JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Zhu, Yongjie; Parviainen, Tiina; Heinilä, Erkka; Parkkonen, Lauri; Hyvärinen, Aapo

Title: Unsupervised representation learning of spontaneous MEG data with nonlinear ICA

Year: 2023

Version: Published version

Copyright: © 2023 The Author(s). Published by Elsevier Inc.

Rights: CC BY-NC-ND 4.0

Rights url: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the original version:

Zhu, Y., Parviainen, T., Heinilä, E., Parkkonen, L., & Hyvärinen, A. (2023). Unsupervised representation learning of spontaneous MEG data with nonlinear ICA. Neuroimage, 274, Article 120142. https://doi.org/10.1016/j.neuroimage.2023.120142

NeuroImage 274 (2023) 120142

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Unsupervised representation learning of spontaneous MEG data with nonlinear ICA

Yongjie Zhu[a,b,*], Tiina Parviainen[c], Erkka Heinilä[c], Lauri Parkkonen[b], Aapo Hyvärinen[a]

[a] Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland
[b] Department of Neuroscience and Biomedical Engineering, Aalto University, 00076 Espoo, Finland
[c] Centre for Interdisciplinary Brain Research, Department of Psychology, University of Jyväskylä, 40014 Jyväskylä, Finland

ARTICLE INFO

Keywords:
Nonlinear independent component analysis (ICA)
Unsupervised learning
Deep generative model
Resting-state network
Non-stationarity
Neurofeedback
Magnetoencephalography (MEG)

ABSTRACT

Resting-state magnetoencephalography (MEG) data show complex but structured spatiotemporal patterns. However, the neurophysiological basis of these signal patterns is not fully known and the underlying signal sources are mixed in MEG measurements. Here, we developed a method based on the nonlinear independent component analysis (ICA), a generative model trainable with unsupervised learning, to learn representations from resting-state MEG data. After being trained with a large dataset from the Cam-CAN repository, the model has learned to represent and generate patterns of spontaneous cortical activity using latent nonlinear components, which reflects principal cortical patterns with specific spectral modes. When applied to the downstream classification task of audio-visual MEG, the nonlinear ICA model achieves competitive performance with deep neural networks despite limited access to labels. We further validate the generalizability of the model across different datasets by applying it to an independent neurofeedback dataset for decoding the subject's attentional states, providing a real-time feature extraction and decoding mindfulness and thought-inducing tasks with an accuracy of around 70% at the individual level, which is much higher than obtained by linear ICA or other baseline methods. Our results demonstrate that nonlinear ICA is a valuable addition to existing tools, particularly suited for unsupervised representation learning of spontaneous MEG activity which can then be applied to specific goals or tasks when labelled data are scarce.

1. Introduction

Over the last decades, spontaneous (or resting-state) brain activity has attracted a great amount of interest in the neuroscience community. It has been demonstrated to be not random but organized into complicated spatiotemporal patterns, measurable for example by magnetoencephalography (MEG) (Brookes et al., 2011; De Pasquale et al., 2010; Vidaurre et al., 2018) and electroencephalography (EEG) (Mantini et al., 2007). While the origins and the electrophysiological basis of resting-state activity remain largely unclear (Brookes et al., 2011; Mantini et al., 2007), from a purely data-driven perspective, they can be related to individual differences (Becker et al., 2020; Sareen et al., 2021) or neurological diseases (Fox et al., 2014; Zhu et al., 2021), for example.

In a data-driven approach, co-activation across anatomically separated regions, manifested as temporal correlations, can be used to delineate functional networks. More sophisticated methods use the non-Gaussian structure of the data in the form of linear independent component analysis (ICA) or blind source separation. For resting-state MEG recordings, linear ICA finds a number of components, which may reflect head motion, non-neuronal physiology, and unconstrained cognition. (Hyvärinen et al., 2010; Vigário et al., 2000). This is an important step forward since ICA can find components in the MEG recordings that would otherwise be difficult to extract. However, these components explain only part of the resting-state activity and may be entangled not only among themselves but with further activity not found by ICA. Linear separation is, in fact, limited and it may be that we need to consider, for example, correlations of the amplitudes in rhythmic activity (Brookes et al., 2011). While for some known nonlinearities, such as power coherence, we may be able to preprocess the data so that linear ICA is enough, the field would benefit from data-driven methods that can uncover and disentangle any hidden spatiotemporal structures from spontaneous electrophysiological data themselves, without having to assume we know the nonlinearities involved. To be effective, such a method should also be able to infer the components or sources from resting-state MEG data, while being able to account for complicated and nonlinear relationships.

The aforementioned requirements point us to deep learning or representation learning with deep neural networks (LeCun et al., 2015). For

* Corresponding author.
E-mail addresses: yongjie.zhu@helsinki.fi (Y. Zhu), aapo.hyvarinen@helsinki.fi (A. Hyvärinen).

https://doi.org/10.1016/j.neuroimage.2023.120142.
Received 27 September 2022; Received in revised form 26 April 2023; Accepted 27 April 2023
Available online 28 April 2023.
1053-8119/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

brain imaging, deep learning has been increasingly applied as a generic class of machine-learning tools to learn features and classifiers from neuroimaging data (Acunzo et al., 2022; Yan et al., 2022; Zubarev et al., 2019). Most applications are within the scope of supervised learning. Typically, a deep neural network model is trained using neuroimaging data as input to give rise to an output that optimally matches the ground truth for a task, such as brain-age prediction (Cole et al., 2017; Jónsson et al., 2019), emotion recognition (Hsu et al., 2022), and pathology detection (Aoe et al., 2019). However, resting-state data do not include any labels or annotations needed for such supervised learning paradigms. They may be obtained after data collection, e.g., by expert labeling, which is often costly and the number of labels typically remains several orders of magnitude smaller than the data points in the MEG data (Banville et al., 2021; Schirrmeister et al., 2017). In addition, it is uncertain to what extent representations learned for a specific task would be generalizable (i.e., transferable) to other tasks. It is also debatable whether deep neural networks with supervised learning are currently superior to more conventional and simpler methods (He et al., 2020).

However, for MEG and EEG, unlabeled data are available in abundance, in particular in the form of resting-state data. This opens up the possibility of using unsupervised versions of deep learning where no labels are needed. Such methods should uncover the underlying nonlinear sources that drive intrinsic brain activity regardless of any task or stimuli. The recently proposed Nonlinear ICA (Hyvärinen and Morioka, 2016), a nonlinear version of ICA, can be an alternative method for finding such hidden nonlinear sources that generate the data with the unsupervised paradigm. While nonlinear ICA is an ill-defined problem in general, it is identifiable and can be estimated, for example with self-supervised learning (SSL), under certain assumptions. SSL is a particular unsupervised learning approach to learning representations from unlabeled data using some additional structure in the data to provide an artificial supervisory signal. Thus one transforms an unsupervised learning problem into a supervised one, called the 'pretext' task. (Jing and Tian, 2020).

Finding such general-purpose features without labels also opens up the way to semi-supervised learning. This means that we learn the features or components from a large unlabeled data set (e.g., a resting-state activity database) and then use those features for a classification task of interest, such as diagnosis or neurofeedback, with a dataset that has limited labels available. Typically, in the terminology of SSL, there is a 'pretext' and a 'downstream' task. Downstream tasks are tasks that people are actually interested in, but with limited or no annotations (labels). The pretext task, on the other hand, is the core of the SSL approach and it must be sufficiently associated with the downstream task so that similar features should be used to carry it out (Banville et al., 2021). Importantly, it must also be possible to generate the annotations for this pretext task using the large unlabeled data alone. In addition to facilitating the downstream task and/or reducing the number of labeled samples necessary, self-supervised learning can also discover more general and robust representations than those learned in supervised learning with specific tasks or goals (Banville et al., 2021; Van den Oord, Li, and Vinyals, 2018).

In this study, we chose to use the recently proposed nonlinear independent component analysis (ICA) estimated by SSL technique (Hyvärinen and Morioka, 2016; Morioka et al., 2021) for unsupervised deep learning. We apply nonlinear ICA to a very large resting-state MEG data set (Shafto et al., 2014; Taylor et al., 2017), without requiring any labels or narrowly focusing on any downstream task. Nonlinear ICA is a generative model capable of learning the identifiable features that generate the data, and it provides a well-defined and interpretable model. Nonlinear ICA is also promising based on the earlier success of linear ICA in neuroimaging data analysis. Briefly, in this study, two kinds of nonlinear ICA models are used, both using self-supervised learning. One is a basic nonlinear ICA version based on time contrastive learning (TCL) building on non-stationarity (Hyvärinen and Morioka, 2016). The other one is independent innovation analysis (IIA) which extends TCL for temporally correlated time series (Morioka et al., 2021). To enable group-level analysis, we further propose group nonlinear ICA for resting-state MEG from multiple subjects based on a multi-task learning scheme. Specifically, nonlinear ICA is trained using a group-shared nonlinear feature extractor which outputs a set of feature values from a morphed parcel time series of each subject, so as to optimize the self-supervised classification performance of subject-specific multinomial logistic regression (MLR) classifiers (Fig. 1B). We then characterize the spatiotemporal and spectral profiles of the latent components learned from resting-state MEG. We show how the representation enables high performance in a simple down-stream classification of visual/auditory decoding task in MEG data from the same population. Lastly, we demonstrate the capability of transferring the representation to a different dataset from a different subject population: we validate the use of the features given by the trained nonlinear ICA model showing how they lead to superior decoding accuracy of attentional states from ongoing MEG data (Zhigalov et al., 2019).

2. Materials and methods

2.1. Overview of the methods

Typically, a generalizable system for representation learning of brain imaging data such as MEG consists of a base module and additional projection modules (Banville et al., 2021; Kim et al., 2021). The base module is trained with unsupervised or self-supervised learning from task-free resting-state (spontaneous) MEG. Therefore, the base module is not tailored to any specific purpose, such as brain age prediction or pathological detection, or any specific task related to cognitive activity. After training, the base module is supposed to be capable of generalization to MEG data in different cognitive task conditions or downstream (classification) tasks, in particular through additional projection modules. The representation learned by the base model can be used by the projection modules which are trained to meet a specific task by supervised learning. It is expected that the base module is designed and trained with a deep architecture to leverage a large number of unannotated (unlabeled) data, whereas the projection modules can be shallow and trained with limited labeled data. Such a semi-supervised scheme makes the learning more efficient since unannotated data are much more abundant than annotated data.

