

# Identifiable Autoregressive Variational Autoencoders for Nonlinear and Nonstationary Spatio-Temporal Blind Source Separation <sup>★</sup>

Mika Sipilä<sup>1</sup> (✉), Klaus Nordhausen<sup>2</sup>, and Sara Taskinen<sup>1</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of Jyväskylä, Finland

<sup>2</sup> Department of Mathematics and Statistics, University of Helsinki, Finland

**Abstract.** The modeling and prediction of multivariate spatio-temporal data involve numerous challenges. Dimension reduction methods can significantly simplify this process, provided that they account for the complex dependencies between variables and across time and space. Nonlinear blind source separation has emerged as a promising approach, particularly following recent advances in identifiability results. Building on these developments, we introduce the identifiable autoregressive variational autoencoder, which ensures the identifiability of latent components consisting of nonstationary autoregressive processes. The blind source separation efficacy of the proposed method is showcased through a simulation study, where it is compared against state-of-the-art methods, and the spatio-temporal prediction performance is evaluated against several competitors on air pollution and weather datasets.

**Keywords:** variational autoencoder · identifiability · multivariate spatio-temporal data · nonlinear ICA

## 1 Introduction

In multivariate spatio-temporal data, the multivariate observations  $\mathbf{x}(\mathbf{s}, t) := \mathbf{x}^t := \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^S$  are collected in various spatial locations  $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^D$  at times  $t \in \mathcal{T} \subset \mathbb{R}$ , where  $\mathcal{X}$  is the domain of  $\mathbf{x}$ ,  $\mathcal{S}$  and  $\mathcal{T}$  are spatial and temporal domains, respectively, and  $D$  is a spatial dimension. Modeling and predicting such data are highly challenging and computationally demanding due to the fact that the spatio-temporal dependency structures, as well as the dependencies between the variables, have to be accounted for. These dependencies are often modeled through  $S \times S$  dimensional covariance function  $\mathbf{C}(\mathbf{x}(\mathbf{s}, t), \mathbf{x}(\mathbf{s}', t'))$ . Modeling the covariance function is especially complicated in case of nonstationary data [20, 21], which means that the covariance function  $\mathbf{C}$  cannot be simplified to stationary form  $\mathbf{C}(\mathbf{x}(\mathbf{s}, t), \mathbf{x}(\mathbf{s}', t')) = \mathbf{C}(\|\mathbf{s} - \mathbf{s}'\|, |t - t'|)$ . Instead, for nonstationary data, the covariance function  $\mathbf{C}$  changes when spatial or temporal locations are shifted.

---

<sup>★</sup> This research was supported by the Research Council of Finland (363261, 453691) and the Vilho, Yrjö and Kalle Väisälä foundation.

Spatio-temporal data modeling can be simplified without restrictive assumptions like stationarity, by using blind source separation. In blind source separation, it is assumed that the observation  $\mathbf{x}$  is generated from the independent latent component  $\mathbf{z}(\mathbf{s}, t) := \mathbf{z}^t := \mathbf{z} \in \mathbb{R}^P$  through a mixing function  $\mathbf{f}$  as

$$\mathbf{x} = \mathbf{f}(\mathbf{z}). \quad (1)$$

Once the latent components are successfully recovered, they can be modeled independently due to their assumed statistical independence. The dependencies among the components of the observed variable vector  $\mathbf{x}$  are therefore presumed to arise exclusively from the mixing function  $\mathbf{f}$ . Blind source separation (BSS) aims to recover the latent components by estimating the mixing and unmixing functions from the observed data.

While most traditional BSS methods, such as spatio-temporal BSS (STBSS) [19], are limited only to linear mixing function  $\mathbf{f}(\mathbf{z}) = \mathbf{A}\mathbf{z}$ , where  $\mathbf{A}$  is a  $S \times P$  matrix, nonlinear BSS variants have also been recently developed. In the nonlinear case however stronger assumptions are needed for identifiability. One such approach for nonlinear BSS assumes, for example, structural sparsity [16]. Other recent developments are mostly for time series, and they solve nonlinear BSS by exploiting either stationary autocorrelation structure or nonstationary variances. For these methods, see [9] and the references therein.

In particular, [13] introduced identifiable variational autoencoder (iVAE) for nonlinear and nonstationary temporal BSS. Later, iVAE have been extended to nonstationary spatial data in [23] and to nonstationary spatio-temporal data in [22]. However, all previous iVAE methods are identifiable only if the latent components possess nonstationary variance, and they do not incorporate previous observations in time in the model. Instead, the previous methods model the nonstationary variance only based on the spatial and temporal location of the observations.

In this paper, we assume that each latent component  $z_i$ , for  $i = 1, \dots, P$ , is generated by a nonstationary autoregressive process defined as follows:

$$z_i(\mathbf{s}, t) = \mu_i(\mathbf{s}, t) + \sum_{r=1}^R \gamma_{i,r}(\mathbf{s}, t) \left( z_i(\mathbf{s}, t-r) - \mu_i(\mathbf{s}, t-r) \right) + \omega_i(\mathbf{s}, t), \quad (2)$$

where  $\mu_i$  is a nonstationary trend function,  $R$  is the autoregressive order,  $\gamma_{i,r}$  is a time- and location-dependent autoregressive coefficient function, and  $\omega_i$  is the innovation term, also varying over location  $\mathbf{s}$  and time  $t$ . A similar model to (2) is considered in [5] in the context of stationary subspace analysis for time series.

We propose an identifiable autoregressive variational autoencoder (iVAEar) which extends the identifiability also to nonstationary autoregressive coefficients. In Section 2, we discuss iVAEar’s model assumptions and identifiability conditions, and in Section 3 we introduce the iVAEar method to estimate the model. We demonstrate iVAEar’s latent component estimation performance through comprehensive simulation studies in Section 4, and illustrate its multivariate spatio-temporal forecasting potential in Section 5. Finally, the paper is con-

cluded in Section 6. All proofs are given in the supplement<sup>3</sup> together with some additional material.

## 2 Autoregressive latent component model and identifiability

In this section, we introduce an autoregressive latent component model and its identifiability results under nonstationary data. We begin by establishing general identifiability conditions for autoregressive latent component models in Definition 1 and Theorems 1 and 2. We then examine specific cases that yield stronger identifiability results: first, we provide general results for the case where  $R = 0$  (Proposition 1), followed by results for Gaussian latent components and Gaussian autoregressive latent components (Propositions 2 and 3, respectively). Note, that although we focus on spatio-temporal data in the paper, all the results and estimation methods apply also for time series data by dropping the spatial location out of the equations.

In original iVAE [13], the main assumption leading to identifiability of the latent component model is that an additional variable  $\mathbf{u} \in \mathcal{U}$ , where  $\mathcal{U}$  is the domain of  $\mathbf{u}$ , is observed so that the latent components  $\mathbf{z}$  have a conditional distribution  $p(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^P p(z_i|\mathbf{u})$ . In all previous iVAE methods,  $\mathbf{u}$  has included information on temporal, spatial, or spatio-temporal location of the observation. In iVAEar, we assume that in addition to spatio-temporal location, we also have the previous  $R$  observations in time,  $\{\mathbf{x}(\mathbf{s}, t-1), \dots, \mathbf{x}(\mathbf{s}, t-R)\} := \mathbf{x}^-$ , as the additional data. The autoregressive assumption leads to the following generative deep latent variable model:

$$p(\mathbf{x}, \mathbf{z}|\mathbf{x}^-; \mathbf{u}) = p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}|\mathbf{z}^-; \mathbf{u}), \quad (3)$$

where  $\mathbf{z}^- = \{\mathbf{z}(\mathbf{s}, t-1), \dots, \mathbf{z}(\mathbf{s}, t-R)\}$  is the set of previous latent components in time. Following [13], the distribution  $p(\mathbf{x}|\mathbf{z})$  is defined as

