Journal of Machine Learning Research 20 (2019) 1–50 Submitted 6/18; Revised 4/19; Published 10/19

Robustifying Independent Component Analysis by Adjusting for Group-Wise Stationary Noise

Niklas Pfister* PFISTER@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zürich
Rämisstrasse 101, 8092 Zürich, Switzerland

Sebastian Weichwald* SWEICHWALD@MATH.KU.DK
Department of Mathematical Sciences, University of Copenhagen
Universitetsparken 5, 2100 Copenhagen, Denmark

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zürich
Rämisstrasse 101, 8092 Zürich, Switzerland

Bernhard Schölkopf BS@TUE.MPG.DE
Empirical Inference Department, Max Planck Institute for Intelligent Systems
Max-Planck-Ring 4, 72076 Tübingen, Germany

Editor: Kenji Fukumizu

Abstract

We introduce coroICA, confounding-robust independent component analysis, a novel ICA algorithm which decomposes linearly mixed multivariate observations into independent components that are corrupted (and rendered dependent) by hidden group-wise stationary confounding. It extends the ordinary ICA model in a theoretically sound and explicit way to incorporate group-wise (or environment-wise) confounding. We show that our proposed general noise model allows to perform ICA in settings where other noisy ICA procedures fail. Additionally, it can be used for applications with grouped data by adjusting for different stationary noise within each group. Our proposed noise model has a natural relation to causality and we explain how it can be applied in the context of causal inference. In addition to our theoretical framework, we provide an efficient estimation procedure and prove identifiability of the unmixing matrix under mild assumptions. Finally, we illustrate the performance and robustness of our method on simulated data, provide audible and visual examples, and demonstrate the applicability to real-world scenarios by experiments on publicly available Antarctic ice core data as well as two EEG data sets. We provide a scikit-learn compatible pip-installable Python package coroICA as well as R and Matlab implementations accompanied by a documentation at https://sweichwald.de/coroICA/.

Keywords: blind source separation, causal inference, confounding noise, group analysis, heterogeneous data, independent component analysis, non-stationary signal, robustness

1. Introduction

The analysis of multivariate data is often complicated by high dimensionality and complex inter-dependences between the observed variables. In order to identify patterns in such

*Authors contributed equally. Most of this work was done while SW was at the Max Planck Institute for Intelligent Systems, Tübingen, Germany.

©2019 Niklas Pfister*, Sebastian Weichwald*, Peter Bühlmann, and Bernhard Schölkopf.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v20/18-399.html.

data it is therefore desirable and often necessary to separate different aspects of the data. In multivariate statistics, for example, principal component analysis (PCA) is a common preprocessing step that decomposes the data into orthogonal principle components which are sorted according to how much variance of the original data each component explains. There are two important applications of this. Firstly, one can reduce the dimensionality of the data by projecting it onto the lower dimensional space spanned by the leading principal components which maximize the explained variance. Secondly, since the principle components are orthogonal, they separate in some sense different (uncorrelated) aspects of the data. In many situations this enables a better interpretation and representation.

Often, however, PCA may not be sufficient to separate the data in a desirable way due to more complex inter-dependences in the multivariate data (see e.g., Section 1.3.3 in Hyvärinen et al. (2002) for an instructive example). This observation motivates the development of independent component analysis (ICA), formally introduced in its current form by Cardoso (1989b) and Comon (1994). ICA is a widely used unsupervised blind source separation technique that aims at decomposing an observed mixture of independent source signals. More precisely, assuming that the observed data is a linear mixture of underlying independent variables, one seeks the unmixing matrix that maximizes the independence between the signals it extracts. There has been a large amount of research on different types of ICA procedures and their interpretations, e.g., Bell and Sejnowski (1995, Infomax) who maximize the entropy, Hyvärinen (1999, fastICA) maximizing the kurtosis or Belouchrani et al. (1997, SOBI) who propose to minimize time-lagged dependences, to name only some of the widespread examples.

ICA has applications in many fields, for example in finance (e.g., Back and Weigend, 1997), the study of functional magnetic resonance imaging (fMRI) data (e.g., McKeown et al., 1998a,b; Calhoun et al., 2003), and notably in the analysis of electroencephalography (EEG) data (e.g., Makeig et al., 1995, 1997; Delorme and Makeig, 2004). The latter is motivated by the common assumption that the signals recorded at EEG electrodes are a (linear) superposition of cortical dipole signals (Nunez and Srinivasan, 2006). Indeed, ICA-based preprocessing has become the de facto standard for the analysis of EEG data. The extracted components are interpreted as corresponding to cortical sources (e.g., Ghahremani et al., 1996; Zhukov et al., 2000; Makeig et al., 2002) or used for artifact removal by dropping components that are dominated by ocular or muscular activity (e.g., Jung et al., 2000; Delorme et al., 2007).

In many applications, the data at hand is heterogeneous and parts of the samples can be grouped by the different settings (or environments) under which the observations were taken. For example, we can group those samples of a multi-subject EEG recording that belong to the same subject. For the analysis and interpretation of such data across different groups, it is desirable to extract one set of common features or signals instead of obtaining individual ICA decompositions for each group of samples separately. Here, we present a novel, methodologically sound framework that extends the ordinary ICA model, respects the group structure and is robust by explicitly accounting for group-wise stationary confounding. More precisely, we consider a model of the form

$$X_i = A \cdot S_i + H_i, \tag{1}$$

where i denotes the sample index, A remains fixed across different groups, S_i is a vector of independent source signals and H_i is a vector of stationary confounding noise variables with fixed covariance within each group (an intuitive example where such a scenario may be encountered in practice is illustrated in Figure 7). Based on this extension to ordinary ICA, we construct a method and an easy to implement algorithm to extract one common set of sources that are robust against confounding within each group and can be used for across-group analyses. The unmixing also generalizes to previously unseen groups.

1.1. Relation to Existing Work

ICA is well-studied with a tremendous amount of research related to various types of extensions and relaxations of the ordinary ICA model. In light of this, it is important to understand where our proposed procedure is positioned and why it is an interesting and useful extension. Here, we look at ICA research from three perspectives and illustrate how our proposed coroICA methodology relates to existing work. First off, in Section 1.1.1 we compare our proposed methodology with other noisy ICA models. In Section 1.1.2, we review ICA procedures based on approximate joint matrix diagonalization. Finally, in Section 1.1.3 we summarize the existing literature on ICA procedures for grouped data and highlight the differences to coroICA.

1.1.1. Noisy ICA Models

The ordinary ICA model assumes that the observed process X is a linear mixture of independent source signals S *without* a confounding term H. Identifiability of the source signals S is guaranteed by assumptions on S such as non-Gaussianity or specific time structures. For coroICA we require—similar to other second-order based methods (cf. Section 1.1.2)—that the source process S is non-stationary. More precisely, we require that either the variance or the auto-covariance of S changes across time. An important extension of the ordinary ICA model is known as noisy ICA (e.g., Moulines et al., 1997) in which the data generating process is assumed to be an ordinary ICA model with additional additive noise. In general, this leads to further identifiability issues. These can be resolved by assuming that the additive noise is Gaussian and the signal sources non-Gaussian (e.g., Hyvärinen, 1999), which enables correct identification of the mixing matrix. Another possibility is to assume that the noise is independent over time, while the source signals are time-dependent[1] (e.g., Choi and Cichocki, 2000b). In contrast, our assumption on the noise term H is much weaker, since we only require it to be stationary and hence in particular allow for time-dependent noise in coroICA. As we show in our simulations in Section 4.2.3 this renders our method robust with respect to confounding noise: coroICA is more robust against time-dependent noise while remaining competitive in the setting of time-independent noise. We refer to the book by Hyvärinen et al. (2002) for a review of most of the existing ICA models and the assumptions required for identifiability.

1. Autocorrelated signals are time-dependent, while the absence of autocorrelation does not necessarily imply time-independence of the signal. We thus use the terms time-dependence and time-independence throughout this article.

1.1.2. ICA based on Approximate Joint Diagonalization

As an extension of PCA, the concept of ICA is naturally connected to the notion of joint diagonalization of covariance-type matrices. One of the first procedures for ICA was FOBI introduced by Cardoso (1989a), which aims to jointly diagonalize the covariance matrix and a fourth order cumulant matrix. Extending on this idea Cardoso and Souloumiac (1993) introduced the method JADE which improves on FOBI by diagonalizing several different fourth order cumulant matrices. Unlike FOBI, JADE uses a general joint matrix diagonalization algorithm which is the de facto standard for all modern approaches. In fact, there is a still-active field that focuses on approximate joint matrix diagonalization, commonly restricted to positive semi-definite matrices, and often with the purpose of improving ICA procedures (e.g., Cardoso and Souloumiac, 1996; Ziehe et al., 2004; Tichavsky and Yeredor, 2009; Ablin et al., 2018).

Both JADE and FOBI are based on the assumption that the signals are non-Gaussian. This ensures that the sources are identifiable given independent and identically distributed observations. A different stream of ICA research departs from this assumption and instead assumes that the data are a linear mixture of independent weakly stationary time-series. This model is often referred to as a second-order source-separation model (SOS). The time structure in these models allows to identify the sources by jointly diagonalizing the covariance and auto-covariance. The first method developed for this setting is AMUSE by Tong et al. (1990) who diagonalize the covariance matrix and the auto-covariance matrix for one fixed lag. The performance of AMUSE is, however, fragile with respect to the exact choice of the lag, which complicates practical application (Miettinen et al., 2012). Instead of only using a single lag, Belouchrani et al. (1997) proposed the method SOBI which uses all lags up to a certain order and jointly diagonalizes all the resulting auto-covariance matrices. SOBI is to date still one of the most commonly employed ICA methods, in particular in EEG analysis.

The SOS model is based on the assumption of weak stationarity of the sources which implies that the signals have fixed variance and auto-covariance structure across time. This assumption can be dropped and the resulting models are often termed non-stationary source separation models (NSS). The non-stationarity can be leveraged to boost the performance of ICA methods in various ways (see Matsuoka et al., 1995; Hyvärinen, 2001; Choi and Cichocki, 2000a,b; Choi et al., 2001; Choi and Cichocki, 2001; Pham and Cardoso, 2001). All aforementioned methods make use of the non-stationarity by jointly diagonalizing different sets of covariance or auto-covariance matrices and mainly differ by how they perform the approximate joint matrix diagonalization. For example, the methods introduced by Choi and Cichocki (2000a,b); Choi et al. (2001) make use of non-stationarity across sources by separating the data into blocks and jointly diagonalizing either the covariance matrices, the auto-covariances or both across all blocks. For our experimental comparisons, we implemented all three of these methods with the slight modification that we use the recent uwedge approximate joint matrix diagonalization procedure due to Tichavsky and Yeredor (2009). We denote the resulting three ICA variants as

- choiICA (var): jointly diagonalize blocks of covariances,
- choiICA (TD): jointly diagonalize blocks of auto-covariances,
- choiICA (var & TD): jointly diagonalize blocks of covariances and auto-covariances.

method	**signal type**	**allowed noise**
choiICA (TD)	varying time-dependence	time-independent
choiICA (var)	varying variance	none
choiICA (var & TD)	varying variance and time-dependence	none
SOBI	fixed time-dependence	time-independent
fastICA[2]	non-Gaussian	none
coroICA	varying time-dependence *and/or* variance	group-wise stationary

Table 1. Important ICA procedures and the signal types they require as well as the noise they can deal with. coroICA is a confounding-robust ICA variant and is the only method for which an identifiability result under time-dependent noise is available.

Depending on the type of matrix which is diagonalized, each procedure detects different types of signals and behaves differently with respect to noise. Choi and Cichocki (2001) suggest a modification of choiICA (TD) in which instead of auto-covariance matrices, differences of auto-correlation matrices are diagonalized. The advantage being that it captures the non-stationarity of a signal more explicitly. Our proposed method similarly aims to use this type of signal but instead of considering the noise-free case, we explicitly formalize a model class that generalizes to noisy settings. Furthermore, we provide an identifiability theorem allowing for group-wise stationary confounding. Such a result has not been proven for the aforementioned method in the noise-free case. For a detailed description of both SOS- and NSS-based methods we refer the reader to the review by Nordhausen (2014) and for recent developments on leveraging non-stationarity for identifiability in non-linear ICA see Hyvärinen and Morioka (2016).

An exhaustive comparison of all methods is infeasible on the one hand due to the sheer amount of different models and methods and on the other hand due to the fact that appropriately maintained and easy adaptable code—for most methods—simply does not exist. Therefore, we focus our comparison on the following representative, modern methods that are most closely related to coroICA: fastICA, SOBI, choiICA (TD), choiICA (var), choiICA (TD & var). The methods and their respective assumptions on the source and noise characteristics are summarized in Table 1.

1.1.3. ICA Procedures for Grouped Data

Applications in EEG and fMRI have motivated the development of a wide variety of blind source separation techniques which are capable of dealing with grouped data, e.g., where groups correspond to different subjects or recording sessions. A short review is given in Hyvärinen (2013) and a detailed exposition in the context of fMRI data is due to Calhoun et al. (2003).

Consider we are given m groups $\{g_1, \dots, g_m\}$ and observe a corresponding data matrix $\mathbf{X}_{g_i} \in \mathbb{R}^{d \times n_i}$ for each group, where d is the number of observed signals and n_i the number of observations. Using this notation, all existing ICA procedures for grouped data can be related to one of three underlying models extending the classical mixing model $\mathbf{X} = A \cdot \mathbf{S}$. The first, often also referred to as "temporal concatenation", assumes that the mixing

2. The fastICA method can be extended to include Gaussian noise (see Hyvärinen, 1999).

remains equal while the sources are allowed to change across groups leading to data of the form

$$(\mathbf{X}_{g_1}, \ldots, \mathbf{X}_{g_m}) = A \cdot (\mathbf{S}_{g_1}, \ldots, \mathbf{S}_{g_m}) \,. \tag{2}$$

The second model, often also referred to as "spatial concatenation", assumes the sources remain fixed ($n_1 = \cdots = n_m$) while the mixing matrices are allowed to change, i.e.,

$$\begin{pmatrix} \mathbf{X}_{g_1} \\ \vdots \\ \mathbf{X}_{g_m} \end{pmatrix} = \begin{pmatrix} A_{g_1} \\ \vdots \\ A_{g_m} \end{pmatrix} \cdot \mathbf{S}. \tag{3}$$

Finally, the third model assumes that both the sources and the mixing remains fixed across groups which implies that for all $k \in \{1, \ldots, m\}$ it holds that

$$\mathbf{X}_{g_k} = A \cdot \mathbf{S}. \tag{4}$$

In all three settings the baseline approach to ICA is to simply apply a classical ICA to the corresponding concatenated or averaged data, i.e., to apply the algorithm to the temporally/spatially concatenated data matrices on the left-hand side of above equations or the average over groups. These ad-hoc approaches are appealing, since they postulate straightforward procedures to solving the problem on grouped data and facilitate interpretability of the resulting estimates. It is these ad-hoc approaches that are implemented as the default behavior in toolboxes like the widely used eeglab for EEG analyses (Delorme and Makeig, 2004).