In such a scheme, nonlinear ICA (Hyvärinen and Morioka, 2016, 2017; Morioka et al., 2021) offers a suitable model for the initial part of the base module, which is composed of feature extractor or additional linear unmixing matrix as Fig. 1 shows. The feature extractor is learnable with self-supervised learning and the linear unmixing matrix is learnable with unsupervised learning without any label data (Fig. 1B). Unsupervised learning with nonlinear ICA can leverage the ever-increasing amount of resting-state MEG data (Larson-Prior et al., 2013; Nisoet al., 2016; Shafto et al., 2014). The latent components or representations extracted from nonlinear ICA can be input to specific projection modules (e.g., linear SVM) to facilitate downstream tasks such as classification of cognitive task and decoding the mental state (Fig. 1C).

2.2. Nonlinear independent component analysis

Nonlinear ICA is a recently proposed unsupervised learning framework based on a nonlinear generalization of the well-known basic ICA (Hyvärinen and Morioka, 2016). It promises a principled approach to representation learning, for example using deep neural network, and attempts to find nonlinear components, in multidimensional data. In the following, we briefly explain the basic principles of nonlinear ICA, but we emphasize that we use algorithms that have been previously published by Hyvärinen and Morioka (2016), Morioka et al. (2021) and refer the reader to those publications for a detailed description.

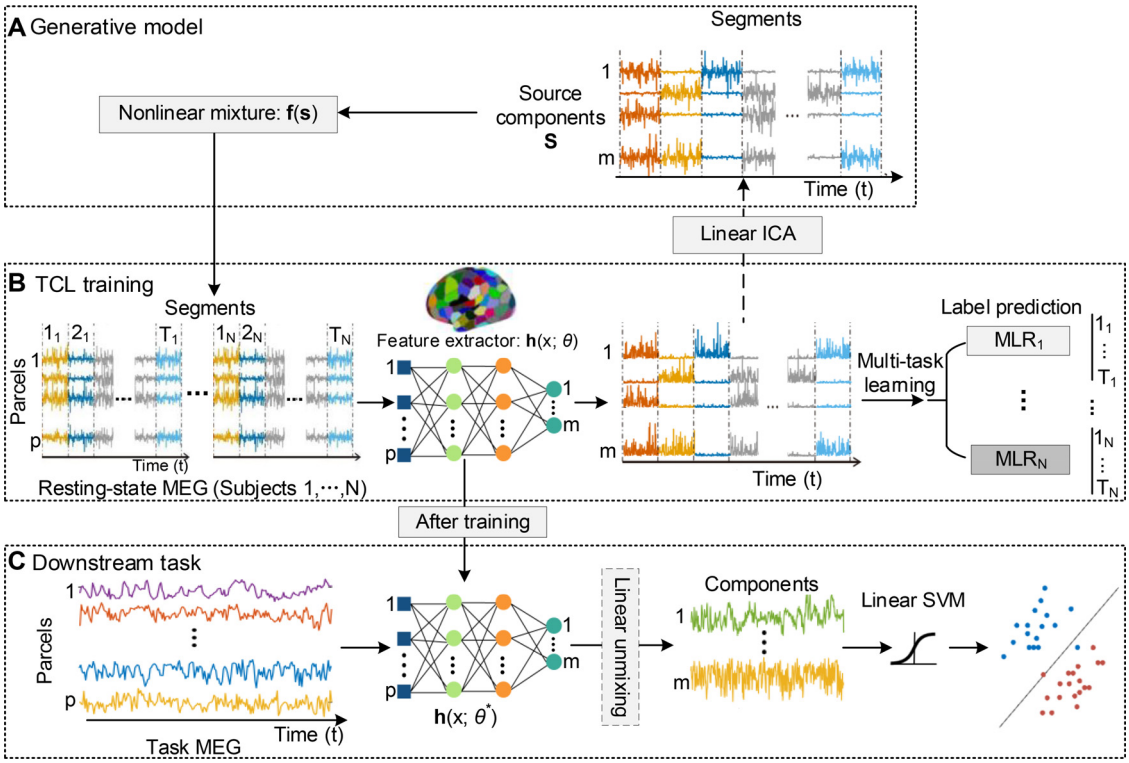

Fig. 1. The proposed approach. **A.** The generative model is basically a nonlinear version of ICA. The observed time series are given by a nonlinear transformation of the components, which are mutually independent, and have segment-wise stationarity, which is a simple form of non-stationarity. **B.** In TCL, we attempt to train a feature extractor $\mathbf{h}(\boldsymbol{x}, \boldsymbol{\theta})$ to be sensitive to the nonstationarity of the data by using a multinomial logistic regression which attempts to discriminate between the segments, labelling each data point with the segment label 1,…,T (as pretext task) for each subject (e.g. $\tau = 1_i, \ldots, T_i$ with subject i). The segments from all subjects are fed into the feature extractor. Note that different subjects are likely to show uninteresting technical differences in group-level analysis. Therefore, we apply a multi-task (multi-subject) learning scheme, which includes a separate top-layer MLR classifier for each subject, but a shared feature extractor (here, a multilayer perceptron, MLP). After training, feature extractor was followed by a linear ICA to resolve the linear indeterminacy and finally obtain components up to point-wise nonlinearities such as squaring. **C.** The feature extractor $\mathbf{h}(\boldsymbol{x}, \boldsymbol{\theta}^*)$ trained on big unlabeled data and the linear unmixing matrix, as a base module, are applied to downstream tasks with label-limited data.

In general, nonlinear ICA assumes a generative model (Fig. 1A):

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) \tag{1}$$

where $\mathbf{x}(t)$ is the observed n-dimensional data point at time t, $\mathbf{f}$ is an invertible and smooth mixing function to be learned from the data, $\mathbf{s}(t)$ is the m-dimensional vector of independent components $s_i(t)$, and m being the number of components.[1] The time series s_i are presumed to be mutually independent. While nonlinear ICA is an ill-defined problem in general (Hyvärinen and Pajunen, 1999), recent work has proposed an identifiable solution using some additional auxiliary information about the data, such as their temporal structure (Hyvärinen and Morioka, 2016, 2017).

The estimation of nonlinear ICA is to find out the underlying sources (components) $\boldsymbol{s}(t)$ in Eq. (1) by learning an unmixing function $\mathbf{g}$ (i.e., $\mathbf{g} = \mathbf{f}^{-1}$) such that

$$\boldsymbol{s}(t) = \mathbf{g}(\boldsymbol{x}(t)), \tag{2}$$

which can be solved (learned) by time-contrastive learning (TCL) method based on the assumption that the sources are independent and temporally non-stationary (see Section 2.2.1).

[1] While the basic form of nonlinear ICA based on time-contrastive learning (TCL) considers $m = n$, it is also possible to have $m<n$ in TCL, as pointed out at the end of Supplementary Material ("Dimension Reduction") of Hyvärinen and Morioka (2016). In practice, this is simply accomplished by having a smaller number of features in the neural network (feature extractor), and no other changes to the procedure are needed (Fig. 1B).

2.2.1. *Time-contrastive learning*

Time-contrastive learning (TCL) is a novel self-supervised learning technique capable of estimating the nonlinear ICA model based on a non-stationarity assumption (Hyvärinen and Morioka, 2016). The starting point in TCL is to assume that each component is non-stationary, which makes the problem well-defined and the model identifiable. The intuitive rationale is that if the components are non-stationary, forcing them to be independent at every segment gives many more constraints than just a single, global constraint of independence as in ordinary linear ICA based on non-Gaussianity. That justifies to some extent why we obtain identifiability. We also note that the definition of the model imposes non-stationarity on the components, but this necessarily implies that the data are non-stationary as well, since the mixing is purely spatial. The non-stationarity is assumed to be much slower than the sampling rate; in other words, the time-series can be divided into segments in each of which the distributions are approximately constant; but crucially, the distribution is different across segments because of the non-stationarity. Accordingly, TCL assumes conditional (segment-wise) independence of the components, instead of marginal independence assumed in ordinary ICA. It was proven that such temporal structure, called time-segment-wise stationarity, enables the estimation of the source signals up to component-wise nonlinearities (Hyvärinen and Morioka, 2016).

Assume the data has been segmented, by a method to be specified. Denote the segment index by $\tau = 1, \ldots, T$, where T is the number of segments, the statistical distribution of each underlying component s_i within each segment can be modelled as an exponential family:

$$\log p_\tau(s_i) = \log q_{i,0}(s_i) + \sum_{j=1}^{k} \lambda_{i,j}(\tau) q_{i,j}(s_i) - \log Z_i(\lambda_{i,1}(\tau), \ldots, \lambda_{i,k}(\tau)), \tag{3}$$

where p_τ is the probability density function (pdf) of segment τ, $q_{i,0}$ is a stationary base density and $q_{i,j}$ with $j \geq 1$ are the sufficient statistics for the exponential family of the component s_i (the index t is dropped for simplicity), and Z_i is the normalization constant. Note that the parameters $\lambda_{i,j}(\tau)$ of source s_i depend on the segment index τ,which creates the nonstationary sources. TCL learns the inverse transformation $\mathbf{g} = \mathbf{f}^{-1}$ (in Eq. (2)) with SSL paradigm. The pre-text task in such SSL is to classify original data points with the corresponding segment indices used as class label, using multinomial logistic regression (MLR). For this reason, TCL adopts a deep neural network including a feature extractor $\mathbf{h}(\boldsymbol{x}(t), \boldsymbol{\theta})$ followed by MLR for label prediction, where $\boldsymbol{\theta}$ is the neural network weights. Thus, it seems intuitively clear that in order to optimally classify observations $\boldsymbol{x}(t)$ into their corresponding segment labels τ, the feature extractor $\mathbf{h}(\boldsymbol{x}(t), \boldsymbol{\theta})$ needs to learn a useful representation of the temporal structure in the underlying distribution of latent sources.

The theory of TCL indicates that the method can learn the latent components $\boldsymbol{s}(t)$ up to pointwise nonlinearities given by the q in Eq. (3) (q is typically squaring or absolute values function) and a linear transformation $\mathbf{A}$; that is, for example, $\boldsymbol{s}(t)^2 = \mathbf{A}\mathbf{h}(\boldsymbol{x}(t))$, where $\mathbf{h}(\boldsymbol{x}(t))$ is the feature extractor to be learned by TCL just like de-mixing the nonlinear part of Eq. (1) and $\mathbf{A}$ is the linear mixing part left in the model (see Fig. 1B). In other words, we can obtain the $\boldsymbol{s}(t)$ up to point-wise squaring by learning the nonlinear unmixing function $\mathbf{g}$ via TCL training (the nonlinear part, $\mathbf{h}(\boldsymbol{x}(t))$) and a further linear ICA (the linear part, $\mathbf{A}$). This is quite surprising since the self-supervised method makes no reference to independent components. A further linear ICA can estimate the remaining linear mixing $\mathbf{A}$ if the number of segments increases to infinity and the data distributions of segments are random in a certain sense. Therefore, the theory proves that TCL (with a further linear ICA) is consistent in the sense of estimation theory: when the amount of data points grows infinitely, the method finds out the right independent components up to the point-wise nonlinearities. This statistical theory assumes that the optimization does not fail by getting trapped in a local optimum; however, this is a typical practical problem in deep learning which is addressed further in our discussion.