$$p(\mathbf{x}|\mathbf{z}) = p_\epsilon(\mathbf{x} - \mathbf{f}(\mathbf{z})), \quad (4)$$

meaning that  $\mathbf{x}$  decomposes into  $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \epsilon$ , where  $\epsilon$  is an independent noise vector. In non-noisy nonlinear BSS (1),  $p_\epsilon$  can be modeled with a zero mean Gaussian distribution with infinitesimal variance. Further, it is assumed that the conditional latent distribution is part of the exponential family:

$$p_{T, \lambda}(\mathbf{z}|\mathbf{z}^-, \mathbf{u}) = \prod_{i=1}^P \frac{Q_i(z_i, z_i^-)}{Z_i(\mathbf{u})} \exp \left[ \sum_{j=1}^k T_{i,j}(z_i, z_i^-) \lambda_{i,j}(\mathbf{u}) \right], \quad (5)$$

where  $Q_i(z_i, z_i^-)$  is a base measure,  $Z_i(\mathbf{u})$  is a normalizing constant,  $\mathbf{T}_i(z_i, z_i^-) = (T_{i,1}(z_i, z_i^-), \dots, T_{i,k}(z_i, z_i^-))^\top$  contains sufficient statistics, and  $\lambda_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}),$

<sup>3</sup> <https://github.com/mikasip/iVAEar>

$\dots, \lambda_{i,k}(\mathbf{u})^\top$  contains the parameters depending on  $\mathbf{u}$ . The dimension  $k$  of each sufficient statistic  $\mathbf{T}_i(z_i, z_i^-)$  and  $\boldsymbol{\lambda}_i(\mathbf{u})$  is assumed to be fixed. The formulation (5) reduces to general exponential family formula if the autoregressive order  $R = 0$ . The exponential family form in (5) includes variables  $z_i$  generated through AR processes with any exponential family innovations if the location  $\mu_i$  and AR coefficients  $\gamma_i^r$  are constant. Some AR processes, such as processes with Gaussian or exponential distributed innovations, fall in this form even with non-stationary location and AR coefficients. The properties of Gaussian AR processes are discussed in more detail later in this section.

Assuming the generative model defined by the equations (3)-(5), and non-linear BSS (1) problem, it is of interest to identify the latent components  $\mathbf{z}$  as well as possible to obtain information about the true generative process behind the observed data. Hence, we next define two identifiability classes that can be obtained with sufficient assumptions. In following, we use the notation  $\{\mathbf{f}^{-1}(\mathbf{x}(s, t-1)), \dots, \mathbf{f}^{-1}(\mathbf{x}(s, t-R))\} := \mathbf{f}^{-1}(\mathbf{x}^-)$  to denote the unmixing function applied to previous  $R$  observations in time individually.

**Definition 1.** Consider the real parameter set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  and the estimated one  $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}})$  of mixing functions, sufficient statistics and natural parameters such that  $p_{\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}}(\mathbf{x}|\mathbf{x}^-, \mathbf{u}) = p_{\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}}(\mathbf{x}|\mathbf{x}^-, \mathbf{u})$  for all  $\mathbf{x}, \mathbf{x}^- \in \mathcal{X}$  and  $\mathbf{u} \in \mathcal{U}$ . If there exists an invertible  $Pk \times Pk$  matrix  $\mathbf{A}$  and a vector  $\mathbf{c}$  so that

$$\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x}), \tilde{\mathbf{f}}^{-1}(\mathbf{x}^-)) = \mathbf{A}\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x}), \mathbf{f}^{-1}(\mathbf{x}^-)) + \mathbf{c} \quad (6)$$

for all  $\mathbf{x}, \mathbf{x}^- \in \mathcal{X}$ , the set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  is identifiable up to an affine transformation. If  $\mathbf{A}$  is a block permutation matrix, then the set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  is identifiable up to block-affine transformation.

The block-affine identifiability is a stronger result, and often desirable. Block-affine identifiability is closely related to permutation and signed scale indeterminacy of  $\mathbf{z}$  of linear BSS. To build intuition about how block-affine identifiability relates to the identifiability of the latent components  $\mathbf{z}$ , we next provide sufficient conditions on the sufficient statistics  $\mathbf{T}$  in the case  $R = 0$  that ensure identifiability of  $\mathbf{z}$  up to permutation and component-wise nonlinearity.

**Proposition 1.** Assume that the set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  is identifiable up to block-affine transformation and that the autoregressive order  $R = 0$ . Further assume:

- (i) A non-noisy BSS model (1), i.e. that  $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$ .
- (ii) There is a function  $\tilde{g}_i : \mathbb{R}^k \rightarrow \mathbb{R}$  for all  $i = 1, \dots, P$  such that  $\tilde{g}_i(\tilde{\mathbf{T}}_i(\tilde{z}_i)) = a_i \tilde{z}_i$ , where  $a_i \neq 0$ .

Then we have that  $\tilde{\mathbf{f}}^{-1}(\mathbf{x}) = \tilde{\mathbf{z}} = \mathbf{P}(g_1(z_1), \dots, g_P(z_P))^\top$ , where  $\mathbf{P}$  is a  $P \times P$  permutation matrix and  $g_1, \dots, g_P$  are component-wise nonlinearities.

Assumption (ii) of Proposition 1 holds for most of common exponential family distributions such as Gaussian, beta, gamma, Pareto, Poisson and exponential distributions, which have sufficient statistic of the form  $T(x) = x$  or

$T(x) = \log(x)$ . If we have a noisy nonlinear BSS instead of non-noisy, there is an additional noise indeterminacy for each component. For the case  $R > 0$  with autoregressive dependencies, similar results can be derived so that the component-wise nonlinearities would depend also on their previous values, i.e., that  $\tilde{\mathbf{f}}^{-1}(\mathbf{x}) = \tilde{\mathbf{z}} = \mathbf{P}(g_1(z_1, z_1^-), \dots, g_P(z_P, z_P^-))^T$ . However, for specific autoregressive models, stronger identifiability results can be obtained. In particular, later in this section we demonstrate that for Gaussian autoregressive latent processes, the latent components can be identified up to permutation, location and scale transformations.

Next, we introduce two theorems that give sufficient conditions to achieve affine or block-affine identifiability. The main identifiability theorem is as follows:

**Theorem 1.** *When the data are generated according the generative model in (3)-(5), and the following holds:*

- (i) *The set  $\{\mathbf{x} \in \mathcal{X} | \rho_\epsilon(\mathbf{x}) = 0\}$  has measure zero, where  $\mathcal{X}$  is a domain of  $\mathbf{x}$  and  $\rho_\epsilon$  is a characteristic function of the density  $p_\epsilon$  in (4).*
- (ii) *The mixing function  $\mathbf{f}$  in (4) is injective.*
- (iii) *The sufficient statistics  $T_{i,j}$  in (5) are differentiable with respect to  $z_i$  almost everywhere, and the functions  $T_{i,1}, \dots, T_{i,k}$  are linearly independent on any subset of  $\mathcal{X}$  with positive measure.*
- (iv) *There exist  $Pk + 1$  distinct points  $\mathbf{u}_0, \dots, \mathbf{u}_{Pk}$  so that the  $Pk \times Pk$  matrix  $\mathbf{L} = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), \dots, \boldsymbol{\lambda}(\mathbf{u}_{Pk}) - \boldsymbol{\lambda}(\mathbf{u}_0))$  is invertible.*

*Then, the set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  is identifiable up to affine transformation.*

While the assumptions (i)-(iii) are not very restrictive, the assumption (iv) is crucial to understand as it restricts the identifiability only to cases where the parameters  $\boldsymbol{\lambda}(\mathbf{u})$  vary enough when  $\mathbf{u}$  changes. Because of this assumption, the latent components are identifiable only when the exponential family parameters are nonstationary.

