Independent Innovation Analysis for Nonlinear Vector Autoregressive Process

Hiroshi Morioka
RIKEN AIP

Hermanni Hälvä
University of Helsinki

Aapo Hyvärinen
University of Helsinki
Université Paris-Saclay, Inria

Abstract

The nonlinear vector autoregressive (NVAR) model provides an appealing framework to analyze multivariate time series obtained from a nonlinear dynamical system. However, the innovation (or error), which plays a key role by driving the dynamics, is almost always assumed to be additive. Additivity greatly limits the generality of the model, hindering analysis of general NVAR processes which have nonlinear interactions between the innovations. Here, we propose a new general framework called independent innovation analysis (IIA), which estimates the innovations from completely general NVAR. We assume mutual independence of the innovations as well as their modulation by an auxiliary variable (which is often taken as the time index and simply interpreted as nonstationarity). We show that IIA guarantees the identifiability of the innovations with arbitrary nonlinearities, up to a permutation and component-wise invertible nonlinearities. We also propose three estimation frameworks depending on the type of the auxiliary variable. We thus provide the first rigorous identifiability result for general NVAR, as well as very general tools for learning such models.

1 INTRODUCTION

Multivariate time series are of considerable interest in a number of domains, such as finance, economics, and engineering. Vector autoregressive (VAR) models have played a central role in capturing the dynamics hidden in such time series (Sims, 1980). VAR models typically attempt to fit a multivariate time series with linear coefficients representing the dependencies of multivariate variables within limited number of lags, and *innovation* (or error) representing new information (impulses) fed to the process at a given time point. Although it has been common practice to maintain a linear functional form to achieve interpretability and tractability, recent studies have provided a growing body of evidence that nonlinearity often exists in time series, and allowing for nonlinearities can be valuable for uncovering important features of dynamics (Jeliazkov, 2013; Kalli and Griffin, 2018; Koop and Korobilis, 2010; Primiceri, 2005; Shen et al., 2019; Teräsvirta, 1994; Tsay, 1998). Many recent studies used a deep learning framework to model nonlinear processes in video (Finn et al., 2016; Lotter et al., 2017; Oh et al., 2015; Srivastava et al., 2015; Villegas et al., 2017; Wichers et al., 2018) or audio (van den Oord et al., 2016), for example, with neural networks.

The innovation plays a key role by driving time series, and it can have a concrete meaning, such as economic shocks in finance, external torques given to a mechanical system, or stimulation in neuroscience experiments. However, its estimation has a serious indeterminacy even with linear models, if only conventional statistical assumptions are made. To facilitate estimation, VAR typically assumes that the innovations are additive, multivariate Gaussian (not necessary uncorrelated), and temporally independent (or serially uncorrelated). A well-known consequence of this is that the innovations cannot be identified: Multiplication of such innovations by any orthogonal matrices will not change distribution of the observed data, which hinders their interpretation. Some studies proposed to incorporate independent component analysis (ICA) framework to guarantee identifiability, by assuming mutual independence of non-Gaussian innovations (Gómez-Herrero et al., 2008; Hyvärinen et al., 2010; Lanne et al., 2017; Moneta et al., 2013). However, those studies assumed linear VAR models, while indeterminacy would presumably be even more serious

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

in general nonlinear VAR (NVAR) models, in which the innovations may not be additive anymore. In fact, a serious lack of identifiability in general nonlinear cases is well-known in nonlinear ICA (NICA) (Hyvärinen and Pajunen, 1999).

We propose a novel VAR analysis framework called independent innovation analysis (IIA), which enables estimation of innovations hidden in unknown general NVAR. We first propose a model which allows for nonlinear interactions between innovations and observations, with very general nonlinearities. IIA can be seen as an extension of recently proposed NICA frameworks (Hälvä and Hyvärinen, 2020; Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019), and guarantees the identifiability of innovations up to permutation and component-wise nonlinearities. The model assumes a certain temporal structure in the innovations, which typically takes the form of nonstationarity, but can be more general. We propose three practical estimation methods for IIA, two of which are self-supervised and can be easily implemented based on ordinary neural network training, and the remaining one uses maximum-likelihood estimation in connection with a hidden Markov model. Our identifiability theory for NVAR is quite different from anything presented earlier, and thus it can contribute as a new general framework for NVAR process.

2 MODEL DEFINITION

2.1 NVAR Model and Demixing Model

We here assume a general NVAR model, which is first order (NVAR(1)) for simplicity:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \mathbf{s}_t), \tag{1}$$

where $\mathbf{f} : \mathbb{R}^{2n} \rightarrow \mathbb{R}^n$ represents an NVAR (mixing) model, and $\mathbf{x}_t = [x_1(t), \dots, x_n(t)]^T$ and $\mathbf{s}_t = [s_1(t), \dots, s_n(t)]^T$ are observations and innovations (or errors) of the process at time point t, respectively. As with ordinary VAR, the innovations are assumed to be temporally independent (serially uncorrelated). Importantly, this model includes potential nonlinear interaction between the observations and innovations, unlike ordinary linear VAR models (Gómez-Herrero et al., 2008; Hyvärinen et al., 2010; Lanne et al., 2017; Moneta et al., 2013) and additive innovation nonlinear models (Shen et al., 2019). We assume that $\mathbf{f}$ is unknown and make minimal regularity assumptions on it. Our goal is to estimate the innovations (latent components) $\mathbf{s}$ only from the observations $\mathbf{x}$ obtained from the unknown NVAR process. The model, learning algorithms, Theorems, and proofs below can be easily extended to higher order models NVAR(p) ($p > 1$) by replacing $\mathbf{x}_{t-1}$ by $[\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-p}]$.