In more detail, the self-supervised TCL algorithm proceeds as follows: (1). Divide a multivariate time series $\mathbf{x}(t)$ into segments, i.e. time windows, indexed by $\tau = 1, \ldots, T$. Any temporal segmentation method can be used, e.g. simple equal-sized bins. For multi-subject learning, time series are labeled separately for each subject, e.g. $\tau = 1_i, \ldots, T_i$ with subject i; (2). Associate each data point with the corresponding segment index τ in which the data point is contained; i.e. the data points in the segment τ are all given the same segment label τ; (3). Learn a feature extractor $\mathbf{h}(\boldsymbol{x}; \boldsymbol{\theta})$ together with MLRs to classify all data points with the corresponding segment labels τ used as class labels, as defined above. These procedures are demonstrated in the Fig. 1B. The purpose of the feature extractor is to extract a feature vector that enables the MLRs to discriminate the segments. Therefore, it seems intuitively clear that the feature extractor needs to learn a useful representation of the temporal structure of the data, in particular the differences of the distributions across segments. Note that different subjects are likely to show uninteresting technical differences for group-level analysis. We thus apply a multi-task (multi-subject) learning scheme, which includes a separate top-layer MLR classifier for each subject, but a shared feature extractor (Fig. 1B). After TCL training is finished, the extracted components are followed by linear ICA, which is applied to disentangle the linear indeterminacy left by the feature extractor part of TCL, finally giving the estimates of the independent components $\mathbf{s}(t)$ up to point-wise nonlinearities such as squaring. In other words, the inverse of function $\mathbf{f}$ (i.e., $\mathbf{g}$ in Eq. (2)) is learned via TCL with further linear ICA, corresponding to feature extractor $\mathbf{h}(\boldsymbol{x}; \boldsymbol{\theta})$ and additional linear unmixing matrix. Such basic nonlinear ICA with TCL estimation is called NICA(TCL) in the current study.

2.2.2. Independent innovation analysis

Independent innovation analysis (IIA) can be considered an extension of the nonlinear ICA framework for a nonlinear vector autoregressive process where instead of the actual times series, their innovations at each time point are decomposed to independent components (Morioka et al., 2021). This general model (a first-order autoregressive process is assumed here merely for simplicity of exposition) can be written as:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t-1),\ \mathbf{s}(t)) \tag{4}$$

where $\mathbf{x}(t-1)$ is the time-delayed time series of $\mathbf{x}(t)$, and $\mathbf{s}(t)$ represents the independent innovations (components).[2] The IIA model in Eq. (4) can be transformed into a form similar to Eq. (1) as:

$$\begin{bmatrix} \boldsymbol{x}(t) \\ \mathbf{x}(t-1) \end{bmatrix} = \tilde{\mathbf{f}}\left(\begin{bmatrix} \mathbf{s}(t) \\ \mathbf{x}(t-1) \end{bmatrix}\right) \tag{5}$$

where $\tilde{\mathbf{f}}$ is the augmented model, which includes the original 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. Thus, this augmented model can then be solved using nonlinear ICA theory. The nonlinear vector autoregressive model provides an appealing framework to analyze multivariate time series obtained from a nonlinear dynamical system. In this study, we use a TCL-based version of IIA (Morioka et al., 2021), which we call NICA(IIA). It can guarantee the identifiability of the innovations with arbitrary nonlinearities, up to a permutation and component-wise invertible nonlinearities under certain assumptions. Note that there is no linear indeterminacy left after TCL for IIA and no further linear ICA is thus required after IIA has been performed (Fig. S1).

2.3. Datasets

We use a large MEG data set from the Cam-CAN repository to train nonlinear ICA models, which can learn an optimal representation of data. We then explore and visualize the latent representation. Next, in an investigation of the transfer of the features to other data sets, we validate the generalization of the representations on an independent neurofeedback dataset.

2.3.1. Cam-CAN dataset

We primarily analyzed data from the open-access Cambridge Center for Aging Neuroscience (Cam-CAN) repository (see Shafto et al. (2014), Taylor et al. (2017)) for details of the dataset and acquisition protocols). Specifically, we used the resting-state and passive audio-visual task MEG data from 652 healthy subjects (male/female = 322/330, mean age = 54.3 ± 18.6, age-range 18–88 years), and their structural (T1-weighted MRI) neuroimaging data for source reconstruction. The MRI images were acquired with a 3T Siemens of scanner with a 32-channel head coil. The MEG data were recorded using a 306-channel Elekta Neuromag Vectorview (102 magnetometers and 204 planar gradiometers; MEGIN Oy, Helsinki, Finland) at a sampling rate of 1 kHz. For the 9-min resting-state measurement, subjects were asked to lie still and remain awake with their eyes closed. We discarded the initial and final 30 s and used the remaining 8 min of data from each subject for further analysis and model training.

In the task MEG, subjects were presented with 120 trials of unimodal stimuli (60 visual stimuli: bilateral/full-field circular checkerboards; 60 auditory stimuli: binaural tones at one of three equiprobable frequencies) at a rate of approximately 1 per second. Following exclusions (e.g.,

[2] Similar to TCL on nonlinear ICA, the basic form of IIA based on TCL considers $m = n$, it is also possible to have $m<n$, and this is simply accomplished by having a smaller number of features in the feature extractor (Fig. S1B).

subjects that did not have both MRI and MEG data, unsatisfactory preprocessing results such as failure to remove cardiac and ocular artifacts, and/or failure to extract the cortical surface for source reconstruction), a final dataset of 610 subjects was retained for further analysis.

2.3.2. Mental-states dataset

As an example of the utility of the nonlinear ICA performed, we analyzed a dataset (also recorded by a similar MEG device) from our earlier study (Zhigalov et al., 2019), where we attempted to decode attentional states – mindfulness or wandering thoughts – from ongoing MEG brain activity. Briefly, we used MEG data from 24 healthy subjects (9 females, 15 males, 27 ± 5.5 years (mean ± SD)), previously recorded in the following experimental paradigm. After a 2-min rest period, participants were instructed to perform one of the tasks while undergoing MEG. The tasks were organized into 2-min blocks in a counterbalanced order and the participants performed each task four times in a single session. The session ended with a 2-min rest block. Two sessions per participant were conducted with a 5-min break between the sessions.

Due to the great difficulty of experimentally inducing mind wandering, the experimental data we used here was based on "simulated mind wandering", which means tasks that create brain activity that is not very different from mind wandering. The tasks were mindfulness meditation (MF), reflection on future planning (FP) and reflection on anxiousness-inducing emotional pictures (EP). In particular, the conditions of reflection on future planning (FP) and reflection on anxiousness-inducing emotional pictures (EP) are simulating mind wandering. In all tasks, subjects were instructed to sit still, fix the gaze on the crosshair, and perform a task after a short (7 s) visual instruction. The visual instruction was shown at the beginning and at the middle of each task to keep subjects' attention. Based on these responses, the subjects' focus was considered reasonably good by (Zhigalov et al., 2019), but we do not analyze them any further.

For the mindfulness meditation task, the participant was instructed to focus attention on the sensations of breathing and move the focus of attention back to the task if mind-wandering occurs. For future planning and anxiety-inducing tasks, the participant individually selected 16 (out of 40) relevant pictures prior to the experiment. In the future planning task, the participant was asked to perform planning related to the picture, presumably following the ensuing chains of thought and keeping his/her mind busy. The anxiety-inducing task was similar to the future planning, but instead of neutral pictures, disturbing, scary, disgusting, or other unpleasant pictures were presented to the participant.

2.4. Preprocessing

Since the Cam-CAN and our mental-states MEG data were both recorded by 306-channel Elekta Neuromag Vectorview, the preprocessing was similar. The MaxFilter software (MEGIN Oy, Helsinki, Finland) with temporal signal space separation (tSSS) was applied to suppress external magnetic interference and compensate for head movements (Taulu and Simola, 2006). Thereafter, the MEG data were processed using the open-source software MNE-Python (Gramfort et al., 2014). MEG data were band-pass filtered to 0.1–40 Hz, and resampled at 256 Hz, as a wider band and higher sampling rate did not significantly improve the performance while increasing computational time. Cardiac and eye movement artifacts were identified using the FastICA algorithm implemented in MNE-Python and automatically classified by comparing the ICA components with the simultaneously recorded electrocardiography (ECG) and electrooculography (EOG) signals (Jas et al., 2018). The FreeSurfer software (http://surfer.nmr.mgh.harvard.edu/) was used for volumetric segmentation of MRI data, cortical surface reconstruction and flattening, cortical parcellation, and neuroanatomical labeling with the Schaefer atlas (Schaefer et al., 2018).

The MNE software was used to create head conductor models and cortically constrained source space based on the anatomical information provided by FreeSurfer, for the MEG-MRI co-registration, and for the preparation of the forward and inverse operators. The sources were constrained within the cortex and assumed to be perpendicular to the local cortical surface. The reconstructed cortical surface was decimated to 4098 evenly distributed vertices per hemisphere with 4.9 mm spacing. Depth-weighted L2-minimum-norm estimate was computed for all current dipoles with a loose orientation of 0.2. The noise covariance matrix was estimated from the empty-room recordings and the inverse solution was noise-normalized. The cortex and thus the source space was divided into 400 Schaefer-parcels for each subject. For each parcel, we performed a principal component analysis to extract spatially orthogonal components that describe the activity, ordered by amount of variance explained. We selected the first principal component as a representation of the parcel's time course of activity. For group-level analysis, the subjects' parcel time series were morphed into a standard atlas, and then temporally concatenated across subjects. For the neurofeedback dataset, the cortical sources were reconstructed using a similar pipeline but with an averaged MRI template since we did not obtain the individual structural MRI of these subjects.

2.5. Feature extractor and nonlinear ICA training

We used a multilayer perceptron (MLP) as the feature extractor that takes a single point in the parcel time-series as input and nonlinearly extracts component activity, for both NICA(TCL) and NICA(IIA). The network consists of concatenated (stacked) hidden layers each followed by nonlinear activation units.