Although identifiability up to a affine transformation might already be useful, in most cases it is desirable to achieve block-affine identifiability. The next theorem gives sufficient conditions for such identifiability.

**Theorem 2.** *Assume that the assumptions of Theorem 1 hold. Further assume:*

- (i) *The dimension of sufficient statistics is  $k \geq 2$ .*
- (ii) *The sufficient statistics  $T_{i,j}$  are twice differentiable with respect to  $z_i$ .*
- (iii) *The mixing function  $\mathbf{f}$  has all second-order cross derivatives.*

*Then, the set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  is identifiable up to block-affine transformation.*

Theorem 2, combined with the additional conditions of Proposition 1, essentially guarantees that latent components can be recovered up to permutation and component-wise nonlinearity. For example, Gaussian distributed latent components with unknown nonstationary mean and variance, with sufficient statistics  $\mathbf{T}_i(z_i) = (z_i, z_i^2)^T$ , fall within Theorem 2. In fact, we can show that for such Gaussian data the identifiability can be further reduced to permutation, scale and location shift, which is in par with identifiability results of linear BSS:

**Proposition 2.** *Assume that the assumptions of Theorem 2 hold and that the data are generated through BSS model (1). Further, assume that the latent components  $z_i$  and the respective estimates  $\tilde{z}_i$  are Gaussian, meaning that  $\mathbf{T}_i(z_i) = (z_i, z_i^2)^\top$  and  $\tilde{\mathbf{T}}_i(\tilde{z}_i) = (\tilde{z}_i, \tilde{z}_i^2)^\top$ . Then we have that  $\tilde{\mathbf{z}} = \mathbf{P}\mathbf{\Lambda}\mathbf{z} + \mathbf{d}$ , where  $\mathbf{P}$  is a permutation matrix and  $\mathbf{\Lambda}$  is a diagonal matrix with non-zero diagonal elements.*

Since our main focus in this paper is on Gaussian autoregressive latent components which always has  $k \geq 2$ , we refer the reader to [13] for  $k = 1$  case, where sufficient conditions are provided for exponential family with  $R = 0$ . When the autoregressive process (3) is assumed for the latent components with Gaussian innovations, we have the following distribution:

$$p(\mathbf{z}|\mathbf{z}^-, \mathbf{u}^t, \dots, \mathbf{u}^{t-R}) = \prod_{i=1}^P \frac{1}{2\pi\sigma_i(\mathbf{u}^t)} \exp \left[ -\frac{\left( z_i - \mu_i(\mathbf{u}^t) - \sum_{r=1}^R (\gamma_{i,r}(\mathbf{u}^t) z_i^{t-r} - \mu_i(\mathbf{u}^{t-r})) \right)^2}{2\sigma^2(\mathbf{u}^t)} \right], \quad (7)$$

where  $\mathbf{u}^t$  denotes the auxiliary variable for the observation  $\mathbf{x}^t$ .

**Proposition 3.** *Assume that the assumptions of Theorem 2 hold and that the data are generated through BSS model (1). Further assume that the latent components  $z_i$  and the respective estimates  $\tilde{z}_i$  are generated through the Gaussian AR process (2) with  $R \geq 1$ . Then we have that  $\tilde{\mathbf{z}} = \mathbf{P}\mathbf{\Lambda}\mathbf{z} + \mathbf{d}$ , where  $\mathbf{P}$  is a permutation matrix,  $\mathbf{\Lambda}$  is a diagonal matrix with non-zero diagonal elements and  $\mathbf{d}$  is a constant vector.*

Proposition 3 gives the main identifiability conditions for the Gaussian autoregressive latent components. In practice, the conditions on the mixing function are not very restrictive. However, condition (iv) of Theorem 1 requires sufficient nonstationarity either in the AR coefficients  $\gamma_{i,r}$  or in the variance  $\sigma_i$ . In Section 3, we introduce an estimation method for estimating the generative model defined by equations (3)-(5).

### 3 Autoregressive identifiable variational autoencoder

The iVAEar method is an autoregressive extension of spatio-temporal iVAE, introduced in [22]. It consists of an encoder  $\mathbf{g}(\mathbf{x}, \mathbf{u})$ , a decoder  $\mathbf{h}(\mathbf{x})$  and an auxiliary function  $\mathbf{w}(\mathbf{u})$ . As the true AR order  $R$  is in general unknown, we use  $W$  to refer to the AR order used in the iVAEar method. The method takes as an input the current observations  $\mathbf{x}$  and their auxiliary data  $\mathbf{u}$ , and the  $W$  previous observations in time and their auxiliary data  $(\mathbf{x}^{t-r}, \mathbf{u}^{t-r})$ ,  $r = 1, \dots, W$ .

The encoder aims to estimate the unmixing function  $\mathbf{q}$ . It maps the observation and auxiliary data pair  $(\mathbf{x}, \mathbf{u})$  into the mean vector  $\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} \in \mathbb{R}^P$  and the variance vector  $\boldsymbol{\sigma}_{\mathbf{z}|\mathbf{x}} \in \mathbb{R}^P$ . For the current observation  $\mathbf{x}$ , the encoder's output is used for reparametrization trick [14] to obtain a new latent representation  $\mathbf{z}'$ . The decoder aims then to estimate the mixing function  $\mathbf{f}$  by trying to construct

the original input  $\mathbf{x}$  from  $\mathbf{z}'$ . For the previous observations  $\mathbf{x}^{t-r}$ , the encoder is used to obtain the corresponding latent component estimates  $\mathbf{u}_{\mathbf{z}|\mathbf{x},\mathbf{u}}^{t-r}$ , which are provided by the mean function  $\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x},\mathbf{u}}(\mathbf{x}^{t-r}, \mathbf{u}^{t-r})$ . These are used to calculate the mean of the Gaussian latent distribution (7).

The auxiliary function  $\mathbf{w}$  aims to estimate the function  $\boldsymbol{\lambda}$  by mapping the auxiliary data  $\mathbf{u}$  into parameters  $\boldsymbol{\mu}_{\mathbf{z}|\mathbf{u}}, \boldsymbol{\sigma}_{\mathbf{z}|\mathbf{u}}, \gamma_{\mathbf{z}|\mathbf{u}}^1, \dots, \gamma_{\mathbf{z}|\mathbf{u}}^W$ , that estimate the true parameters of the autoregressive Gaussian distribution (7). In addition, the auxiliary function is used to obtain the mean estimates  $\boldsymbol{\mu}_{\mathbf{z}|\mathbf{u}}^{t-r}$  based on the auxiliary data  $\mathbf{u}^{t-r}$  of the previous observations.