Several procedures have been proposed tailored to specific applications that extend on these baselines by employing additional assumptions. The most prominent such extensions are tensorial methods that have found popularity in fMRI analysis. They express the group index as an additional dimension (the data is thus viewed as a $\mathbb{R}^{d \times n \times m}$ tensor) and construct an estimate factorization of the tensor representation. Many of these procedures build on the so called PARAFAC (parallel factor analysis) model (Harshman, 1970). Recasting the tensor notation, this model is of the form (3) with $A_{g_k} = A \cdot D_{g_k}$ for all groups and for diagonal matrices $D_{g_1}, \ldots, D_{g_m}$. As can be seen from this representation, the PARAFAC model allows the mixing matrices to change across groups while they are constrained to be the same up to different scaling of the mixing matrix columns (intuitively, across groups the source dimensions are allowed to project with different strengths onto the observed signal dimensions). Given that the matrices $D_{g_1}, \ldots, D_{g_m}$ are sufficiently different it is possible to estimate this model uniquely without further assumptions. However, in the case that some of these diagonal matrices are equal identifiability is lost. In such cases Beckmann and Smith (2005) suggest to additionally require that the individual components of the sources be independent. This is comparable to the case where uncorrelatedness may not be sufficient for the separation of sources while independence is.

The coroICA procedure also allows for grouped-data but aims at inferring a fixed mixing matrix A, i.e., a model as given in (2) is considered. In contrast to vanilla concatenation procedures, our methodology naturally incorporates changes across groups by allowing and adjusting for different stationary confounding noise in each group. We argue why this leads to a more robust procedure and also illustrate this in our simulations and real data

experiments. More generally, our goal is to learn an unmixing which allows to generalize to new and previously unseen groups; think for example about learning an unmixing based on several different training subjects and extending it to new so far unseen subjects. Such tasks can appear in brain-computer interfacing applications and can also be of relevance more broadly in feature learning for classification tasks where classification models are to be transferred from one group/domain to another. Since our aim is to learn a fixed mixing matrix A that is confounding-robust and readily applicable to new groups, coroICA cannot naturally be compared to models that are based on spatial concatenation (3) or fixed sources *and* mixings (4); these methods employ fundamentally different assumptions on the model underlying the data generating process, the crucial difference being that we allow the sources and their time courses to change between groups.

1.2. Our Contribution

One strength of our methodology is that it explicates a statistical model that is sensible for data with group structure and can be estimated efficiently, while being supported by provable identification results. Furthermore, providing an explicit model with all required assumptions enables a constructive discussion about the appropriateness of such modeling decisions in specific application scenarios. The model itself is based on a notion of invariance against confounding structures from groups, an idea that is also related to invariance principles in causality (Haavelmo, 1944; Peters et al., 2016); see also Section 3 for a discussion on the relation to causality.

We believe that coroICA is a valuable contribution to the ICA literature on the following grounds:

- We introduce a methodologically sound framework which extends ordinary ICA to settings with grouped data and confounding noise.
- We prove identifiability of the unmixing matrix under mild assumptions, importantly, we explicitly allow for time-dependent noise thereby lessening the assumptions required by existing noisy ICA methods.
- We provide an easy to implement estimation procedure.
- We illustrate the usefulness, robustness, applicability, and limitations of our newly introduced coroICA algorithm as well as characterize the advantage of coroICA over existing ICAs: The source separation by coroICA is more stable across groups since it explicitly accounts for group-wise stationary confounding.
- We provide an open-source scikit-learn compatible ready-to-use Python implementation available as coroICA from the Python Package Index repository as well as R and Matlab implementations and an intuitive audible example which is available at https://sweichwald.de/coroICA/.

2. Methodology

We consider a general noisy ICA model inspired by ideas employed in causality research (see Section 3). We argue below that it allows to incorporate group structure and enables joint inference on multi-group data in a natural way. For the model description, let $S_i = (S_i^1, \ldots, S_i^d)^\top \in \mathbb{R}^{d\times 1}$ and $H_i = (H_i^1, \ldots, H_i^d)^\top \in \mathbb{R}^{d\times 1}$ be two independent vector-valued sequences of random variables where $i \in \{1, \ldots, n\}$. The components $S_i^1, \ldots, S_i^d$ are

assumed to be mutually independent for each i while, importantly, we allow for any weakly stationary noise H. Let $A \in \mathbb{R}^{d\times d}$ be an invertible matrix. The d-dimensional data process $(X_i)_{i\in\{1,\dots,n\}}$ is generated by the following noisy linear mixing model

$$X_i = A \cdot S_i + H_i, \quad \text{for all } i \in \{1,\dots,n\}. \tag{5}$$

X is a linear combination of source signals S and confounding variables H. In this model, both S and H are unobserved. One aims at recovering the mixing matrix A as well as true source signals S from observations of X. Without additional assumptions, the confounding H makes it impossible to identify the mixing matrix A. Even with additional assumptions it remains a difficult task (see Section 1.1.1 for an overview of related ICA models). Given the mixing matrix A it is straightforward to recover the confounded source signals $\widetilde{S}_i = S_i + A^{-1} \cdot H_i$.

Throughout this paper, we denote by $\mathbf{X} = (X_1,\dots,X_n) \in \mathbb{R}^{d\times n}$ the observed data matrix and similarly by $\mathbf{S}$ and $\mathbf{H}$ the corresponding (unobserved) source and confounding data matrices. For a finite data sample generated by this model we hence have

$$\mathbf{X} = A \cdot \mathbf{S} + \mathbf{H}.$$

In order to distinguish between the confounding H and the source signals S we assume that the two processes are sufficiently different. This can be achieved by assuming the existence of a group structure such that the covariance of the confounding H remains stationary within a group and only changes across groups.

Assumption 1 (group-wise stationary confounding) *There exists a collection of m disjoint groups $\mathcal{G} = \{g_1,\dots,g_m\}$ with $g_k \subseteq \{1,\dots,n\}$ and $\cup_{k=1}^m g_k = \{1,\dots,n\}$ such that for all $g \in \mathcal{G}$ the process $(H_i)_{i\in g}$ is weakly stationary.*

Under this assumption and given that the source signals change enough within groups, the mixing matrix A is identifiable (see Section 2.2). Similar to existing ICA methods discussed in Section 1.1.2, we propose to estimate the mixing matrix A by jointly diagonalizing empirical estimates of dependence matrices. In contrast to existing methods, we explicitly allow and adjust for the confounding H. The process of finding a matrix V that simultaneously diagonalizes a set of matrices is known as joint matrix diagonalization and has been studied extensively (e.g., Ziehe et al., 2004; Tichavsky and Yeredor, 2009). In Section 2.3, we show how to construct an estimator for V based on approximate joint matrix diagonalization.

The key step in adjusting for the confounding is to make use of the assumption that in contrast to the signals S the confounding H remains stationary within groups. Depending on the type of signal in the sources one can consider different sets of matrices. Here, we distinguish between two types of signals.

Variance signal In case of a variance signal, the variance process of each signal source $\mathrm{Var}(S_i^j)$ changes over time. These changes can be detected by examining the covariance matrix $\mathrm{Cov}(X_i)$ over time. For $V = A^{-1}$ and using (5) it holds for all $i \in \{1,\dots,n\}$ that

$$V\,\mathrm{Cov}(X_i)V^\top = \mathrm{Cov}(S_i) + V\,\mathrm{Cov}(H_i)V^\top.$$

Since the source signal components S_i^j are mutually independent, the covariance matrix $\mathrm{Cov}(S_i)$ is diagonal. Moreover, due to Assumption 1 the covariance matrix of the confounding H is constant, though not necessarily diagonal, within each group. This implies for all groups $g \in \mathcal{G}$ and for all $k, l \in g$ that

$$V\left(\mathrm{Cov}(X_k) - \mathrm{Cov}(X_l)\right)V^\top = \mathrm{Cov}(S_k) - \mathrm{Cov}(S_l) \tag{6}$$

is a diagonal matrix.

Time-dependence signal In case of a time-dependence signal, the time-dependence of each signal source S_i^j changes over time, i.e., for fixed τ, $\mathrm{Cov}(S_i^j, S_{i-\tau}^j)$ changes over time. These changes lead to changes in the auto-covariance matrices $\mathrm{Cov}(X_i, X_{i-\tau})$. Analogous to the variance signal it holds for all $i \in \{\tau+1, \ldots, n\}$ that

$$V\,\mathrm{Cov}(X_i, X_{i-\tau})V^\top = \mathrm{Cov}(S_i, S_{i-\tau}) + V\,\mathrm{Cov}(H_i, H_{i-\tau})V^\top.$$

Since the source signal components S_i^j are mutually independent, the auto-covariance matrix $\mathrm{Cov}(S_i, S_{i-\tau})$ is diagonal and due the stationarity of H (see Assumption 1) the auto-covariance $\mathrm{Cov}(H_i, H_{i-\tau})$ is constant within each group. This implies for all groups $g \in \mathcal{G}$, for all $k, l \in g$ and for all τ that

$$V\left(\mathrm{Cov}(X_k, X_{k-\tau}) - \mathrm{Cov}(X_l, X_{l-\tau})\right)V^\top = \mathrm{Cov}(S_k, S_{k-\tau}) - \mathrm{Cov}(S_l, S_{l-\tau}) \tag{7}$$

is a diagonal matrix.

For both signal types, we can identify V by simultaneously diagonalizing differences of (auto-)covariance matrices. Details and identifiability results are given in Section 2.3. The two signal types considered differ from both, the more classical settings of non-Gaussian time-independent signals as considered for example by fastICA, and the stationary signals with fixed time-dependence assumed for SOBI (cf. Table 1). Owing to the non-stationarity of the signal we can allow for more general forms of noise.

2.1. Motivating Examples

To get a better understanding of our proposed ICA model in (5), we illustrate two different aspects: the group structure and the noise model.

Noise model coroICA can be viewed as a noisy ICA, where the noise is allowed to be group-wise non-stationary. This generalizes existing noisy ICA methods, which, to the best of our knowledge, all assume that the noise is independent over time with various additional assumptions. The following example illustrates the intuition behind our model via a toy-application to natural images.

Example 1 (unmixing noisy images) *We provide an illustration of how our proposed method compares to other ICA approaches under the presence of noise. Four images, each* 450×300 *pixels and with three RGB color channels, are used to construct four sources* S^1, S^2, S^3, S^4 *as follows.*[3] *Every color channel is converted to a one dimensional vector by*

3. The images are freely available from Pexels GmbH (2018).

cutting each image into 15×10 *equally sized patches (i.e., each patch consists of* 30×30 *pixels) and concatenating the row-wise vectorized patches. This procedure preserves the local structure of the image. We concatenate the three color channels and consider them as separate groups for our model. Thus, each of the four sources* $S^1, \ldots, S^4$ *consists of* $n = 3 \cdot 450 \cdot 300 = 405.000$ *observations, that is, three groups of* 135.000 *observations corresponding to the RGB color channels. Next, we construct locally dependent noise that differs across color channels. Here, locally dependent means that the added noise is similar (and dependent) for pixels which are close to each other. This results in four noise processes* $H^1, \ldots, H^4$*. We combine the sources with the noise and apply a random mixing matrix* A *to obtain the following observed data*

$$X = A \cdot S + H.$$

The recast noisy images $\widetilde{S} = S + A^{-1}H$ *are illustrated in the first row and the recast observed mixtures* X *in the second row of Figure 1. The last three rows are the resulting reconstructions of three different ICA procedures,* coroICA (TD)*, fastICA and choiICA (TD). As expected, fastICA as a noise-free ICA method, appears frail to the noise in the images. While choiICA (TD) is able to adjust for independent noise, it is unable to properly adjust for the spatial dependence of the noise process and thus leads to undesired reconstruction results. In contrast,* coroICA (TD) *is able to recover the noisy images. It is the noise and its characteristics that break the two competing ICA methods, since all three methods are able to unmix the images in the noise-free case (not shown here).*

The noise model we employ is motivated by recent advances in causality research where the group-wise stationary noise can be interpreted as unobserved confounding factors in linear causal feedback models. We describe this in more detail with an explicit example application to Antarctic ice core data in Section 3.

Group structure A key aspect of our model is that it aims to leverage group-structure to improve the stability of the umixing under the presence of group-wise confounding. Here we refer to the following notion of stability: A stable unmixing matrix extracts the same set of independent sources when applied to the different groups; it is robust against the confounding that varies across groups and introduces dependences. A standard ICA method is not able to estimate the correct unmixing $V = A^{-1}$, if the data generating process follows our confounded ICA model in (5). These methods extract signals that are not only corrupted by the group-wise confounding but also are mixtures of the independent sources and are thus not stable in the aforementioned sense. This is illustrated by the "America's Got Talent Duet Problem" (cf. Example 2), an extension and alteration of the classical "cocktail party problem".

Example 2 (America's Got Talent Duet Problem) *Consider the problem of evaluating two singers at a duet audition individually. This requires to listen to the two voices separately, while the singers perform simultaneously. There are two sound sources in the audition room (the two singers) and additionally several noise sources which corrupt the recordings at the two microphones (or the jury member's two ears). A schematic of such a setting is illustrated in Supplement B, Figure 12. The additional noise comes from an audience and two open windows. One can assume that this noise satisfies our Assumption 1*

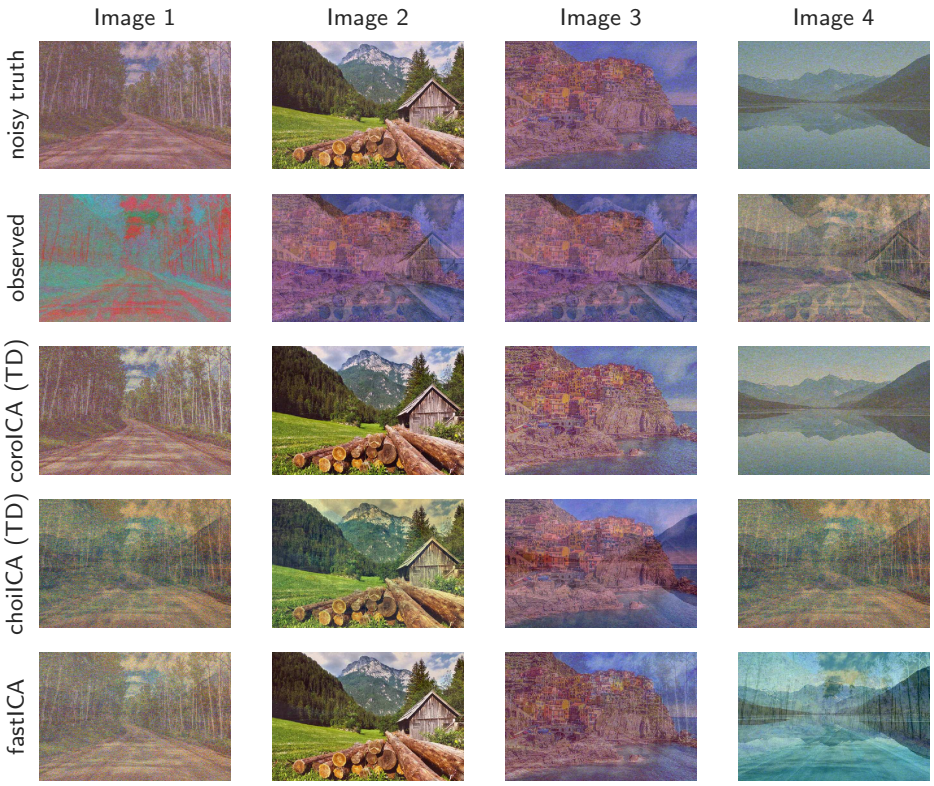


Figure 1. Images accompanying example 1. The top row shows noisy unmixed images, the second row shows mixed images, and the last three rows show unmixed and rescaled images resulting from an application of coroICA (TD), choiICA (TD) and fastICA (cf. Table 1). Here, only coroICA (TD) is able to correctly unmix the images and recover the original (noise-corrupted) images.

on a single group. The sound stemming from the audience can be seen as an average of many sounds, hence remaining approximately stationary over time. Typical sounds from an open window also satisfy this assumption, for example sound from a river or a busy road. Our methodology, however, also allows for more complicated settings in which the noise shifts at known points in times, for example if someone opens or closes a window or starts mowing the lawn outside. In such cases we use the known time blocks of stationary noise as groups and apply coroICA (var) on this grouped data. An example with artificial sound data related to this setting is available at `https://sweichwald.de/coroICA/`*. We show that coroICA (var) is able to recover useful sound signals with the two voices being separated into different dimensions and thus allows to listen to them individually. In contrast, existing ICAs applied to the time concatenated data fail to unmix the two singers.*

2.2. Identifiability

Identifiability requires that the source signals S change sufficiently strong within groups. The precise notion of a strong signal depends on the type of signal. As discussed previously,

we consider two types of non-stationary signals (i) variance signals and (ii) time-dependence signals. Depending on the signal type we formalize two slightly different assumptions that characterize source signals that ensure identifiability. Firstly, in the case of a variance signal, we have the following assumption.