To estimate the innovation, we propose a new framework called IIA, which learns the inverse (demixing) of the NVAR (mixing) model from the observations in data-driven manner, based on some statistical assumptions on the innovations. The theory is related to ICA (Hyvärinen, 1999), which estimates a demixing from *instantaneous* mixtures of latent components, i.e., $\mathbf{x}_t = \mathbf{f}_{\mathrm{ICA}}(\mathbf{s}_t)$, where $\mathbf{f}_{\mathrm{ICA}} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is usually a linear function. However, IIA includes a recurrent structure of the observations in the model (Eq. 1), which makes IIA theoretically distinct from ordinary ICA. Nevertheless, in the following we leverage the recently developed theory of NICA (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019).

We start by transforming the NVAR model to something similar to NICA. This leads us to consider the following augmented NVAR (mixing) model

$$\begin{bmatrix}\mathbf{x}_t \\ \mathbf{x}_{t-1}\end{bmatrix} = \tilde{\mathbf{f}}\left(\begin{bmatrix}\mathbf{s}_t \\ \mathbf{x}_{t-1}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{f}(\mathbf{x}_{t-1}, \mathbf{s}_t) \\ \mathbf{x}_{t-1}\end{bmatrix}, \tag{2}$$

where $\tilde{\mathbf{f}} : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ is the augmented model, which includes the original NVAR model $\mathbf{f}$ in the half of the space, and an identity mapping of $\mathbf{x}_{t-1}$ in the remaining subspace. Importantly, this augmentation does not impose any particular constraint on the original model. We only assume that this augmented model is invertible (i.e. bijective; while $\mathbf{f}$ itself cannot be invertible) as well as sufficiently smooth, but we do not constrain it in any other way. The estimation of the innovation $\mathbf{s}$ can then be achieved by learning the inverse (demixing) of the augmented NVAR model $\tilde{\mathbf{f}}$:

$$\begin{bmatrix}\mathbf{s}_t \\ \mathbf{x}_{t-1}\end{bmatrix} = \tilde{\mathbf{g}}\left(\begin{bmatrix}\mathbf{x}_t \\ \mathbf{x}_{t-1}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1}) \\ \mathbf{x}_{t-1}\end{bmatrix}, \tag{3}$$

where $\tilde{\mathbf{g}} : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ is the augmented demixing model of the (true) augmented NVAR model $\tilde{\mathbf{f}}$, and $\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1}) \in \mathbb{R}^n$ is the sub-space of the demixing model representing a mapping from two temporally consecutive observations to the innovation at the corresponding timing. This is simply a deduction from Eq. 2, and does not impose any additional assumptions on the original model.

2.2 Innovation Model with Auxiliary Variable

The estimation of the demixing model in an unsupervised (or self-supervised) manner needs some assumptions on the innovations. Although some studies guaranteed the identifiability by assuming mutual independence of the innovations in linear VAR models (Hyvärinen et al., 2010; Lanne et al., 2017; Moneta et al., 2013), it would not be enough in nonlinear cases, as can be seen in well-known indeterminacy of NICA with i.i.d. components (Hyvärinen and Pajunen, 1999). Thus, we here adopt the framework recently proposed for NICA (Hälvä and Hyvärinen, 2020; Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019); we assume that the distribution of the innovation is time-dependent, and modulated through an observable (or unobservable, Section 3.3) auxiliary information about the innovation, represented by a random variable $\mathbf{u}_t$ for each data point t. In practice, $\mathbf{u}_t$ can simply be time-index t to represent data-point-specific modulations or a time-segment-index to represent segment-wise modulations, thus incorporating information about nonstationarity. More specifically, we assume the followings:

A1. Each s_i is statistically dependent on some m-dimensional random auxiliary variable $\mathbf{u}$, but conditionally independent of the other s_j, and has a univariate exponential family distribution conditioned on $\mathbf{u}$ (we omit data index t here):

$$p(\mathbf{s}|\mathbf{u}) = \prod_{i=1}^{n} \frac{Q_i(s_i)}{Z_i(\mathbf{u})} \exp\left[\sum_{j=1}^{k} q_{ij}(s_i)\lambda_{ij}(\mathbf{u})\right], \quad (4)$$

where Q_i is the base measure, Z_i is the normalizing constant, k is the model order, q_{ij} is the sufficient statistics, and $\lambda_{ij}(\mathbf{u})$ is a parameter (scalar function) depending on $\mathbf{u}$.[1]

This model is related to the assumption of Gaussian innovations in ordinary VAR, but requires more specific properties represented by conditional independence and sufficient probabilistic modulation, determined by an auxiliary variable $\mathbf{u}$. Note that exponential families have universal approximation capabilities, so this assumption is not very restrictive (Sriperumbudur et al., 2017).

3 LEARNING ALGORITHMS

Depending on the type of the auxiliary variable $\mathbf{u}$ in the innovation model (see A1), we can develop three learning algorithms; The first one (IIA-GCL; Section 3.1) is for general cases of observable $\mathbf{u}$, the second one (IIA-TCL; Section 3.2) is for specific type of observable $\mathbf{u}$ underlying within a finite number of classes, and the last one (IIA-HMM; Section 3.3) is for unobservable $\mathbf{u}$ represented by hidden Markov chain.