The encoder, the decoder and the auxiliary function are modeled using deep neural networks with parameters  $\boldsymbol{\theta}_g, \boldsymbol{\theta}_h, \boldsymbol{\theta}_w$ , that refer to the weights and biases of encoder, decoder and auxiliary function, respectively. The parameters  $\boldsymbol{\theta} = (\boldsymbol{\theta}_g, \boldsymbol{\theta}_h, \boldsymbol{\theta}_w)^\top$  of the neural networks are optimized by minimizing the lower bound of the data log-likelihood, or evidence lower bound (ELBO):

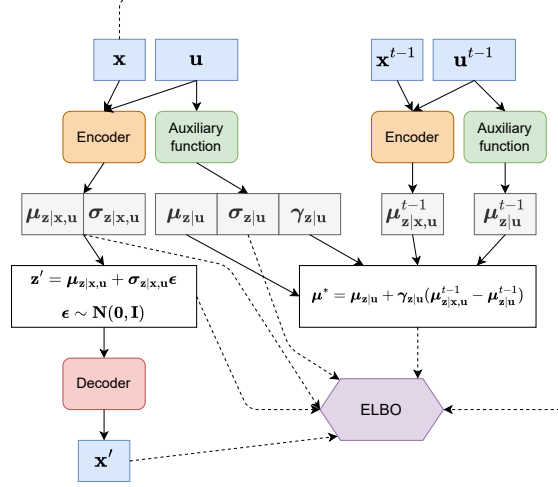
$$\text{ELBO} = E_{q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x},\mathbf{u})}(\log p_{\boldsymbol{\theta}_h}(\mathbf{x}|\mathbf{z}) + \log p_{\boldsymbol{\theta}_w}(\mathbf{z}|\mathbf{z}^-, \mathbf{u}) - \log q_{\boldsymbol{\theta}_g}(\mathbf{z}|\mathbf{x}, \mathbf{u})), \quad (8)$$

where the first part,  $p_{\boldsymbol{\theta}_h}(\mathbf{x}|\mathbf{z})$ , controls the reconstruction accuracy and the second part,  $\log p_{\boldsymbol{\theta}_w}(\mathbf{z}|\mathbf{z}^-, \mathbf{u}) - \log q_{\boldsymbol{\theta}_g}(\mathbf{z}|\mathbf{x}, \mathbf{u})$ , is the Kullback-Leibler divergence, which tries to keep the variational distribution  $\log q_{\boldsymbol{\theta}_g}(\mathbf{z}|\mathbf{x}, \mathbf{u})$  close to the prior distribution  $\log p_{\boldsymbol{\theta}_w}(\mathbf{z}|\mathbf{z}^-, \mathbf{u})$ . Because Gaussian autoregressive latent data is assumed (7), the distributions  $p_{\boldsymbol{\theta}_w}$ ,  $q_{\boldsymbol{\theta}_g}$  and  $p_{\boldsymbol{\theta}_h}$  are assumed Gaussian, ensuring that the estimated components follow the same distribution (7). Specifically, we set  $p_{\boldsymbol{\theta}_w} = N(\mathbf{z}|\boldsymbol{\mu}^*, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}|\mathbf{u}}))$ , where  $\boldsymbol{\mu}^* = \boldsymbol{\mu}_{\mathbf{z}|\mathbf{u}} + \sum_{i=1}^R \gamma_{\mathbf{z}|\mathbf{u}}^{t-r}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x},\mathbf{u}}^{t-R} - \boldsymbol{\mu}_{\mathbf{z}|\mathbf{u}}^{t-R})$ ,  $q_{\boldsymbol{\theta}_g} = N(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x},\mathbf{u}}, \text{diag}(\boldsymbol{\sigma}_{\mathbf{z}|\mathbf{x},\mathbf{u}}))$  and  $p_{\boldsymbol{\theta}_h} = N(\mathbf{x}|\mathbf{x}', \beta \mathbf{I})$ , where  $\beta > 0$  is a small constant that represents the variance of (4). By decreasing  $\beta$ , the weight of the reconstruction loss is increased in the loss function similarly as in  $\beta$ -VAE [8]. The whole iVAEar framework is illustrated in  $R = 1$  case in Figure 1. For more details of iVAE framework, see [13, 22, 23].

For iVAEar, we construct the auxiliary data following [22] based on either spatial and temporal segmentation or spatial and temporal radial basis functions. In segmentation based algorithm, the spatial domain is divided into equally sized two dimensional square segments, and the temporal domain into equally sized one dimensional segments. The auxiliary variable then gives the spatial and temporal segments corresponding to the observation. In radial basis function based algorithm, multiple spatial and temporal node points are selected from spatial and temporal domains. The auxiliary variable, i.e. radial basis functions, are then constructed based on distance between the location of the observation and each of the node points. Segmentation based iVAEar is denoted iVAEar\_s and radial basis function based iVAEar is denoted iVAEar\_r in the rest of the paper. For further details of constructing the auxiliary data, see [22].

If the underlying latent components satisfy the assumptions of Theorem 1 or Theorem 2, then we have the following consistency result.

**Theorem 3.** *Assume that the Theorem 1 or Theorem 2 hold. Further assume that the family of the variational distributions  $q_{\boldsymbol{\theta}_g}(\mathbf{z}|\mathbf{x}, \mathbf{x}^-, \mathbf{u})$  contains the dis-*



**Fig. 1.** Schematic presentation of iVAEar method in  $R = 1$  case.

tribution  $p_{\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}}(\mathbf{z}|\mathbf{x}, \mathbf{x}^-, \mathbf{u})$ . Then iVAEar learns the true set  $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$  up to the identifiability classes given by Theorems 1 and 2 in the limit of infinite data.

In AR Gaussian latent data case, when also  $q\theta_g$  is Gaussian, then by Proposition 3, iVAEar estimates the true latents  $\mathbf{z}$  up to permutation, signed scale and location shift in the limit of infinite data.

The auxiliary function of iVAEar enables the method to be used for spatio-temporal interpolation or forecasting purposes. Particularly, iVAEar\_r method provides smooth estimates of the spatio-temporal functions  $\mu_i(\mathbf{u}^t)$ ,  $\gamma_{i,r}(\mathbf{u}^t)$  and  $\sigma_i(\mathbf{u}^t)$ ,  $i = 1, \dots, P$ ,  $r = 1, \dots, R$ . These can be used to predict the latent components to new spatio-temporal locations, after which the predictions can be transformed into observation space by using the decoder of the trained iVAEar. The prediction capabilities of iVAEar are illustrated later in Section 5.

## 4 Simulations

The simulations of this paper are two-fold; in the first part in Section 4.1, various simulations are performed under the assumption that the true AR order  $R$  is known. In the second part in Section 4.2, the performance of iVAEar\_r is studied under the assumption that the true autoregressive order  $R$  is unknown. The implementations of all iVAE and iVAEar variants together with the code to simulate the data in all considered settings and to reproduce the case study of Section 5, are available in GitHub<sup>4,5</sup>.

<sup>4</sup> <https://github.com/mikasip/NonlinearBSS>

<sup>5</sup> <https://github.com/mikasip/iVAEar>



#### 4.1 Main simulations

In this section, simulation studies are used to compare the performances of iVAEar\_r and iVAEar\_s against segmentation and radial basis function based spatio-temporal iVAE methods, iVAEs and iVAEr, respectively, as proposed in [22], STBSS, and symmetric FastICA (FICA) with hyperbolic tangent non-linearity [10]. In simulations, we generate the latent spatio-temporal fields  $\mathbf{z}$  and a mixing function  $\mathbf{f}$ . We are particularly interested in performance in settings, where the variance and/or the AR coefficients of the latent fields  $\mathbf{z}$  are varying in space and in time. Hence, we select one setting with nonstationary AR1 coefficient, one with nonstationary variance and one with both AR1 coefficient and variance nonstationary. In addition, each of the settings is considered with and without nonstationary spatio-temporal trend function. Next, we give all the simulation details and explain how  $\mathbf{z}$  and  $\mathbf{f}$  are generated.