Assumption 2 (signals with independently changing variance) *For each pair of components $p, q \in \{1, \dots, d\}$ we require the existence of three (not necessarily unique) groups $g_1, g_2, g_3 \in \mathcal{G}$ and three corresponding pairs $l_1, k_1 \in g_1$, $l_2, k_2 \in g_2$ and $l_3, k_3 \in g_3$ such that the two vectors*

$$\begin{pmatrix} \mathrm{Var}\left(S^p_{l_1}\right) - \mathrm{Var}\left(S^p_{k_1}\right) \\ \mathrm{Var}\left(S^p_{l_2}\right) - \mathrm{Var}\left(S^p_{k_2}\right) \\ \mathrm{Var}\left(S^p_{l_3}\right) - \mathrm{Var}\left(S^p_{k_3}\right) \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \mathrm{Var}\left(S^q_{l_1}\right) - \mathrm{Var}\left(S^q_{k_1}\right) \\ \mathrm{Var}\left(S^q_{l_2}\right) - \mathrm{Var}\left(S^q_{k_2}\right) \\ \mathrm{Var}\left(S^q_{l_3}\right) - \mathrm{Var}\left(S^q_{k_3}\right) \end{pmatrix}$$

are neither collinear nor equal to zero.

In case of time-dependence signals we have the analogous assumption.

Assumption 3 (signals with independently changing time-dependence) *For each pair of components $p, q \in \{1, \dots, d\}$ we require the existence of three (not necessarily unique) groups $g_1, g_2, g_3 \in \mathcal{G}$ and three corresponding pairs $l_1, k_1 \in g_1$, $l_2, k_2 \in g_2$ and $l_3, k_3 \in g_3$ for which there exists $\tau \in \{1, \dots, n\}$ such that the two vectors*

$$\begin{pmatrix} \mathrm{Cov}\left(S^p_{l_1}, S^p_{l_1-\tau}\right) - \mathrm{Cov}\left(S^p_{k_1}, S^p_{k_1-\tau}\right) \\ \mathrm{Cov}\left(S^p_{l_2}, S^p_{l_2-\tau}\right) - \mathrm{Cov}\left(S^p_{k_2}, S^p_{k_2-\tau}\right) \\ \mathrm{Cov}\left(S^p_{l_3}, S^p_{l_3-\tau}\right) - \mathrm{Cov}\left(S^p_{k_3}, S^p_{k_3-\tau}\right) \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \mathrm{Cov}\left(S^q_{l_1}, S^q_{l_1-\tau}\right) - \mathrm{Cov}\left(S^q_{k_1}, S^q_{k_1-\tau}\right) \\ \mathrm{Cov}\left(S^q_{l_2}, S^q_{l_2-\tau}\right) - \mathrm{Cov}\left(S^q_{k_2}, S^q_{k_2-\tau}\right) \\ \mathrm{Cov}\left(S^q_{l_3}, S^q_{l_3-\tau}\right) - \mathrm{Cov}\left(S^q_{k_3}, S^q_{k_3-\tau}\right) \end{pmatrix}$$

are neither collinear nor equal to zero.

Intuitively, these assumptions ensure that the signals are not changing in exact synchrony across components, which removes degenerate types of signals. In particular, they are satisfied in the case that the variance or auto-covariance processes change pair-wise independently over time. Whenever one of these assumptions is satisfied, the mixing matrix A is uniquely identifiable.

Theorem 1 (identifiability of the mixing matrix)
Assume the data process $(X_i)_{i \in \{1,\dots,n\}}$ satisfies the model in (5) and Assumption 1 holds. If additionally either Assumption 2 or Assumption 3 is satisfied, then A is unique up to permutation and rescaling of its columns.

Proof A proof is given in Supplement A. ■

2.3. Estimation

In order to estimate V from a finite observed sample $\mathbf{X} \in \mathbb{R}^{d \times n}$, we first partition each group into subgroups. We then compute the empirical (auto-)covariance matrices on each

subgroup. Finally, we estimate a matrix that simultaneously diagonalizes the differences of these empirical (auto-)covariance matrices using an approximate joint matrix diagonalization technique. This procedure results in three methods depending on which type of matrices we diagonalize. Similar to our notation for the different versions of choiICAs we denote these methods by coroICA (var) if we diagonalize differences of covariances, coroICA (TD) if we diagonalize differences of auto-covariances, and coroICA (var & TD) if we diagonalize both differences of covariance and auto-covariances.

More precisely, for each group $g \in \mathcal{G}$, we first construct a partition $\mathcal{P}_g$ consisting of subsets of g such that each $e \in \mathcal{P}_g$ satisfies that $e \subseteq g$ and $\cup_{e \in \mathcal{P}_g} e = g$. This partition $\mathcal{P}_g$ should be granular enough to capture the changes in the signals described in Assumption 2 or 3. We propose partitioning each group based on a grid such that the separation between grid points is large enough for a reasonable estimation of the covariance matrix and at the same time small enough to capture variations in the signals. In our experiments, we observed robustness with respect to the exact choice; only too small partitions should be avoided since otherwise the procedure is fragile due to poorly estimated covariance matrices. More details on the choice of the partition size are given in Remark 2. Depending on whether a variance or time-dependence signal or a hybrid thereof is considered, we fix time lags $\mathrm{T} \subset \mathbb{N}_0$.

Next, for each group $g \in \mathcal{G}$, each distinct pair $e, f \in \mathcal{P}_g$, and each $\tau \in \mathrm{T}$ we define the matrix

$$M_{e,f}^{g,\tau} := \widehat{\mathrm{Cov}}_\tau(\mathbf{X}_e) - \widehat{\mathrm{Cov}}_\tau(\mathbf{X}_f),$$

where $\widehat{\mathrm{Cov}}_\tau(\cdot)$ denotes the empirical (auto-)covariance matrix for lag τ and $\mathbf{X}_e$ is the data matrix restricted to the columns corresponding to the subgroup e. Assumption 1 ensures that $V M_{e,f}^{g,\tau} V^\top$ is approximately diagonal. We are therefore interested in finding an invertible matrix V which approximately jointly diagonalizes the matrices in the set

$$\mathcal{M}^{\mathrm{all}} := \left\{ M_{e,f}^{g,\tau} \,\middle|\, g \in \mathcal{G} \text{ and } e, f \in \mathcal{P}_g \text{ and } \tau \in \mathrm{T} \right\}. \tag{8}$$

The number of matrices in this set grows quadratically in the number of partitions. This can lead to large numbers of matrices to be diagonalized. Another option that reduces the computational load is to compare each partition to its complement, which leads to the following set of matrices

$$\mathcal{M}^{\mathrm{comp}} := \left\{ M_{e,\bar{e}}^{g,\tau} \,\middle|\, g \in \mathcal{G} \text{ and } e \in \mathcal{P}_g \,(\text{with } \bar{e} := g \setminus e) \text{ and } \tau \in \mathrm{T} \right\} \tag{9}$$

or to compare only neighboring partitions as in

$$\mathcal{M}^{\mathrm{neighbor}} := \left\{ M_{e,\mathrm{neighbor}(e)}^{g,\tau} \,\middle|\, g \in \mathcal{G} \text{ and } e \in \mathcal{P}_g \text{ and } \tau \in \mathrm{T} \right\}, \tag{10}$$

where neighbor(e) is the partition to the right of e.

The task of jointly diagonalizing a set of matrices is a well-studied topic in the literature and is referred to as approximate joint matrix diagonalization. Many solutions have been proposed for different assumptions made on the matrices to be diagonalized. In this paper, we use the uwedge algorithm[4] introduced by Tichavsky and Yeredor (2009). The basic idea

4. As a byproduct of our work, we are able to provide a new stable open-source Python/R/Matlab implementation of the uwedge algorithm which is also included in our respective coroICA packages.

behind uwedge is to find a minimizer of a proxy for the loss function

$$\ell(V) = \sum_{M \in \mathcal{M}^*} \left(\sum_{k \neq l} \left[V M V^\top \right]_{k,l}^2 \right),$$

over the set of invertible matrices, where in our case $\mathcal{M}^* \in \{\mathcal{M}^{\text{all}}, \mathcal{M}^{\text{comp}}, \mathcal{M}^{\text{neighboring}}\}$.

The full estimation procedure based on the set $\mathcal{M}^{\text{neighbouring}}$ defined in (9) is made explicit in the pseudo code in Algorithm 1 (where ApproximateJointDiagonalizer stands for a general approximate joint diagonalizer; here we use uwedge).

Remark 2 (choosing the partition and the lags) *Whenever there is no obvious partition of the data, we propose to partition the data into equally sized blocks with a fixed partition size. The decision on how to choose a partition size should be driven by type of non-stationary signal one expects and the dimensionality of the data. For example, in the case of a variance signal the partition should be fine enough to capture areas of high and low variance, while at the same time being coarse enough to allow for sufficiently good estimates of the covariance matrices. That said, for applications to real data sets the signals are often of various length implying that there is a whole range of partition sizes which all work well. In cases with few data points, it can then be useful to consider several grids with different partition sizes and diagonalize across all resulting differences simultaneously. This somewhat removes the dependence of the results on the exact choice of a partition size and increases the power of the procedure. We employ this approach in Section 3.1. In general, the lags* T *should be chosen as* $\mathrm{T} = \{0\}$, $\mathrm{T} \subset \mathbb{N}$, *or* $\mathrm{T} \subset \mathbb{N}_0$, *depending on whether a variance signal, time-dependence signal, or a hybrid thereof is considered. For time-dependence signal, we recommend to determine up to which time-lag the autocorrelation of the observed signals has sufficiently decayed, and use all lags up to that point.*

2.4. Assessing the Quality of Recovered Sources

Assessing the quality of the recovered sources in an ICA setting is an inherently difficult task, as is typical for unsupervised learning procedures. The unidentifiable scale and ordering of the sources as well as the unclear choice of a performance measure render this task difficult. Provided that ground truth is known, several scores have been proposed, most notably the Amari measure introduced by Amari et al. (1995) and the minimum distance (MD) index due to Ilmonen et al. (2010). Here, we use the MD index, which is defined as

$$\mathrm{MD}(\hat{V}, A) = \frac{1}{\sqrt{p-1}} \inf_{C \in \mathcal{C}} \|C\hat{V}A - \mathrm{Id}\|,$$

where the set $\mathcal{C}$ consists of matrices for which each row and column has exactly one nonzero element. Intuitively, this score measures how close $\hat{V}A$ is to a rescaled and permuted version of the identity matrix. One appealing property of this score is that it can be computed efficiently by solving a linear sum assignment problem. In contrast to the Amari measure, the MD index is affine invariant and has desirable theoretical properties (see Ilmonen et al., 2010).

Algorithm 1: coroICA

input : data matrix $\mathbf{X}$
group index $\mathcal{G}$ (user selected)
group-wise partition $(\mathcal{P}_g)_{g\in\mathcal{G}}$ (user selected)
lags $\mathrm{T} \subset \mathbb{N}_0$ (user selected)

initialize empty list $\mathcal{M}$
for $g \in \mathcal{G}$ **do**
  for $e \in \mathcal{P}_g$ **do**
    for $\tau \in \mathrm{T}$ **do**
      append $\widehat{\mathrm{Cov}}_\tau(\mathbf{X}_e) - \widehat{\mathrm{Cov}}_\tau(\mathbf{X}_{\mathrm{neighbour}(e)})$ to list $\mathcal{M}$
    end
  end
end
$\widehat{V} \leftarrow \mathrm{ApproximateJointDiagonalizer}(\mathcal{M})$
$\widehat{\mathbf{S}} \leftarrow \widehat{V}\mathbf{X}$

output: unmixing matrix $\widehat{V}$
sources $\widehat{\mathbf{S}}$

We require a different performance measure for our real data experiments where the true unmixing matrix is unknown. Here, we check whether the desired independence (after adjustment for the constant confounding) is achieved by computing the following covariance instability score (CIS) matrix. It measures the instability of the covariance structure of the unmixed sources $\widehat{\mathbf{S}}$ and is defined for a each groups $g \in \mathcal{G}$ and a corresponding partition $\mathcal{P}_g$ (see Section 2.3) by

$$\mathrm{CIS}(\widehat{\mathbf{S}}, \mathcal{P}_g) \coloneqq \frac{1}{|\mathcal{P}_g|} \sum_{e\in\mathcal{P}_g} \left(\frac{\widehat{\mathrm{Cov}}(\widehat{\mathbf{S}}_e) - \widehat{\mathrm{Cov}}(\widehat{\mathbf{S}}_{\mathrm{neighbour}(e)})}{\widehat{\sigma}_{\widehat{\mathbf{S}}_g} \cdot \widehat{\sigma}_{\widehat{\mathbf{S}}_g}^\top} \right)^2,$$

where $\widehat{\sigma}_{\widehat{\mathbf{S}}} \in \mathbb{R}^{d\times 1}$ is the empirical standard deviation of $\widehat{\mathbf{S}}$ and the fraction is taken element-wise. The CIS matrix is approximately diagonal whenever $\widehat{\mathbf{S}}$ can be written as the sum of independent source signals $\mathbf{S}$ and confounding $\mathbf{H}$ with fixed covariance. This is condensed into one scalar that reflects how stable the sources' covariance structure is by averaging the off-diagonals of the CIS matrix

$$\mathrm{MCIS}(\widehat{\mathbf{S}}, \mathcal{P}_g)^2 \coloneqq \frac{1}{d(d-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{d} \Big[\mathrm{CIS}(\widehat{\mathbf{S}}, \mathcal{P}_g)\Big]_{i,j}.$$

The differences taken in the CIS score extract the variance signals such that the mean covariance instability score (MCIS) can be understood as a measure of independence between the recovered variance signal processes. High values of MCIS imply strong dependences beyond stationary confounding between the signals. Low values imply weak dependences. MCIS is a reasonable score whenever there is a *variance signal* (as described in Section 2)

in sources and is a sensible evaluation metric of ICA procedures in such cases. In case of time-dependence signal (as described in Section 2), one can define an analogous score based on the auto-covariances. Here, we restrict ourselves to the variance signal case as for all our applications this appeared to constitute the dominant part of the signal.

