geir Ottersen, and Martin Visbeck. *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, volume 134. 01 2003. ISBN 0-87590-994-9. doi: 10.1029/GM134.

Wilbur Y. Chen and Huug Van den Dool. Sensitivity of teleconnection patterns to the sign of their primary action center. *Monthly Weather Review*, 131(11):2885 – 2899, 2003. doi: 10.1175/1520-0493(2003)131<2885:SOTPTT>2.0.CO;2. URL `https://journals.ametsoc.org/view/journals/mwre/131/11/1520-0493_2003_131_2885_sotptt_2.0.co_2.xml`.

Kingtse C. Mo. Relationships between low-frequency variability in the southern hemisphere and sea surface temperature anomalies. *Journal of Climate*, 13(20):3599 – 3610, 2000. doi: 10.1175/1520-0442(2000)013<3599:RBLFVI>2.0.CO;2. URL `https://journals.ametsoc.org/view/journals/clim/13/20/1520-0442_2000_013_3599_rblfvi_2.0.co_2.xml`.

David W. J. Thompson and Susan Solomon. Interpretation of recent southern hemisphere climate change. *Science*, 296(5569):895–899, 2002. doi: 10.1126/science.1069270. URL `https://www.science.org/doi/abs/10.1126/science.1069270`.

Teng Liu, Dean Chen, Lan Yang, Jun Meng, Zanchenling Wang, Josef Ludescher, Jingfang Fan, Saini Yang, Deliang Chen, Jürgen Kurths, et al. Teleconnections among tipping elements in the earth system. *Nature Climate Change*, 13(1):67–74, 2023.

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

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.

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

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

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

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.

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

Xavier-Andoni Tibau. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science 1: e12.*, 2022.

Fabrizio Falasca, Pavel Perezhogin, and Laure Zanna. Data-driven dimensionality reduction and causal inference for spatiotemporal climate fields. *Physical Review E*, 109(4), April 2024. ISSN 2470-0053. doi: 10.1103/physreve.109.044202. URL `http://dx.doi.org/10.1103/PhysRevE.109.044202`.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022b.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. In *Forty-first International Conference on Machine Learning*, 2024.

Philippe Brouillard, Sebastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. 2024.

Julien Boussard, Chandni Nagda, Julia Kaltenborn, Charlotte Emilie Elektra Lange, Philippe Brouillard, Yaniv Gurwicz, Peer Nowack, and David Rolnick. Towards causal representations of climate model data. *arXiv preprint arXiv:2312.02858*, 2023.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Jeremy R Manning, Rajesh Ranganath, Kenneth A Norman, and David M Blei. Topographic factor analysis: a bayesian model for inferring brain networks from neural data. *PloS one*, 9(5):e94914, 2014.

Amirreza Farnoosh and Sarah Ostadabbas. Deep markov factor analysis: Towards concurrent temporal and spatial analysis of fmri data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17876–17888. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/951124d4a093eeae83d9726a20295498-Paper.pdf.

Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596, 2009.

Clayton Sanford, Anna Kwa, Oliver Watt-Meyer, Spencer Clark, Noah Brenowitz, Jeremy McGibbon, and Christopher Bretherton. Improving the predictions of ml-corrected climate models with novelty detection, 2022. URL https://arxiv.org/abs/2211.13354.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

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

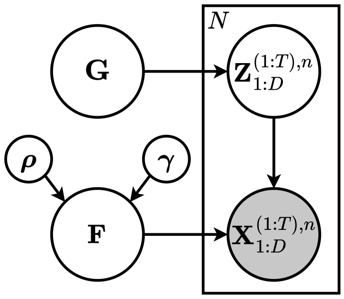
David Pollard. *A user's guide to measure theoretic probability*. Number 8. Cambridge University Press, 2002.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

A Appendix

A.1 Theory

A.1.1 ELBO Derivation



$$\begin{aligned}
\mathbf{Z}_d^{(t)} &= f_d\left(\mathrm{Pa}_G^d(<t), \mathrm{Pa}_G^d(t)\right) + \eta_d^{(t)} \\
\boldsymbol{\rho}_d &\sim U[0,1]^K, \boldsymbol{\gamma}_d \sim \mathcal{N}(0, I) \\
\mathbf{F}_d &= [\mathrm{RBF}_d(x_\ell; \boldsymbol{\rho}_d, \boldsymbol{\gamma}_d)]_{\ell=1}^L, \quad x_\ell \in \mathcal{G} \\
\mathbf{X}_\ell &= g_\ell([\mathbf{FZ}]_\ell) + \varepsilon_\ell \\
\varepsilon_\ell &\sim \mathcal{N}(0, \sigma_\ell^2 I)
\end{aligned}$$

Figure 6: Probabilistic graphical model for SPACY and the generative equations. Shaded circles are observed and hollow circles are latent.

Proposition 1. *The data generation model described in Figure 2 admits the following evidence lower bound (ELBO):*

$$\begin{aligned}
\log p_\theta\left(\mathbf{X}^{(1:T),1:N}\right) \geq \sum_{n=1}^{N} \Bigg\{ &\mathbb{E}_{q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})q_\phi(\mathbf{G})q_\phi(\mathbf{F})} \Bigg[\log p_\theta\left(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n}, \mathbf{F}\right) \\
&+ \left[\log p_\theta\left(\mathbf{Z}^{(1:T),n}|\mathbf{G}\right) - \log q_\phi\left(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n}\right)\right]\Bigg]\Bigg\} + \mathbb{E}_{q_\phi(\mathbf{G})}[\log p(\mathbf{G}) - \log q_\phi(\mathbf{G})] \\
&+ \mathbb{E}_{q_\phi(\mathbf{F})}[\log p(\mathbf{F}) - \log q_\phi(\mathbf{F})] = \textit{ELBO}(\theta, \phi)
\end{aligned}$$

Proof. We begin with the log-likelihood of the observed data:

$$\log p_\theta\left(\mathbf{X}^{(1:T),1:N}\right) = \log \int p_\theta\left(\mathbf{X}^{(1:T),1:N}, \mathbf{Z}^{(1:T),1:N}, \mathbf{G}, \mathbf{F}\right)\, d\mathbf{Z}\, d\mathbf{G}\, d\mathbf{F}$$

We multiply and divide by the variational distribution $q_\phi\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{X}^{(1:T),1:N}\right) q_\phi\left(\mathbf{G}\right) q_\phi\left(\mathbf{F}\right)$ to create an evidence lower bound (ELBO) using Jensen's inequality:

$$\begin{aligned}
&\log p_\theta\left(\mathbf{X}^{(1:T),1:N}\right) \\
&= \log \int \frac{q_\phi\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{X}^{(1:T),1:N}\right) q_\phi\left(\mathbf{G}\right) q_\phi\left(\mathbf{F}\right)}{q_\phi\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{X}^{(1:T),1:N}\right) q_\phi\left(\mathbf{G}\right) q_\phi\left(\mathbf{F}\right)} p_\theta\left(\mathbf{X}^{(1:T),1:N}, \mathbf{Z}^{(1:T),1:N}, \mathbf{G}, \mathbf{F}\right)\, d\mathbf{Z}\, d\mathbf{G}\, d\mathbf{F} \\
&\geq \mathbb{E}_{q_\phi\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{X}^{(1:T),1:N}\right)q_\phi(\mathbf{G})q_\phi(\mathbf{F})}\left[\log \frac{p_\theta\left(\mathbf{X}^{(1:T),1:N}, \mathbf{Z}^{(1:T),1:N}, \mathbf{G}, \mathbf{F}\right)}{q_\phi\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{X}^{(1:T),1:N}\right) q_\phi\left(\mathbf{G}\right) q_\phi\left(\mathbf{F}\right)}\right]. \qquad (9)
\end{aligned}$$

By the assumptions of the data generative process,

$$p_\theta\left(\mathbf{X}^{(1:T),1:N}, \mathbf{Z}^{(1:T),1:N}, \mathbf{G}, \mathbf{F}\right) = p_\theta\left(\mathbf{X}^{(1:T),1:N}|\mathbf{Z}^{(1:T),1:N}, \mathbf{F}\right) p_\theta\left(\mathbf{Z}^{(1:T),1:N}|\mathbf{G}\right) p\left(\mathbf{F}\right) p\left(\mathbf{G}\right)$$