3.1 General Contrastive Learning Framework (IIA-GCL)

In the general case with observable and possibly continuous-valued $\mathbf{u}$, we develop a general contrastive learning (GCL) framework for IIA, based on the recently proposed NICA framework (Hyvärinen et al., 2019). In IIA-GCL, we train a feature extractor and a logistic regression classifier, which discriminates a real dataset composed of the true observations of $(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{u}_t)$, from a version where randomization is performed on $\mathbf{u}$. Thus we define two datasets in which a data point t is written as follows, respectively:

$$\tilde{\mathbf{x}}_t = (\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{u}_t) \text{ vs. } \tilde{\mathbf{x}}_t^* = (\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{u}^*), \quad (5)$$

where $\mathbf{u}^*$ is a random value from the distribution of $\mathbf{u}$, but independent of $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$, created in practice by random permutation of the empirical sample of $\mathbf{u}$. We learn a nonlinear logistic regression system using a regression function of the form

$$\begin{aligned} r(\tilde{\mathbf{x}}_t) = &\sum_{i=1}^{n}\sum_{j=1}^{k} \psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1}))\mu_{ij}(\mathbf{u}_t) + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t) \\ &+ \alpha(\mathbf{u}_t) + \beta(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) + \gamma(\mathbf{x}_{t-1}), \end{aligned} \quad (6)$$

which gives the posterior probability of the first class $\tilde{\mathbf{x}}$ as $1/(1+\exp(-r(\tilde{\mathbf{x}}_t))$. The scalar-valued functions ψ_{ij}, h_i, μ_{ij}, ϕ, α, β, and γ take some specific combinations of $\mathbf{x}_t$, $\mathbf{x}_{t-1}$, and $\mathbf{u}_t$ as input, which are designed to match to the difference of the log-pdfs of $(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{u}_t)$ in the two datasets, given the innovation model Eq. 4 (see Supplementary Material A). The universal approximation capacity (Hornik et al., 1989) is assumed for those functions; they would typically be learned by neural networks. This learning framework and the regression function are based on the following Theorem, proven in Supplementary Material A:

Theorem 1. *Assume the following:*

1. *We obtain observations and auxiliary variable* $\mathbf{u}$ *from an NVAR model (Eq. 1), whose augmented model (Eq. 2) is invertible and sufficiently smooth.*

2. *The latent innovations of the process are temporally independent, follow the assumption A1 with* $k \geq 2$*, and the sufficient statistics* q_{ij} *are twice differentiable.*

3. *(Assumption of Variability) There exist* $nk+1$ *distinct points* $\mathbf{u}_0, \dots, \mathbf{u}_{nk}$ *such that the matrix*

$$\mathbf{L} = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), \dots, \boldsymbol{\lambda}(\mathbf{u}_{nk}) - \boldsymbol{\lambda}(\mathbf{u}_0)) \quad (7)$$

of size $nk \times nk$ *is invertible, where* $\boldsymbol{\lambda}(\mathbf{u}) = (\lambda_{11}(\mathbf{u}), \dots, \lambda_{nk}(\mathbf{u}))^T \in \mathbb{R}^{nk}$.

[1] The k is assumed to be minimal, meaning that we cannot rewrite the form with a smaller $k' < k$. The parameters are assumed that for each i, $(\exists(\lambda_{i1}(\mathbf{u}), \dots, \lambda_{ik}(\mathbf{u}))|\forall s_i, \sum_{j=1}^{k} q_{ij}(s_i)\lambda_{ij}(\mathbf{u}) = \text{const}) \implies (\lambda_{i1}(\mathbf{u}), \dots, \lambda_{ik}(\mathbf{u})) = 0$. These conditions are required for the distribution to be strongly exponential (Khemakhem et al., 2020), which is not very restrictive, and satisfied by all the usual exponential family distributions.

4. *We train a nonlinear logistic regression system with universal approximation capability to discriminate between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^*$ in Eq. 5 with regression function in Eq. 6.*

5. *The augmented function $\tilde{\mathbf{h}}(\mathbf{x}_t, \mathbf{x}_{t-1}) = [\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1}), \mathbf{x}_{t-1}] : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is invertible.*

6. *The scalar functions ψ_{ij} in Eq. 6 are twice differentiable, and for each i, the following implication holds: $(\exists \boldsymbol{\theta} \in \mathbb{R}^k | \forall y, \sum_{j=1}^k \psi_{ij}(y)\theta_j = const) \implies \boldsymbol{\theta} = 0$.*

Then, in the limit of infinite data, $\mathbf{h}$ in the regression function provides a consistent estimator of the IIA model: The functions $h_i(\mathbf{x}_t, \mathbf{x}_{t-1})$ give the independent innovations, up to permutation and scalar (component-wise) invertible transformations.

This Theorem guarantees the convergence (consistency) of the learning algorithm. It immediately implies the identifiability of the innovations, up to a permutation and component-wise invertible nonlinearities. This kind of identifiability for innovations is stronger than any previous results in the literature. The estimation is based on the learning of nonlinear logistic regression function, and thus can be easily implemented based on ordinary neural network training. The Assumption of Variability requires the auxiliary variable $\mathbf{u}$ to have a sufficiently strong and diverse effect on the distributions of the innovations. The assumptions on the NVAR model are not too restrictive, and supposed to be satisfied in many applications. The temporal independence of the innovations is the ordinary assumption for VAR. The assumption 6 indicates that ψ_{ij} are not functionally redundant; any ψ_{ij} cannot be represented by a linear combination of $\psi_{il \neq j}$. Although the assumptions of the nonlinear functions to be trained (assumptions 5 and 6) are not trivial, we assume they are only necessary to have a rigorous theory, and immaterial in any practical implementation.