In case of *variance signals* the MCIS appears natural and appropriate as independence measure: It measures how well the individual variance signals (and hence the relevant information) are separated. To get a better intuition, let $A = (a_1, \ldots, a_d) \in \mathbb{R}^{d\times d}$ denote the mixing and $V = (v_1, \ldots, v_d)^\top \in \mathbb{R}^{d\times d}$ the corresponding unmixing matrix (i.e., $V = A^{-1}$, a_i are columns of A and v_i are rows of V). Then it holds that,

$$\begin{aligned}\operatorname{Cov}(X_i)v_j^\top &= A\operatorname{Cov}(S_i)A^\top v_j^\top + \operatorname{Cov}(H_i)v_j^\top \\ &= A\operatorname{Cov}(S_i)e_j^\top + A\operatorname{Cov}(H_i)e_j^\top \\ &= a_j \operatorname{Var}(S_i^j) + A\operatorname{Cov}(H_i)e_j^\top \end{aligned} \tag{11}$$

Under our group-wise stationary confounding assumption (Assumption 1) this implies that within all groups $g \in \mathcal{G}$, it holds for all $l, k \in g$ that

$$\left(\operatorname{Cov}(X_l) - \operatorname{Cov}(X_k)\right) v_j^\top = a_j \left(\operatorname{Var}(S_l^j) - \operatorname{Var}(S_k^j)\right). \tag{12}$$

This equation holds also in the confounding-free case and it reflects the contribution of the signal (in terms of variance signal) of the j-th recovered source S^j to the the variance signal in all components of the observed multivariate data X.

While in the population case the equality in (12) is satisfied exactly, this is no longer the case when the (un-)mixing matrix is estimated on finite data. Consider two subsets $e, f \in g$ for some group $g \in \mathcal{G}$, then using the notation from Section 2.3 and denoting by $\widehat{v}_j$ and $\widehat{a}_j$ the estimates of v_j and a_j, respectively, it holds that

$$\begin{aligned} M^g_{e,f}\widehat{v}_j^\top &= \big[\widehat{\operatorname{Cov}}(\mathbf{X}_e) - \widehat{\operatorname{Cov}}(\mathbf{X}_f)\big]\widehat{v}_j^\top \\ &= \widehat{A}\big[\widehat{\operatorname{Cov}}(\widehat{S}_e) - \widehat{\operatorname{Cov}}(\widehat{S}_f)\big]\widehat{A}^\top\widehat{v}_j^\top \\ &= \widehat{A}\big[\widehat{\operatorname{Cov}}(\widehat{S}_e) - \widehat{\operatorname{Cov}}(\widehat{S}_f)\big]e_j^\top \\ &\approx \widehat{a}_j(\operatorname{Var}(S_e^j) - \operatorname{Var}(S_f^j)). \end{aligned} \tag{13}$$

The approximation is close only if the empirical estimate $\widehat{V}$ correctly unmixes the j-th source. Essentially, MCIS measures the extent to which this approximation holds true for all components simultaneously across the subsets specified by the partition $\mathcal{P}_g$. It is also possible to consider individual components by assessing how closely the following proportionality is satisfied

$$\sum_{M\in\mathcal{M}^*} \operatorname{sign}(\widehat{v}_j M \widehat{v}_j^\top) M\widehat{v}_j^\top \propto \widehat{a}_j. \tag{14}$$

In EEG experiments, this can also be assessed visually by comparing the topographic maps corresponding to columns of A with so-called activation maps corresponding to the left-hand side in (14). More details on this are provided in Section 4.3.3.

3. Causal Perspective

Our underlying noisy ICA model (5) and the assumption on the noise (Assumption 1) are motivated by causal structure learning scenarios. ICA is closely linked to the problem of identifying structural causal models (SCMs) (see Pearl, 2009; Imbens and Rubin, 2015; Peters et al., 2017). Shimizu et al. (2006) were the first to make this connection explicit and used ICA to infer causal structures. To make this more precise consider the following linear SCM

$$X_i = B \cdot X_i + \widetilde{S}_i, \tag{15}$$

where X_i are observed covariates and $\widetilde{S}_i$ are noise terms. An SCM induces a corresponding causal graph over the involved variables by drawing an edge from variables on the right-hand side to the one on the left-hand side of (15). Moreover, we can define noise interventions (Pearl, 2009) by allowing the distributions of the noise terms $\widetilde{S}_i$ to change for different i. In the language of ICA, this means that the signals $\widetilde{S}_i$ encode the different interventions (over time) on the noise variables. Assuming that the matrix $\mathrm{Id} - B$ is invertible, we can rewrite (15) as

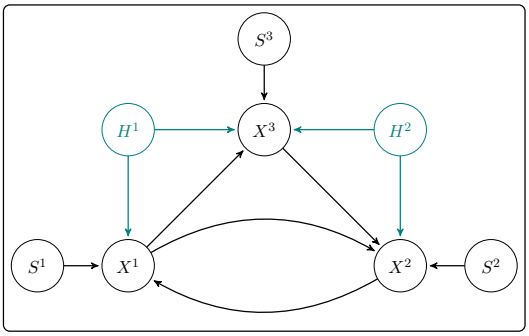
$$X_i = (\mathrm{Id} - B)^{-1}\widetilde{S}_i,$$

which can be viewed as an ICA model with mixing matrix $A = (\mathrm{Id} - B)^{-1}$. Instead of taking the noise term $\widetilde{S}_i$ as independent noise sources one can also consider $\widetilde{S}_i = S_i + H_i$. In that case the linear SCM in (15) describes a causal model between the observed variables X_i in which hidden confounding is allowed. This is illustrated in Figure 2, which depicts a 3 variable SCM with feedback loops and confounding. Learning a causal model as in (15) with ICA is generally done by performing the following two steps.

(i) **(ICA)** The matrix $(\mathrm{Id} - B)$ is inferred by ICA up to an undefined scale and permutation of its rows by using an appropriate ICA procedure. This step is often infeasible in the presence of confounding H since existing ICA methods only allow noise under restrictive assumptions (cf. Table 1).

(ii) **(identify B)** There are essentially two assumptions that one can make in order for this to work. The first is to assume the underlying causal model has an acyclic structure as in Shimizu et al. (2006). In such cases the matrix B needs to be permuted to an upper triangular matrix. The second option is to allow for feedback loops in the causal model but restrict the types of feedback to exclude infinite loops as in Hoyer et al. (2008) and Rothenhäusler et al. (2015).

When performing step (i) there are two important modeling assumptions that are made when selecting the ICA procedure: (a) the type of allowed signals (types of interventions) and (b) the type of allowed confounding. For the classic ICA setting with non-Gaussian source signals and no noise this translates to the class of linear non-Gaussian models, such as Linear Non-Gaussian Acyclic Models (LiNGAMs) introduced by Shimizu et al. (2006). While such models are a sensible choice in a purely observational setting (i.e., no samples from interventional settings) they are somewhat misspecified in terms of (a) when data from different interventional settings or time-continuous intervention shifts are observed (see Remark 3). In those settings, it is more natural to use ICA methods that are tailored to sequential shifts as for example choiICA or coroICA. Moreover, most common ICA methods



$$
\begin{aligned}
X^1 &\leftarrow b_{1,2}X^2 && + S^1 + H^1 \\
X^2 &\leftarrow b_{2,1}X^1 + b_{2,3}X^3 && + S^2 + H^2 \\
X^3 &\leftarrow b_{3,1}X^1 && + S^3 + H^1 + H^2
\end{aligned}
$$

Figure 2. Illustration of an SCM with (including colored nodes H^1, H^2) and without (excluding colored nodes) confounding.

consider noise-free mixing, which from a causal perspective implies that no hidden confounding is allowed. While noisy ICA weakens this assumption, existing methods only allow for time-independent or even iid noise, which again greatly restricts the type of confounding. In contrast, our proposed coroICA allows for any type of block-wise stationary confounding, hence greatly increasing the class of causal models which can be inferred. This is attractive for causal modeling as it is a priori unknown whether hidden confounding exists. Therefore, our proposed procedure allows for robust causal inference under general confounding settings. In Section 3.1, we illustrate a potential application to climate science and how the choice of ICA can have a strong impact on the estimates of the causal parameters.

Remark 3 (relation between interventions and non-stationarity) *A causal model does not only describe the observational distribution but also the behavior of the data generating model under all of the allowed interventions. Here, we restrict the allowed interventions to distribution shifts in the source signals, that either change the distribution block-wise (e.g., abruptly changing environmental conditions) or continuously (e.g., continuous shifts in the environmental conditions). Any such shifts are by definition synonymous with the process S_i being non-stationary. In our proposed causal model (15) the non-stationarity of the signal therefore corresponds to shifts in the environmental conditions which can be utilized, using coroICA, to infer the underlying causal structure. From this perspective, the causal inference procedure we propose here is a method based on interventional data rather than plainly observational data, while the interventions are not exactly known.*

3.1. Application to Climate Science

To motivate the foregoing causal model we consider a prominent example from climate science: the causal relationship between carbon dioxide concentration (CO_2) and temperature (T). More precisely, we consider Antarctic ice core data that consists of temperature and carbon dioxide measurements of the past 800'000 years due to Bereiter et al. (2015, carbon dioxide) and Jouzel et al. (2007, temperature). We combined both temperature and carbon dioxide data and recorded measurements every 500 years by a cubic interpolation of the raw data. The data is shown in Figure 3 (right). Oversimplifying, one can model this data

as an SCM with time-lags as follows

$$\begin{pmatrix} \log(\mathrm{CO}_2)_t \\ \mathrm{T}_t \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & \beta \\ \alpha & 0 \end{pmatrix}}_{=B_0} \begin{pmatrix} \log(\mathrm{CO}_2)_t \\ \mathrm{T}_t \end{pmatrix} + \sum_{k=1}^{p} B_k \begin{pmatrix} \log(\mathrm{CO}_2)_{t-k} \\ \mathrm{T}_{t-k} \end{pmatrix} + \widetilde{S}_t, \tag{16}$$

where $\widetilde{S}_t = S_t + H_t$ with S_t component-wise independent non-stationary source signals and H_t a stationary confounding process. Vector-valued linear time-series models of this type are referred to as structural auto regressive models (SVARs) (see e.g., Lütkepohl, 2005). They have been previously analyzed in the confounding free-case by Hyvärinen et al. (2010), using an ICA based causal inference approach. A graphical representation of such a model is shown in Supplement B.2, Figure 13. In this example, we can think of the source signals S_t as being two independent summaries of important factors that affect both temperature and carbon dioxide and vary over time, e.g., environmental catastrophes like volcano eruptions and large wildfires, sunspot activity or ice-coverage. These variations can be considered as changing environmental conditions or interventions (see Remark 3). On the other hand the stationary confounding process H_t can be thought of as factors which affect both temperature and carbon dioxide in a constant fashion over time, for example this could be effects due the shifts in the earth's rotation axis.

Assuming that this was the true underlying causal model, we could use it to predict what happens under interventions. From a climate science perspective an interesting intervention is given by doubling the concentration of CO_2 and determining the resulting instantaneous (faster than 1000 years) effect on the temperature. This effect is commonly referred to as equilibrium climate sensitivity (ECS) due to CO_2 which is loosely defined as the change in degrees temperature associated with a doubling of the concentration of carbon dioxide in the earth's atmosphere. In the fifth assessment report of the United Nations Intergovernmental Panel on Climate Change it has been stated that "there is high confidence that ECS is extremely unlikely less than 1 °C and medium confidence that the ECS is likely between 1.5 °C and 4.5 °C and very unlikely greater than 6 °C" (Intergovernmental Panel on Climate Change, 2014, Chapter 10). Since the measurement frequency in our model is quite low (500 years) and we model the logarithm of carbon dioxide the ECS corresponds to

$$\mathrm{ECS} = \log(2)\alpha.$$

Estimating the model in (16) can be done by first fitting a vector auto-regressive model of the time lags using OLS resulting in a vector of residuals

$$R_t = \begin{pmatrix} \log(\mathrm{CO}_2)_t \\ \mathrm{T}_t \end{pmatrix} - \begin{pmatrix} \widehat{\log(\mathrm{CO}_2)_t} \\ \widehat{\mathrm{T}_t}. \end{pmatrix}$$

Then, one can apply the two-step causal inference procedure described in Section 3 to

$$R_t = B_0 R_t + \widetilde{S}_t.$$

Since we are in a two-dimensional setting, step (ii) (i.e., identifying the causal parameters α and β from the estimated mixing matrix) only requires to assume that feedback loops

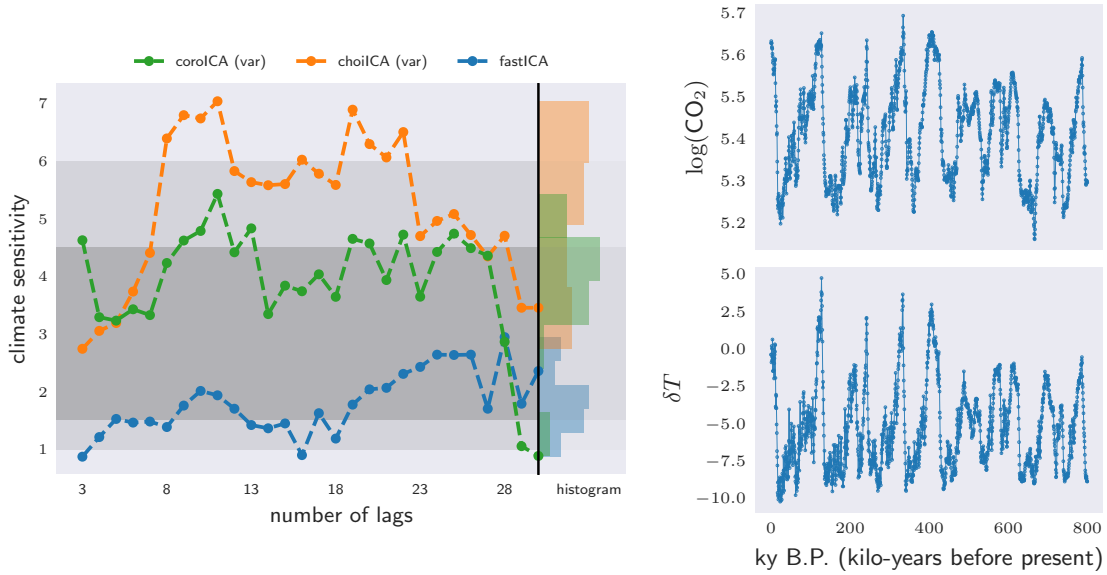


Figure 3. (left) Estimated equilibrium climate sensitivity (ECS) for different ICAs depending on the number of lags included into the SVAR model. The light gray and dark gray overlay indicate likely and very likely value ranges, respectively, for the true value of climate sensitivity as per the fifth assessment report of the United Nations Intergovernmental Panel on Climate Change (cf. Section 3.1). The differences across procedures illustrate that the choice of ICA has a large effect on the estimation. (right) Interpolated time-series data, which we model with an SVAR model.

do not blow-up, which translates into B_0 having spectral norm less than one. Given that the signal is sufficiently strong (i.e., there are sufficient interventions on both CO_2 and T), it is possible to recover the causal parameters by trying both potential permutations of the sources with subsequent scaling and assessing whether the aforementioned condition is satisfied.