In all simulations, we set the observed dimension  $S = 6$  and the latent dimension  $P = 6$ . The number of spatial locations is  $n_s = 100$  and the number of time points is  $n_t = 500$ . The spatial locations  $\mathbf{s}_1, \dots, \mathbf{s}_{n_s}$  are generated uniformly in the domain  $[0, 1] \times [0, 1]$ , and the observations over time are set at times  $t = 1, \dots, n_t$ . The latent spatio-temporal fields are generated using the following vector AR process. Assume the spatial field at time  $t$  to be  $\boldsymbol{\delta}(t) = (\delta(\mathbf{s}_1, t), \dots, \delta(\mathbf{s}_{n_s}, t))$ . By using the vector AR process we have then

$$\boldsymbol{\delta}(t) = \sum_{r=1}^R \rho_r \mathbf{K}_r(t) \boldsymbol{\delta}(t-r) + \boldsymbol{\epsilon}_{\boldsymbol{\delta}}(t), \quad (9)$$

where  $r = 1, \dots, R$  is the order of AR process,  $\rho_r$  is the baseline AR coefficient for the  $r$ th order,  $\mathbf{K}_r(t)$  is a spatial kernel matrix for time  $t$ , which determines the temporal correlation with spatial locations, and  $\boldsymbol{\epsilon}_{\boldsymbol{\delta}}(t)$  is a  $n_s$ -dimensional Gaussian noise vector with spatial covariance function  $C(\epsilon_{\boldsymbol{\delta}}(\mathbf{s}, t), \epsilon_{\boldsymbol{\delta}}(\mathbf{s}', t))$ ,  $\mathbf{s}, \mathbf{s}' \in \{\mathbf{s}_1, \dots, \mathbf{s}_{n_s}\}$ . If the kernel matrices  $\mathbf{K}_r(t)$  are diagonal, the generated data have separable spatio-temporal covariance function, i.e., data do not have any spatio-temporal interaction. For the simulations, we set  $R = 1$ . As spatial covariance function for time  $t$  we use variance modulated Matern covariance function

$$C(\epsilon_{\boldsymbol{\delta}}(\mathbf{s}, \mathbf{s}', t)) = \sigma(\mathbf{s}, t) \sigma(\mathbf{s}', t) \frac{1}{2^{\nu-1} \Gamma(\nu)} \left( \frac{\|\mathbf{s} - \mathbf{s}'\|}{\phi} \right)^{\nu} K_{\nu} \left( \frac{\|\mathbf{s} - \mathbf{s}'\|}{\phi} \right), \quad (10)$$

where  $\sigma$  modulates the variance based on the time and spatial location,  $K_{\nu}$  is a modified Bessel function of second kind, and  $\phi$  and  $\nu$  are range and shape parameters, respectively. The common Matern parameters for all settings are provided in the supplementary material. In the simulations, we consider data with and without trend. The spatio-temporal trend is generated as composition of cyclical and liner trends as follows:

$$\mu(s_1, s_2, t) = \theta_{s_1} s_1 + \theta_{s_2} s_2 + \theta_t t + \alpha \sin(\omega_{s_1} s_1 + \omega_{s_2} s_2 + \omega_t t + \omega_c). \quad (11)$$

The parameters are generated so that  $\theta_{s_1}, \theta_{s_2} \sim \text{Unif}(-3, 3)$ ,  $\theta_t \sim \text{Unif}(-0.01, 0.01)$ ,  $\omega_{s_1}, \omega_{s_2} \sim \text{Unif}(0.2, 4)$ ,  $\omega_t \sim \text{Unif}(0.01, 0.1)$ ,  $\omega_c \sim \text{Unif}(0, 2\pi)$  and  $\alpha \sim \text{Unif}(-2, 2)$ .

**Setting 1.** The latent fields have constant variance  $\sigma(\mathbf{s}, t) = 1$  and varying AR1 coefficients over space and time. The kernel matrix  $\mathbf{K}_1(t)$  is a diagonal matrix with AR1 coefficients  $\gamma(\mathbf{s}_1, t), \dots, \gamma(\mathbf{s}_{n_s}, t)$  in the diagonal for each spatial location  $\mathbf{s}_1, \dots, \mathbf{s}_{n_s}$ . The parameters  $\gamma(\mathbf{s}, t)$  are generated as

$$\gamma(\mathbf{s}, t) = \cos\left(\frac{2\pi tb}{n_t} - c(\mathbf{s})\right), \quad (12)$$

where  $b$  is a scale parameter and  $c(\mathbf{s})$  is a shift parameter. To obtain variability in space, we generate the shift parameters  $c(\mathbf{s})$  from the Gaussian distribution  $N(0, 0.3)$  with Matern spatial covariance function with parameters  $\phi_c, \nu_c$ . The Matern parameters for shift are  $\phi_{c_1}, \nu_{c_1} = (0.25, 5)$ ,  $\phi_{c_2}, \nu_{c_2} = (0.15, 2)$ ,  $\phi_{c_3}, \nu_{c_3} = (0.1, 3)$ ,  $\phi_{c_4}, \nu_{c_4} = (0.3, 4)$ ,  $\phi_{c_5}, \nu_{c_5} = (0.2, 1)$  for the latent components  $z_1, \dots, z_5$ . The scale parameters  $b$  are generated from  $\text{Unif}(1, 10)$  and the baseline AR1 parameters  $\rho_r$  are generated from  $\text{Unif}(0.6, 0.99)$  for each latent component.

**Setting 2.** The zero-mean latent fields  $z_i^*$  are generated as in Setting 1. Then, the final latent fields are obtained as  $z_i(\mathbf{s}, t) = z_i^*(\mathbf{s}, t) + \mu_i(\mathbf{s}, t)$ , where  $\mu_i(\mathbf{s}, t)$  is generated as in (11).

**Setting 3.** The latent fields have constant AR1 coefficients and varying variance over space and time. The kernel matrix is  $\mathbf{K}_1(t)$  is identity matrix for all  $t$ . The spatial domain is divided randomly into 5 clusters and the time domain into 10 segments providing 50 spatio-temporal segments  $S_1, \dots, S_{50}$ , each having their own standard deviation  $\sigma_1, \dots, \sigma_{50}$ . The function  $\sigma$  is then  $\sigma(\mathbf{s}, t) = \sum_{k=1}^{50} \mathbb{1}((\mathbf{s}, t) \in S_k) \sigma_k$ , where  $\mathbb{1}$  is an indicator function giving 1, if the location  $(\mathbf{s}, t)$  belongs in segment  $S_k$ , otherwise it gives 0. The baseline AR1 parameters  $\rho_r$  are generated from  $\text{Unif}(0.1, 0.9)$  for each latent component.

**Setting 4.** The zero-mean latent fields  $z_i^*$  are generated as in Setting 3. Then, the final latent fields are obtained as  $z_i(\mathbf{s}, t) = z_i^*(\mathbf{s}, t) + \mu_i(\mathbf{s}, t)$ , where  $\mu_i(\mathbf{s}, t)$  is generated as in (11).

**Setting 5.** The latent fields have varying variances and varying AR1 coefficients over space and time. The fields are generated by combining settings 1 and 2. That is, we have an identical situation to Setting 2, but the function  $\sigma$  is defined as in Setting 4.

**Setting 6.** The zero-mean latent fields  $z_i^*$  are generated as in Setting 5. Then, the final latent fields are obtained as  $z_i(\mathbf{s}, t) = z_i^*(\mathbf{s}, t) + \mu_i(\mathbf{s}, t)$ , where  $\mu_i(\mathbf{s}, t)$  is generated as in (11).

These simulation settings are considered to investigate how different types of nonstationarities affect the performance of the algorithms. The Settings 1 and 2 do not have any nonstationarity in variance, but do have nonstationary AR1 coefficient, meaning that the identifiability results hold for iVAEar methods, but not for iVAEs and iVAEr. In Settings 3-6 the variance is nonstationary, and hence the identifiability holds for all iVAE methods. Nonetheless, these settings are of interest when comparing performances when there are additional stationary or nonstationary autocorrelation present. Nonstationary trend is considered in Settings 2, 4 and 6 to see if that affects the performance.