NICA(TCL) settings: We set the number of layers as $L = 3$ since our preliminary study showed that a three-layer network gives optimal classification accuracies during the training. We fixed the number of nodes as 80 in the first hidden layer, and 40 in the second hidden layer. The number of nodes in the output layer was equal to the number of components. We tried different values for the number of nodes in the output layer to examine the effect of the number of components. We used rectified linear unit (ReLU) as the activation function in the middle layers, and an adaptive Maxout unit exclusively for the output layer. Maxout unit was constructed by taking the maximum across two affine fully-connected weight groups. To prevent overfitting, we applied dropout and batch normalization to hidden layers. The MLRs follow the feature extractor. Its goal is to predict segment labels from the activities of components extracted by the feature extractor.

NICA(IIA) settings: In addition to MLP for feature extractor ($\mathrm{feature_MLP}$), we used another MLP that took the time delayed ($x_{t-1:t-3}$) of the parcel time series as the input, which was called ϕ_MLP, since IIA included a recurrent structure of the observations in the model (Morioka et al., 2021). Here, a third-order autoregression model was assumed based on our preliminary experiments (Fig. S7). Regarding the high temporal resolution of M/EEG signals (fs=256), we fixed the time lag between two consecutive samples to 8 (31 ms), which means $x(t-1)$ took the value by shifting 8 samples of x (e.g., $x(t-8)$). The architectures of the $\mathrm{feature_MLP}$ and ϕ_MLP were the same as that in NICA(TCL) except that the point-wise square nonlinearity was applied in the last layer. In IIA model, the MLRs took the inputs that were the weighted squared sums of the output units of the $\mathrm{feature_MLP}$ and ϕ_MLP, respectively (see Fig. S1 for demonstration).

To perform the time-contrastive learning, we segmented the parcel time series (data size: 400 parcels × number of time points) into equal size of 3 s (768 data points) for each subject. We first tested the effect of the segment length on the downstream task and 3-s segment length seemed to provide a relatively optimal classification accuracy in the audio-visual MEG data. (Fig. S2). The feature extractor took 400×1 values at single time points as one sample, and was followed MLRs that gave the segment label. Since different subjects were likely to have uninteresting technical differences in terms of the group-level training, we applied a multi-task (multi-subject) learning scheme, which included a separate top-layer MLR classifier (subject-specific) for each subject, but a shared feature extractor (Fig. 1B).

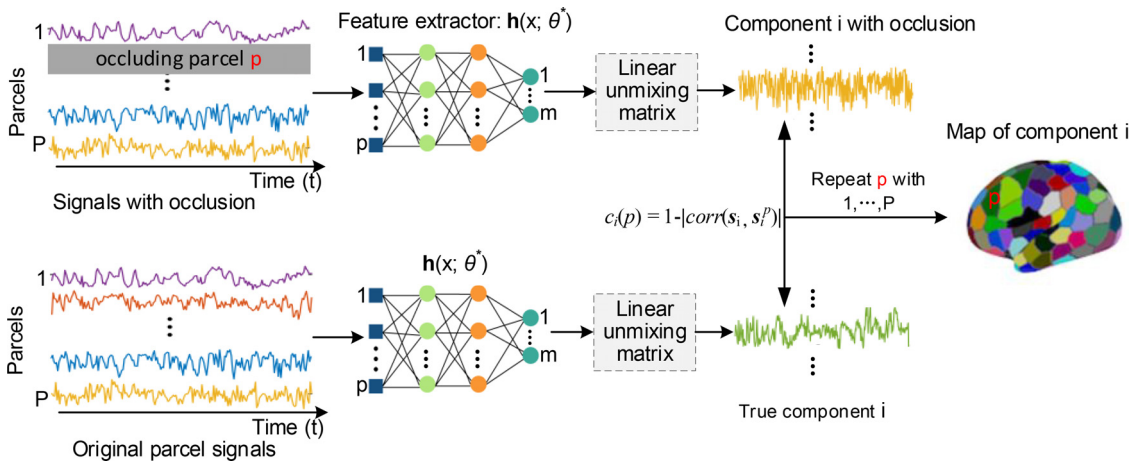

Fig. 2. Illustration of the occlusion sensitivity analysis. The p-th parcel signal was set to zero and the time series were feed into well-trained feature extractor $\mathbf{h}(\boldsymbol{x}, \boldsymbol{\theta}^*)$ (followed by an unmixing matrix from linear ICA if the model was NICA(TCL)) and obtained the components with occlusion. The true components were obtained by the original time series. The maps were computed according to Eq. (6).

The networks (feature extractor and the MLRs) were trained by stochastic gradient descent (SGD) based on back-propagation, which is commonly used in deep learning studies. The other training parameters were set as follows: Initial learning rate of 0.01, learning rate decay of 0.1, momentum of SGD of 0.9, and mini-batch size of 256. The initial weights of each layer were randomly drawn from a uniform distribution. The loss function is the cross entropy between the predicted and true segment label. After TCL training is finished, fastICA with the default setting in scikit-learn (Pedregosa et al., 2011) is applied to the outputs of the feature extractor in the NICA(TCL) model, but not used in the NICA(IIA) model. The training was performed only from resting-state MEG data in the Cam-CAN repository; the audio-visual MEG data were later used to evaluate the generalizability of the trained network. Furthermore, the neurofeedback data were used to test the ability to transfer the nonlinear ICA models to a new dataset, as an example of the potential applications.

2.6. Visualizing the spatial, spectral and temporal profiles of representations

For examining the resulting components (i.e., after linear ICA on the output of feature extractor for NICA(TCL) and output of feature extractor for NICA(IIA)), we extracted components from the dataset by applying the trained feature extractor to resting-state sessions. We then performed spectral analysis on the components across the segments to see their spectral profiles. The power spectral density (PSD) was estimated for each segment by a discrete Fourier transform, after standardizing the signal within the segment to have zero-mean and unit variance. To examine the spatial characteristics of the components, spatial profiles were obtained by examining the sensitivity of feature extractor outputs to the occlusion parcel signals. This method is commonly used in deep learning studies for computer vision to visualize the input specificities of the neural network by systematically occluding different regions of the input image with a grey square, and monitoring the output of the neural networks (Zeiler and Fergus, 2014). The idea is that the probability of the correct class would drop significantly when the actual object was occluded, in an image classification task. One could visualize the heatmap of the probability of the correct label as a function of the occluded position in the image space. Different from the image classification context, however, we here attempted to examine the effect of each parcel signal on the components of the time series, and thus computed the correlation between the 'true' components (outputs without occlusion) and the components with occlusion, instead of the classification probability with occlusion. Specifically, for each component, we obtained the cortical heatmap of the correlation as a function of the parcel position. The correlation with the 'true' components would drop significantly when the occluded parcel was important for the disentanglement of the corresponding component. Intuitively, the cortical heatmaps of the contribution were calculated as:

$$c_i(p) = 1 - \left|corr\left(\boldsymbol{s}_i, \boldsymbol{s}_i^p\right)\right| \tag{6}$$

where $\boldsymbol{s}_i$ is the i-th component without occlusion, and $\boldsymbol{s}_i^p$ is the occluded component by occluding the p-th parcel input (set to zero). $corr()$ indicates the Pearson correlation, and $|()|$ is the absolute value (Fig. 2). $\boldsymbol{c}_i$ presents the spatial pattern of i-th component, which might be considered as a nonlinear version of co-activation patterns for linear ICA.

We also visualized the segment-wise band-limited power (variances) to see the temporal non-stationarities.

2.7. Downstream task of classification from audio-visual MEG

We used the trained model for classification of the stimulus modality in the audio-visual task MEG data to evaluate the performance of the generalizability to the downstream task. Specifically, the audio-visual MEG data were cropped into epochs from −300 ms to 500 ms after each stimulus onset. After source reconstruction, the parcel time series were fed into the feature extractor to obtain the components (Fig. 1C). Thereafter, the classification was performed using a linear support vector machine (SVM) classifier as implemented in scikit-learn (Pedregosa et al., 2011), which was trained on the stimulus labels and sliding-window-averaged components (width=20 and stride=8 samples) obtained for each epoch. Note that the features input to the classifier are obtained by sliding-window-averaged components. We here perform epoch-wise decoding (not sample-wise decoding), which is not sensitive to the length of the pre-stimulus period and can significantly improve the accuracy of the classification for event-related data (Schirrmeister et al., 2017). The performance was evaluated by the generalizability of a classifier across subjects, i.e., one-subject-out cross-validation (OSO–CV).

2.8. Comparison with fully supervised and other self-supervised learning

Nonlinear ICA models were compared to baseline approaches on the downstream tasks including other self-supervised learning (SSL) and purely supervised learning methods. For the baseline SSL, we adopted the temporal context prediction tasks Relative Positioning (RP) for feature learning in the pretext task (Banville et al., 2021), which is closely related to nonlinear ICA. We adopted StagerNet, a four-layer convolutional neural network, as the feature extractor (embedder) in SSL

(Banville et al., 2021). For the purely supervised learning, we used the EEGNet as the deep learning architecture (Lawhern et al., 2018), as it has shown to perform well for epoch-wise classification across participants. SSL(RP) was trained on resting-state data like nonlinear ICA. For the main hyperparameters of SSL(RP), window length, τ_{pos} and τ_{neg} (controlling the size of the "positive" and "negative" contexts, respectively), we first tested the performance of the downstream task by setting different parameter values and chose the values with close-to-optimal classification accuracies based on the instructions in the original paper with plain cross-validation. We finally set window length to 2 s, τ_{pos}= 5 s and τ_{neg}= 120 s. The purely supervised case was directly trained on the audio-visual task data, i.e., it had access to the labeled data.

2.9. Generalization to mental-states dataset

We validated the generalization or transfer of the nonlinear ICA components using neurofeedback data as an independent dataset (Zhigalov et al., 2019), where we attempted to decode attentional states (mindfulness or wandering thoughts) from ongoing brain activity measured by MEG. Similar to the downstream task on audio-visual data, we extracted the nonlinear components by applying the feature extractor trained on resting-state data from Cam-CAN repository on neurofeedback signals (Fig. 1C). For classification, we performed epoch-wise decoding, which means the averaged squared activities of components during each non-overlapping 2 s epoch were used as feature vectors for the linear SVM classifier. We further used linear feature extraction methods (linear ICA and PCA) and fully supervised deep learning for comparison. The FastICA and PCA methods were adopted as the traditional linear baseline feature extractor, and 15 components were also extracted for a fair comparison of the nonlinear ICA. The linear components were divided into 4-s epochs with 75% overlap and the epochs were Fourier-transformed and the spectra divided into four frequency bands: delta (1–3 Hz), theta (4–7 Hz), alpha (8–12 Hz), and low-beta (13–24 Hz). The amplitude spectra averaged (across frequencies) inside these four frequency bands were then used as features for classification, which is a baseline method provided by our prior study (Zhigalov et al., 2019). For the fully supervised method, we adopted the StagerNet as the supervised deep learning architecture, which was shown to suit well for window-wise classification of ongoing data in sleep staging (Banville et al., 2021; Chambon et al., 2018). Different from the 30 s window-wise decoding in sleep staging, we here set a 4-s window as one sample since we had a relatively short data duration.