Further, note that $\mathbf{X}^{(1:T),1:N}$ are conditionally independent given $\mathbf{F}, \mathbf{Z}^{(1:T),1:N}$. Also, $\mathbf{X}^{(1:T),n}$ is conditionally independent of $\mathbf{Z}^{(1:T),m}$ given $\mathbf{Z}^{(1:T),n}, \mathbf{F}$ for $m \neq n$. This implies that:

$$p_\theta\left(\mathbf{X}^{(1:T),1:N} | \mathbf{Z}^{(1:T),1:N}, \mathbf{F}\right) = \prod_{n=1}^{N} p_\theta\left(\mathbf{X}^{(1:T),n} | \mathbf{Z}^{(1:T),n}, \mathbf{F}\right).$$

Similarly, $\mathbf{Z}^{(1:T),1:N}$ are conditionally independent given $\mathbf{G}$, which implies

$$p_\theta\left(\mathbf{Z}^{(1:T),1:N} | \mathbf{G}\right) = \prod_{n=1}^{N} p_\theta\left(\mathbf{Z}^{(1:T),n} | \mathbf{G}\right).$$

Substituting these terms back into (9) and grouping terms according to the variables $\mathbf{Z}, \mathbf{G}, \mathbf{F}$ yields the ELBO.

$$\begin{aligned} \log p_\theta\left(\mathbf{X}^{(1:T),1:N}\right) \geq \sum_{n=1}^{N} \Bigg\{ & \mathbb{E}_{q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n}) q_\phi(\mathbf{G}) q_\phi(\mathbf{F})} \Bigg[\log p_\theta\left(\mathbf{X}^{(1:T),n} | \mathbf{Z}^{(1:T),n}, \mathbf{G}, \mathbf{F}\right) \\ & + \left(\log p_\theta\left(\mathbf{Z}^{(1:T),n} | \mathbf{G}\right) - \log q_\phi\left(\mathbf{Z}^{(1:T),n} | \mathbf{X}^{(1:T),n}\right)\right) \Bigg] \Bigg\} \\ & + \mathbb{E}_{q_\phi(\mathbf{G})}\left[\log p(\mathbf{G}) - \log q_\phi(\mathbf{G})\right] \\ & + \mathbb{E}_{q_\phi(\mathbf{F})}\left[\log p(\mathbf{F}) - \log q_\phi(\mathbf{F})\right] \equiv \text{ELBO}(\theta, \phi). \end{aligned}$$

□

A.1.2 Identifiability

Definition 1 (Linearly Independent Family)**.** Let $\mathcal{F}$ be a family of real-valued, parametric functions $\mathcal{F} = \left\{f_\psi : \mathbb{R}^K \to \mathbb{R}\right\}$. $\mathcal{F}$ is said to be a linearly independent family if, for any finite set $\{\psi_1, ..., \psi_n\}$, we have

$$\sum_{k=1}^{n} \alpha_k f_{\psi_k} = 0 \implies \alpha_k = 0 \quad \forall k \in [n]. \tag{10}$$

Definition 2 (Spatial Factor Process)**.** Let $\mathcal{G} = [0,1]^K$ be a $K-$dimensional grid, and let $\mathbf{Z} \in \mathbb{R}^{D \times T}$. Suppose $\mathcal{F} = \{F_{\psi_1}, ..., F_{\psi_D}\}$ is a finite linearly independent family. We define a Spatial Factor Process SFP$(\mathbf{Z}, \mathcal{F}, p_\varepsilon)$, denoted by $\mathbf{X} : \mathcal{G} \to \mathbb{R}^T$, as follows:

For each location $x \in \mathcal{G}$ in the grid,

$$\mathbf{X}(x) = \mathbf{F}_x^\top \mathbf{Z} + \varepsilon_x \tag{11}$$

where

$$\mathbf{F}_x = \begin{bmatrix} F_{\psi_1}(x) \\ \vdots \\ F_{\psi_D}(x) \end{bmatrix},$$

and $\varepsilon_x \sim p_\varepsilon(\cdot)$ is a normally distributed noise term

A Spatial Factor Process (SFP) extends the concept of a gridded time series to an infinite resolution. Instead of observing the time series on a finite set of grid points, we assume that a time series can be observed at every location within a bounded K-dimensional grid, $\mathcal{G} = [0,1]^K$. In the above definition, $\mathbf{Z}$ represents a (fixed) realization of a $D-$dimensional time series of length T.

We now show that SFPs are identifiable, i.e., if the distributions of two SFPs are equal, then their corresponding parameters $\mathbf{Z}$ and $\mathbf{F}$ are also equal (upto permutation).

Theorem 3 (Identifiability of SFPs)**.** *If we have two SFPs $\mathbf{X} = SFP(\mathbf{Z}, \mathcal{F}, p_\varepsilon)$ and $\mathbf{Y} = SFP(\tilde{\mathbf{Z}}, \tilde{\mathcal{F}}, q_\varepsilon)$ where none of the rows of $\mathbf{Z}$ or $\tilde{\mathbf{Z}}$ are all zero, such that $p(\mathbf{X}(x)) = p(\mathbf{Y}(x))$ for every $x \in \mathcal{G}$, then $\mathbf{Z} = P\tilde{\mathbf{Z}}$ and $\mathcal{F} = \tilde{\mathcal{F}}$ for some permutation matrix P.*

Proof. Note that, for every $\mathbf{v} \in \mathbb{R}^T$,

$$\begin{aligned} & p(\mathbf{X}(x) = \mathbf{v}) = p(\mathbf{Y}(x) = \mathbf{v}) \\ \implies & p_\varepsilon \left(\varepsilon_x = \mathbf{v} - \mathbf{y}\right) = q_\varepsilon \left(\tilde{\varepsilon}_x = \mathbf{v} - \tilde{\mathbf{y}}\right) \end{aligned}$$

where $\mathbf{y} = \mathbf{F}_x^\top \mathbf{Z}$ and $\tilde{\mathbf{y}} = \tilde{\mathbf{F}_x}^\top \tilde{\mathbf{Z}}$. Since p_ε and q_ε are normally distributed, this can only be true when

$$\begin{aligned} & \mathbf{y} = \tilde{\mathbf{y}} \\ \implies & \mathbf{F}_x^\top \mathbf{Z} = \tilde{\mathbf{F}_x}^\top \tilde{\mathbf{Z}} \quad \forall x \in \mathcal{G} \\ \implies & \sum_{j=1}^{D} F_{\psi_j}(x) z_{jt} = \sum_{j=1}^{D} F_{\tilde{\psi_j}}(x) \tilde{z}_{jt} \quad \forall x \in \mathcal{G}, t \in [T] \\ \implies & \sum_{j=1}^{D} F_{\psi_j}(x) z_{jt} - \sum_{j=1}^{D} F_{\tilde{\psi_j}}(x) \tilde{z}_{jt} = 0 \quad \forall x \in \mathcal{G}, t \in [T] \end{aligned} \tag{12}$$

Suppose $\{\psi_1, \dots, \psi_D\} \cap \{\tilde{\psi_1}, \dots, \tilde{\psi_D}\} = \varnothing$. Then, this would imply that $z_{jt} = \tilde{z_{jt}} = 0 \quad \forall j, t$, which is a contradiction since we assume that none of the time series are all 0. This implies that $\{\psi_1, \dots, \psi_D\} \cap \{\tilde{\psi_1}, \dots, \tilde{\psi_D}\} \neq \varnothing$. Assume $V = \left\{(i, j) : \psi_i = \tilde{\psi_j}\right\}$ and define $I = \{i : \exists j \text{ such that } (i, j) \in V\}$, $J = \{j : \exists\, i \text{ such that } (i, j) \in V\}$. Define the function $\mathcal{V} : I \to J, \quad \mathcal{V}(i) = j$ such that $(i, j) \in V$. Then (12) can be written as:

$$\sum_{\substack{j=1 \\ j \notin I}}^{D} F_{\psi_j}(x) z_{jt} - \sum_{\substack{j=1 \\ j \notin J}}^{D} F_{\tilde{\psi_j}}(x) \tilde{z}_{jt} + \sum_{\substack{j=1 \\ j \in I}}^{D} F_{\psi_j}(x) \left(z_{jt} - \tilde{z}_{\mathcal{V}(j)t}\right) = 0 \quad \forall x \in \mathcal{G}, t \in [T].$$

If $I \neq \varnothing$, then $z_{jt} = 0 \; \forall j \notin I$ due to the linear independence of F_{ψ_j}, which contradicts our assumption of non-zero time series. Therefore, we must have that $\{\psi_1, \dots, \psi_D\} = \{\tilde{\psi_1}, \dots, \tilde{\psi_D}\}$, and $z_{jt} = \tilde{z}_{\mathcal{V}(j)t} \; \forall j, t$. □

We now consider the identifiability of the parameters from the observational distribution. To this end, we first introduce a useful lemma. We adapt the arguments from Lemma 3 in Boussard et al. [2023] with some modifications.