**Mixing function.** The observations  $\mathbf{x}$  are obtained by applying a linear or nonlinear mixing function  $\mathbf{f}_L$  to the generated latent components  $\mathbf{z}$ . The function  $\mathbf{f}_L$  is generated using multilayer perceptron (MLP) following, e.g. [11–13]. The parameter  $L$  denotes the number of mixing layers used in MLP. Each layer  $i$  consists of a  $P \times P$  mixing matrix  $\mathbf{B}_i$  and an activation function  $\psi_i$ . The matrices  $\mathbf{B}_i$  are normalized to have unit length rows and columns in order to avoid vanishing of any of the latent components in the mixing process. The mixing function  $\mathbf{f}_L$  can be then defined recursively as

$$\mathbf{f}_L(\mathbf{z}) = \begin{cases} \psi_L(\mathbf{B}_L \mathbf{z}), & L = 1, \\ \psi_L(\mathbf{B}_L \mathbf{f}_{L-1}(\mathbf{z})), & L \in \{2, 3, \dots\}, \end{cases}$$

where the activation function  $\psi_L$  is linear for the first layer and exponential linear unit (ELU), given as

$$\psi_i(x) = \begin{cases} x, & x \geq 0, \\ \exp(x) - 1, & x < 0, \end{cases}$$

for the other layers. This results  $\mathbf{f}_1$  with one layer being linear mixing, and when  $L$  increases, the mixing function becomes increasingly nonlinear.

**Performance index.** The performance of the algorithms is measured using the mean correlation coefficient (MCC), which is also used for example in [7, 12, 22, 23]. MCC is a function of correlation matrix  $\mathbf{\Omega} = \text{Cor}(\mathbf{z}, \hat{\mathbf{z}})$  of the true and estimated latent components. MCC measures how similar the optimal permutation of  $\mathbf{\Omega}$  is to  $P \times P$  identity matrix, and is calculated as

$$\text{MCC}(\mathbf{\Omega}) = \frac{1}{P} \sup_{\mathbf{P} \in \mathcal{P}} \text{tr}(\mathbf{P} \text{abs}(\mathbf{\Omega})), \quad (13)$$

where  $\mathcal{P}$  is a set of all possible  $P \times P$  permutation matrices,  $\text{tr}(\cdot)$  is the trace of a matrix and  $\text{abs}(\cdot)$  denotes taking elementwise absolute values of a matrix. The values of MCC vary in range  $[0, 1]$ , where 1 is the optimal value, meaning that estimated components  $\hat{\mathbf{z}}$  correlate perfectly with the true components  $\mathbf{z}$ .

**Model specifications.** All iVAE models have 3 hidden layers with 128 units in encoder, decoder and auxiliary functions. All hidden layers use leaky rectified unit (ReLU) activation function [18]. iVAEar\_r and iVAEr are set up with spatial resolution levels  $H = (2, 9)$  and temporal resolution levels  $G = (9, 17, 37)$ . In iVAEar1\_s and iVAEs,  $10 \times 10$  spatial segmentation is used by producing 100 equally sized segments, and temporal domain is divided into 100 segments, each of which contains 5 consecutive time points. For details of constructing the radial basis function based and segmentation based auxiliary variables, see [22]. All models are trained for 60 epochs with batch size of 64, and use the learning rate of 0.001 with polynomial decay of second-order over 10000 training steps, where the learning rate after the first 10000 training steps is 0.0001. STBSS uses two spatial ring kernels (0, 0.15) and (0.15, 0.3), and time lag of 1. These parameters were selected by training STBSS with multiple different parameters

in each setting, and selecting the parameters that provided the best results on average. For more about STBSS and its parameters, see [19].

**Simulation results.** The results of the simulations are provided in Figure 2. Overall, the best results, especially in nonlinear scenarios, are obtained by iVAEar\_r, followed by iVAEar\_s in all settings. Nonstationary trend (Settings 2, 4 and 6) results in worse performance for all of the methods compared to settings where the trend is not present (Settings 1, 3 and 6).

In Setting 1, where only AR1 coefficient is nonstationary, the latent components are successfully recovered only by iVAEar\_r and iVAEar\_s under nonlinear mixing. Under linear mixing, FICA performs nearly as well as iVAEar\_r and iVAEar\_s. STBSS is the fourth best performing method, followed by iVAEs and iVAEr.

In Setting 2, where also nonstationary trend is added, iVAEar\_r and iVAEar\_s are the only methods with decent performance, although their performance also drops considerably in nonlinear settings.

In Setting 3 with nonstationary variance, all of the methods perform relatively well. FICA and all iVAE based methods perform almost equally well under the linear mixing, but under the nonlinear mixing, FICA's performance suffers more. iVAEar based methods perform better than their iVAE counterparts, which is probably due to the fact that there are still stationary autocorrelation present in the latent components.

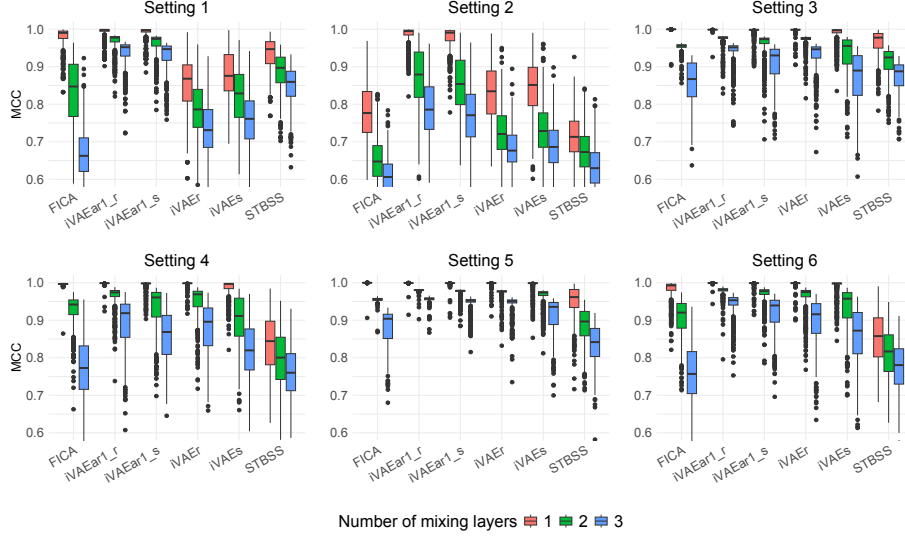
In Setting 4, where the nonstationary trend is included into scenario of Setting 3, all of the methods lose performance. However, iVAEar\_r still maintains its performance nearly as well as in Setting 3, being clearly the best method.

In Settings 5 and 6, where the variance and the AR1 coefficient are nonstationary, the results are very similar to the results of Settings 3 and 4, but the performances of FICA and iVAE methods are consistently slightly better due to the stronger nonstationarity. iVAE based methods maintain their performances better in nonlinear cases, and all of the methods perform slightly worse when the nonstationary trend is included.

Overall, autoregressive iVAE methods bring considerable improvement in performance as compared to the existing nonlinear STBSS methods. Based on the results, the methods can successfully estimate the latent components if there is either nonstationarity in autocorrelation or in variance. Nonstationary trend seems to be more challenging to tackle for the methods. Radial basis function based iVAEar, iVAEar\_r, is the best performing method in all of the settings, and is the recommended choice for nonlinear nonstationary STBSS problems.

## 4.2 Sensitivity for AR order mismatch

In this section, we study how sensitive the best performing method, iVAEar\_r, is for AR order mismatch. The data are generated from Settings 1 and 5 with the true AR orders  $R = 1$  and  $R = 3$ , and the latent components  $\mathbf{z}$  are estimated using the iVAEar\_r with AR orders  $W = 1, 3, 5$ , denoted iVAEar1\_r, iVAEar3\_r and iVAEar5\_r.