2.9.1. Classification methods: individual vs. group classification

We used two scenarios to train and test the SVM classifiers. In the first scenario of "individual classifier", we trained the classifier using individual data from the first session and tested the classifier using data from the second session. In the second scenario of "group classifier", we trained the classifier using data from both sessions and all subjects except one "testing" subject and tested the classifier using the testing subject's data from the second session. The second scenario is more challenging, essentially providing information on the generalizability of the classifier across subjects. We applied the features extracted from all methods to investigate whether and how it is possible to discriminate (decode) between mindfulness meditation (MF), future planning (FP), and reflection on anxious-inducing emotional pictures (EP) tasks.

2.9.2. Statistical analysis of classification accuracies between methods

To assess the statistical differences between the classification accuracies for different tasks or for different decoding methods, we applied the Wilcoxon signed-rank test with FDR correction. Specifically, for each subject, we have one classification accuracy value for each method since we use leave one out cross-validation for the SVM classifiers. All the subjects' accuracies were fed into the Wilcoxon rank sum test.

3. Results

3.1. Nonlinear ICA learns representations that show spectra-specific brain patterns

The feature extractors learned component-specific, spatiotemporal elements, so as to disentangle the MEG data into latent nonlinear components. The learning was based on logistic regression in a "self-supervised" scheme where class labels of time series fragments were defined based on temporal segments. To understand what kind of representation was captured from data by NICA(IIA), we visualized the spatiotemporal and spectral patterns for each component (Fig. 3). The spatial patterns were computed by occlusion sensitivity analysis (Fig. 3A), which quantified the contribution to the disentanglement of the corresponding component. The obtained spatial patterns could be considered as co-activation patterns in the nonlinear case. We also computed representative frequency spectra by taking the average of the spectra of the segment-wise components (Fig. 3B). The temporal profiles were obtained by taking the segment-wise band-limited power (variances), exposing the temporal non-stationarities (Fig. 3C). Here, we just show 5 of the 15 components for simplicity; the others are shown in the supplementary material (Fig. S3) and the components extracted by linear ICA are also shown in the supplementary material (Fig. S9) for comparison. These representative patterns, learned from resting-state data, seem to exhibit frequency-specific co-activation patterns similar to previously reported brain networks (Brookes et al., 2011; Vidaurre et al., 2018). Components I, II and IV had similar spectral modes that peaks in around 25 Hz. The brain activities of I and II have similar but distinct spatial distributions, especially around primary somatosensory and motor cortices, which represented the beta-specific motor networks reported earlier (O'Neill et al., 2017; Ramkumar et al., 2014). Component IV shows a distribution spanning frontal, parietal and right temporal cortices exhibiting a beta-dominated spectrum, which may be related to the beta-dependent fronto-parietal networks (Brookes et al., 2011). Components III and V have a spectrum peaking at approximately 10 Hz. Component III exhibits local maxima in the primary and secondary frontal areas, and posterior parietal cortices, similar to IV but different spectral mode. Component V shows a strong visual-cortex pattern with an alpha-dominated spectral features. Additionally, the temporal profiles of the components demonstrate fluctuations in time (Fig 3C), which is consistent with the non-stationarity assumption of the time-contrastive learning.

3.2. Nonlinear ICA facilitates downstream task with limited labeled data

To examine whether nonlinear ICA trained on resting-state session data reduces the need for labeled task-session MEG data, we applied the trained feature extractors to audio-visual task MEG data. We compared their performance on this downstream task to one of the various established approaches such as fully supervised learning, while varying the number of labeled samples available. Downstream task performance was evaluated by training linear SVM models on labeled samples, where the training set contained at least one and up to all existing labeled examples. Additionally, fully supervised models were trained directly on labeled data only. We also included traditional linear methods, linear ICA and Principal Component Analysis (PCA), as feature extractors for comparison. The same number of components as nonlinear ICA was set and the linear unmixing and loading matrix were trained on unlabeled resting-state data. One can see that the linear ICA and PCA as feature extractors underperformed the deep-learning methods. Fig. 4A demonstrates the impact of the number of labeled data on downstream performance. First, when using well-trained feature extractor in nonlinear ICA for the downstream tasks, we observed important above-chance performance, 88.3% and 85.5% test accuracy (2-class, chance level=50%) for NICA(IIA) and NICA(TCL), separately. Additionally, we observed that the performance of nonlinear ICA outperformed alternative approaches

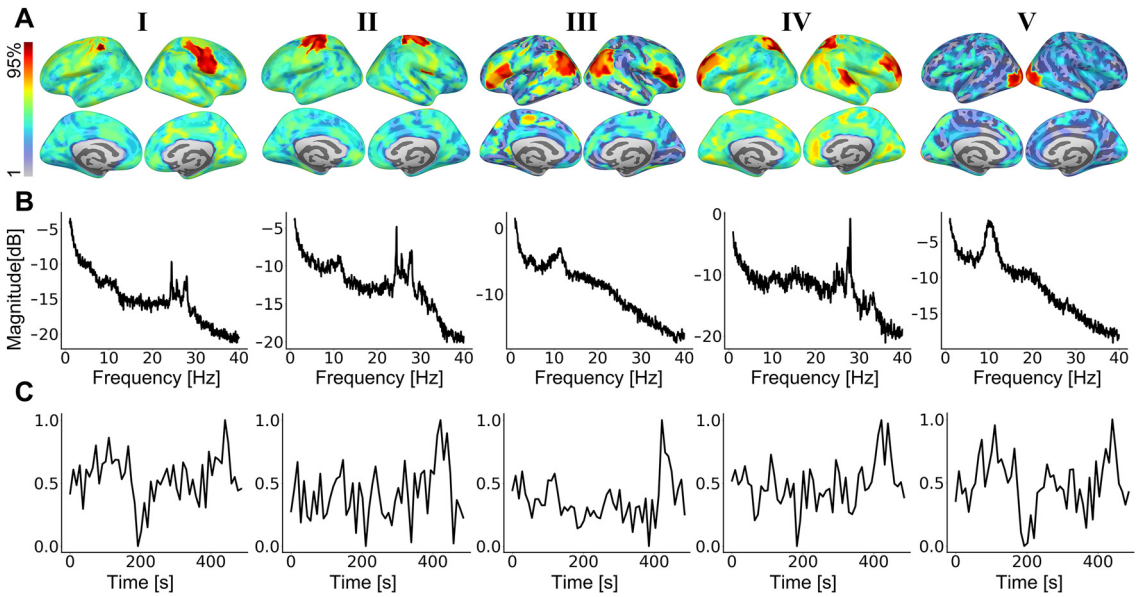

Fig. 3. Representations learned by nonlinear ICA, which seem to represent frequency-specific spatial patterns. Here, we show five components (the remaining components are given in the supplementary material Fig. S3). **A:** Spatial patterns similar to the RSNs obtained by fMRI or MEG. Such spatial profiles were obtained by examining the contribution of each parcel to the components (see Section 2.6). **B:** Averaged segment-wise spectra of the components. **C:** Segment-wise band-limited power (or variances), showing the temporal non-stationarity, (we just show 500 s length for demonstration).

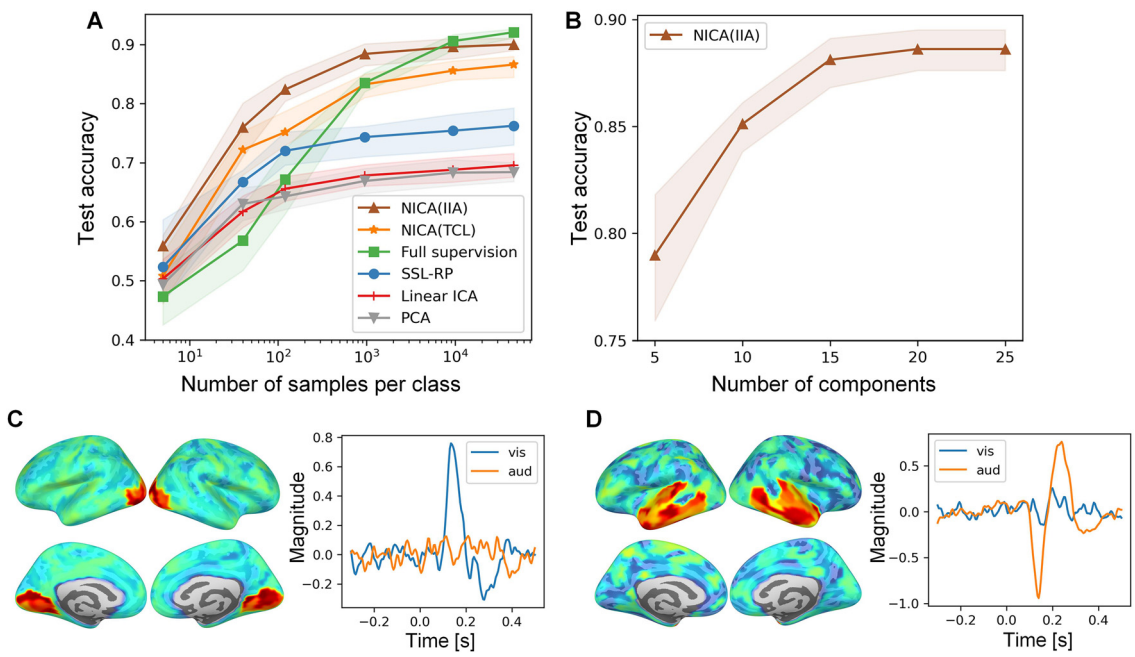

Fig. 4. Downstream tasks on visual-auditory session data with limited labels. **A.** Impact of the number of labeled samples per class on downstream performance and comparison with a self-supervised paradigm based on relative positioning task (SSL-RP) and a full supervised deep model as well as linear ICA and PCA. **B.** The effects of component numbers in NICA(IIA). **C-D.** The spatial and temporal profiles of the top components contributing the classification. The temporal profiles of the components averaged separately for auditory (orange) and visual trials (blue). 0 s is the onset of the stimulus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

including the fully supervised model in low-labeled data regimes. The performance gap between nonlinear ICA and full supervision was as high as 16.8 percentage points when the number of labeled examples was less than 100. It remained in favor of nonlinear ICA models up to around 1000 examples per class, at which point full supervision finally began to exceed the performance of nonlinear ICA. These results show that nonlinear ICA with training on unlabeled data was general-purpose enough to facilitate classification problems based on stimuli-induced data; it systematically outperformed or equaled other methods in low-to-medium labeled data regimes and remained competitive in a high labeled data regime. Note that all the analysis that follows is based on NICA(IIA), since it outperformed NICA(TCL) here.