3.2 Time-Contrastive Learning Framework (IIA-TCL)

In the special case in which $\mathbf{u}_t$ is observable and integer within a finite number of classes $[1, T]$, we can also develop a TCL-based framework for the estimation (Hyvärinen and Morioka, 2016). This special case includes time-segment-wise stationary process in which $\mathbf{u}_t$ represents the time segment index at time t.

Instead of the two-class logistic regression used in IIA-GCL, IIA-TCL uses a multinomial logistic regression (MLR) classifier for the learning. More specifically, we learn a nonlinear MLR using a softmax function which represents the posterior distribution of $\mathbf{u}$, by the form

$$p(\mathbf{u}_t = \tau | \mathbf{x}_t, \mathbf{x}_{t-1}) = \frac{\exp(\sum_{i=1}^n \sum_{j=1}^k z_{ij\tau})}{\sum_{l=1}^T \exp(\sum_{i=1}^n \sum_{j=1}^k z_{ijl})},$$
$$z_{ijl} = w_{ijl}\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1})) + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = l) + b_l, \quad (8)$$

where $w_{ij\tau}, b_\tau$ are the class-specific weight and bias parameters of the MLR, and ψ_{ij}, h_i, and ϕ are again scalar-valued functions assumed to have the universal approximation capacity. This functional form is designed based on the innovation model given by Eq. 4 (see Supplementary Material B). This learning framework and the regression function are justified on the following Theorem, proven in Supplementary Material B:

Theorem 2. *Assume the following:*

1. *We obtain observations and auxiliary variable $\mathbf{u}$ from an NVAR model (Eq. 1), whose augmented model (Eq. 2) is invertible and sufficiently smooth.*

2. *The latent innovations of the process are temporally independent, follow the assumption A1 with $k \geq 2$, and the sufficient statistics q_{ij} are twice differentiable.*

3. *The auxiliary variable $\mathbf{u}$ is an integer in $[1, T]$, with T the number of values it takes (classes).*

4. *The modulation matrix of size $nk \times (T-1)$*

$$\mathbf{L} = (\boldsymbol{\lambda}(2) - \boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(T) - \boldsymbol{\lambda}(1)) \quad (9)$$

   has full row rank nk, where $\boldsymbol{\lambda}(\tau) = (\lambda_{11}(\mathbf{u} = \tau), \ldots, \lambda_{nk}(\mathbf{u} = \tau))^T \in \mathbb{R}^{nk}$.

5. *We train a multinomial logistic regression with universal approximation capability to predict the class label (auxiliary variable) $\mathbf{u}_t$ from $(\mathbf{x}_t, \mathbf{x}_{t-1})$ with regression function in Eq. 8.*

6. *The augmented function $\tilde{\mathbf{h}}(\mathbf{x}_t, \mathbf{x}_{t-1}) = [\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1}), \mathbf{x}_{t-1}] : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is invertible.*

7. *The scalar functions ψ_{ij} in Eq. 8 are twice differentiable, and for each i, the following implication holds: $(\exists \boldsymbol{\theta} \in \mathbb{R}^k | \forall y, \sum_{j=1}^k \psi_{ij}(y)\theta_j = const) \implies \boldsymbol{\theta} = 0$.*

Then, in the limit of infinite data in each class, $\mathbf{h}$ in the regression function provides a consistent estimator of the IIA model: The functions $h_i(\mathbf{x}_t, \mathbf{x}_{t-1})$ give the independent innovations, up to permutation and scalar (component-wise) invertible transformations.

Many of the assumptions are the same as those in IIA-GCL, except for the specifics of the innovation model (assumptions 3 and 4) and the learning algorithm (assumption 5). The estimation is based on self-supervised nonlinear MLR, and thus can be easily implemented based on ordinary neural network training, like IIA-GCL. Although the estimation methods are different, the identifiability result implied here by IIA-TCL is the same as above by IIA-GCL. Note that here the limit of infinite data takes the form that each class (value of T) has an infinite number of data points. In practice, each class is thus required to have a sufficient number of samples, so T needs to be much smaller than the total number of data points; this would be natural if T is a segment index (see Fig. 1b for the empirical result of this point).

3.3 Hidden Markov Model Framework (IIA-HMM)

Next, we consider a special case where no $\mathbf{u}$ is observed, and no segmentation is imposed as in TCL. Instead, we assume the nonstationarity is described by hidden states following a discrete-time Markov model (Hälvä and Hyvärinen, 2020). This framework does not require $\mathbf{u}_t$ to be observable unlike the previous two frameworks, and thus can learn the model in an "purely unsupervised" manner. It is essentially like TCL but the segmentation is inferred as part of the learning process.

We assume the following temporal structure for $\mathbf{u}$;

A2. The latent auxiliary variable $\mathbf{u}_t \in \{1, \dots, C\}$ represents a hidden random states at each time point, and it is described by a Markov chain governed by a time-invariant transition-probability matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$, where $A_{i,j}$ denotes the probability of transitioning from state i to j.