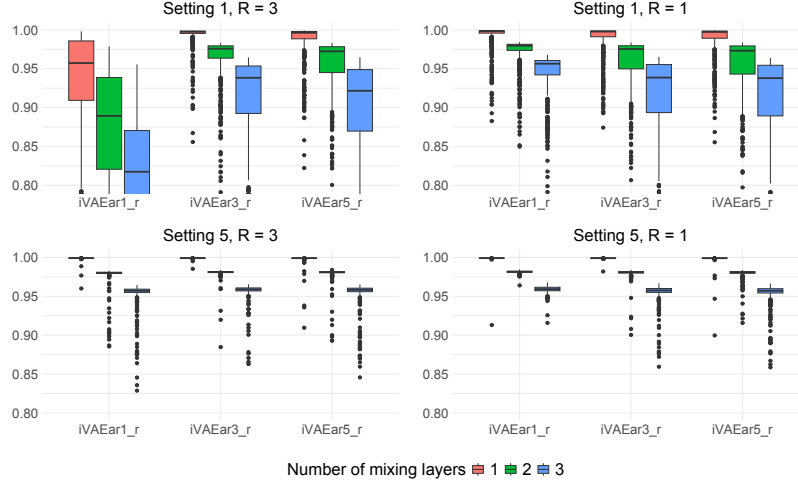
**Fig. 2.** Mean correlation coefficients from 500 trials for Settings 1-6. The y-axis shows MCC (optimal value = 1), while the x-axis represents different methods. Box colors indicate the number of mixing layers in the mixing function.

In  $R = 1$  scenario, the settings are identical to Settings 1 and 5 of the Section 4.1. In  $R = 3$  scenario, the data are generated as in Settings 1 and 5, but the AR coefficients  $\gamma_r(\mathbf{s}_i, t)$ ,  $r = 1, \dots, R$ ,  $i = 1, \dots, n_s$ , are generated as in (12). The coefficients are then multiplied by constants  $d_r$ , where  $d_r \sim \text{Unif}(0, 1)$ , to create varying magnitudes to the components. The baseline AR coefficients are set to  $\rho_r = 1$ ,  $r = 1, \dots, R$ . To guarantee the weak-sense stationarity of the AR process, defined in Definition 2 (supplementary material), the AR coefficients are scaled as follows:

$$\gamma_r(\mathbf{s}_i, t) = \frac{\gamma_r(\mathbf{s}_i, t)}{\max_{i,t}(|\gamma_r(\mathbf{s}_i, t)| + |\gamma_r(\mathbf{s}_i, t)| + |\gamma_r(\mathbf{s}_i, t)|) + 0.01}, \quad (14)$$

for each latent component  $j = 1, \dots, P$ . This procedure guarantees  $|\gamma_r(\mathbf{s}_i, t)| + |\gamma_r(\mathbf{s}_i, t)| + |\gamma_r(\mathbf{s}_i, t)| < 1$  for all  $r = 1, \dots, R$ ,  $i = 1, \dots, n_s$ , which is a sufficient condition for fulfilling the weak-sense stationarity.

The results are presented in Figure 3. In the case where only AR coefficients are nonstationary, the best performance is achieved when the true AR order  $W = R$  is used in the model. Based on the results, it is safer to use larger  $W$  as the performance drops only by little when  $W > R$ . The performance drops more significantly when too small  $W$  is used in the model. In the case where also variance is nonstationary, the effect of incorrect AR order is negligible, although the correct AR order still produces the best performance. In general, based on the results, it is safer to use  $W = 3$  or  $W = 5$  in the model rather than  $W = 1$ .

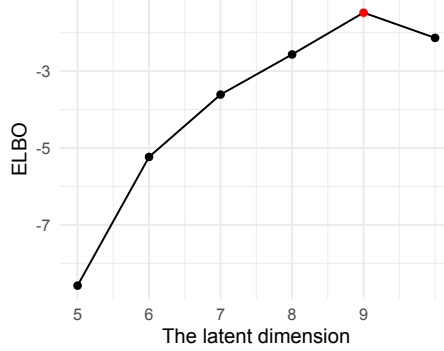


**Fig. 3.** Mean correlation coefficients of 500 trials for Setting 1 (top) and Setting 5 (bottom) with  $R = 1$  and  $R = 3$ . The y-axis shows MCC (optimal value = 1), while the x-axis represents different methods. Box colors indicate the number of mixing layers in the mixing function.

## 5 Case study

We apply the iVAEar\_r and iVAEar\_s methods to an air pollution dataset [1] to predict future values and compare their accuracy against iVAEr, spatio-temporal kriging [3], ARIMA [2] and vector ARIMA (VARIMA) [17]. Spatio-temporal kriging considers both spatial and temporal dependencies, while ARIMA models only temporal structures, making predictions separately for each station. Both kriging and ARIMA fit models univariately and do not account for cross-variable dependencies. In contrast, VARIMA models cross-dependencies between the variables through multivariate autoregressive process, but does modeling individually for each station. iVAEar\_r and iVAEr incorporate cross-variable dependencies through latent component decomposition and spatio-temporal trends. Additionally, iVAEar\_r estimates autoregressive structures of latent components for improved prediction.

The data consist of hourly air pollution and weather measurements from 64 stations in Athens, Greece, spanning 2020–2023. We use daily observations at 12 PM, resulting in  $n_t = 1124$ . The data include seven weather variables (wind speed U, wind speed V, dew point temperature, soil temperature, air temperature, relative humidity, precipitation) and four air pollution variables (PM10, PM2.5, NO2, O3). Precipitation is removed due to its predominantly zero values, yielding  $S = 10$ . Six stations lacking complete data are excluded, leaving  $n_s = 58$ . The remaining 162 missing observations are imputed using CUTOFF [4]. The last 24 time points serve as test data, while the first 1100 are used for training.



**Fig. 4.** ELBO for different latent dimensions.

We estimate the latent dimension  $P$  by fitting iVAEar\_r models with  $P = 5, \dots, 10$ , selecting the best model using knee-point detection and profile AIC (pAIC) [22]. ELBOs for different latent dimensions are shown in Figure 4. Both methods indicate  $P = 9$  as optimal, which is used in final models.

For forecasting with iVAEar\_r, iVAEar\_s and iVAEr, auxiliary data must remain within the training data bounds. Hence, we use seasonal periods  $t_s = 1, \dots, 365$  instead of absolute time  $t = 1, \dots, 1124$  and introduce a one-hot encoded year factor to allow inter-year variability. Spatial resolution levels are set to  $H = (2, 9)$ , learning rate to 0.0001, variance parameter  $\beta = 0.02$ , batch size to 64, and training spans 40 epochs. A hyperparameter search optimizes temporal resolution for iVAEar\_r and iVAE\_r, segmentation sizes for iVAEar\_s, hidden units in the auxiliary function, and autoregressive order for iVAEar\_r and iVAEar\_s. The best parameters are selected by leaving 10 last time points of the training data for validation. Selected parameters are  $G = (9, 17)$ ,  $n_{\theta_w} = 16$ , and  $R = 2$  for iVAEar\_r,  $G = (9, 17)$ ,  $n_{\theta_w} = 16$  for iVAE\_r and spatial segment size of 5000, temporal segment size of 5 and  $R = 3$  for iVAEar\_s.

For ARIMA, VARIMA and kriging, seasonal trends are removed as these methods assume seasonal stationarity. Seasonality is modeled as

$$x_i(\mathbf{s}, t) = \beta_{0,i} + \beta_{1,i} \cos(2\pi t/365) + \beta_{2,i} \sin(2\pi t/365) + x_{res,i}(\mathbf{s}, t),$$

where residuals  $x_{res,i}$  are predicted using ARIMA and kriging. Kriging uses product-sum covariance models, while ARIMA selects the best model for each station via corrected AIC with AR orders  $0, \dots, 5$ , MA orders  $0, \dots, 5$ , and integration determined by the KPSS test [15]. In VARIMA, we select the best model for each station based on AIC. The options are models with  $AR = 1, \dots, 8$  and  $MA = 0$ , or a model with  $AR = 1$  and  $MA = 1$ . Integration order was selected to be 1 for whole data, from options 1 or 0, based on better validation accuracy. VARIMA models with larger number of parameters caused numerical instability, and were hence not considered.