As with many other ICA algorithms, it was also challenging for nonlinear ICA to determine the number of components, and choosing different number of components might affect the results to some extent. We here examined the impact of different numbers of extracted components in NICA(IIA) on the downstream classification task. As can be seen in Fig. 4B, the performance starts to reach the optimal level when the

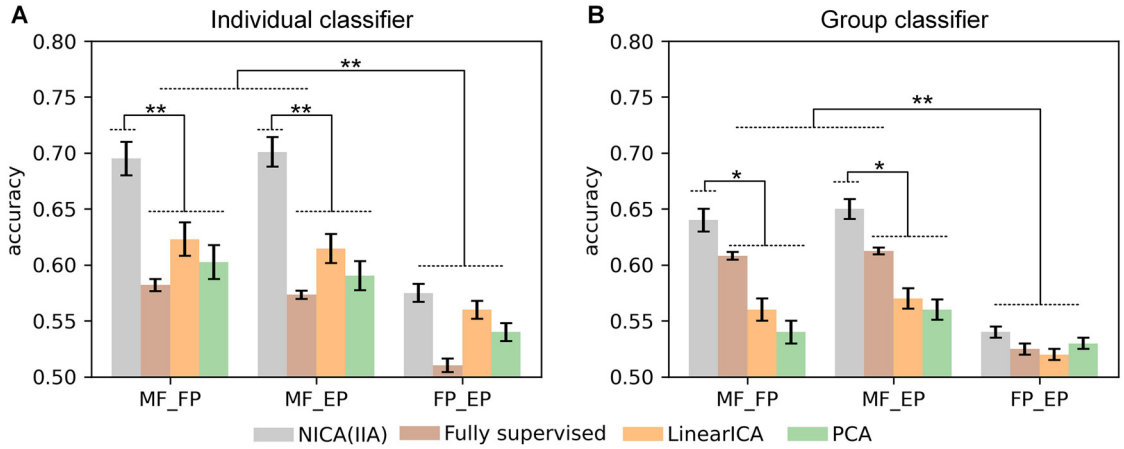

Fig. 5. Decoding attentional states using the trained NICA (IIA) model from resting-state MEG of the Cam-CAN. **A.** Classification accuracies averaged across subjects for individual classification. **B.** And group classification. MF_FP denotes mindfulness meditation vs. future planning task, MF_EP denotes mindfulness meditation vs. reflection of anxious-inducing emotional pictures task, and FP_EP denotes future planning task vs. reflection of anxious-inducing emotional pictures task. Error bars represent standard error of mean ($p < 0.01$ *, $p < 0.001$ **).

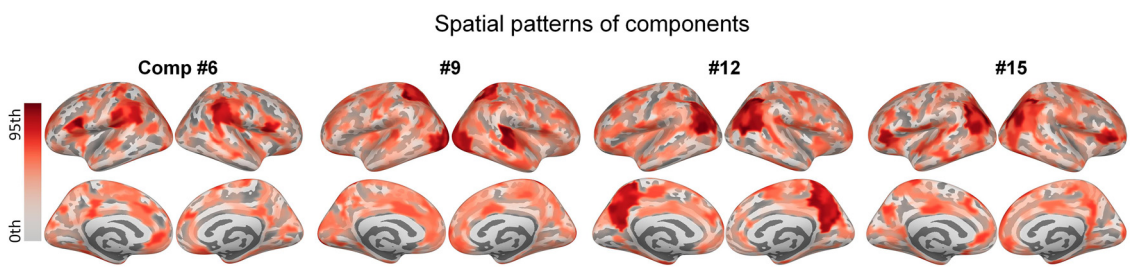

Fig. 6. Spatial patterns of components extracted from neurofeedback data with largest 4 contributions to the SVM classifier.

component number is set to 15, and after that no significant increase in the classification accuracy is observed. Therefore, we set the number of components to 15 in all the experiments.

We also present two components that showed top contributions in the SVM classification in Fig. 4C-D and the spatiotemporal patterns imply that they were related to the stimulus-specific dynamics of the brain. We visualize the temporal dynamics of the components by averaging trials separately for auditory and visual stimuli. We observe that one of the spatial patterns was strongly activated in the visual cortex and averaged visual trials had an obvious peak emerging 150 ms after the stimuli while no peaks for auditory trails (Fig. 4C). The other one demonstrates a strong activation around temporal cortices and the averaged auditory epochs exhibited stimuli-related peaks (Fig. 4D). These results suggest that nonlinear ICA of resting-state data could capture meaningful stimulus-related features which are useful for the downstream tasks.

3.3. Nonlinear ICA model generalizes to new data set

To further investigate the transfer, or generalization of nonlinear ICA model trained with big data on Cam-CAN repository to an independent dataset, we applied the well-trained NICA(IIA) to the neurofeedback dataset presented in our previous study (Zhigalov et al., 2019), to decode attentional states from the MEG data (Fig 4). Specifically, we applied the features or components extracted from NICA(IIA) to discriminate between mindfulness meditation (MF), future planning (FP) and reflection on anxious-inducing emotional pictures (EP) tasks. We also compared with the linear approaches including linear ICA and principal component analysis (PCA), and full supervised learning trained directly on the data itself (Fig. 5).

The NICA(IIA) with individual classifier provided accuracies well above chance-level, 0.68 ± 0.016 (MF vs. FP), 0.71 ± 0.011 (MF vs. EP) and 0.59 ± 0.013 (FP vs. EP), significantly larger than the full supervision and linear methods ($p < 0.001$). The accuracies MF vs. FP and MF vs. EP were significantly larger ($p < 0.001$, individual classifier; $p < 0.001$, group classifier) than the accuracy FP vs. EP, suggesting that FP and EP have similar neuronal correlates. The accuracies for the fully supervised learning-based individual classifier were relatively low, 0.582 ± 0.005 (MF vs. FP), 0.573 ± 0.004 (MF vs. EP) and 0.51 ± 0.006 (FP vs. EP), even lower than the linear-based classifier due to the limited number of labels.

The NICA(IIA) with group classifier offered the accuracies, 0.62 ± 0.012 (MF vs. FP), 0.65 ± 0.01 (MF vs. EP) and 0.55 ± 0.007 (FP vs. EP), also significantly larger that the full supervision and linear methods ($p < 0.01$). In addition, the fully supervised deep learning scored 0.604 + 0.003 (MF vs. FP), 0.61 ± 0.004 (MF vs. EP) and 0.53 ± 0.005 (FP vs. EP); this exceeded the linear-based methods. The slightly increased number of labels did not make the fully supervised method comparable in performance with NICA(IIA). The group-level classifiers actually provided slightly lower accuracies compared to the individual classifier for all the approaches, including nonlinear ICA, showing that it is difficult to generalize the classifier for decoding attentional states from different subjects.

3.4. Brain activity patterns during mindfulness

To demonstrate the brain patterns during mindfulness, we here present four components (Fig. 6) with greatest contributions to the classification based on the SVM classification coefficients in the group classifier (Fig. S4). These components' spatial patterns were associated with both mindfulness and wandering thought tasks, showing spatial profiles (Fig. 6) related to brain areas of default mode networks (DMN), dorsal attention (DAN) and cingulo-opercular networks (CON). The spatial profile of component 6 exhibits brain distribution similar to CON. The spatial pattern of component 9 shows a rhythmic activity in supe-

rior parietal, intraparietal and visual areas overlapping with DAN. The sub-region of DMN, reported with activation during mind wandering (Groot et al., 2021), was also observed in spatial patterns of component 12 and 15. The spatial pattern of component 12 had a high rhythmic activity in posterior DMN areas, while component 15 presented spatial activation in sub-areas of lateral DMN. These results show that the nonlinear ICA pre-trained on large resting-state data can transfer to a different dataset recorded with a different experiment design from different subjects, and uncover specific task-related brain patterns.

4. Discussion

We presented a group nonlinear ICA framework for unsupervised representation learning from big spontaneous MEG data which facilitates analysis of downstream tasks with limited labeled data. Experimental results demonstrated that nonlinear ICA can uncover and disentangle the nonlinear components underlying spontaneous electrophysiological brain activity. Those feature extractors successfully generalize to downstream classification tasks even from an independent dataset. We expect that the proposed approach will provide new insight into resting-state brain networks and their temporal dynamics, in addition to its utility for brain decoding.

We observed that features extracted by nonlinear ICA outperformed fully supervised learning on audio-visual task MEG when the number of labeled samples were less than 100 per class (Fig. 4A). Specifically, NICA(IIA) retained the outperformance until up to 10,000 samples per class, where full supervision finally began to exceed it, however by a 1.8–3.2% margin only. These results demonstrate that unsupervised learning based on nonlinear ICA has the potential to be an effective tool for improving classification accuracy when annotations are time-consuming or expensive, which is very common in practice. For instance, annotating sleep bio-signal recordings requires a trained technician to visually look through hours of data and to label short epochs one-by-one. Clinical recordings, such as for epilepsy diagnosis or detection of brain lesions, must likewise be reviewed by neurologists (Banville et al., 2021). Moreover, in some cases, knowing exactly what the participants were thinking or doing in cognitive neuroscience experiments (e.g., movie-watching or music listening) can be challenging, making it hard to obtain accurate labels. In tasks related to mental imagery or attention, for instance, the subjects might not be following instructions or the process under study might be difficult to quantify objectively (e.g., meditation, emotions). Meanwhile, a huge amount of resting-state (spontaneous or task-free) data has been made public. Thus, nonlinear ICA can be applied to such semi-supervised scenarios by well leveraging those unlabeled data.

Our results showed that nonlinear ICA methods outperformed SSL(RP) method (Fig. 4A), which was applied to EEG data of a few channels as self-supervised learning based on temporal context prediction tasks inspired by the autocorrelations (temporal dependencies) of time series (Banville et al., 2021). We speculate this is because the nonlinear ICA with time-contrastive learning is based on the temporal non-stationarity of the time series, which might be more in line with the properties of the MEG/EEG data than the temporal dependencies assumption. We further observed that NICA(IIA) always outperformed NICA(TCL) in the downstream tasks, since the temporal dependences and non-stationarity of multivariate time series were both considered in the NICA(IIA) model. A model with such a combination of temporal structures could be a better fit for the data, thus providing better performance.