We applied this procedure based on coroICA (var) to the data in order to estimate climate sensitivity and compared it with results obtained when using fastICA or choiICA (var). The results are given in Figure 3. We believe the results illustrate two important aspects. Firstly, the choice of the lags has a strong effect on the estimation of the causal effect parameters, particularly for boundary cases. If it is chosen too small the remaining time-dependence in the data can obscure the signal. If it is chosen too big part of the signal starts being removed. Choosing an appropriate number of lags is therefore crucial. One option would be to apply an information criterion (AIC or BIC) for this. Secondly, the results illustrate that the choice of ICA has a large impact on the estimated causal effect parameters. More specifically, both the assumed signal as well as the assumed confounding have an impact on the estimation. Compare the results between fastICA (non-Gaussian signal) and choiICA/coroICA (variance signal) for the former and observe the differences between fastICA/choiICA (no confounding) and coroICA (adjusted for stationary confounding) for the latter. The choice of the ICA algorithm should therefore be driven by the assumptions

(both on signal type and confounding) one is willing to employ on the underlying model. Considering a variance signal and adjusting for confounding, coroICA appears to lead to estimates of equilibrium climate sensitivity that are more closely in line with the highly likely bands previously identified by the United Nations Intergovernmental Panel on Climate Change. This observation is only indicative as all three methods yield highly variable results and also the panel's highly likely band rests on certain assumptions that may become refuted at some later point. coroICA can be considered a conservative choice if no assumptions on confounding can be made, while noise-free methods may outperform if indeed there were no confounding factors.

4. Experiments

In this section, we analyze empirical properties of coroICA. To this end, we first illustrate the performance of coroICA as compared to time-concatenated versions of (noisy) ICA variants on simulated data with and without confounding. We also compare on real data and outline potential benefits of using our method when analyzing multi-subject EEG data.

4.1. Competing Methods

In all of our numerical experiments, we apply coroICA as outlined in Algorithm 1, where we partition each group based on equally spaced grids and run a fixed number of $10 \cdot 10^3$ iterations of the uwedge approximate joint diagonalizer. Unless specified otherwise, coroICA refers to coroICA(var) (i.e., the variance signal based version) and we explicitly write coroICA(var), coroICA(TD) and coroICA(var & TD) whenever appropriate to avoid confusion. We compare with all of the methods in Table 1. Since no Python implementation was publicly available, we implemented the choiICAs and SOBI methods ourselves also based on a fixed number of $10 \cdot 10^3$ iterations of the uwedge approximate joint diagonalizer. For fastICA we use the implementation from the scikit-learn Python library due to Pedregosa et al. (2011) and use the default parameters.

For the simulation experiments in Section 4.2, we also compare to random projections of the sources, where the unmixing matrix is simply sampled with iid standard normal entries. The idea of this comparison is to give a baseline of the unmixing problem and enhance intuition about the scores' behavior. In order to illustrate the variance in this method, we generally sample 100 random projections and show the results for each of them. A random mixing does not lead to interpretable sources, thus we do not compare with random projections in the EEG experiments in Section 4.3.

4.2. Simulations

In this section, we investigate empirical properties of coroICA in well-controlled simulated scenarios. First off, we show that we can recover the correct mixing matrix given that the data is generated according to our model (5) and Assumptions 1 and 2 hold, while the other ICAs necessarily fall short in this setting (cf. Section 4.2.1). Moreover, in Section 4.2.2 we show that even in the absence of any confounding (i.e., when the data follows the ordinary ICA model and $H \equiv 0$ in our model) we remain competitive with all competing ICAs. Finally, in Section 4.2.3 we analyze the performance of coroICA for various types of signals

and noise settings. Our first two simulation experiments are based on block-wise shifting variance signals, which we describe in Data Set 1 and our third simulation experiment is based on GARCH type models described in Data Set 2.

4.2.1. Dependence on Confounding Strength

For this simulation experiment, we sample data according to Data Set 1 and choose to simulate $n = 10^5$ (dimension $d = 22$) samples from $m = 10$ groups where each group contains $n/m = 10^4$ observations. Within each group, we select a random partition consisting of $|\mathcal{P}_g| = 10$ subsets while ensuring that these have the same size on average. We fix the signal strength to $c_1 = 1$ and consider the behavior of coroICA (trained on half of the groups with an equally spaced grid of 10 partitions per group) for different confounding strengths $c_1 = \{0.125, 0.25, 0.5, 1, 1.5, 2, 2.5, 3\}$. The results for 1000 repetitions are shown in Figure 4. To allow for a fair comparison we take the same partition size for choiICA (var).

Data Set 1: Block-wise shifting variance signals

For our simulations we select m equally sized groups $\mathcal{G} \coloneqq \{g_1, \ldots, g_m\}$ of the data points $\{1, \ldots, n\}$ and for each group $g \in \mathcal{G}$ construct a partition $\mathcal{P}_g$. Then, we sample a model of the form

$$X_i = A \cdot (S_i + C \cdot H_i)\,,$$

where the values on the right-hand side are sampled as follows:

- $A, C \in \mathbb{R}^{d\times d}$ are sampled with iid entries from $\mathcal{N}(0,1)$ and $\mathcal{N}(0, \frac{1}{d})$, respectively.
- For each $g \in \mathcal{G}$ the variables $H_i \in \mathbb{R}^d$ are sampled from $\mathcal{N}(0, \sigma_g^2 \operatorname{Id}_d)$, where the σ_g^2 are sampled iid from $\operatorname{Unif}(0.1, b_1)$.
- For each $g \in \mathcal{G}$ and $e \in \mathcal{P}_g$ the variables $S_i \in \mathbb{R}^d$ are sampled from $\mathcal{N}(0, \eta_e^2 \operatorname{Id}_d)$, where the η_e^2 are sampled iid from $\operatorname{Unif}(0.1, b_2)$.

The parameters b_1 and b_2 are selected in such a way that the expected confounding strength $c_1 = \mathbb{E}(\sigma_g^2)$ and variance signal strength $c_2 \coloneqq \mathbb{E}(|\eta_e^2 - \eta_f^2|)$ are as dictated by the respective experiment. Due to the uniform distribution this reduces to

$$b_1 = 2c_1 - 0.1 \quad \text{and} \quad b_2 = 3c_2 + 0.1.$$

The results indicate that in terms of the MD index the competitors all become worse as the confounding strength increases. All competing ICAs systematically estimate an incorrect unmixing matrix. coroICA on the other hand only shows a very small loss in precision as confounding increases; the small loss is expected due to the decreasing signal to noise ratio. In terms of MCIS, the behavior is analogous but slightly less well resolved; with increasing confounding strength the unmixing estimation of all competing ICAs is systematically biased resulting in bad separation of sources and high MCIS scores both out-of-sample and in-sample.

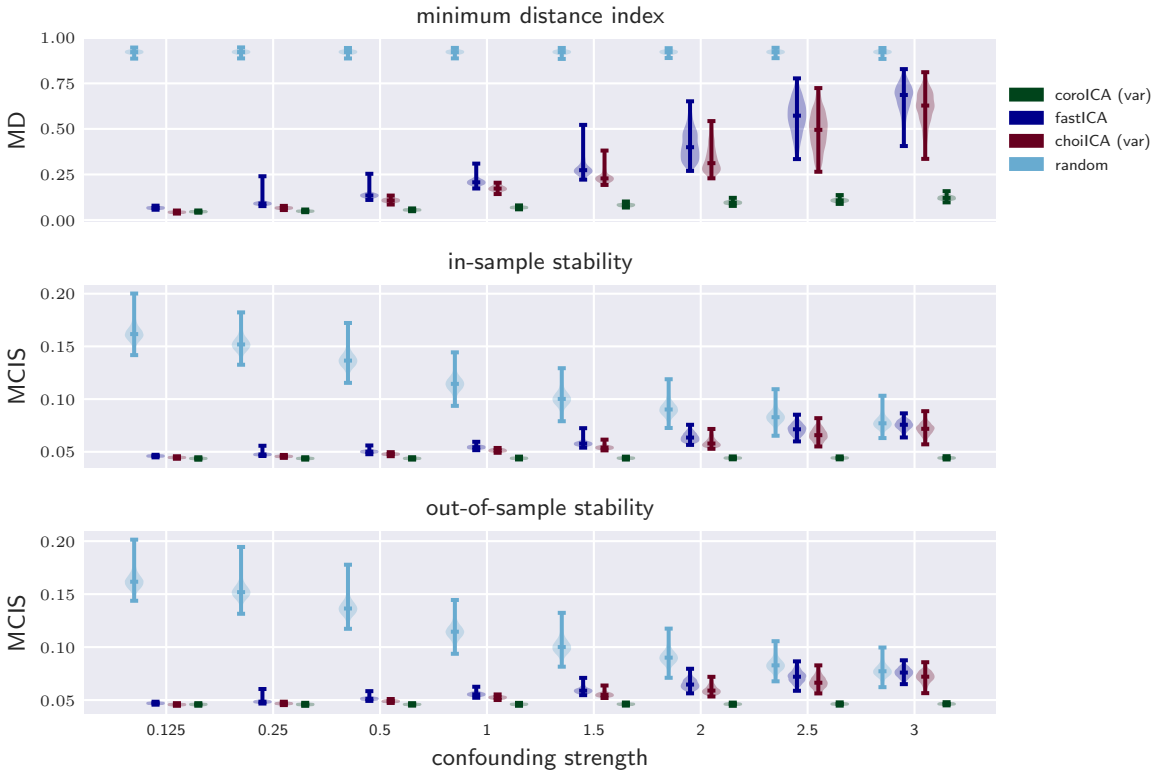


Figure 4. Results of the simulation experiment described in Section 4.2.1. Plot shows performance measures (MD: small implies close to truth; MCIS: small implies stable) for fixed signal strength and various confounding strengths. The difference between the competing ICAs and coroICA is more prominent for higher confounding strengths where the estimates of the competing ICAs are increasingly different from the true unmixing matrix and the sources become increasingly unstable.

4.2.2. Efficiency in Absence of Group Confounding

For this simulation experiment, we sample data according to Data Set 1 and choose to simulate $n = 2 \cdot 10^4$ (dimension $d = 22$) samples from $m = 10$ groups where each group contains $n/m = 2 \cdot 10^3$ observations. Within each group, we then select a random partition consisting of $|\mathcal{P}_g| = 10$ subsets while ensuring that these have the same size on average. This time, to illustrate performance in the absence of confounding, we fix the confounding strengths $c_1 = 0$ and consider the behavior of coroICA (applied to half of the groups with an equally spaced grid of 10 partitions per group) for different signal strengths $c_2 = \{0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$. The results for 1000 repetitions are shown in Figure 5. Again, choiICA (var) is applied with the same partition size.

The results indicate that overall coroICA performs competitive in the confounding-free case. In particular, there is no drastic negative hit on the performance of coroICA as compared to choiICA (var) in settings where the data follows the ordinary ICA model. The

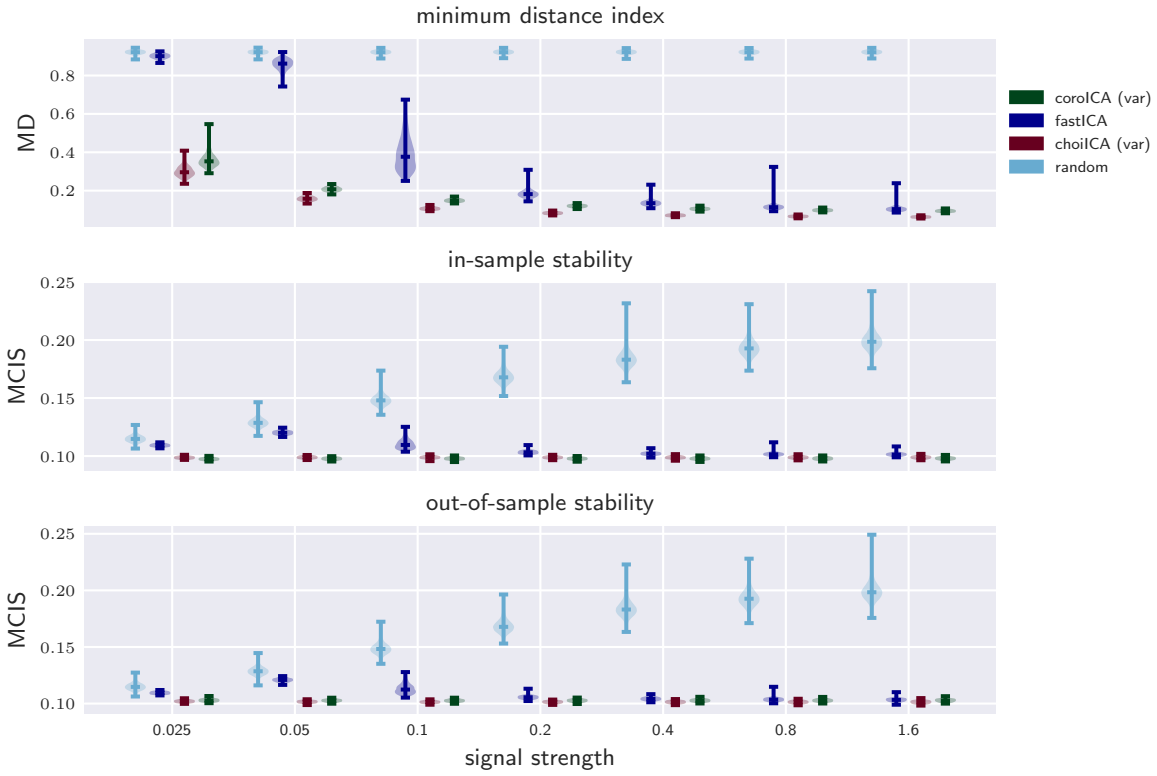


Figure 5. Results of the simulation experiment described in Section 4.2.2. Plot shows performance measures (MD: small implies close to truth; MCIS: small implies stability) for data generated without confounding and for various signal strengths. These results are reassuring, as they indicate that when applied to data that follows the ordinary ICA model, coroICA still performs competitive to competing ICAs even though it allows for a richer model class.

slight advantage compared to fastICA in this setting is due to the signal type which favors ICA methods that focus on variance signals.

4.2.3. Comparison with Other Noisy ICA Procedures

To get a better understanding of how our proposed ICA performs for different signal and noise types, we compare it on simulated data as described in Data Set 2. We illustrate the different behavior with respect to the different types of signal by applying all three of our proposed coroICA procedures (coroICA (var), coroICA (TD) and coroICA (var & TD)) and compare them to the corresponding choiICA variants which do not adjust for confounding (choiICA (var), choiICA (TD) and choiICA (var & TD)). While all coroICA procedures can deal with any type of stationary noise, choiICA (TD) only works for time-independent noise and choiICA (var) and choiICA (var & TD) cannot handle any type of noise at all (see Table 1). Additionally, we also compare with fastICA to assess its performance in the various noise settings. The results are depicted in Figure 6.

Data Set 2: GARCH simulation

For this simulation we consider different settings of the confounded mixing model

$$X_t = AS_t + H_t.$$

More precisely, we consider the following three different GARCH type signals: (i) changing variance, (ii) changing time-dependence, and (iii) both changing variance and changing time-dependence. For each of these signal types we consider two types of confounding (noise) terms: (a) time-independent and (b) time-dependent auto-regressive noise. For both we construct d independent processes $\tilde{H}^1, \ldots, \tilde{H}^d$ and then combine them with a random mixing matrix C as follows

$$H_t = C \cdot \tilde{H}_t.$$

Full details are given in Supplement B.3.