From the NVAR observation model with the hidden Markov chain $\mathbf{u}_t$ generating the innovations for each data point t, the likelihood is given by, using the probability transformation formula,

$$p(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T; \mathbf{A}, \boldsymbol{\theta}) = p(\mathbf{x}_0) \prod_{t=1}^{T} |\mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\ \times \sum_{\mathbf{u}_1, \dots, \mathbf{u}_T} \pi_{\mathbf{u}_1} p(\mathbf{s}_1 | \mathbf{u}_1; \boldsymbol{\theta}) \prod_{t=2}^{T} \mathbf{A}_{\mathbf{u}_{t-1}, \mathbf{u}_t} p(\mathbf{s}_t | \mathbf{u}_t; \boldsymbol{\theta}) \quad (10)$$

where $\boldsymbol{\theta} = \{\lambda, \mathbf{g}\}$, λ denotes the parameters of the innovation model with omitting subscripts (Eq. 4), $\mathbf{g}$ is the demixing model, whose augmented model is $\tilde{\mathbf{g}}$ (Eq. 3), $\boldsymbol{\pi} = (\pi_1, \dots, \pi_C)$ is the stationary distribution of the latent state $\mathbf{u}$, $p(\mathbf{x}_0)$ is the marginal distribution of $\mathbf{x}_0$, and $\mathbf{J}\tilde{\mathbf{g}}$ denotes the Jacobian of $\tilde{\mathbf{g}}$. The summation (marginalization) is taken over all possible combinations of $\mathbf{u}_1, \dots, \mathbf{u}_T$.

Unlike the previous two frameworks which are based on self-supervised learning, the estimation of the model has to be done by a maximum-likelihood framework since $\mathbf{u}_t$ is unobservable here. For example, EM algorithm can be deployed when the innovation model was chosen from a well-known family such that the normalizing constant is tractable. The algorithm basically follows that of Hälvä and Hyvärinen (2020), with some differences coming from the autoregressive structure in the observations; the demixing model $\tilde{\mathbf{g}}$ has the augmented structure defined in Eq. 3, and the marginal distribution model of the observation $p(\mathbf{x}_0)$ is required. The E-step finds the optimal sequence of the latent states $(\mathbf{u}_1, \dots, \mathbf{u}_T)$, and M-step updates the parameters of the model so as to maximize the lower bound. Since a closed-form of the update for $\mathbf{g}$ is not available in many cases, a gradient ascent update is taken instead. Although the gradient of the determinant of the Jacobian $|\mathbf{J}\tilde{\mathbf{g}}|$ is generally considered to be difficult, recent developments of autograd packages, such as JAX, makes it possible to calculate them numerically up to moderate dimensions (Hälvä and Hyvärinen, 2020). Moreover, it can be computed using the recently proposed relative gradient method (Gresele et al., 2020). The identifiability of this framework is discussed in Supplementary Material C.

4 EXPERIMENTS

4.1 Simulation 1: IIA-GCL for Artificial Dynamics with Nonstationary Innovations

Data Generation We generated data from an artificial NVAR process with nonstationary innovations. The innovations were randomly generated from a Gaussian distribution by modulating its mean and standard deviation across time t, i.e., $\mathbf{u}_t = t$. The modulations were designed to be temporally smooth and continuous. The dimensions of the observations and innovations (n) were 20. As the NVAR model, we used a multilayer perceptron we call NVAR-MLP, which takes a concatenation of $\mathbf{x}_{t-1}$ and $\mathbf{s}_t$ as an input, then outputs $\mathbf{x}_t$. The goal of this simulation is to estimate the innovations $\mathbf{s}$ only from the observable time series $\mathbf{x}$, without knowing the parameters of the NVAR-MLP. See Supplementary Material D for more details of the experimental settings.

Training Considering the innovation model with $\mathbf{u}_t = t$, we here used IIA-GCL for the estimation of the latent innovations. We adopted MLPs as the nonlinear scalar functions in Eq. 6. The nonlinear regression function was trained by back-propagation with a momentum term so as to discriminate the real dataset from its $\mathbf{u}_t$-randomized version. For comparison, we also applied NICA based on GCL (NICA-GCL; Hyvärinen et al. (2019)), an NVAR with additive innovation model (AD-NVAR), and variational autoencoder (VAE; Kingma and Welling (2014)) to the same data.

Result The IIA-GCL framework could reconstruct the innovations reasonably well even for the nonlinear mixture cases ($L > 1$) (Fig. 1a). We can see that a larger amount of data make it possible to achieve higher performance, and higher complexity of the NVAR model makes learning more difficult. AD-NVAR performed well for the linear mixture case ($L = 1$) because the additive innovation model is equivalent to the general NVAR model in the linear case; however, it was much worse in the nonlinear case. As expected, the other methods performed worse than IIA-GCL because their model did not match well to the NVAR generation model.

4.2 Simulation 2: IIA-TCL for Artificial Dynamics with Nonstationary Innovations

Training Next, to evaluate the IIA-TCL framework, we applied it to the same data used in Simulation 1. For IIA-TCL, we first divided the time series into 256 equally-sized segments, and used the segment label as the auxiliary variable $\mathbf{u}_t$; i.e., we assume that the data are segment-wise stationary, which should be approximately true because the modulations were designed to be temporally smooth and continuous. The training and evaluation methods follow those in Simulation 1. For comparison, we also applied NICA based on TCL (NICA-TCL; Hyvärinen and Morioka (2016)). See Supplementary Material E for more details.