Since the nonlinear mixing function in nonlinear ICA was typically approximated by a neural network such as multi-layer perception (MLP), it is very difficult to understand the spatial patterns of the relevant neural activity, especially when compared to linear ICA (where the columns of the mixing matrix give the spatial or co-activation patterns). We here adopted a visualization technique from deep learning studies for computer vision, called occlusion sensitivity (Zeiler and Fergus, 2014). Instead of the corrected label probability used on image classification tasks, we here computed the effect (weight) of the channel (parcel) of the input time series on each component. This procedure enables a visualization analogous to the linear ICA. The obtained spatial patterns could be considered as nonlinear co-activation patterns. However, other visualization methods for neural networks exist and we consider them an interesting topic for future research.

Based on the spatial visualization method, we demonstrated that the representations learned with nonlinear ICA capture and disentangle oscillatory brain patterns related to the frequency-specific resting-state functional networks (Fig. 3). Such networks have been reported elsewhere by using other approaches, such as Hidden Markov models (Baker et al., 2014; Vidaurre et al., 2018a, b; Vidaurre et al., 2016) and fourier-ICA (Ramkumar et al., 2014). In the nonlinear ICA model, the feature extractor models inverse inference of the sources from the brain activity in a data-driven manner. For example, we observed the beta oscillatory activities in the bilateral somatosensory and motor cortices, and alpha-dominated rhythmic activities in visual areas (Fig. 3). Early MEG studies indicate that the around 20 Hz oscillation is generated precentrally and appears more related to motor than somatosensory processing whereas the 10 Hz oscillation is post-central and associated with the processing of tactile information, although both rhythms are modulated by movement and tactile stimulation (Hari and Salmelin, 1997; Ramkumar et al., 2014). Interestingly, when the trained nonlinear ICA model was applied to task MEG data, task/stimulus-related co-activation patterns were uncovered: For audio-visual data, a bilateral auditory component and a visual component had high contributions to the classification, but with different latencies of the temporal courses (Fig. 4C-D).

Our results further demonstrated that latent space of the nonlinear ICA can be transferred to a mental-states MEG dataset for decoding the attentional states. It outperformed the baseline methods including linear decomposition methods such as PCA and linear ICA (as used in our previous study Zhigalov et al. (2019)), as well as conventional supervised deep learning (Fig. 5). In contrast to linear features, the nonlinear ICA features provide complementary information that is made accessible to purely linear classification methods, resulting in more accurate classification. Compared to fully supervised deep learning, the advantage of nonlinear ICA is that it was trained from a big task-free database, while supervised deep learning suffered from the limited number of labels. This point is further demonstrated by comparing the accuracies of individual classifiers and group classifiers: the accuracies of the individual classifiers (training on individual subjects) based on full supervision were even lower than those of the linear methods, presumably due to the catastrophically low number of data points used by each classifier.

We observed a variety of components with specific spatial patterns that contribute to decoding mindfulness meditation, implicating that several neuronal mechanisms may underlie mindfulness state (Fig. 6). Most spatial patterns were related to sub-regions of DMN, DAN and CON, which is consistent with an fMRI study on mindfulness meditation that suggests that mind wandering may evoke multiple high-order cognitive networks such as default and executive network regions (Christoff et al., 2009). For example, the spatial patterns of component 6 spanned the inferior frontal gyrus and the precuneus overlapping with CON regions, which have been shown to be consistently activated when individuals engaged in demanding mental activity (Christoff et al., 2009). Component 9 demonstrated high rhythmic activity in superior parietal, intraparietal, and visual areas overlapping with DAN, which has been shown to be more active during task-free states than during a wide range of states involving goal-directed task performance (Kucyi, 2018). It should be noted that the connection between these classifier weights and the neural correlates is not straight-forward. Interpretation of the weights can lead to wrong conclusions regarding the origin of neural signals of interest, since significant nonzero weights may also be associated with task-irrelevant signals (Haufe et al., 2014). We also visualized the four largest coefficients for individual subjects. The results showed that the components associated with the largest coeffi-

cients were highly individual (Fig. S5), which makes the generalization of the classifier coefficients over subjects impractical. We found relatively low classification accuracies in decoding between future planning (FP) and reflection on anxiousness-inducing emotional pictures (EP) tasks (Fig. 5), which suggests similarity of the brain activities during these tasks. Although these tasks are behaviorally quite distinct and might be different regarding the amplitudes of the evoked responses (Olofsson et al., 2008; Zhigalov et al., 2019), they may be similar with regard to task-nonspecific cortical processes associated with attentional states. Consequently, the neurofeedback paradigm that focuses on ongoing neuronal activity may be insensitive to this difference.

The overall individual-level classification accuracy for attentional states was around 70%, which might be relatively low to be useful in a neurofeedback system. However, for a few subjects, the accuracies were around 80% or more, which might constitute sufficient improvements in decoding attentional states in mindfulness meditation to be practically useful. Furthermore, this relatively low decoding accuracies might result from the fact that many of the participants did not have previous experiences in mindfulness meditation: The neurofeedback might perform much better after the participants gained more experience. Generalization across participants was even more difficult, presumably due to the large individual differences already noted by (Zhigalov et al., 2019).

Transfer learning is a topic of great interest in applications of machine learning on brain imaging. From that viewpoint, we here explored the performance of the nonlinear ICA model transferred both across different paradigms (experimental design) and different subjects (from a different dataset) (Zhang et al., 2018). For the audio-visual classification, the task data were recorded from the same population as the training data, but under different experimental paradigms. For the mindfulness classification, the neurofeedback data were collected for a different purpose, at a different site, and from different subjects, but using the same type of device. In this case, the Cam-CAN dataset can be thought of as a "secondary" dataset to assist in the analysis of a small "primary" dataset (i.e., the neurofeedback dataset). Usually, it is quite challenging to transfer models across subjects due to individual differences (Zhang et al., 2018). It might seem even more challenging to transfer the models to different tasks; nevertheless, the ability to use the same image features in different tasks has been one of the main success stories in deep learning for computer vision (Simonyan and Zisserman, 2014). In the brain imaging case, a further point is whether the transfer is between measurements with similar devices or not: The same type of measuring equipment might have similar nonlinear mixing systems. In future research, we would like to examine the possibility to transfer nonlinear ICA across datasets measured by different devices.

Regarding the methodological considerations, three core hyperparameters require setting: the window length, the number of independent components, and the network architecture. A judicious selection of window length is important, and represents a trade-off between temporal resolution and the stability of the model training. In principle, TCL could be performed on different time scales depending on the length of the window chosen; but in practice, each window must also contain enough data points, which makes analysis of very short time scales challenging. The selection of the number of independent components is a fundamental question for all ICA methodologies. In the absence of theoretically motivated methods to hyperparameter selection, we opted instead to repeat the analysis for different values, and cross-validate. Specifically, the models were trained within the whole resting-state MEG data, and we tuned those two hyperparameters of models (the number of components and window length) in the whole audio-visual MEG data with plain cross-validation. The value of 15 components were finally chosen and seemed to give a good trade-off between the dimension of the model and downstream classification accuracy. Finally, the neural network (NN) architectures are related to the degree of nonlinearity of the transform. In the present work, we selected 3-layer MLP components based on our previous experience, which gives a relatively reasonable degree of nonlinearity. However, how the degree of nonlinearity and the NN structures (e.g., convolutional NN) affect the results are still interesting questions and we leave them for future work.

As with many other ICA algorithms, the determination of the model order, i.e., the selection of the number of components, is indeed a crucial issue in TCL for the estimation of nonlinear ICA. In our preliminary experiments, the repetition analysis for different values demonstrated that on the one hand, if we decreased the number of components, the classification accuracies of the downstream task get lower, suggesting that the transferability of the model to a new dataset gets weakened; and on the other hand, if we increase the number of components, no significant increase in classification accuracy after 15 components is observed, implying the number of underlying nonlinear independent components may no longer increase. The experiments suggest that the setting $m = 15$ was reasonable in the current study, considering that the components were properly demixed and the transferability of models to a new dataset. However, the best setting may vary across datasets due to different experimental paradigms.

For very high-dimensional data, an obvious problem is that the number of parameters in the model escalates rapidly if we aim to analyze a big number of latent sources (high model order) simultaneously. This may hinder the estimation in practice even with reasonable sample sizes and computational capabilities. Although, in theory, the identifiability of nonlinear ICA is not an issue in high dimensions, empirically we don't have much evidence beyond the model order of the current study. Thus, this might be a possible issue for practical applications, where a large number of latent components (high model order) needs to be estimated.

Regarding the model order of the autoregressive model in IIA, determining the order for a nonlinear autoregressive model is a challenging task compared to linear models because the nonlinear relationships between the dependent variable and its lagged values are not as straightforward to identify. In the current study, without standard methods available, we tried some values to look at the impact on the performance of downstream tasks (Fig. S7). This preliminary test implies that a third-order model gives relatively optimal performance for classification. It should be noted that this experiment just provides a basic reference but it is not able to accurately estimate the underlying model order. In addition, the order may also vary across the different data modalities. For example, the fMRI data with slowly varying features may have different model orders from the M/EEG dataset with highly varying rates. Future work should therefore seek other approaches to determine the validity of the autoregressive model order, particularly if the present method was to be used for fMRI data.

In addition, the model here is supposed to be noise-free. We point out that the most commonly used linear ICA methods assume a noise-free mixing; Probabilistic ICA by (Beckmann and Smith, 2004) does start by assuming a noisy mixing, but in the end the ICA algorithm used (after PCA) assumes that the mixing is noise-free. The estimation of noisy version of NICA is very complicated (Hälvä et al., 2021), which we will leave the analysis for future work.

The training for NICA models just like any deep learning approach needs a lot of computational sources. For example, training of the feature extractor by TCL took about 8 h (64 GB Memory, NVIDIA Tesla P100 GPU). Although such computational cost is not very expensive by deep learning standards, it is still more costly compared to traditional methods (e.g., linear ICA). Also, unlike linear ICA in neuroimaging community, training NICA models needs massive data, which makes its application mainly focus on relatively big data. In addition, it is rather complicated and difficult to interpret and visualize spatial patterns of the nonlinear components due to the nonlinear mixing function. Although we here adopted an occlusion sensitivity analysis method for visualization, more attention to such interpretation would be still warranted.