Lemma 1 (Denoising $\mathbf{X}$). *Assume we have two models $\mathbf{X}$ and $\tilde{\mathbf{X}}$ with spatial factors $\mathbf{F}_x$ and $\tilde{\mathbf{F}}_x$ respectively. Assume that the observational distributions of $\mathbf{X}(x)$ and $\tilde{\mathbf{X}}(x)$ are equal, i.e., the following property holds:*

For any finite set of grid points $\{x_1, \dots, x_n\} \in \mathcal{G}$, we have

$$p\left(\mathbf{X}(x_1) = \chi_1, \dots, \mathbf{X}(x_n) = \chi_n\right) = p\left(\tilde{\mathbf{X}}(x_1) = \chi_1, \dots, \tilde{\mathbf{X}}(x_n) = \chi_n\right) \tag{13}$$

for all values of $(\chi_1, \dots, \chi_n) \in \mathbb{R}^T \times \mathbb{R}^n$. Then we have that the following holds:

Given any set of points $\{x'_1, \dots, x'_k\}$, we have that $p\left(\mathbf{Y}(x'_1), \dots, \mathbf{Y}(x'_k)\right) = p\left(\tilde{\mathbf{Y}}(x'_1), \dots, \tilde{\mathbf{Y}}(x'_k)\right)$, where $\mathbf{Y}(x) := \mathbf{F}_x^\top \mathbf{Z}$ and $\tilde{\mathbf{Y}}(x) := \tilde{\mathbf{F}}_x^\top \tilde{\mathbf{Z}}$.

Proof. Pick n distinct grid points $\{x_1, \dots, x_n\} \subseteq \mathcal{G}$ such that $\{x'_1, \dots, x'_k\} \cap \{x_1, \dots, x_n\} = \phi$ and $n + k > D$. Let $L = n + k$. Then, we can use the same argument as in Lemma 3 in Boussard et al. [2023] on the distributions of $\{\mathbf{X}(x_1), \dots, \mathbf{X}(x_n), \mathbf{X}(x'_1), \dots, \mathbf{X}(x'_k)\}$ and $\left\{\tilde{\mathbf{X}}(x_1), \dots, \tilde{\mathbf{X}}(x_n), \tilde{\mathbf{X}}(x'_1), \dots, \tilde{\mathbf{X}}(x'_k)\right\}$, which we repeat for the sake of completeness.

Let $\mathbb{P}_{\mathbf{X}(x_1),\dots,\mathbf{X}(x_n),\mathbf{X}(x'_1),\dots,\mathbf{X}(x'_k)}$ and $\mathbb{P}_{\tilde{\mathbf{X}}(x_1),\dots,\tilde{\mathbf{X}}(x_n),\tilde{\mathbf{X}}(x'_1),\dots,\tilde{\mathbf{X}}(x'_k)}$ denote the probability measures corresponding to the densities

$$p(\mathbf{X}(x_1), \dots, \mathbf{X}(x_n), \mathbf{X}(x'_1), \dots, \mathbf{X}(x'_k)) := \int p(\mathbf{X}(x_1), \dots, \mathbf{X}(x_n), \mathbf{X}(x'_1), \dots, \mathbf{X}(x'_k), \mathbf{Z}) \, d\mathbf{Z},$$

$$p(\tilde{\mathbf{X}}(x_1), \dots, \tilde{\mathbf{X}}(x_n), \tilde{\mathbf{X}}(x'_1), \dots, \tilde{\mathbf{X}}(x'_k)) := \int p(\tilde{\mathbf{X}}(x_1), \dots, \tilde{\mathbf{X}}(x_n), \tilde{\mathbf{X}}(x'_1), \dots, \tilde{\mathbf{X}}(x'_k), \tilde{\mathbf{Z}}) \, d\tilde{\mathbf{Z}},$$

respectively.

It is given that:

$$\mathbb{P}_{\mathbf{X}(x_1),\ldots,\mathbf{X}(x_n),\mathbf{X}(x_1'),\ldots,\mathbf{X}(x_k')} = \mathbb{P}_{\tilde{\mathbf{X}}(x_1),\ldots,\tilde{\mathbf{X}}(x_n),\tilde{\mathbf{X}}(x_1'),\ldots,\tilde{\mathbf{X}}(x_k')}$$

Define $\mathbf{Y}(x) := \mathbf{F}_x^\top \mathbf{Z}$ and $\tilde{\mathbf{Y}}(x) := \tilde{\mathbf{F}}_x^\top \tilde{\mathbf{Z}}$, where $\mathbf{Z} \sim p(\mathbf{Z})$ and $\tilde{\mathbf{Z}} \sim p(\tilde{\mathbf{Z}})$. Let $\mathcal{Y} = (\mathbf{Y}(x_1),\ldots,\mathbf{Y}(x_n),\mathbf{Y}(x_1'),\ldots,\mathbf{Y}(x_k'))$ and $\tilde{\mathcal{Y}} = \left(\tilde{\mathbf{Y}}(x_1),\ldots,\tilde{\mathbf{Y}}(x_n),\tilde{\mathbf{Y}}(x_1'),\ldots,\tilde{\mathbf{Y}}(x_k')\right)$. Let $\mathbb{P}_{\mathbf{Y}(x)}$ and $\mathbb{P}_{\tilde{\mathbf{Y}}(x)}$ be the distributions of $\mathbf{Y}(x)$ and $\tilde{\mathbf{Y}}(x)$, respectively. We have:

$$\mathbf{X}(x) = \mathbf{Y}(x) + \varepsilon_x, \quad \tilde{\mathbf{X}}(x) = \tilde{\mathbf{Y}}(x) + \tilde{\varepsilon}_x,$$

where $\varepsilon_x \sim \mathcal{N}(0, \sigma^2 I_T)$ and $\tilde{\varepsilon}_x \sim \mathcal{N}(0, \tilde{\sigma}^2 I_T)$.

Denote $\boldsymbol{\varepsilon} = \left(\varepsilon_{x_1},\ldots,\varepsilon_{x_n},\varepsilon_{x_1'},\ldots,\varepsilon_{x_k'}\right)$ and $\tilde{\boldsymbol{\varepsilon}} = \left(\tilde{\varepsilon}_{x_1},\ldots,\tilde{\varepsilon}_{x_n},\tilde{\varepsilon}_{x_1'},\ldots,\tilde{\varepsilon}_{x_k'}\right)$.

By the additive structure of the model, the equality of measures becomes a convolution equation:

$$\mathbb{P}_{\mathcal{Y}} * \mathbb{P}_{\boldsymbol{\varepsilon}} = \mathbb{P}_{\tilde{\mathcal{Y}}} * \mathbb{P}_{\tilde{\boldsymbol{\varepsilon}}},$$

where $\mathbb{P}_{\boldsymbol{\varepsilon}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\varepsilon}}}$ represent the measures of the Gaussian noise terms, and $*$ denotes convolution.

Applying the Fourier transform $\mathscr{F}$ to both sides and using the fact that the Fourier transform of a convolution is the product of the Fourier transforms [Pollard, 2002],

$$\begin{aligned} &\mathscr{F}\left(\mathbb{P}_{\mathcal{Y}} * \mathbb{P}_{\boldsymbol{\varepsilon}}\right) = \mathscr{F}\left(\mathbb{P}_{\tilde{\mathcal{Y}}} * \mathbb{P}_{\tilde{\boldsymbol{\varepsilon}}}\right) \\ \implies &\mathscr{F}\left(\mathbb{P}_{\mathcal{Y}}\right)\mathscr{F}\left(\mathbb{P}_{\boldsymbol{\varepsilon}}\right) = \mathscr{F}\left(\mathbb{P}_{\tilde{\mathcal{Y}}}\right)\mathscr{F}\left(\mathbb{P}_{\tilde{\boldsymbol{\varepsilon}}}\right). \end{aligned}$$

Given that the Fourier transform of a zero-mean Gaussian random vector with covariance $\sigma^2 I_{LT}$ is $e^{-\frac{\sigma^2}{2}\omega^\top\omega}$, we can rewrite the above as:

$$\mathscr{F}\left(\mathbb{P}_{\mathbf{Y}(x)}\right)(\omega)e^{-\frac{\sigma^2}{2}\omega^\top\omega} = \mathscr{F}\left(\mathbb{P}_{\tilde{\mathbf{Y}}(x)}\right)(\omega)e^{-\frac{\tilde{\sigma}^2}{2}\omega^\top\omega}, \quad \forall\omega \in \mathbb{R}^T.$$