Result IIA-TCL performed better than NICA-TCL (Fig. 1b). In addition, even though the innovation model matches IIA-GCL better than IIA-TCL (the modulations are temporally smooth and continuous, and thus not segment-wise stationary), IIA-TCL achieved slightly better performances than IIA-GCL (note that the performances of IIA-GCL is the same as those in Fig. 1a because we used the same data); this finding is consistent with the comparison of NICA-GCL and NICA-TCL by Hyvärinen et al. (2019). As with IIA-GCL, a larger number of data points leads to higher performance (i.e. the method seems to converge), and again, higher complexity of the NVAR models makes learning more difficult. See also Supplementary Material F in the two dimensional case to visually see the difference of the estimation performances.

4.3 Simulation 3: IIA-HMM for Artificial Dynamics with Hidden Markov Process

Data Generation We generated data from an artificial NVAR process with hidden Markov model. The innovations were generated based on hidden Markov chain with modulating the mean and the variance of Gaussian distribution for each state. The observations were then obtained by the same method described in Simulation 1, using the generated innovations. The dimensions of the observations and innovations (n) were 5, and the number of latent states (C) was 11. See Supplementary Material G for more details.

Training We used here EM algorithm to maximize the likelihood for estimating the parameters of the demixing model and the innovation process, as in Hälvä and Hyvärinen (2020). For comparison, we also applied NICA-HMM (Hälvä and Hyvärinen, 2020), IIA-TCL, and AD-NVAR.

Result IIA-HMM performed better than the other baseline methods (Fig. 1c), except for the most complex case ($L = 5$) possibly because of the difficulty of the optimization due to the larger number of parameters. The worse performances of IIA-TCL are likely to be due to the inconsistency between the artificial temporal segments used for the training and the actual sequence of the hidden states, and also much smaller number of latent states compared to the number of the artificial segments. AD-NVAR did not perform well even for the linear case ($L = 1$) because the innovations are not necessarily marginally independent each other this time. As with the previous frameworks, a larger number of data points leads to higher performance, and again, higher complexity of the NVAR models makes learning more difficult.

4.4 Experiments on Real Brain Imaging Data

To evaluate the applicability of IIA to real data, we applied it on multivariate time series of electrical activities of the human brain, measured by magnetoencephalography (MEG). In particular, we used a dataset measured during auditory or visual presentations of words (Westner et al., 2018). Although ICA is often used to analyze brain imaging data, relying on the assumption of mutual independence of the hidden components, the event-related components (such as event-related potentials; ERPs) are not likely to be independent because they may have similar temporal patterns time-locked to the stimulation. However, the innovations generating the components should still be independent because they would be generated by different brain sources, which motivates us to use IIA rather than ICA (see Supplementary Material H for

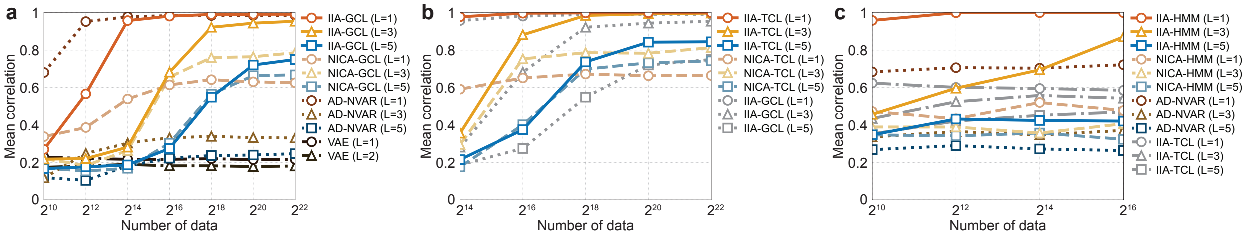

Figure 1: (Simulation) Estimation of the latent innovations from unknown artificial NVAR process by IIA. (**a**) (Simulation 1; IIA-GCL) Mean absolute correlation coefficients between innovations and their estimates by IIA-GCL (solid lines), with different settings of the complexity (number of layers L) of the NVAR models and data points. For comparison: NICA based on GCL (NICA-GCL, dashed line), NVAR with additive innovation model (AD-NVAR, dotted line), and variational autoencoder (VAE, dash-dot line). IIA-GCL generally has higher correlations than the baseline methods. (**b**) (Simulation 2; IIA-TCL) Estimation performances by the IIA-TCL framework (solid lines), evaluated by the same data used in Simulation 1. For comparison: NICA based on TCL (NICA-TCL, dashed line) and IIA-GCL shown in **a** (dotted line). (**c**) (Simulation 3; IIA-HMM) Estimation performances by the IIA-HMM framework (solid lines). For comparison: NICA based on HMM (NICA-HMM, dashed line), NVAR with additive innovation model (AD-NVAR, dotted line), and IIA-TCL (dash-dot line).

the details of the data and settings).

Data and Preprocessing We used a publicly available MEG dataset (Westner et al., 2018). Briefly, the participants were presented with a random word selected from 420 unrelated German nouns either visually or auditorily, for each trial. MEG signals were measured from twenty healthy volunteers by a 148-channel magnetometer (219.1±22.4 trials for each subject; 2,207 auditory and 2,174 visual trials in total for all subjects). We band-pass filtered the data between 4 Hz and 125 Hz (sampling frequency = 300 Hz). The dimension of the data was reduced to 30 by PCA.

IIA Settings We used IIA-TCL for the training, by assuming a third-order NVAR model (NVAR(3)) and the segment-wise-stationarity of the latent innovations. The trial data were divided into 84 equally sized segments of length of 8 samples (26.7 ms), and the segment label was used as the auxiliary variable $\mathbf{u}_t$. The same segment labels were given across the trials; however, considering the possible stimulus-specific dynamics of the brain, we assigned different labels for the auditory and visual trials. In total, there are 168 segments (classes) to be discriminated by MLR. We used MLPs for the nonlinear scalar functions (Eq. 8), and fixed the number of components to 5. We fixed the time interval between two consecutive samples to 3 (10 ms).