Like any deep learning method, nonlinear ICA suffers from the problem of local minima: any run of the learning algorithm is not guaranteed to find the best solution. While in theory, this problem seems insurmountable, typical deep learning practice is to simply accept the

minimum obtained after using the largest reasonable amount of computational capacity. This is what we did in this paper, and thus there is a possibility that better results might be obtained by another run. To examine the run-to-run variability, we performed the experiment with different random seeds and computed the similarity (correlation) between the spatial maps and the component time series. The results demonstrate the similarities are more than 0.8 for most components, suggesting the decomposition is stable (Figs. S6&8).

To conclude, we present nonlinear ICA for unsupervised representation learning of cortical resting-state MEG activity in a data-driven manner. Our results suggest that nonlinear ICA model is able to capture and disentangle the generative components underlying resting-state activity, characterizing the spontaneous oscillatory patterns. Features extracted by nonlinear ICA outperformed fully supervised learning on audio-visual task MEG when the number of labeled samples were limited. As an initial example of a neurofeedback application, nonlinear ICA trained on large open access dataset was successfully transferred to a new data set for an attentional state classification task.

Data & code statement

The data used in the manuscript (Zhu et al. Unsupervised representation learning of spontaneous MEG data with Nonlinear ICA) are from the Cambridge Centre for Ageing and Neuroscience repository (Cam-CAN; https://www.cam-can.org/). The analysis code is available from the authors upon request.

CRediT authorship contribution statement

Yongjie Zhu: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Tiina Parviainen**: Data curation, Writing – review & editing. **Erkka Heinilä:** Data curation, Writing – review & editing. **Lauri Parkkonen:** Conceptualization, Writing – review & editing. **Aapo Hyvärinen:** Conceptualization, Methodology, Writing – review & editing, Project administration, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We wish to thank the reviewers and editors for the useful comments to improve the paper a lot. We thank Dr. Hiroshi Morioka for the useful discussion at the beginning of the project. L.P. was funded in part by the European Research Council (No. 678578). A.H. was supported by a Fellowship from CIFAR, and the Academy of Finland. The authors acknowledge the computational resources provided by the Aalto Science-IT project, and also wish to thank the Finnish Grid and Cloud Infrastructure (FGCI) for supporting this project with computational and data storage resources.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.120142.

References

Acunzo, D.J., Low, D.M., Fairhall, S.L., 2022. Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. Neuroimage, 119005.

Aoe, J., Fukuma, R., Yanagisawa, T., Harada, T., Tanaka, M., Kobayashi, M., … Kishima, H., 2019. Automatic diagnosis of neurological diseases using MEG signals with a deep neural network. Sci. Rep. 9 (1), 1–9.

Baker, A.P., Brookes, M.J., Rezek, I.A., Smith, S.M., Behrens, T., Smith, P.J.P., Woolrich, M., 2014. Fast transient networks in spontaneous human brain activity. Elife 3, e01867.

Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-.A., Gramfort, A., 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. J. Neural Eng. 18 (4), 046020.

Becker, R., Vidaurre, D., Quinn, A.J., Abeysuriya, R.G., Jones, O.P., Jbabdi, S., Woolrich, M.W., 2020. Transient spectral events in resting state MEG predict individual task responses. Neuroimage 215, 116818.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23 (2), 137–152.

Brookes, M.J., Woolrich, M., Luckhoo, H., Price, D., Hale, J.R., Stephenson, M.C., … Morris, P.G., 2011. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. Proc. Natl. Acad. Sci. 108 (40), 16783–16788.

Chambon, S., Galtier, M.N., Arnal, P.J., Wainrib, G., Gramfort, A., 2018. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Trans. Neural Syst. Rehabil. Eng. 26 (4), 758–769.

Christoff, K., Gordon, A.M., Smallwood, J., Smith, R., Schooler, J.W., 2009. Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. Proc. Natl. Acad. Sci. 106 (21), 8719–8724.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. Neuroimage 163, 115–124.

De Pasquale, F., Della Penna, S., Snyder, A.Z., Lewis, C., Mantini, D., Marzetti, L., … Romani, G.L., 2010. Temporal dynamics of spontaneous MEG activity in brain networks. Proc. Natl. Acad. Sci. 107 (13), 6040–6045.

Fox, M.D., Buckner, R.L., Liu, H., Chakravarty, M.M., Lozano, A.M., Pascual-Leone, A, 2014. Resting-state networks link invasive and noninvasive brain stimulation across diverse psychiatric and neurological diseases. Proc. Natl. Acad. Sci. 111 (41), E4367–E4375.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M.S., 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460.

Groot, J.M., Boayue, N.M., Csifcsák, G., Boekel, W., Huster, R., Forstmann, B.U., Mittner, M., 2021. Probing the neural signature of mind wandering with simultaneous fMRI-EEG and pupillometry. Neuroimage 224, 117412.

Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., Hyvarinen, A., 2021. Disentangling identifiable features from noisy data with structured nonlinear ICA. In: Advances in Neural Information Processing Systems (NeurIPS2021), Vol. 34, pp. 1624–1633.

Hari, R., Salmelin, R., 1997. Human cortical oscillations: a neuromagnetic view through the skull. Trends Neurosci. 20 (1), 44–49.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., … Yeo, B.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206, 116276.

Hsu, S.-.H., Lin, Y., Onton, J., Jung, T.-.P., Makeig, S., 2022. Unsupervised learning of brain state dynamics during emotion imagination using high-density EEG. Neuroimage, 118873.

Hyvärinen, A., Morioka, H., 2016. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. Advances in Neural Information Processing Systems (NIPS2016), Vol. 29.

Hyvärinen, A., Morioka, H., 2017. Nonlinear ICA of temporally dependent stationary sources. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS2017), pp. 460–469.

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

Hyvärinen, A., Ramkumar, P., Parkkonen, L., Hari, R., 2010. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. Neuroimage 49 (1), 257–271.

Jas, M., Larson, E., Engemann, D.A., Leppäkangas, J., Taulu, S., Hämäläinen, M., Gramfort, A., 2018. A reproducible MEG/EEG group study with the MNE software: recommendations, quality assessments, and good practices. Front. Neurosci. 530.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4037–4058.

Jónsson, B.A., Bjornsdottir, G., Thorgeirsson, T., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D., … Ulfarsson, M., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat. Commun. 10 (1), 1–10.

Kim, J.-.H., Zhang, Y., Han, K., Wen, Z., Choi, M., Liu, Z., 2021. Representation learning of resting state fMRI with variational autoencoder. Neuroimage 241, 118423.

Kucyi, A., 2018. Just a thought: how mind-wandering is represented in dynamic brain connectivity. Neuroimage 180, 505–514.

Larson-Prior, L.J., Oostenveld, R., Della Penna, S., Michalareas, G., Prior, F., Babajani-Feremi, A., … Di Pompeo, F., 2013. Adding dynamics to the human connectome project with MEG. Neuroimage 80, 190–201.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J. Neural Eng. 15 (5), 056013.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., Corbetta, M., 2007. Electrophysiological signatures of resting state networks in the human brain. Proc. Natl. Acad. Sci. 104 (32), 13170–13175.

Morioka, H., Hälvä, H., Hyvärinen, A., 2021. Independent innovation analysis for nonlinear vector autoregressive process. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS2021), pp. 1549–1557.
Niso, G., Rogers, C., Moreau, J.T., Chen, L.-.Y., Madjar, C., Das, S., … Jolicoeur, P., 2016. OMEGA: the open MEG archive. Neuroimage 124, 1182–1187.
O'Neill, G.C., Tewarie, P.K., Colclough, G.L., Gascoyne, L.E., Hunt, B.A., Morris, P.G., … Brookes, M.J., 2017. Measurement of dynamic task related functional networks using MEG. Neuroimage 146, 667–678.
Olofsson, J.K., Nordin, S., Sequeira, H., Polich, J., 2008. Affective picture processing: an integrative review of ERP findings. Biol. Psychol. 77 (3), 247–265.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Ramkumar, P., Parkkonen, L., Hyvärinen, A., 2014. Group-level spatial independent component analysis of Fourier envelopes of resting-state MEG data. Neuroimage 86, 480–491.
Sareen, E., Zahar, S., Van De Ville, D., Gupta, A., Griffa, A., Amico, E., 2021. Exploring MEG brain fingerprints: evaluation, pitfalls, and interpretations. Neuroimage 240, 118331.
Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-.N., Holmes, A.J., … Yeo, B.T., 2018. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb. Cortex 28 (9), 3095–3114.
Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., … Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. 38 (11), 5391–5420.
Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., … Dalgleish, T., 2014. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurol. 14 (1), 1–25.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
Taulu, S., Simola, J., 2006. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. 51 (7), 1759.
Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., … Henson, R.N., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144, 262–269.
Van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv e-prints, arXiv:1807.03748.
Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A.J., Alfaro-Almagro, F., Smith, S.M., Woolrich, M.W., 2018a. Discovering dynamic brain networks from big data in rest and task. Neuroimage 180, 646–656.
Vidaurre, D., Hunt, L.T., Quinn, A.J., Hunt, B.A., Brookes, M.J., Nobre, A.C., Woolrich, M.W., 2018b. Spontaneous cortical activity transiently organises into frequency specific phase-coupling networks. Nat. Commun. 9 (1), 1–13.
Vidaurre, D., Quinn, A.J., Baker, A.P., Dupret, D., Tejero-Cantero, A., Woolrich, M.W., 2016. Spectrally resolved fast transient brain states in electrophysiological data. Neuroimage 126, 81–95.
Vigário, R., Sarela, J., Jousmäki, V., Hämäläinen, M., Oja, E., 2000. Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans. Biomed. Eng. 47 (5), 589–593.
Yan, W., Qu, G., Hu, W., Abrol, A., Cai, B., Qiao, C., … Calhoun, V.D., 2022. Deep learning in neuroimaging: promises and challenges. IEEE Signal Process. Mag. 39 (2), 87–98.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. Paper Presented at the European Conference on Computer Vision.
Zhang, H., Chen, P.-.H., Ramadge, P., 2018. Transfer learning on fMRI datasets. Paper Presented at the International Conference on Artificial Intelligence and Statistics.
Zhigalov, A., Heinilä, E., Parviainen, T., Parkkonen, L., Hyvärinen, A., 2019. Decoding attentional states for neurofeedback: mindfulness vs. wandering thoughts. Neuroimage 185, 565–574.
Zhu, Y., Wang, X., Mathiak, K., Toiviainen, P., Ristaniemi, T., Xu, J., … Cong, F., 2021. Altered eeg oscillatory brain networks during music-listening in major depression. Int. J. Neural Syst. 31 (03), 2150001.
Zubarev, I., Zetter, R., Halme, H.-.L., Parkkonen, L., 2019. Adaptive neural network classifier for decoding MEG signals. Neuroimage 197, 425–434.