We now aim to show that $\sigma^2 = \tilde{\sigma}^2$. Assume, without loss of generality, that $\sigma^2 < \tilde{\sigma}^2$. Dividing both sides by $e^{-\frac{\sigma^2}{2}\omega^\top\omega}$ yields:

$$\mathscr{F}\left(\mathbb{P}_{\mathcal{Y}}\right)(\omega) = \mathscr{F}\left(\mathbb{P}_{\tilde{\mathcal{Y}}}\right)(\omega)e^{-\frac{\tilde{\sigma}^2-\sigma^2}{2}\omega^\top\omega}, \quad \forall\omega \in \mathbb{R}^{LT}.$$

Here, $e^{-\frac{\tilde{\sigma}^2-\sigma^2}{2}\omega^\top\omega}$ is the Fourier transform of a Gaussian distribution with covariance $(\tilde{\sigma}^2 - \sigma^2)I_{LT}$. However, note that the left-hand side is the Fourier transform of a distribution supported on the column span of $\mathbf{F}_x$, which lies in a D-dimensional subspace of $\mathbb{R}^{LT}$. In contrast, the right-hand side corresponds to a distribution with full support in $\mathbb{R}^{LT}$, as it involves the convolution of $\mathbb{P}_{\tilde{\mathcal{Y}}}$ with a LT-dimensional Gaussian random variable. This is a contradiction, as the supports of the distributions on both sides must match.

Thus, we must have $\sigma^2 = \tilde{\sigma}^2$.

Finally, with $\sigma^2 = \tilde{\sigma}^2$, we conclude that:

$$\begin{aligned} \mathscr{F}\left(\mathbb{P}_{\mathcal{Y}}\right) &= \mathscr{F}\left(\mathbb{P}_{\tilde{\mathcal{Y}}}\right), \\ \mathbb{P}_{\mathcal{Y}} &= \mathbb{P}_{\tilde{\mathcal{Y}}}. \end{aligned}$$

Marginalizing out the variables $\mathbf{Y}(x_1),\ldots,\mathbf{Y}(x_n)$ and $\tilde{\mathbf{Y}}(x_1),\ldots,\tilde{\mathbf{Y}}(x_n)$ yields the desired result.

□

Theorem 4 (Identifiability of the latents). *Suppose we have two spatial factor processes $\mathbf{X}(x)$ and $\tilde{\mathbf{X}}(x)$ with spatial factors $\mathbf{F}_x$ and $\tilde{\mathbf{F}}_x$ respectively, generated from linearly independent families $\mathcal{F} = \{f_{\psi_1},\ldots,f_{\psi_D}\}$ and $\tilde{\mathcal{F}} = \left\{f_{\tilde{\psi}_1},\ldots,f_{\tilde{\psi}_D}\right\}$ respectively.*

Suppose the observational distributions of $\mathbf{X}(x)$ and $\tilde{\mathbf{X}}(x)$ are equal, i.e., the following property holds:

For any finite set of grid points $\{x_1, \dots, x_n\} \in \mathcal{G}$, we have

$$p\left(\mathbf{X}(x_1) = \chi_1, \dots, \mathbf{X}(x_n) = \chi_n\right) = p\left(\tilde{\mathbf{X}}(x_1) = \chi_1, \dots, \tilde{\mathbf{X}}(x_n) = \chi_n\right) \tag{14}$$

for all values of $(\chi_1, \dots, \chi_n) \in \mathbb{R}^T \times \mathbb{R}^n$.

Then the latent variable distribution is identifiable up to transformation by an invertible matrix.

Proof. Since (14) holds, we can apply Lemma 1, by which we have that:

$$p\left(\mathbf{Y}(x_1) = \mathbf{y}_1, \dots, \mathbf{Y}(x_n) = \mathbf{y}_n\right) = p\left(\tilde{\mathbf{Y}}(x_1) = \mathbf{y}_1, \dots, \tilde{\mathbf{Y}}(x_n) = \mathbf{y}_n\right) \tag{15}$$

$\forall \left(\mathbf{y}_1, \dots, \mathbf{y}_n\right) \in \mathbb{R}^T \times \mathbb{R}^n$ where $\mathbf{Y}(x) = \mathbf{F}_x^\top \mathbf{Z}$.

Since the family $\mathcal{F}$ is linearly independent, we can pick D points $\{x_1, \dots, x_D\}$ from $\mathcal{G}$ such that

$$\mathfrak{F} = \begin{bmatrix} f_{\psi_1}(x_1) & \cdots & f_{\psi_D}(x_1) \\ \vdots & & \vdots \\ f_{\psi_1}(x_D) & \cdots & f_{\psi_D}(x_D) \end{bmatrix} = \begin{bmatrix} \text{——} & \mathbf{F}_{x_1}^\top & \text{——} \\ \vdots & \vdots & \vdots \\ \text{——} & \mathbf{F}_{x_D}^\top & \text{——} \end{bmatrix}$$

is full rank [2].

Similarly, we can pick D points $\{\tilde{x}_1, \dots, \tilde{x}_D\}$ from $\mathcal{G}$ such that

$$\tilde{\mathfrak{F}} = \begin{bmatrix} f_{\tilde{\psi}_1}(\tilde{x}_1) & \cdots & f_{\tilde{\psi}_D}(\tilde{x}_1) \\ \vdots & & \vdots \\ f_{\tilde{\psi}_1}(\tilde{x}_D) & \cdots & f_{\tilde{\psi}_D}(\tilde{x}_D) \end{bmatrix} = \begin{bmatrix} \text{——} & \tilde{\mathbf{F}}_{\tilde{x}_1}^\top & \text{——} \\ \vdots & \vdots & \vdots \\ \text{——} & \tilde{\mathbf{F}}_{\tilde{x}_D}^\top & \text{——} \end{bmatrix}$$

is full rank.

Define:

$$\mathcal{Y} = \begin{bmatrix} \mathbf{Y}(x_1) \\ \vdots \\ \mathbf{Y}(x_D) \end{bmatrix}, \quad \tilde{\mathcal{Y}} = \begin{bmatrix} \tilde{\mathbf{Y}}(x_1) \\ \vdots \\ \tilde{\mathbf{Y}}(x_D) \end{bmatrix}.$$

Let $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_D)$ and $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_D)$.

Observe that

$$\mathcal{Y} = \mathfrak{F}\mathbf{Z}$$
$$\tilde{\mathcal{Y}} = \tilde{\mathfrak{F}}\tilde{\mathbf{Z}}$$

Using the formula for transformation of random variables,

$$p\left(\mathcal{Y} = \mathbf{y}\right) = \left|\det\left(\mathfrak{F}\right)\right| p\left(\mathbf{Z} = \mathfrak{F}^{-1}\mathbf{y}\right), \forall \mathbf{y} \in \mathbb{R}^{D \times T}$$
$$p\left(\tilde{\mathcal{Y}} = \mathbf{y}\right) = \left|\det\left(\tilde{\mathfrak{F}}\right)\right| p\left(\tilde{\mathbf{Z}} = \tilde{\mathfrak{F}}^{-1}\mathbf{y}\right), \forall \mathbf{y} \in \mathbb{R}^{D \times T}$$

Applying (15) for the points $\{x_1, \dots, x_D\}$, we can obtain

$$\left|\det\left(\mathfrak{F}\right)\right| p\left(\mathbf{Z} = \mathfrak{F}^{-1}\mathbf{y}\right) = \left|\det\left(\tilde{\mathfrak{F}}\right)\right| p\left(\tilde{\mathbf{Z}} = \tilde{\mathfrak{F}}^{-1}\mathbf{y}\right), \forall \mathbf{y} \in \mathbb{R}^{D \times T}$$
$$\implies p\left(\mathbf{Z} = \mathfrak{F}^{-1}\mathbf{y}\right) = \frac{\left|\det\left(\tilde{\mathfrak{F}}\right)\right|}{\left|\det\left(\mathfrak{F}\right)\right|} \times p\left(\tilde{\mathbf{Z}} = \tilde{\mathfrak{F}}^{-1}\mathbf{y}\right), \forall \mathbf{y} \in \mathbb{R}^{D \times T}.$$

[2] See for example `https://math.stackexchange.com/questions/3516189/prove-existence-of-evaluation-points-such-that-the-matrix-has-nonzero-determinan`

Making the substitution $\mathbf{z} = \mathfrak{F}^{-1}\mathbf{y}$ and writing $\mathcal{M} = \tilde{\mathfrak{F}}^{-1}\mathfrak{F}$ yields:

$$p\left(\mathbf{Z} = z\right) = \frac{\left|\det\left(\tilde{\mathfrak{F}}\right)\right|}{\left|\det\left(\mathfrak{F}\right)\right|} \times p\left(\tilde{\mathbf{Z}} = \mathcal{M}\mathbf{z}\right),\ \forall \mathbf{y} \in \mathbb{R}^{D\times T}.$$

for the invertible matrix $\mathcal{M}$. Thus, we can recover the latent distribution up to transformation via an invertible matrix. □

A.2 Implementation Details

A.2.1 Loss Terms

We explain how we implement the various loss terms in Equation (7).