Evaluation Methods For evaluation, we performed classification of the stimulus modality (auditory or visual) by using the estimated innovations. The classification was performed using a linear support vector machine (SVM) classifier trained on the stimulation label and sliding-window-averaged innovations obtained for each trial. The performance was evaluated by the generalizability of a classifier across subjects, i.e., one-subject-out cross-validation (OSO-CV). For comparison, we also evaluated NICA-TCL (Hyvärinen and Morioka, 2016) and AD-NVAR(3). We omitted $L = 1$ for IIA-TCL because of the instability of training. We visualized the spatial characteristics of each innovation component by estimating the optimal (maximal and minimal) input $\mathbf{x}_t$ while fixing $\mathbf{x}_{t-1:t-3}$ to zero.

Results Figure 2a shows the decoding accuracies of the stimulus categories, across different methods and the number of layers for each model. The performances by IIA-TCL with nonlinear models ($L \geq 2$) were significantly higher than the other baseline methods ($p < 0.05$; Wilcoxon signed-rank test, FDR correction), which indicates the importance of the modeling of the MEG signals by NVAR, especially with the nonlinear (non-additive) interactions of the innovations.

The left panels of Fig. 2b show the temporal patterns of the innovations during the auditory and visual stimuli. Some components have clear differences between the stimulus modalities, which implies that those components are related to the stimulus-specific dynamics of the brain; e.g., C1 and C2 represent auditory- and visual-relevant innovations, respectively. Such stimulus-specificity can be also seen from the spatial characteristics of the components; C1 is strongly activated by the MEG signals around auditory areas of the brain, while C2 is more activated by the visual areas. C3 seems to represent stimulus-evoked activities on the parietal region caused by both categories. Those results show that IIA-TCL extracted reasonable com-

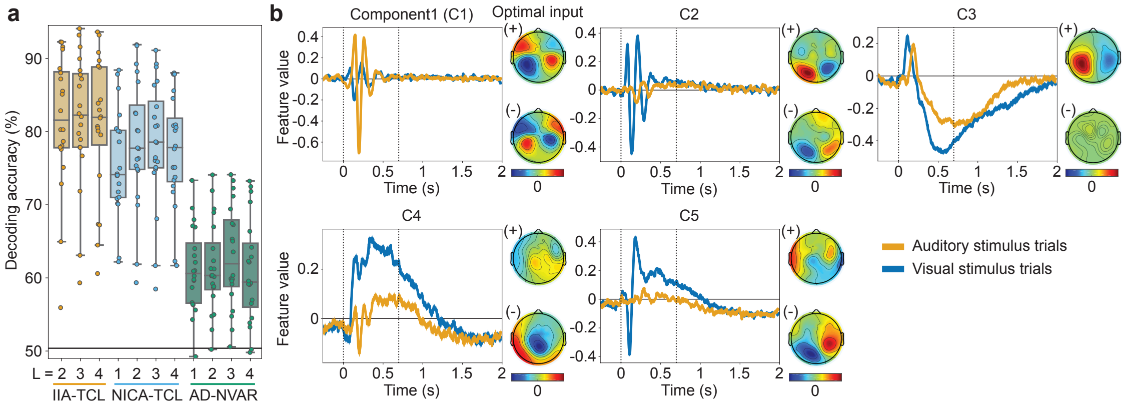

Figure 2: IIA-TCL on the electrical activity data measured by MEG from the human brain during auditory or visual stimuli of German nouns. (**a**) Decoding accuracies of the stimulus category predicted from the innovations extracted by IIA-TCL and the other baseline methods. The performance was measured by one-subject-out cross-validation (OSO-CV), with changing the number of layers L for each method. Each point represents a testing accuracy on a target subject. The black horizontal line indicates the chance level. (**b**) The temporal pattern and the spatial specificity of each component trained by IIA-TCL ($L = 3$). (Left) The temporal patterns of the components averaged separately for auditory (orange) and visual trials (blue). 0 s is the onset of the stimulus, and the latter vertical line represents the average duration of the stimuli. (Right) The spatial topographies of the optimal input (MEG signal; top view) which maximizes (+) and minimizes (−) the component.

ponents (innovations) relevant to the external stimuli automatically from the data in a data-driven manner.

5 DISCUSSION

IIA can be seen as a generalization of the recently proposed NICA frameworks (Hälvä and Hyvärinen, 2020; Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019), with the important difference that observations can have recurrent temporal structure. The theory strictly includes NICA as a special case, since the main assumptions can be satisfied even if the NVAR model (Eq. 1) does not actually depend on $\mathbf{x}_{t-1}$, which corresponds to the instantaneous nonlinear mixture model of NICA: $\mathbf{x}_t = \mathbf{f}_{\text{ICA}}(\mathbf{s}_t)$. This connection can be also seen by comparing the regression functions; by omitting the dependencies of Eqs. 6 and 8 on $\mathbf{x}_{t-1}$, we can obtain the same algorithms as NICA (Hyvärinen and Morioka (2016) with k=1, and Hyvärinen et al. (2019)). This indicates that the regression functions of IIA can learn NICA models as a special case. See Supplementary Material F for the empirical comparison in the two dimensional case.