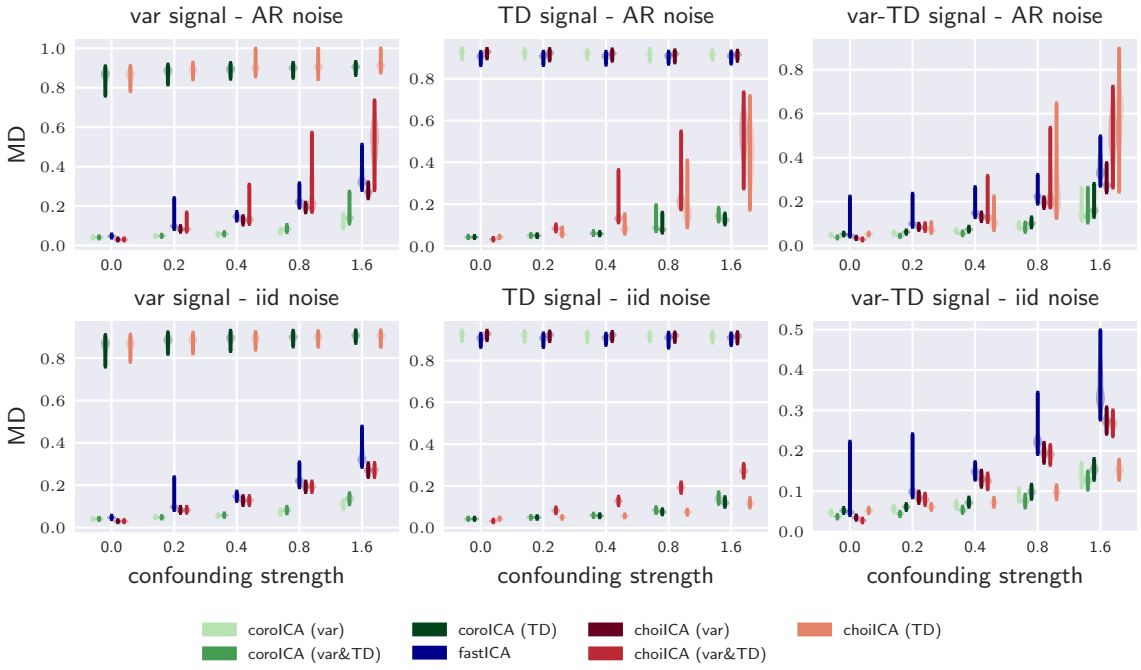


Figure 6. Results of the simulation experiment described in Section 4.2.3 and Data Set 2. Plots show performance (MD: small implies close to truth) for data generated with auto-regressive (AR) or iid noise and for var, TD, and var & TD signal as described in Data Set 2. coroICA (var & TD) is able to estimate the correct mixing in all of the considered settings, while others break whenever the more restrictive signal/noise assumptions are not met.

In all settings the most general method coroICA (var & TD) is able to estimate the correct mixing. The two signal specific methods coroICA (TD) and coroICA (var) are also able to accurately estimate the mixing in settings where a corresponding signal exists. It is also

worth noting that they slightly outperform coroICA (var & TD) in these settings. In contrast, when comparing with the choiICA variants, coroICA is in general able to outperform the corresponding method. Only in the setting of a changing time-dependence with time-independent noise, choiICA (TD) is able to slightly outperform coroICA (TD).

4.2.4. Summary of the Performance of coroICA

In summary, coroICA performs well on a larger model class consisting of both the group-wise confounded as well as the confounding-free case. An advantage over all competing ICAs is gained in confounded settings (as shown in Section 4.2.1) while there is at most a small disadvantage in the unconfounded case (cf. Section 4.2.2). This suggests that whenever the data is expected to contain at least small amounts of stationary noise or confounding, one may be better off using coroICA as the richer model class will guard against wrong results. The results in Section 4.2.3 further underline the robustness of our proposed method to various types of noise (and signals) for which other methods break. Again, even in settings that satisfy the assumptions of the more tailored methods coroICA remains competitive.

4.3. EEG Experiments

ICA is often applied in the analysis of EEG data. Here, we illustrate the potential benefit and use of coroICA for this. Specifically, we consider a multi-subject EEG experiment as depicted in Figure 7. The goal is to find a single mixing matrix that separates the sources simultaneously on all subjects. Our proposed model allows that the EEG recordings for each subject have a different but stationary noise term H. We illustrate the applicability

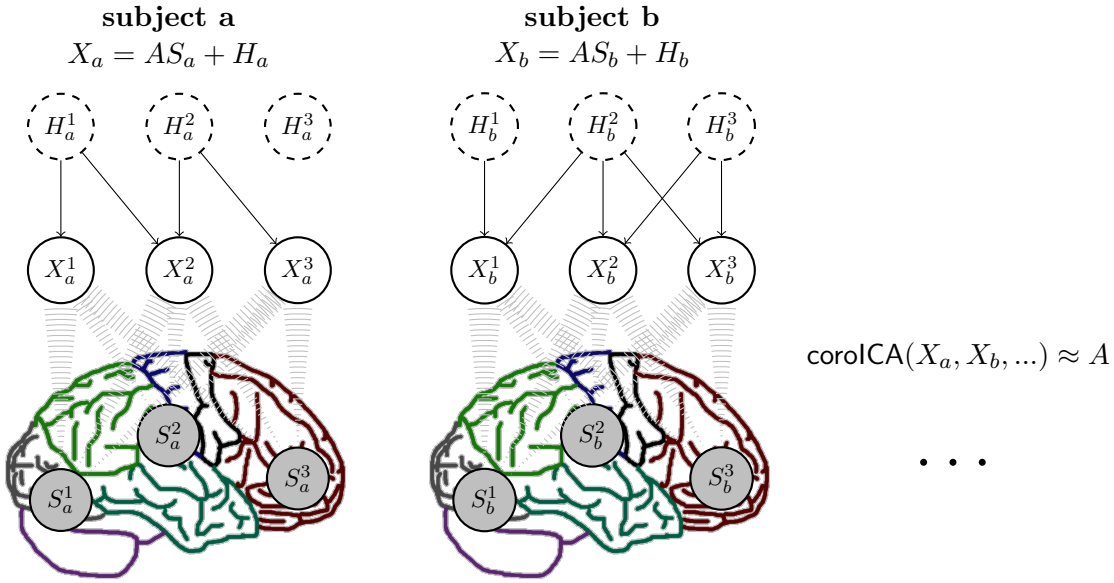


Figure 7. Illustration of a multi-subject EEG recording. For each subject, EEG signals X are recorded which are assumed to be corrupted by subject-specific (but stationary) noise terms H. The goal is to recover a single mixing matrix A that separates signals well across all subjects.

of our method to this setting based on two publicly available EEG data sets.

Data Set 3: CovertAttention data

This data set is due to Treder et al. (2011) and consists of EEG recordings of 8 subjects performing multiple trials of covertly shifting visual attention to one out of 6 cued directions. The data set contains recordings of

- 8 subjects,
- for each subject there exist 6 runs with 100 trials,
- each recording consists of 60 EEG channels recorded at 1000 Hz sampling frequency, while we work with the publicly available data that is downsampled to 200 Hz.

Since visual inspection of the data revealed data segments with huge artifacts and details about how the publicly available data was preprocessed was unavailable to us, we removed outliers and high-pass filtered the data at 0.5 Hz. In particular, along each dimension we set those values to the median along its dimension that deviate more than 10 times the median absolute distance from this median. We further preprocess the data by re-referencing to common average reference (car) and projecting onto the orthogonal complement of the null component. For our unmixing estimations, we use the entire data, i.e., including intertrial breaks.

For classification experiments (cf. Section 4.3.2) we use, in line with Treder et al. (2011), the 8–12 Hz bandpass-filtered data during the 500–2000 ms window of each trial, and use the log-variance as bandpower feature (Lotte et al., 2018). The classification analysis is restricted to valid trials (approximately 311 per subject) with the desired target latency as described in Treder et al. (2011).

Results on the CovertAttention Data Set 3 are presented here, while the results of the analogous experiments on the BCICompIV2a Data Set 4 are deferred to Supplement C. For both data sets, we compare the recovered sources of coroICA with those recovered by competing ICA methods. Since ground truth is unknown we report comparisons based on the following three criteria:

stability and independence
: We use MCIS (cf. Section 2.4) to assess the stability and independence of the recovered sources both in- and out-of-sample.

classification accuracy
: For both data sets there is label information available that associates certain time windows of the EEG recordings with the task the subjects were performing at that time. Based on the recovered sources, we build a classification pipeline relying on feature extraction and classification techniques that are common in the field (Lotte et al., 2018). The achieved classification accuracy serves as a proxy of how informative and suitable the extracted signals are.

topographies

For a qualitative assessment, we inspect the topographic maps of the extracted sources, as well as the corresponding power spectra and a raw time-series chunk. This is used to illustrate that the sources recovered by coroICA do not appear random or implausible for EEG recordings and are qualitatively similar to what is expected from other ICAs. Furthermore, we provide an overview over all components achieved on Data Set 3 by SOBI, fastICA, and coroICA in the Supplementary Section D, where components are well resolved when the corresponding topographic map and activation map are close to each other (cf. Section 2.4).

4.3.1. Stability and Independence

We aim to probe stability not only in-sample but also verify the expected increase in stability when applying the unmixing matrix to data of new unseen subjects, i.e., to new groups of samples with different confounding specific to that subject. In order to assess stability and independence of the recovered sources in terms of the MCIS both in- and out-of-sample and for different amounts of training samples, we proceed by repeatedly splitting the data into a training and a test data set. More precisely, we construct all possible splits into training and test subjects for any given number of training subjects. For each pair of training and test set, we fit an unmixing matrix using coroICA and all competing methods described in Section 4.1. We then compute the MCIS on the training and test data for each method separately and collect the results of each training-test split for each number of training subjects.

Results obtained on the CovertAttention data set (with equally spaced partitions of ≈15 seconds length) are given in Figure 8 and the results for the BCICompIV2a data set (with equally spaced partitions of ≈15 seconds length) are shown in Supplement C.1, Figure 14. For both data sets the results are qualitatively similar and support the claim that the unmixing obtained by coroICA is more stable when transferred to new unseen subjects. While for the competing ICAs the instability on held-out subjects does not follow a clear decreasing trend with increasing number of training subjects, coroICA can successfully make use of additional training subjects to learn a more stable unmixing matrix.

Due to the characteristics and low signal-to-noise ratio in EEG recordings, the evaluation based on the absolute MCIS score is less well resolved than what we have seen in the simulations before. For this reason we additionally provide a more focused evaluation by considering the MCIS fraction: the fraction of the MCIS achieved on a subject by the respective competitor method divided by the MCIS achieved on that subject by coroICA when trained on the same subjects. Thus, this score compares MCIS on a per subject basis, where values greater than 1 indicate that the respective competing ICA method performed worse than coroICA. Figure 9 shows the results on the CovertAttention Data Set 3 confirming that coroICA can successfully incorporate more training subjects to derive a better unmixing of signals.

4.3.2. Classification based on Recovered Sources

While the results in the previous section indicate that coroICA can lead to more stable separations of sources in EEG than the competing methods, in scenarios with an unknown

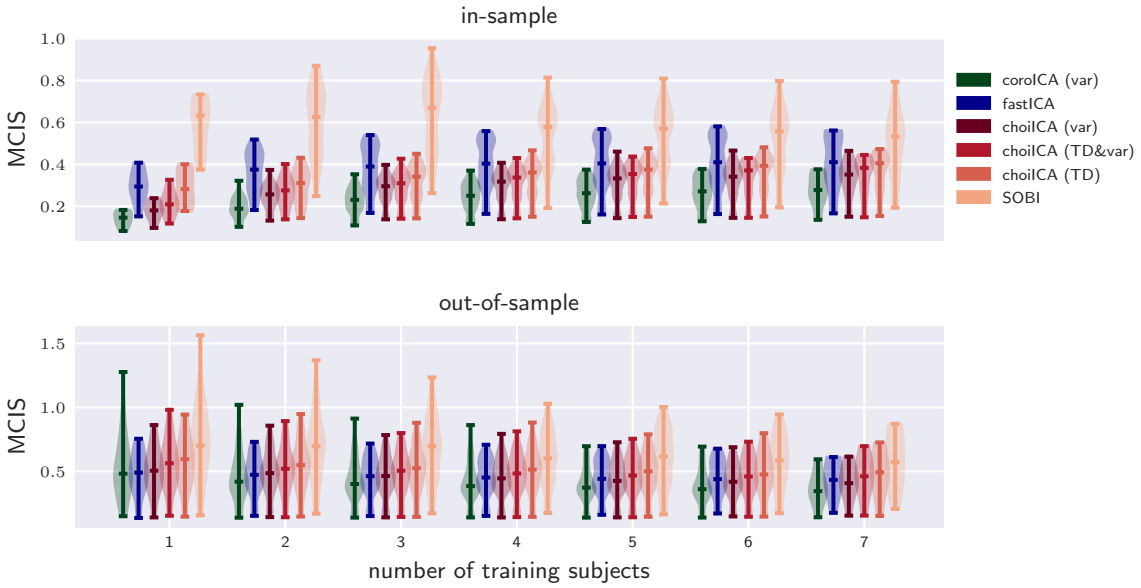


Figure 8. Experimental results for comparing the stability of sources (MCIS: small implies stable) trained on different numbers of training subjects (cf. Section 4.3.1), here on the CovertAttention Data Set 3, demonstrating that coroICA, in contrast to the competing ICA methods, can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.

ground truth the stability of the recovered sources cannot serve as the sole determining criterion for assessing the quality of recovered sources. In addition to asking whether the recovered sources are stable and independent variance signals, we hence also need to investigate whether the sources extracted by coroICA are in fact reasonable or meaningful. In the "America's Got Talent Duet Problem" (cf. Example 2) this means that each of the recovered sources should only contain the voice of one (independent) singer (plus some confounding noise that is not the other singer). For EEG data, this assessment is not as easy. Here, we approach this problem from two angles: (a) in this section we show that the recovered sources are informative and suitable for common EEG classification pipelines, (b) in Section 4.3.3 we qualitatively assess the extracted sources based on their power spectra and topographic maps.