The first term $\log p_\theta\left(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n}, \mathbf{F}\right)$ in Equation (7) represents the conditional likelihood of the observed data $\mathbf{X}^{(1:T),n}$ conditioned on $\mathbf{Z}^{(1:T),n}$ and $\mathbf{F}$. This is calculated as the mean squared error (MSE) between the recovered and original time series:

$$\log p_\theta\left(\mathbf{X}^{(1:T),n}|\mathbf{Z}^{(1:T),n}, \mathbf{F}\right) = \sum_{\ell=1}^{L} \left\|\mathbf{X}_\ell^{(1:T),n} - \widehat{\mathbf{X}}_\ell^{(1:T),n}\right\|^2$$

where $\widehat{\mathbf{X}}_\ell^{(t),n} = g_\ell\left([\mathbf{F}\mathbf{Z}]_\ell^{(t)}\right)$ is the reconstructed time-series from the spatial factor $\mathbf{F}$ and latent time series $\mathbf{Z}$ sampled from the variational distributions.

The term $\log p_\theta\left(\mathbf{Z}^{(1:T),n}|\mathbf{G}\right)$ denotes the conditional likelihood of the latent time-series given the sampled graph $\mathbf{G}$.

For SPACY-L, this is implemented as follows:

$$\begin{aligned}\log p_\theta\left(\mathbf{Z}^{(1:T),n}\middle|\mathbf{G}\right) &= \sum_{t=L}^{T}\sum_{d=1}^{D} \log p_\theta\left(\mathbf{Z}_d^{(t),n}\middle|\mathrm{Pa}_{\mathbf{G}}^d(\leq t)\right)\\ &= \sum_{t=L}^{T}\sum_{d=1}^{D}\left[\mathbf{Z}_d^{(t),n} - \sum_{k=0}^{\tau}\sum_{j=1}^{D}\left(\mathbf{G}\circ W\right)_{j,d}^k \times \mathbf{Z}_j^{(t-k),n}\right]^2.\end{aligned}$$

For SPACY-NL, the equation follows from the conditional spline flow model employed in Durkan et al. [2019], Gong Wenbo and Nick [2022]. The conditional spline flow model handles more flexible noise distributions, and can also model history-dependent noise. The structural equations are modeled as follows:

$$\mathbf{Z}_d^{(t)} = f_d\left(\mathrm{Pa}_{\mathbf{G}}^d(<t), \mathrm{Pa}_{\mathbf{G}}^d(t)\right) + w_d\left(\mathrm{Pa}_{\mathbf{G}}^d(<t)\right),$$

where $f_d\left(\mathrm{Pa}_{\mathbf{G}}^d(<t), \mathrm{Pa}_{\mathbf{G}}^d(t)\right)$ takes the form presented in (2). The spline flow model uses hypernetwork that predicts parameters for the conditional spline flow model, with embeddings $\mathscr{E}$, and hypernetworks ξ_η and λ_η. The only difference is that the output dimension of ξ_η is different, being equal to the number of spline parameters.

The noise variables $\eta_d^{(t)}$ are described using a conditional spline flow model,

$$p_{w_d}(w_d(\eta_d^{(t)}) \mid \mathrm{Pa}_{\mathbf{G}}^d(<t)) = p_\eta(\eta_d^{(t)})\left|\frac{\partial(w_d)^{-1}}{\partial\eta_d^{(t)}}\right|, \tag{16}$$

with $\eta_d^{(t)}$ modeled as independent Gaussian noise.

The marginal likelihood becomes:

$$\log p_\theta\left(\mathbf{Z}^{(1:T),n}\middle|\mathbf{G}\right) = \sum_{t=\tau}^{T}\sum_{d=1}^{D}\log p_\theta\left(\mathbf{Z}_d^{(t),n}\middle|\mathrm{Pa}_{\mathbf{G}}^d(<t),\mathrm{Pa}_{\mathbf{G}}^d(t)\right)$$
$$= \sum_{t=\tau}^{T}\sum_{d=1}^{D}\log p_{w_d}\left(u_d^{(t),n}\middle|\mathrm{Pa}_{\mathbf{G}}^d(<t)\right) \tag{17}$$

where $u_d^{(t),n} = \mathbf{Z}_d^{(t),n} - f_d\left(\mathrm{Pa}_{\mathbf{G}}^d(<t),\mathrm{Pa}_{\mathbf{G}}^d(t)\right)$.

The prior distribution $p(\mathbf{G})$ is modeled as follows:

$$p(\mathbf{G}) \propto \exp\left(-\alpha\left\|\mathbf{G}^{(0:T)}\right\|^2 - \sigma h\left(\mathbf{G}^0\right)\right). \tag{18}$$

The first term is a sparsity prior and $h\left(\mathbf{G}_0\right)$ is the acyclicity constraint from [Zheng et al., 2018].

The terms $\mathbb{E}_{q_\phi(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n})}\left[-\log q_\phi\left(\mathbf{Z}^{(1:T),n}|\mathbf{X}^{(1:T),n}\right)\right], \mathbb{E}_{q_\phi(\mathbf{G})}[-\log q_\phi(\mathbf{G})]$ and $\mathbb{E}_{q_\phi(\mathbf{F})}[-\log q_\phi(\mathbf{F})]$ represent the entropies of the variational distributions and are evaluated in closed form, since their parameters are modeled as samples from Gaussian and Bernoulli distributions.

Finally, the prior term $p(F)$ is evaluated based on the assumed generative distribution mentioned in Equation (3).

A.2.2 Spatial Factors

To capture more complex spatial structures, we model the scale $\boldsymbol{\gamma}_d$ by introducing two additional parameter matrices $\mathbf{A}$ and $\mathbf{B}$. The matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and the vector $\mathbf{B} = \begin{bmatrix} e \\ g \end{bmatrix}$ together influence the covariance structure of the RBF. Specifically, the covariance matrix Σ is constructed as:

$$\Sigma = \mathbf{A}\mathbf{A}^T + \exp(\mathbf{B}), \tag{19}$$

This covariance structure enables the RBF to capture anisotropic scaling in different directions. The matrix $\mathbf{A}\mathbf{A}^T$ provides a base covariance matrix, while the exponential of $\mathbf{B}$ controls the overall scale of each spatial factor. As a result, the RBF kernel, which determines the spatial factor $\mathbf{F}$, is defined as:

$$\mathbf{F}_d^\ell = \exp\left(-\frac{1}{2}\|x_\ell - \boldsymbol{\rho}_d\|_{\Sigma^{-1}}^2\right), \tag{20}$$

where $\|x_\ell - \boldsymbol{\rho}_d\|_{\Sigma^{-1}}^2 = (x_\ell - \boldsymbol{\rho}_d)^T\Sigma^{-1}(x_\ell - \boldsymbol{\rho}_d)$ represents a Mahalanobis distance, allowing the RBF to have a more sophisticated shape that depends on the learned covariance Σ.

A.2.3 Training Details

We train the SPACY model for 700 epochs, using an 80/20 training and validation split to evaluate the validation likelihood during training.

Freezing Latent Causal Modules. To stabilize the training and ensure accurate causal discovery, we freeze the parameters of the latent SCM and causal graph, and only train the spatial factors and encoder for the first 200 epochs. This allows the spatial factor parameters to be learned without interference from incorrect causal relationships in the latent space. Once these modules are unfrozen after 200 epochs, the complete forward model and variational distribution parameters are trained jointly for the remaining 500 epochs.

This approach ensures that the inferred latent representations are sufficiently robust before learning the causal structure of the latent SCM.

A.2.4 Evaluation Details

The mean correlation coefficient (MCC) is adapted as a measure of alignment between the inferred and true latent variables, widely used in causal representation learning works [Yao et al., 2022a,b].