Applying IIA on time series has some practical advantages compared to NICA. First, autoregressive structures are generally inherent in any kinds of dynamics, and their explicit modeling is beneficial for the estimation. Second, innovations are usually more independent mutually than the processes generated by them, because the independence of processes implies the independence of their innovations, but not vice versa, as argued in the linear case by Hyvärinen (1998). Thus, innovations are likely to give a better fit to any model assuming independence of the latent variables.

While IIA estimates innovations from the observed time series, the NVAR model $\mathbf{f}$ is left unknown, unlike in ordinary VAR analyses. In practice, we can estimate $\mathbf{f}$ after IIA as a post-processing, by fitting a nonlinear function which outputs $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$ and the estimated $\mathbf{s}_t$. Since IIA guarantees the estimation of $\mathbf{s}$ up to a permutation and element-wise invertible nonlinearities, this should be possible if the model to be fitted has universal approximation capability.

6 CONCLUSION

We proposed independent innovation analysis (IIA) as a new general framework to nonlinearly extract innovations hidden in a time series. In contrast to the common simplifying assumption of additive innovations, IIA can deal with a general NVAR model in which innovations are not additive. Any general nonlinear interactions between the innovations and the observations are allowed. To guarantee identifiability, IIA requires some assumptions on the innovations, in particular mutual independence conditionally on an auxiliary variable which also needs to modulate the distributions of the innovations. A typical case would be nonstationary innovations mutually independent at each time point.

We proposed three practical estimation methods. Two of them were based on a self-supervised training of a nonlinear feature extractor by (multinomial) logistic regression. They can thus be easily implemented by ordinary neural network training. The third one is a "purely unsupervised" framework based on maximum-likelihood estimation, specifically applicable when an auxiliary segmentation variable is unobservable (or in practice, we do not want to impose some simple segmentation). The consistency of the estimation is guaranteed up to a permutation and component-wise invertible nonlinearity, which implies the strongest identifiability proof of general NVAR in the literature, by far. IIA can be seen as a generalization of recently proposed NICA frameworks, and includes them as special cases.

Experiments on real brain imaging data by MEG showed distinctive components relevant to the external-stimulus categories. This result suggests a wide applicability of the method to different kinds of time series such as video, econometric, and biomedical data, in which innovation plays an important role.

Acknowledgements

This research was supported in part by JSPS KAKENHI 18KK0284, 19K20355, and 19H05000, and JST PRESTO JPMJPR2028. A.H. was funded by a Fellow Position from CIFAR, the DATAIA convergence institute as part of the "Programme d'Investissement d'Avenir" (ANR-17-CONV-0003) operated by Inria, and the Academy of Finland (project #330482).

References

C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 64–72. 2016.

G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J.L. Cantero. Measuring directional coupling between eeg sources. *NeuroImage*, 43(3):497 – 508, 2008.

L. Gresele, G. Fissore, A. Javaloy, B. Schölkopf, and A. Hyvärinen. Relative gradient optimization of the jacobian term in unsupervised deep learning. In *Advances in Neural Information Processing Systems (NeurIPS2020)*, 2020.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012.

H. Hälvä and A. Hyvärinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. volume 124 of *Proceedings of Machine Learning Research*, pages 939–948. PMLR, 2020.

T. Hastie, R Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, NY, 2001.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

A. Hyvärinen. Independent component analysis for time-dependent stochastic processes. In *ICANN 98*, pages 135–140, 1998.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10(3):626–634, 1999.

A. Hyvärinen. Blind source separation by nonstationarity of variance: a cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, 2001.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 3765–3773. 2016.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.*, 12(3):429 – 439, 1999.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010.

A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *AISTATS*, pages 859–868, 2019.

I. Jeliazkov. *Nonparametric Vector Autoregressions: Specification, Estimation, and Inference*, volume 32, pages 327–359. Emerald Group Publishing Limited, 2013.

M. Kalli and J. E. Griffin. Bayesian nonparametric vector autoregressive models. *Journal of Econometrics*, 203(2):267 – 282, 2018.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

D. P Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR 2014*, 2014.

G. Koop and D. Korobilis. Bayesian multivariate time series methods for empirical macroeconomics. *Found. Trends Econ.*, 3(4):267–358, 2010.

M. Lanne, M. Meitz, and P. Saikkonen. Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics*, 196(2):288 – 304, 2017.

W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *ICLR 2017*, 2017.

A. Moneta, D. Entner, P. O. Hoyer, and A. Coad. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

J. Oh, X. Guo, H. Lee, R. L Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems (NIPS) 28*, pages 2863–2871. 2015.

G. E. Primiceri. Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852, 2005.

Y. Shen, G. B. Giannakis, and B. Baingana. Nonlinear structural vector autoregressive models with application to directed brain networks. *IEEE Transactions on Signal Processing*, 67(20):5325–5339, 2019.

C. A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.

B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 843–852, 2015.

T. Teräsvirta. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89 (425):208–218, 1994.

R. S. Tsay. Testing and modeling multivariate threshold models. *Journal of the American Statistical Association*, 93(443):1188–1202, 1998.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR 2017*, 2017.

B. U. Westner, S. S. Dalal, S. Hanslmayr, and T. Staudigl. Across-subjects classification of stimulus modality from human meg high frequency activity. *PLOS Computational Biology*, 4(3):1–14, 2018.

N. Wichers, R. Villegas, D. Erhan, and H. Lee. Hierarchical long-term video prediction without supervision. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 6038–6046, 2018.