In both data sets there are labeled trials, i.e., segments of data during which the subject covertly shifts attention to one of six cues (cf. Data Set 3) or performs one of four motor imagery tasks (cf. Data Set 4). Based on these, one can try to predict the trial label given the trial EEG data. To mimic a situation where the sources are transferred from other subjects, we assess the informativeness of the extracted sources in a leave-k-subjects-out fashion as follows. We estimate an unmixing matrix on data from all but k subjects, compute bandpower features for each extracted signal and for each trial (as described in Data Set 3 and Data Set 4), and on top of those we train an ensemble of 200 bootstrapped shrinkage linear discriminant analysis classifiers where each is boosted by a random forest classifier on the wrongly classified trials. This pipeline (signal unmixing, bandpower-feature

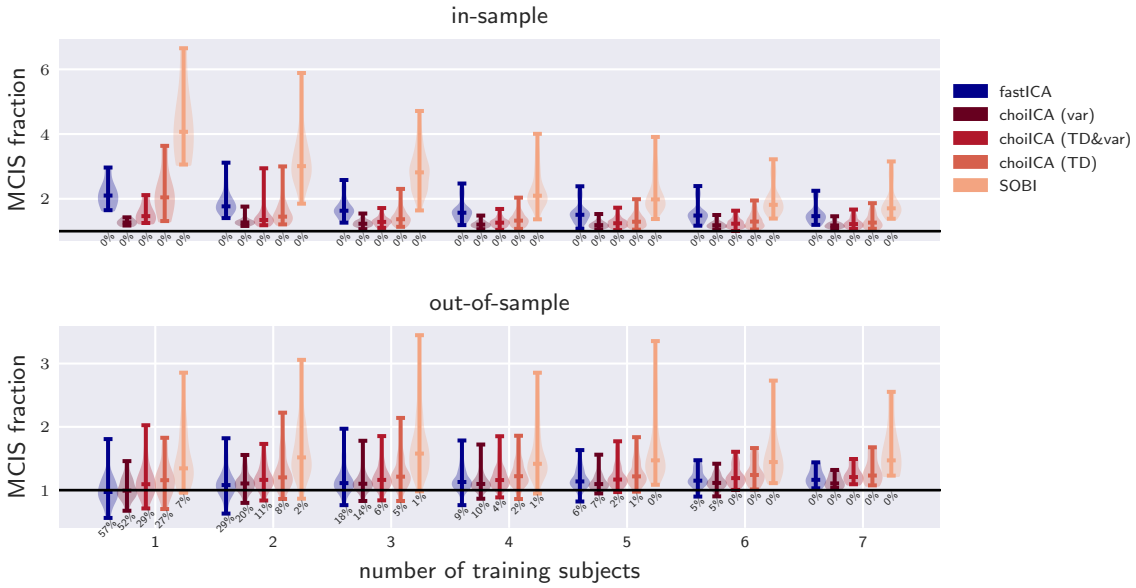


Figure 9. Experimental results for comparing the stability of sources of the competing methods relative to the stability obtained by coroICA (MCIS fraction: above 1 implies less stable than coroICA) trained on different numbers of training subjects (cf. Section 4.3.1), here on the CovertAttention Data Set 3, demonstrating that coroICA can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.

computation, trained ensemble classifier), is then used to predict the trials on the k held-out subjects.

The results are reported in Figure 10 and Supplement C.2, Figure 16 which show for each number of training subjects, the accuracies achieved on the respective held-out subjects when using the unmixing obtained on the remaining subjects by either coroICA or one of the competitor methods. The results on both data sets support the claim that the sources recovered by coroICA are not only stable but in addition also capture meaningful aspects of the data that enable competitive classification accuracies in fully-out-of-sample classification. The mean improvement in classification accuracy of coroICA over the other methods increases with increasing number of training subjects. This behavior is expected since it is difficult to disambiguate signal from subject-specific confounding for few training subjects, while coroICA is expected to learn an unmixing which better adjusts for the confounding with more training subjects.

It is worth noting that these classification results depend heavily on the employed classification pipeline subsequent to the source separation. Here, our goal is only to show that coroICA does indeed separate the data into informative sources. In practice, and when only classification accuracy matters, one might also consider using a label-informed source separation (Dähne et al., 2014), employ common spatial patterns (Koles et al., 1990) or use decoding techniques based on Riemannian geometry (Barachant et al., 2012).

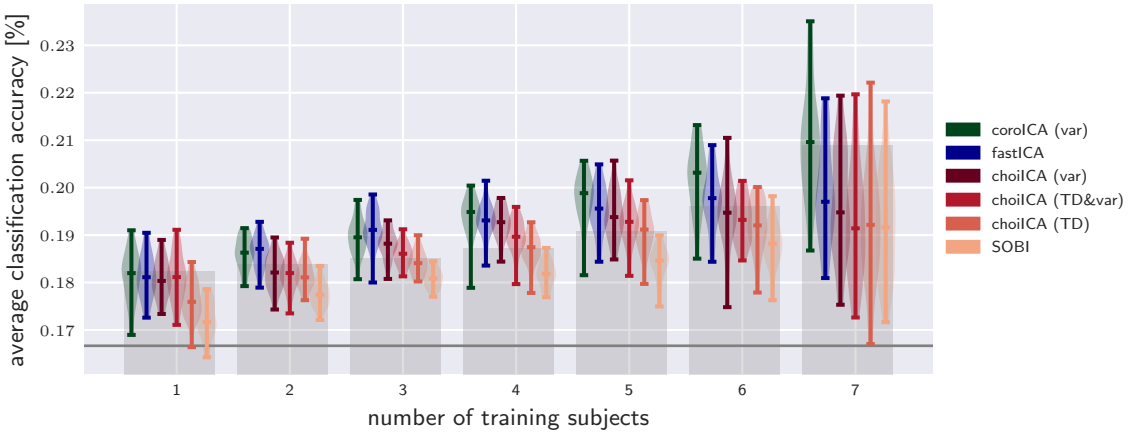


Figure 10. Classification accuracies on held-out subjects (cf. Section 4.3.2), here on the CovertAttention Data Set 3. Gray regions indicate a 95% confidence interval of random guessing accuracies.

4.3.3. Topographic Maps

The components that coroICA extracts from EEG signals are stable (cf. Section 4.3.1) and meaningful in the sense that they contain information that enables classification of trial labels, which is a common task in EEG studies (cf. Section 4.3.2). In this section, we complement the assessment of the recovered sources by demonstrating that the results obtained by coroICA lead to topographies, activation maps, power spectra and raw time-series that are similar to what is commonly obtained during routine ICA analyses of EEG data when the plausibility and nature of ICA components is to be judged.

Topographies are common in the EEG literature to depict the relative projection strength of extracted sources to the scalp sensors. More precisely, the column-vector a_j of $A = V^{-1}$ that specifies the mixing of the j-th source component is visualized as follows. A sketched top view of the head is overlayed with a heatmap where the value at each electrodes' position is given by the corresponding entry in a_j. These topographies are indicative of the nature of the extracted sources, for example the dipolarity of source topographies is a criterion invoked to identify cortical sources (Delorme et al., 2012) or the topographies reveal that the source mainly picks up changes in the electromagnetic field induced by eye movements. Another way to visualize an extracted source is an activation map, which is commonly obtained by depicting the vector $\widehat{\mathrm{Cov}}(X)v_j^\top$ (where v_j is j-th row of unmixing matrix V) and shows for each electrode how the signal observed at that electrode covaries with the signal extracted by v_j (Haufe et al., 2014). Besides inspecting the raw time-series data, another criterion invoked to separate cortical from muscular components is the log power spectrum. For example, a monotonic increase in spectral power starting at around 20 Hz is understood to indicate muscular activity (Goncharova et al., 2003) and peaks in typical EEG frequency ranges are used to identify brain-related components.[5].

5. These are commonly employed criteria which are also advised in the eeglab tutorial (Delorme and Makeig, 2004, https://sccn.ucsd.edu/wiki/Chapter_09:_Decomposing_Data_Using_ICA) and the neurophys-

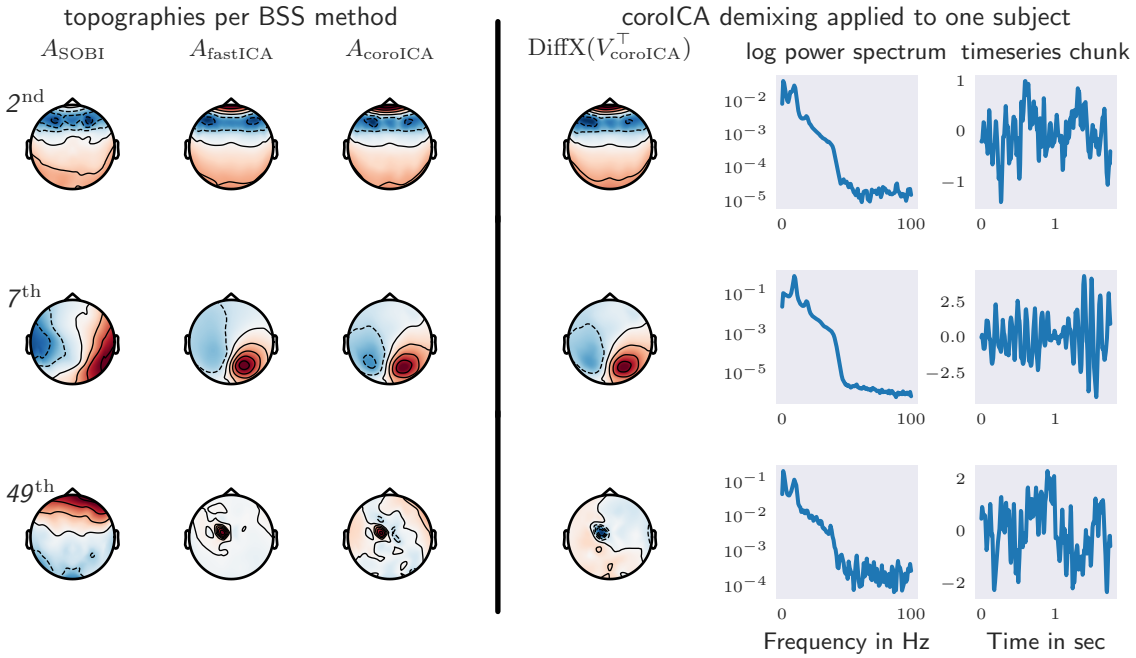


Figure 11. Visualization of exemplary EEG components recovered on the CovertAttention Data Set 3. On the left the topographies of three components are shown where the mixing matrix is the inverse of the unmixing matrix obtained by SOBI (A_{SOBI}), the unmixing matrix obtained by fastICA (A_{fastICA}) and that of coroICA (var) (A_{coroICA}). On the right we depict, for a randomly chosen subject, the activation maps (cf. Section 4.3.3 and 2.4), the log power spectra, and randomly chosen chunks of the raw time-series data corresponding to the respective coroICA (var) components. Components extracted by coroICA (var) are qualitatively similar to those of the commonly employed ICA procedures; see Section 4.3.3 for details.

In Figure 11, we depict the aforementioned criteria for three exemplary components extracted by coroICA on the CovertAttention Data Set 3. Following the discussion in Section 2.4 we show the activation maps as

$$\text{DiffX}(v_j^{\top}) = \sum_{\mathcal{M}\in\mathcal{M}^*} \text{sign}(v_j M v_j^{\top}) M v_j^{\top},$$

which captures variance changing signal and allows to asses the quality of a recovered source by comparison to the topographic map a_j (cf. Equation 2.4). Here, the idea is to demonstrate that coroICA components are qualitatively similar to components extracted by commonly employed SOBI-ICA or fastICA. Therefore, we choose to display one example of an ocular component (2^{nd} where the topography is indicative of eye movement), a cortical component (7^{th} where the dipolar topography, the typical frequency peak at around 8–12 Hz, and the amplitude modulation visible in the raw time-series are indicative of the

iological biomarker toolbox wiki (Hardstone et al., 2012, https://www.nbtwiki.net/doku.php?id=tutorial:how_to_use_ica_to_remove_artifacts).

cortical nature), and an artifactual component (51^{st} where the irregular topography and the high frequency components indicate an artifact). For comparison, we additionally show for each component the topographies of the components extracted by SOBI-ICA or fastICA by matching the recovered source which most strongly correlates with the one extracted by coroICA. The components extracted by coroICA closely resemble the results one would obtain from a commonly employed ICA analysis on EEG data.

For completeness, we provide an overview over all components extracted on Data Set 3 by SOBI, fastICA, and coroICA (var) in the Supplementary Section D. Components are well resolved when the corresponding topographic map and activation map are close to each other (cf. Section 2.4), which, by visual inspection, appears to be more often the case for coroICA than for the competing methods.

5. Conclusion

In this paper, we propose a method for recovering independent sources corrupted by group-wise stationary confounding. It extends ordinary ICA to an easily interpretable model, which we believe is relevant for many practical problems as is demonstrated in Section 3.1 for climate data and Section 4.3 for EEG data. We give explicit assumptions under which the sources are identifiable in the population case (cf. Section 2.2). Moreover, we introduce a straightforward algorithm for estimating the sources based on the well-understood concept of approximate joint matrix diagonalization. As illustrated in the simulations in Section 4.2, this estimation procedure performs competitive even for data from an ordinary ICA model, while additionally being robust and able to adjust for group-wise stationary confounding. For real data, we show that the coroICA model indeed performs reasonably on EEG data and leads to improvements in comparison to commonly employed approaches, while at the same time preserving an enhanced interpretation of the recovered sources.

Acknowledgments

The authors thank Vinay Jayaram, Nicolai Meinshausen, Jonas Peters, Gian Thanei, the action editor Kenji Fukumizu, and anonymous reviewers for helpful discussions and constructive comments. NP and PB were partially supported by the European Research Commission grant 786461 CausalStats - ERC-2017-ADG.

Supplementary material

The supplementary material consists of the following appendices.

A Identifiability Proof

B Complementary Material

C EEG Experiments on Data Set 4

D All Topographies on Data Set 3

Appendix A. Identifiability Proof

The proof is based on Theorem 1 from Kleinsteuber and Shen (2013). For completeness, we introduce some of the notation therein and state their result with adapted notation to ease following our proof of Theorem 1. We begin by defining the empirical correlation between two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ as

$$\widehat{\mathrm{Corr}}(\mathbf{v}, \mathbf{w}) := \begin{cases} \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}, & \text{if } \mathbf{v} \neq 0 \text{ and } \mathbf{w} \neq 0, \\ 1, & \text{otherwise.} \end{cases}$$

Moreover, for a collection of $(d \times d)$-real diagonal matrices $\{Z_1, \ldots, Z_m\}$, we define the following collinearity measure

$$\rho(Z_1, \ldots, Z_m) := \max_{1 \leq k < l \leq d} |\widehat{\mathrm{Corr}}(\mathbf{z}_k, \mathbf{z}_l)|, \tag{17}$$

where $\mathbf{z}_j := (z_1(j) \ldots, z_m(j))$ and $z_i(j)$ is the j-th diagonal element of the matrix Z_i. Using this notation we can state the uniqueness result due to Kleinsteuber and Shen (2013, Theorem 1) as follows.

Theorem 4 (Kleinsteuber and Shen (2013, Theorem 1)) *Let $D_i \in \mathbb{R}^{d \times d}$, for $i \in \{1, \ldots, m\}$ be diagonal, and let $M \in \mathbb{R}^{d \times d}$ be an invertible matrix so that $M^\top D_i M$ is diagonal as well. Then M is essentially, up to scaling and permutation of its columns, unique if and only if $\rho(D_1, \ldots, D_m) < 1$.*

Using this result we prove Theorem 1.