Here, MCC is computed as the mean of the correlation coefficients between each pair of true and inferred latent variables, providing a balanced metric that captures how well the inferred variables match the true underlying causal structure.

To evaluate the accuracy of inferred causal graphs and representations, we match the nodes of the inferred graph to the ground truth using a permutation-invariant approach. Specifically, we apply the Hungarian algorithm to find the optimal permutation of nodes that aligns the inferred graph's adjacency matrix with the ground truth, minimizing the discrepancies between them. This optimal permutation is then used to calculate both the F1 Score and the Mean Correlation Coefficient (MCC), providing consistent node alignment across these metrics.

A.3 Synthetic Experiment

This section provides more details about how we set up and run experiments using SPACY on synthetic datasets.

A.3.1 Dataset generation

We experiment with several configurations of synthetic data. The latent time series are generated using either (1) a linear structural causal model (SCM) with randomly initialized weights and additive Gaussian noise, or (2) a nonlinear SCM, where the structural equations are modeled by randomly initialized MLPs, combined with additive history-dependent conditional-spline noise.

The spatial decoder, represented by the function g_ℓ, is configured either as linear or nonlinear, depending on the experiment setting. For nonlinear scenarios, we use randomly initialized MLPs. We generate $N = 100$ samples of data, with $T = 100$ time length each and represented on a grid of size 100×100. This brings the total data dimension to $100 \times 100 \times 100 \times 100$. We vary the number of nodes ($D = 10, 20$ and 30) in each setting.

For ground-truth latents, we generate two separate sets of synthetic datasets: a linear dataset with independent Gaussian noise and a nonlinear dataset with history-dependent noise modeled using conditional splines Durkan et al. [2019]. We generate one random graph (specifically, Erdős-Rényi graphs) and treat them as ground-truth causal graphs.

Linear Latent SCM. We model the data as:

$$\mathbf{Z}_d^{(t)} = \sum_{k=0}^{\tau} \sum_{d'=1}^{D} (\mathbf{G} \circ W)^k_{d',d} \times \mathbf{Z}^d_{t-k} + \eta^t_d \tag{21}$$

with $\eta^t_d \in \mathcal{N}(0, 0.5)$. Each entry of the matrix W is drawn from $U[0.1, 0.5] \cup U[-0.5, -0.1]$

Non-linear Latent SCM. We model the data as:

$$\mathbf{Z}_d^{(t)} = f_d\left(\text{Pa}^d_G(< t), \text{Pa}^d_G(t)\right) + \eta_d^{(t)}$$

where f_d are randomly initialized multi-layer perceptions (MLPs), and the random noise $\eta_d^{(t)}$ is generated using history-conditioned quadratic spline flow functions [Durkan et al., 2019].

Spatial Factors To generate the spatial factor matrices $\mathbf{F}$, we first sample the centers $\boldsymbol{\rho}_d$ of the RBF kernels uniformly over the grid while enforcing a minimum distance constraint to ensure separation between centers. Specifically, the minimum distance between any two centers is set to be $\frac{1}{10}$ of the grid dimension. The scales $\boldsymbol{\gamma}_d$ are sampled to define the extent of each RBF kernel, drawn uniformly from the range $U[3, 6]$. With these parameters, each entry of the spatial factor matrix $\mathbf{F}^\ell_d$ is determined by the RBF kernel as follows:

$$\mathbf{F}^\ell_d = \exp\left(-\frac{||x_\ell - \boldsymbol{\rho}_d||^2}{\exp(\boldsymbol{\gamma}_d)}\right),$$

where x_ℓ denotes the spatial coordinates of the ℓ^{th} grid point, $\boldsymbol{\rho}_d$ is the center, and $\boldsymbol{\gamma}_d$ is the scale of the d^{th} latent variable.

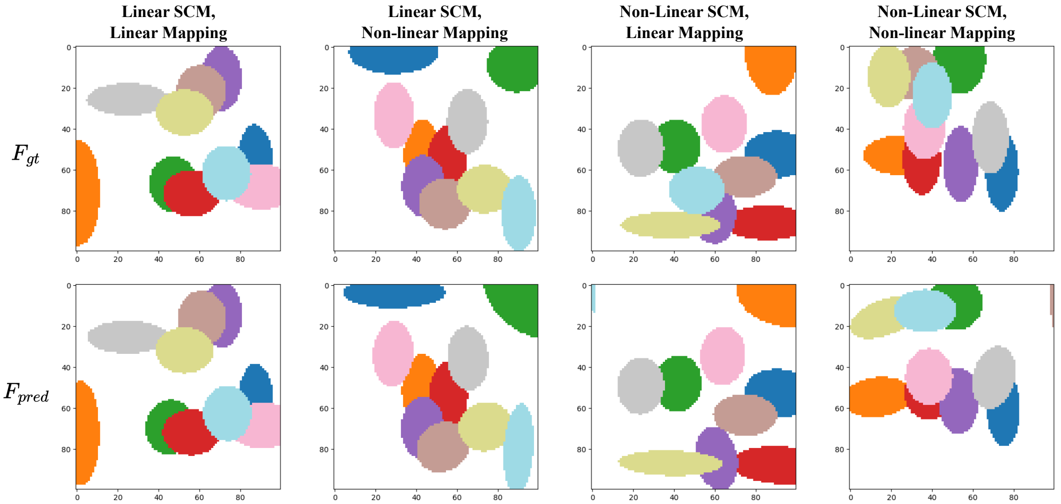


Figure 7: Visualization of the ground-truth and inferred spatial factors for different combinations of linear and nonlinear functions for SCMs and spatial mappings (top row: ground-truth, bottom row: predicted/inferred). We demonstrate the visualization when latent dimension $D = 10$

.

Spatial Mapping For the generation of $\mathbf{X}_\ell$, we pass the product of the spatial factors and the latent time series through a non-linearity g_ℓ:

$$\mathbf{X}_\ell = g_\ell\left([\mathbf{FZ}]_\ell\right) + \varepsilon_\ell, \quad \varepsilon_\ell \sim \mathcal{N}(0, \sigma_\ell^2 I) \tag{22}$$

where g_ℓ is the spatial mapping. It is implemented as a randomly initialized multi-layer perception (MLP) with the embedding of dimension 1 in the non-linear map setting, or as an identity function in the linear map setting. ε_ℓ is the grid-wise Gaussian noise added.

Baselines For all baselines, we used the default hyperparameter values. We use the Mapped-PCMCI implementation from [Tibau, 2022][3]. For Linear-Response we refer to the implementation from [Falasca et al., 2024][4]For LEAP and TDRL, we implemented the encoder and decoder using convolution neural networks as this choice best fits our data modality. For LEAP we used the CNN encoder and decoder architecture from the mass-spring system experiment [5]. For TDRL we used the CNN encoder and decoder architecture from the modified cartpole environment experiment [6].

A.3.2 Qualitative Results

Figure 7 shows a comparison between the ground truth and inferred spatial factors F on synthetic datasets. Overall, we observe that the modes from inferred spatial factors align well with the ground truth modes in terms of location and scales with minor deviations in shape. In the non-linear SCM dataset, the model shows some slight errors with at most 1 missing mode. However, the model still maintains the overall spatial representation recovery. This is also reflected by the quantitative results as performance falls slightly short for non-linear SCM.

A.3.3 Visualization details

In this section, we describe how we visualize the spatial factors for both synthetic and global temperature experiments. We represent the spatial influence of different modes on a grid by highlighting the areas where certain modes are active. The method identifies significant regions in the grid by applying a threshold based on a chosen percentile of the weights (for example, 95%). This thresholding helps to isolate areas where a mode's spatial influence is particularly strong, creating a mask that highlights these regions.

[3] Mapped-PCMCI: `https://github.com/xtibau/savar`

[4] Linear-Response: `https://github.com/FabriFalasca/Linear-Response-and-Causal-Inference`

[5] LEAP: `https://github.com/weirayao/leap`

[6] TDRL: `https://github.com/weirayao/tdrl`

These masked regions are then combined to generate a comprehensive view of how all modes influence the spatial grid. The visualization distinguishes the areas affected by different modes, allowing for easy identification of their spatial patterns and overlaps. This approach allows for a clear visual interpretation of the complex spatial structure represented by the modes, facilitating the analysis of their respective influences and interactions.

For complex spatial factors and graphs, we use a merging process that simplifies the causal global dynamics by combining nodes based on the proximity of node centers. The process identifies merging clusters in the grid by applying a threshold based on a chosen percentile of all the pair-wise distances (for example, lower 5%), and merging nodes that fall below the threshold.