Prediction accuracy is measured using mean squared error (MSE):

$$MSE(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \frac{1}{n} \sum_{j=1}^n (x_{i,j} - \hat{x}_{i,j})^2,$$

where  $\mathbf{x}_i$  contains true values and  $\hat{\mathbf{x}}_i$  predicted ones. Combined accuracy is assessed via weighted MSE (wMSE):

$$wMSE(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{S} \sum_{i=1}^S \frac{MSE(\mathbf{x}_i, \hat{\mathbf{x}}_i)}{\sigma^2(\mathbf{x}_i)},$$

where  $\sigma^2(\mathbf{x}_i)$  is the variance of the deseasonalized variable.

Table 1 presents forecasting results. iVAEar\_r outperforms competitors based on wMSE. VARIMA has the second best combined performance, and the best performance for predicting wind speeds. ARIMA has the second worst combined performance but excels for PM10, PM2.5, and NO2. Kriging and iVAEr perform similarly, with kriging being slightly better overall and excelling in soil temperature predictions. iVAEar\_r achieves the lowest errors for dew point temperature, air temperature and O3. iVAEar\_s has the best prediction performance for relative humidity, and has high accuracy for O3 as well, but its high errors on soil temperature, air temperature and NO2 makes it the worst method when considering the overall performance. Notably, O3 and relative humidity predictions benefit significantly from incorporating cross-variable dependencies, underscoring the advantage of iVAEar\_r, iVAEar\_s and VARIMA over univariate models. However, iVAEar\_s shows inconsistent performance in prediction and is suboptimal for this task. Its segmentation-based auxiliary variables lead to a highly non-continuous estimate of the trend function, which hinders the model’s ability to generalize to future data. Therefore, iVAEar\_r is the preferred method for forecasting purposes.

**Table 1.** Mean squared errors for predictions in time.

	Wind Speed U	Wind Speed V	Dewpoint Temp	Soil Temp	Temp Rel. Humidity	PM10	PM2.5	NO2	O3	wMSE
iVAEar_r	1.57	6.16	<b>3.42</b>	1.08	<b>3.44</b>	64.15	81.31	30.42	106.87	<b>93.09</b>
iVAEar_s	1.70	7.25	4.70	8.75	7.22	<b>47.29</b>	81.56	30.71	200.20	94.91
iVAEr	2.05	11.36	4.24	0.60	4.60	96.40	84.60	31.89	114.69	174.50
Kriging	1.71	8.41	4.73	<b>0.44</b>	5.49	131.21	82.92	43.65	104.03	141.89
ARIMA	1.79	6.22	4.85	3.08	9.22	119.62	<b>75.15</b>	<b>27.26</b>	<b>97.98</b>	190.36
VARIMA	<b>1.54</b>	<b>5.75</b>	4.11	1.64	8.29	65.68	75.70	35.80	99.06	121.81

## 6 Conclusions and Discussion

We have proposed a novel autoregressive iVAE method for nonlinear spatio-temporal BSS, extending identifiability results to cases with nonstationary autoregressive coefficients. Our simulation studies demonstrate superior latent com-



ponent estimation compared to state-of-the-art methods, and real-world applications to air pollution and weather datasets show that iVAEar achieves significantly improved multivariate spatio-temporal prediction accuracy. Furthermore, we establish strong identifiability results, particularly for autoregressive Gaussian latent components.

A limitation of iVAEar is its reliance on a strict autoregressive assumption in time, making it optimal for separable spatio-temporal processes. Future work should explore extensions to nonseparable models and to more general graph structured data. As the identifiability under nonstationary AR coefficients was studied in this paper mainly for Gaussian innovations, the robustness of the method against innovations from other distributions should be studied in future.

In prediction tasks, careful hyperparameter selection and validation are necessary to prevent overfitting, and auxiliary variables must be chosen to ensure compatibility between training and test data. Additionally, iVAEar can be combined with univariate spatio-temporal prediction methods such as graphLSTM [6], allowing latent components to be predicted separately before reconstructing the observed data.

As iVAEar can be used for both time series and spatio-temporal data, it is a valuable method for latent component estimation and multivariate prediction across various fields, including environmental sciences, meteorology, and neuroscience, where applications often involve multiple temporal or spatio-temporal variables representing the same underlying phenomenon.

**Acknowledgments.** We gratefully acknowledge support from the Vilho, Yrjö, and Kalle Väisälä Foundation (MS), the Research Council of Finland (grants 453691 to ST and 363261 to KN), and the HiTEc COST Action (CA21163) (KN, ST). We also thank CSC – IT Center for Science, Finland, for providing computational resources. Finally, we thank the three anonymous referees for their insightful feedback.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Angelis, G.F., Emvolidis, A., Theodorou, T.I., Zamichos, A., Drosou, A., Tzavaras, D.: Regional datasets for air quality monitoring in European cities. In: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. pp. 6875–6880 (2024)
2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, 5th edn. (2015)
3. Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley (2011)
4. Feng, L., Nowak, G., O’Neill, T., Welsh, A.: CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology* **519**, 3591–3605 (2014)
5. Flumian, L., Matilainen, M., Nordhausen, K., Taskinen, S.: Stationary subspace analysis based on second-order statistics. *Journal of Computational and Applied Mathematics* **436**, 115379 (2024)

6. Gao, X., Li, W.: A graph-based LSTM model for PM2.5 forecasting. *Atmospheric Pollution Research* **12**(9), 101150 (2021)
7. Hälvä, H., Hyvärinen, A.: Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In: *Conference on Uncertainty in Artificial Intelligence*. pp. 939–948. PMLR (2020)
8. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.:  $\beta$ -VAE: Learning basic visual concepts with a constrained variational framework. In: *International conference on learning representations* (2017)
9. Hyvärinen, A., Khemakhem, I., Morioka, H.: Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns* **4**(10) (2023)
10. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10**(3), 626–634 (1999)
11. Hyvärinen, A., Morioka, H.: Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems* **29** (2016)
12. Hyvärinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. pp. 859–868. PMLR (2019)
13. Khemakhem, I., Kingma, D., Monti, R., Hyvärinen, A.: Variational autoencoders and nonlinear ICA: A unifying framework. In: *International Conference on Artificial Intelligence and Statistics*. pp. 2207–2217. PMLR (2020)
14. Kingma, D.P., Welling, M.: Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* (2013)
15. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* **54**(1-3), 159–178 (1992)
16. Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K.E., Le Priol, R., Lacoste, A., Lacoste-Julien, S.: Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In: *Conference on Causal Learning and Reasoning*. pp. 428–484. PMLR (2022)
17. Lütkepohl, H.: *New introduction to multiple time series analysis*. Springer Science & Business Media (2005)
18. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: *Proceedings of the ICML Workshop on Deep Learning for Audio, Speech, and Language Processing* (2013)
19. Muehlmann, C., De Iaco, S., Nordhausen, K.: Blind recovery of sources for multivariate space-time random fields. *Stochastic Environmental Research and Risk Assessment* **37**, 1593–1613 (2023)
20. Porcu, E., Furrer, R., Nychka, D.: 30 years of space-time covariance functions. *WIREs Computational Statistics* **13**(2), e1512 (2021)
21. Salvaña, M.L.O., Genton, M.G.: Nonstationary cross-covariance functions for multivariate spatio-temporal random fields. *Spatial Statistics* **37**, 100411 (2020)
22. Sipilä, M., Cappello, C., De Iaco, S., Nordhausen, K., Taskinen, S.: Modelling multivariate spatio-temporal data with identifiable variational autoencoders. *Neural Networks* **181**, 106774 (2025)
23. Sipilä, M., Nordhausen, K., Taskinen, S.: Nonlinear blind source separation exploiting spatial nonstationarity. *Information Sciences* **665**, 120365 (2024)