A.4 Scalability

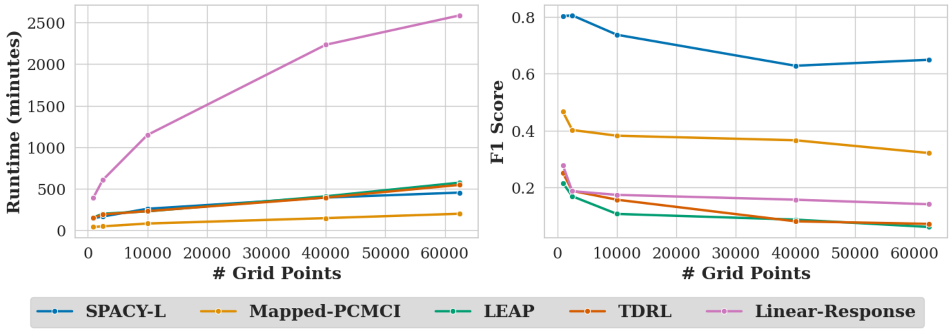
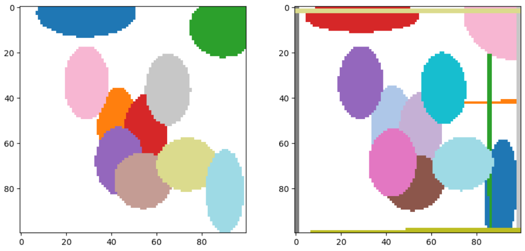


Figure 8: Comparison of runtime (in minutes) and F1 score across different grid sizes. The left plot shows how the runtime increases with grid size, while the right plot displays the corresponding F1 scores for causal discovery. Average over 5 runs reported.

We also measure the scalability of SPACY with increasing grid-size. For this experiment, we used the dataset with linear SCM and linear spatial mapping. Figure 8 demonstrates the scalability and performance of SPACY compared to the baseline methods as the grid size L increases. The runtime plot indicates that, while all methods experience an increase in runtime with increasing grid size, SPACY strikes a good balance, exhibiting moderate growth in computational time while maintaining strong causal discovery performance. Although Mapped-PCMCI is the most efficient in terms of runtime, it underperforms in causal discovery. LEAP and TDRL show similar or higher computational costs than SPACY but fail to match its performance. Linear-Response, in particular, scales poorly in terms of runtime with increasing grid size.

A.5 Ablation Study



(a) Ground truth modes

(b) Inferred modes

D^*	SPACY-L $(D = D^*)$	SPACY-L $(D = D^* + 10)$
10	0.623 ± 0.06	$\mathbf{0.642 \pm 0.07}$
20	$\mathbf{0.752 \pm 0.03}$	0.549 ± 0.03
30	$\mathbf{0.596 \pm 0.05}$	0.529 ± 0.06

(c) Causal discovery performance (F1-score)

Figure 9: Overview of the results for over-specification ablation study. (a) Visualization of the ground-truth location and scale of the spatial modes. (b) Visualization of the inferred location and scale when we over-specify the number of nodes. (c) Causal discovery performance after matching and eliminating nodes. Average over 5 seeds reported

Over-specifying D. SPACY requires specifying the number of latent variables D as a hyperparameter. In practice, the exact number of underlying factors is often unknown. We examine the effect of overspecifying D by setting it to $D^* + 10$, where D^* represents the true number of nodes used

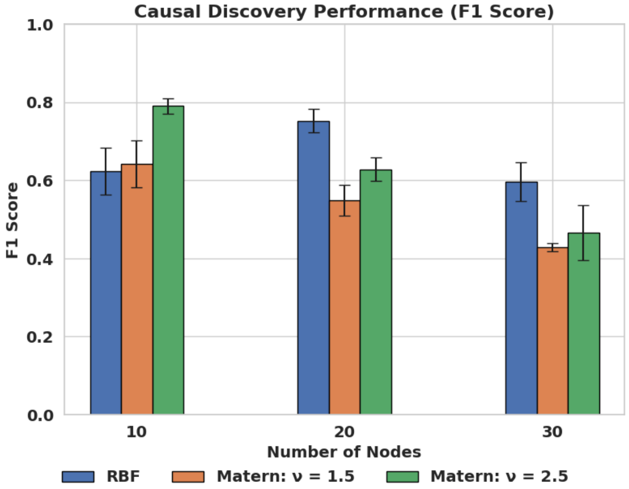


Figure 10: The causal discovery performance (F1 score) of SPACY using different kernel functions as spatial factors. Average of 5 seeds reported

to generate the data. We use the synthetic dataset with grid dimensions 100×100, linear SCM and non-linear mapping.

Figure 9 illustrates the results of our experiment. When $D^* = 10$, despite over-specifying the number of nodes, the inferred spatial modes' general locations align well with the ground truth. The presence of additional modes does not significantly detract from the accuracy of detecting the primary spatial modes. This suggests that SPACY maintains robust learning of spatial representations even when D exceeds the true number of spatial factors. This observation also holds true when comparing the causal discovery performance using the F1 score.

Different Kernels We experiment with different kernel functions in to test SPACY's robustness to the choice of kernel used in modeling the spatial factors. We use the synthetic dataset with linear SCM and nonlinear spatial mapping.

The Matérn kernel is a generalization of the RBF kernel that introduces an additional parameter ν controlling the smoothness of the function. The Matérn kernel is defined as:

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{r}{\ell} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{r}{\ell} \right),$$

where:

- $r = \|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance between points $\mathbf{x}$ and $\mathbf{x}'$,
- ℓ is the length scale,
- $\nu > 0$ controls the smoothness of the function,
- $\Gamma(\cdot)$ is the gamma function,
- $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind.

For specific values of ν, the Matérn kernel simplifies to closed-form expressions:

- **When $\nu = 1.5$:**
$$k^{1.5}_{\text{Matérn}}(r) = \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right).$$
- **When $\nu = 2.5$:**
$$k^{2.5}_{\text{Matérn}}(r) = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r}{\ell}\right).$$

By adjusting ν, the Matérn kernel can model functions with varying degrees of smoothness, providing more flexibility than the RBF kernel.

We replace the RBF kernel in SPACY with the Matérn kernel using $\nu = 1.5$ and $\nu = 2.5$. The inferred spatial modes' general locations and scales align well with the ground truth across all kernel settings (illustrated in Figure 11). This consistency demonstrates that SPACY's spatial representations are robust to the choice of kernel function.

Figure 10 presents the F1-Score and MCC for SPACY using the RBF kernel and both Matérn kernel settings. The results show that SPACY achieves similar performance with the Matérn kernels compared to the RBF kernel, indicating that the variational inference framework effectively generalizes across different kernel functions.

We also visualize the spatial factors learned by SPACY. From the visualization in 11 when $D = 10$, despite changing the kernel function type, the modes from inferred spatial factors align well with the ground truth in terms of location and scale. This suggests that SPACY is robust to the kernel choice in modeling the spatial factors.

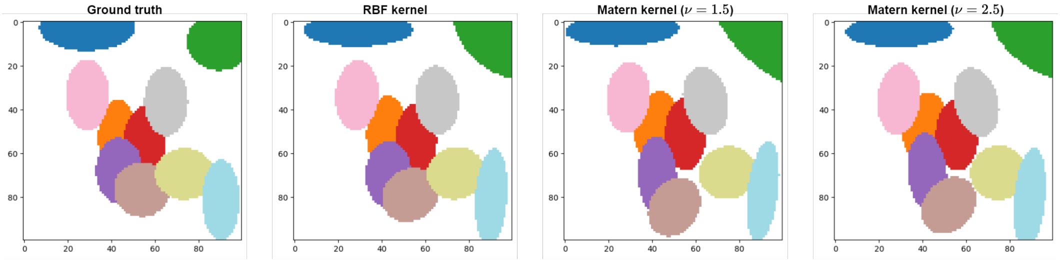


Figure 11: Overview of the visualization of the spatial factor when using different kernel functions. We compare inferred spatial factors using RBF, Matern Kernel ($\nu = 1.5$), and Matern Kernel ($\nu = 2.5$) with the ground truth spatial factors

A.6 Hyperparameter Details

In this section, we list the hyperparameters choices for SPACY in our experiments.

Dataset	**Synthetic-L** ($D = 10, 20, 30$)	**Synthetic-NL**	**Global Temperature**
Hyperparameter			
Matrix LR	10^{-3}	10^{-3}	10^{-3}
SCM LR	10^{-3}	10^{-3}	10^{-3}
Spatial Encoder LR	10^{-3}	10^{-3}	10^{-3}
Spatial Factor LR	10^{-2}	10^{-2}	10^{-2}
Spatial Decoder LR	10^{-3}	10^{-3}	10^{-3}
Batch Size	100	100	100
# Outer auglag steps	60	60	60
# Max inner auglag steps	6000	6000	6000
ξ_f, λ_f embedding dim	none	64	none
Sparsity factor α	10	10	10
Spline type	None	Quadratic	None
g_ℓ embedding dim	32	32	32

Table 1: Table showing the hyperparameters used with SPACY.

For SPACY, we used an augmented Lagrangian training procedure to enforce the acyclicity constraint in the model [Zheng et al., 2018]. We closely follow the procedure employed by Gong Wenbo and Nick [2022] for scheduling the learning rates (LRs) across different modules of our model. For the Synthetic-L, Synthetic-NL, and Global Temperature datasets, the outer augmented Lagrangian (auglag) steps are set to 60, with a maximum of 6000 inner auglag steps.

We used the rational spline flow model described in Durkan et al. [2019]. We use the quadratic or linear rational spline flow model in all our experiments, both with 8 bins. The MLPs ξ_f and λ_f have

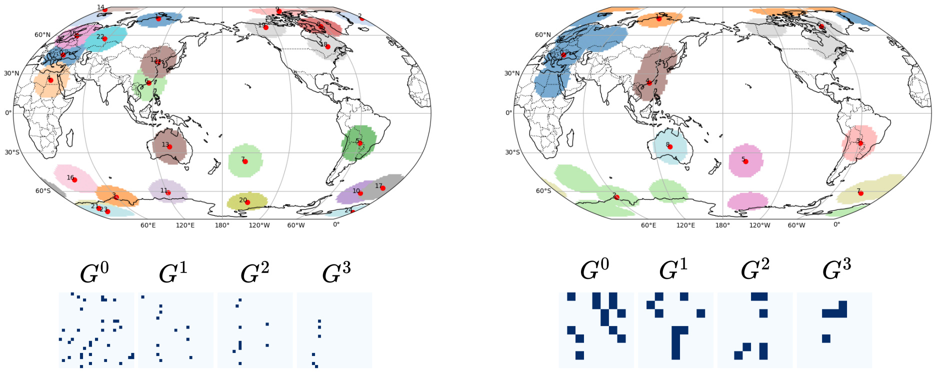


Figure 12: Visualization of (left) the learned spatial factors and causal graph (right) the learned spatial factors and causal graph after merging based on proximity and graph links.

2 hidden layers each and LeakyReLU activation functions, where e is the embedding dimension. We also use layer normalization and skip connections. Table 1 summarizes the hyperparameters used for training.

A.7 Global Temperature

The **Global Temperature Dataset** is a comprehensive, mixed real-simulated dataset encompassing monthly global temperature data spanning the years 1999 to 2001. It contains 7531 simulated samples, each with a time sequence of 24 months, covering the entire globe at a fine spatial resolution. The grid size is 145 × 192, which corresponds to a spatial division of approximately 1.24° latitude and 1.875° longitude. This spatial resolution allows the dataset to provide detailed global coverage, capturing temperature variations across diverse geographical regions. The resulting data dimensions are $7531 \times 24 \times 145 \times 192$, representing the total number of samples, the temporal sequence, and the spatial grid, respectively.

To facilitate causal analysis of complex climate phenomena beyond seasonal patterns, we applied a deseasonalization procedure. This normalization process involved computing the monthly mean for each month across all years and then adjusting the data accordingly (for example, normalizing all January data by the mean of all January values). This approach aims to remove regular seasonal influences and identify potential drivers of change within the climate system.

For our analysis, we employ the SPACY method to uncover latent representations within the data. These representations capture regions of similar weather properties and help identify causal links between these regions and weather phenomena occurring elsewhere. We use the SPACY-L variant for our experiment. Specifically, we use 25 latent variables (denoted as $D = 25$) and a maximum lag of three months ($\tau = 3$).

Figure13 demonstrates the visualization of the individual modes inferred by Varimax-PCA. Some nodes/components exhibit clear spatial patterns that are interpretable in terms of physical or location-based information. However, multiple components are more diffuse and have less interpretable locations. For instance, it may be hard to attribute physical location for node $13, 14, 19, 25$. There are also clusters of nodes that show similar spatial features, such as node $4, 6$, suggesting they capture similar underlying components.

The visualization of the modes and causal graph deduced by Mapped-PCMCI is shown in Figure 14. While the locality pattern can be observed in important regions such as Australia, Africa, and East Asia, many of the inferred modes appear diffused across the map. This suggests that the underlying spatial structure is not cleanly partitioned into distinct, interpretable modes.

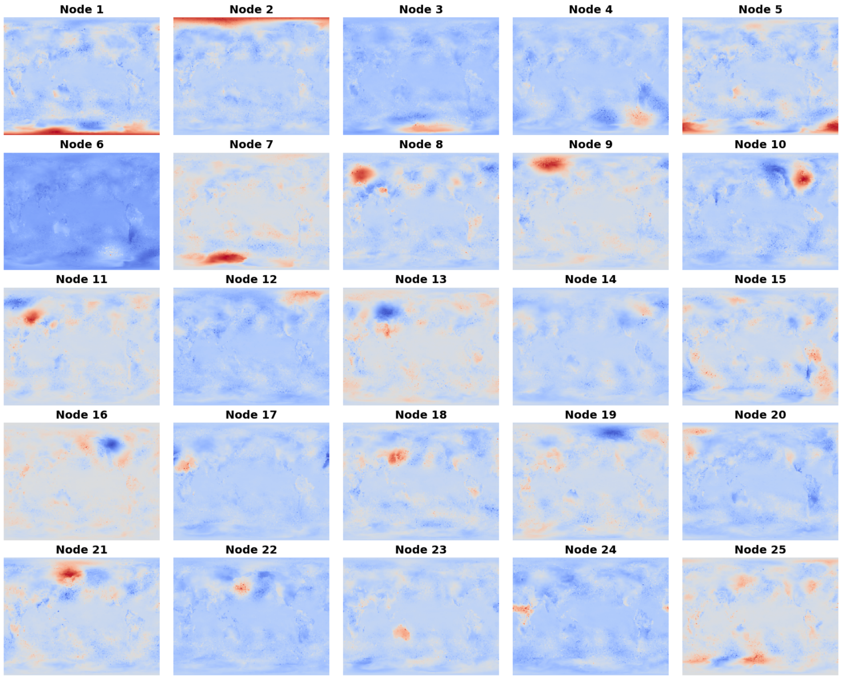


Figure 13: Visualization of the spatial nodes inferred by Varimax-PCA from the Global Temperature Dataset

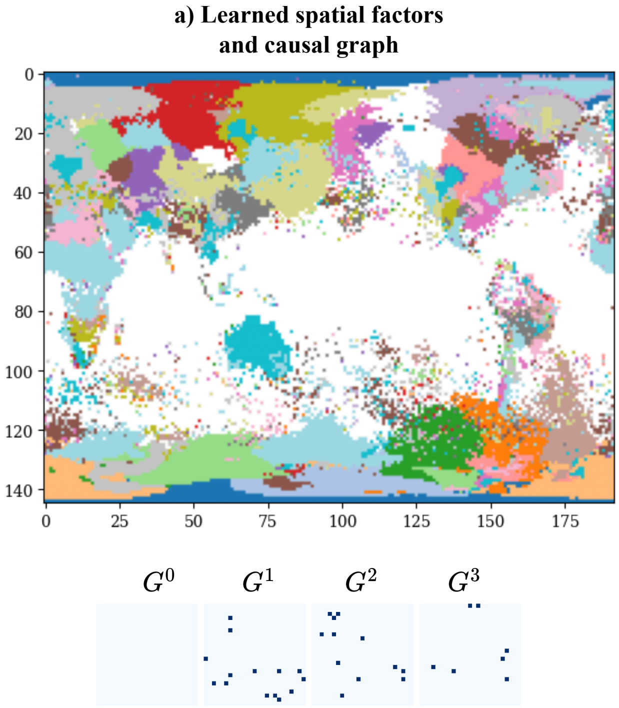


Figure 14: Visualization of the spatial factor inferred by Varimax-PCA and causal graph inferred by PCMCI$^+$, following the procedure in section A.3.3