Proof The theorem is proven by the correct invocation of Theorem 4. We first define the unmixing matrix $V := A^{-1}$ and introduce the sets of matrices

$$\mathcal{D}_{\mathrm{var}} := \{V(\mathrm{Cov}(X_k) - \mathrm{Cov}(X_l))V^\top \,|\, g \in \mathcal{G} \text{ and } k, l \in g\}.$$

and

$$\mathcal{D}_{\mathrm{TD}} := \{V(\mathrm{Cov}(X_k, X_{k-\tau}) - \mathrm{Cov}(X_l, X_{l-\tau}))V^\top \,|\, g \in \mathcal{G} \text{ and } k, l \in g\}.$$

Due to the assumed ICA model and Assumption 1, all matrices in the sets $\mathcal{D}_{\text{var}}$ and $\mathcal{D}_{\text{TD}}$ are diagonal (cf. (6) and (7)). Moreover, for $g \in \mathcal{G}$ and $k, l \in g$ it holds that

$$\begin{aligned} &V(\operatorname{Cov}(X_k) - \operatorname{Cov}(X_l))V^\top \\ &\quad = \operatorname{Cov}(S_k) - \operatorname{Cov}(S_l) \\ &\quad = \operatorname{diag}(\operatorname{Var}(S_k^1) - \operatorname{Var}(S_l^1), \ldots, \operatorname{Var}(S_k^d) - \operatorname{Var}(S_l^d)) \end{aligned}$$

and

$$\begin{aligned} &V(\operatorname{Cov}(X_k, X_{k-\tau}) - \operatorname{Cov}(X_l, X_{l-\tau}))V^\top \\ &\quad = \operatorname{Cov}(S_k, S_{k-\tau}) - \operatorname{Cov}(S_l, S_{l-\tau}) \\ &\quad = \operatorname{diag}(\operatorname{Cov}(S_k^1, S_{k-\tau}^1) - \operatorname{Cov}(S_l^1, S_{l-\tau}^1), \ldots, \operatorname{Cov}(S_k^d, S_{k-\tau}^d) - \operatorname{Cov}(S_l^d, S_{l-\tau}^d)). \end{aligned}$$

Next, we define for all $j \in \{1, \ldots, d\}$ the vectors

$$\begin{aligned} \mathbf{z}_j &= \left(\left(\operatorname{Var}(S_k^j) - \operatorname{Var}(S_l^j) \right)_{k,l \in g} \right)_{g \in \mathcal{G}} \\ \text{or} \quad \mathbf{z}_j &= \left(\left(\operatorname{Cov}(S_k^j, S_{k-\tau}^j) - \operatorname{Cov}(S_l^j, S_{l-\tau}^j) \right)_{k,l \in g} \right)_{g \in \mathcal{G}}, \end{aligned}$$

depending on whether a variance signal or time-dependence signal is being considered, respectively. Then, Assumption 2 or Assumption 3 implies for all distinct pairs $p, q \in \{1, \ldots, d\}$ that

$$|\widehat{\operatorname{Corr}}(\mathbf{z}_p, \mathbf{z}_q)| = \frac{|\mathbf{z}_p \cdot \mathbf{z}_q|}{\|\mathbf{z}_p\| \|\mathbf{z}_q\|} < 1.$$

Hence, for either $\mathcal{D} = \mathcal{D}_{\text{var}}$ or $\mathcal{D} = \mathcal{D}_{\text{TD}}$ it holds that $\rho(\mathcal{D}) < 1$, where ρ is defined in (17). Since the identity matrix satisfies that $\operatorname{Id} D \operatorname{Id}^\top$ is diagonal for all $D \in \mathcal{D}$, we can invoke Theorem 4 to conclude that any matrix $M \in \mathbb{R}^{d \times d}$ for which $MDM^\top$ is diagonal for all $D \in \mathcal{D}$, is equal to the identity matrix up to scaling and permutation of its columns. Next, we consider the two signal types separately.

- **variance signal:** If there is a variance signal that satisfies Assumption 2, assume there exists an invertible matrix $\widetilde{A}$ such that for all $g \in \mathcal{G}$ and all $k, l \in g$ it holds that

  $$\widetilde{A}^{-1}(\operatorname{Cov}(X_k) - \operatorname{Cov}(X_l))(\widetilde{A}^{-1})^\top = \operatorname{Cov}(S_k) - \operatorname{Cov}(S_l).$$

  Then, it also holds that

  $$(V\widetilde{A}) \underbrace{(\operatorname{Cov}(S_k) - \operatorname{Cov}(S_l))}_{\in \mathcal{D}_{\text{var}}} (V\widetilde{A})^\top = V(\operatorname{Cov}(X_k) - \operatorname{Cov}(X_l))V^\top,$$

  which is diagonal.

- **time-dependence signal:** If there is a time-dependence signal that satisfies Assumption 3, assume there exists an invertible matrix $\widetilde{A}$ such that for all $g \in \mathcal{G}$ and all $k, l \in g$ it holds that

  $$\widetilde{A}^{-1}(\operatorname{Cov}(X_k, X_{k-\tau})) - \operatorname{Cov}(X_l, X_{l-\tau}))(\widetilde{A}^{-1})^\top = \operatorname{Cov}(S_k, S_{k-\tau}) - \operatorname{Cov}(S_l, S_{l-\tau}).$$

Then, it also holds that

$$(V\widetilde{A})\underbrace{(\mathrm{Cov}(S_k, S_{k-\tau}) - \mathrm{Cov}(S_l, S_{l-\tau}))}_{\mathcal{D}_{\mathrm{TD}}}(V\widetilde{A})^\top = V(\mathrm{Cov}(X_k, X_{k-\tau}))-\mathrm{Cov}(X_l, X_{l-\tau}))V^\top,$$

which is diagonal.

Using the above reasoning, either of the two cases—depending on whether Assumption 2 or 3 holds—shows that $V\widetilde{A}$ is equal to the identity matrix up to permutation and rescaling of its columns. Moreover, this implies that $\widetilde{A}$ is equal to A up to scaling and permutation of its columns. This completes the proof of Theorem 1. ■

Appendix B. Complementary Material

B.1. America's Got Talent

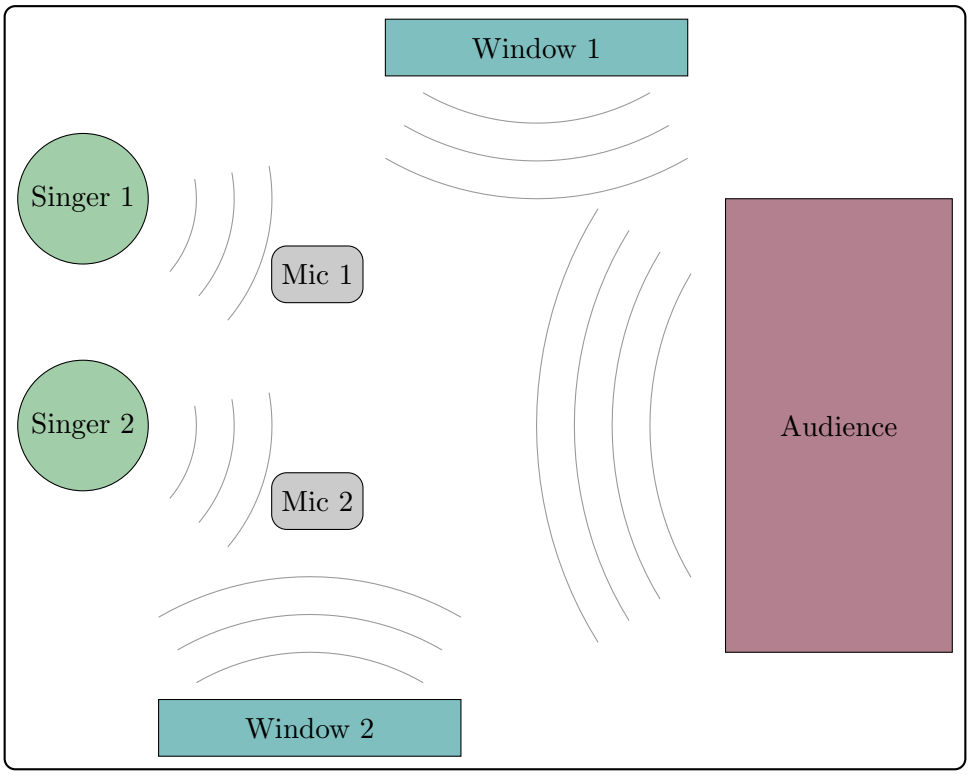


Figure 12. Schematic of the "America's Got Talent Duet Problem" described in Example 2. The sound from the windows and audience is taken to be confounding noise which has fixed covariance structure over given time blocks. The challenge is to recover the sound signals from the individual singers given the recordings of the two microphones.

B.2. Causality Example

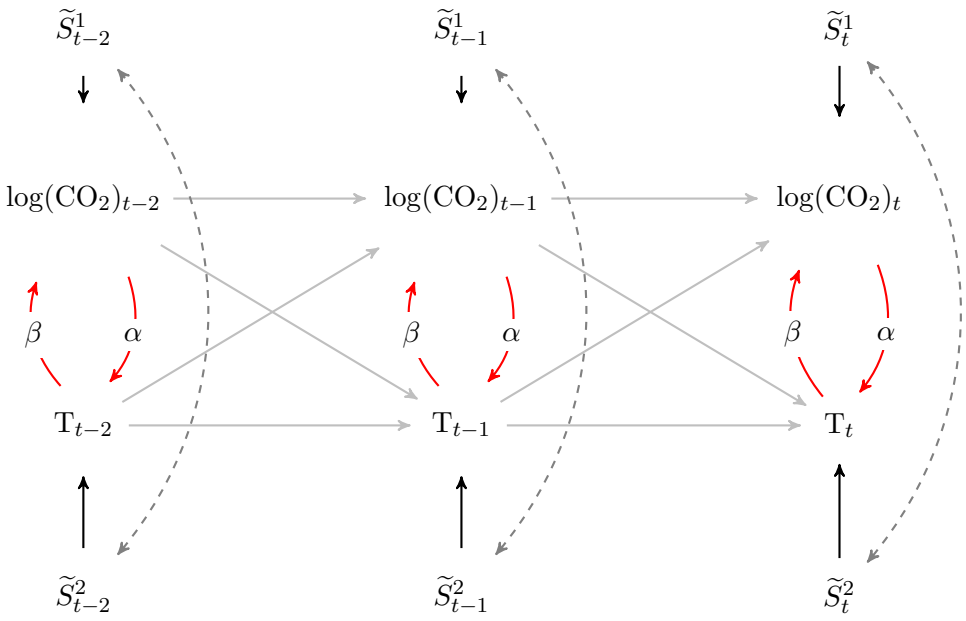


Figure 13. Graphical representation of the causal feedback model between carbon dioxide (CO_2) and temperature (T). The dashed line corresponds to stationary confounding.

B.3. Simulations

The GARCH model that we simulate from in Section 4.2.3 is specified as follows. We simulate sources $S^1, \ldots, S^d$ from the following GARCH-type model

$$\begin{aligned} \sigma_i^2 &= a_1 + a_2 \cdot (S_{i-1}^j)^2 + a_3 \cdot \sigma_{i-1}^2 \\ S_i^j &= b_1 S_{i-1}^j + \cdots + b_p S_{i-p}^j + \sigma_i \varepsilon_i, \end{aligned}$$

where the ε_i are independent and standard normal. Moreover, the noise terms $H^1, \ldots, H^d$ are assumed to be either given by the following AR-process

$$H_i^j = c_1 H_{i-1}^j + \cdots + c_q H_{i-q}^j + \nu_i,$$

where ν_i are independent standard normal, q is uniformly distributed on $\{1, \ldots, 10\}$ and c_i independent $\mathcal{N}\left(0, 1/(i+1)^2\right)$ or simply as iid $\mathcal{N}(0,1)$ random variables. The final data is then constructed according to the following equation

$$X_i = A \cdot S_i + \widetilde{H}_i,$$

where $\widetilde{H}_i = ACH_i$ and $A, C \in \mathbb{R}^{d \times d}$ are sampled with iid entries from $\mathcal{N}(0,1)$ and $\mathcal{N}(0, \frac{1}{d})$, respectively. To illustrate, the effect of the signal type we consider the following three settings.

- **Setting 1 (time-independent with changing variance)**
  Set $a = (0.005, 0.026, 0.97)$ such that the variance changes over time and $p = 0$ to ensure time-independent signals. Based on these settings we sample $n = 200000$ observations.

- **Setting 2 (varying time-dependence structure with constant variance)**
  Set $a = (1, 0, 0)$ such that the variance is fixed to 1. Then, sample p 100 times uniformly from $\{1, \ldots, 10\}$ and b_i independent from $\mathcal{N}\left(0, 1/(i+1)^2\right)$ and simulate 2000 observations for each of the 100 parameter settings, leading to a total of $n = 200000$ observations.

- **Setting 3 (varying time-dependence structure with changing variance)**
  Set $a = (0.005, 0.026, 0.97)$ such that the variance changes over time. Then, we sample p 100 times uniformly from $\{1, \ldots, 10\}$ and b_i independent from $\mathcal{N}\left(0, 1/(i+1)^2\right)$ and simulate 2000 observations for each of the 100 parameter settings, leading to a total of $n = 200000$ observations.

Appendix C. EEG Experiments on Data Set 4

Analogous to Sections 4.3.1 and 4.3.2 we conducted experiments on the BCICompIV2a Data Set 4, the results of which are presented in the subsequent sections.

Data Set 4: BCICompIV2a data

This data set is due to Tangermann et al. (2012, Section 5) and consists of EEG recordings of 9 subjects performing multiple trials of 4 different motor imagery tasks. The data set contains recordings of

- 9 subjects, each recorded on 2 different days,
- for each subject and day there exist 6 runs with 48 trials,
- each recording consists of 22 EEG channels recorded at 250 Hz sampling frequency,
- and is bandpass filtered between 0.5 and 100 Hz and is 50 Hz notch filtered.

For our analysis we only use the trial-data, i.e., the concatenated segments of seconds 3–6 of each trial (corresponding to the motor imagery part of the trials (Tangermann et al., 2012)). We further preprocess the data by re-referencing to common average reference (car) and projecting onto the orthogonal complement of the null component.

As features for classification experiments (cf. Section 4.3.2) on this data set we use bandpower in the 8–30 Hz band as measured by the log-variance of the 8–30 Hz bandpass-filtered trial data (Lotte et al., 2018).

C.1. Stability and Independence

Figure 14. Experimental results for comparing the stability of sources (MCIS: small implies stable) trained on different numbers of training subjects (cf. Section 4.3.1), here on the BCICompIV2a Data Set 4, demonstrating that corοICA, in contrast to the competing ICA methods, can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.

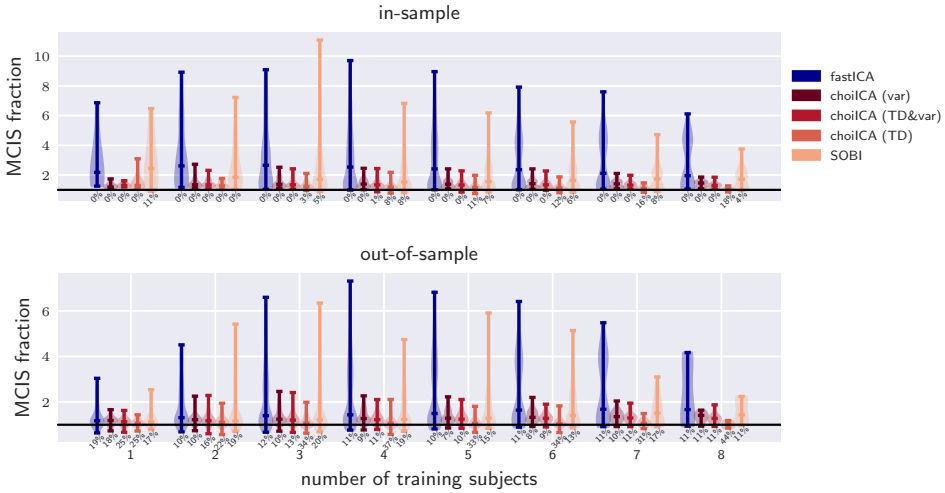


Figure 15. Experimental results for comparing the stability of sources of competitors relative to the stability obtained by corοICA (MCIS fraction: > 1 implies less stable than corοICA) trained on different numbers of training subjects (cf. Section 4.3.1), here on the BCICompIV2a Data Set 4, demonstrating that corοICA can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.

C.2. Classification based on Recovered Sources

Figure 16. Classification accuracies on held-out subjects (cf. Section 4.3.2), here on the BCICompIV2a Data Set 4. Gray regions indicate a 95% confidence interval of random guessing accuracies.

Appendix D. All Topographies and Activation Maps on Data Set 3

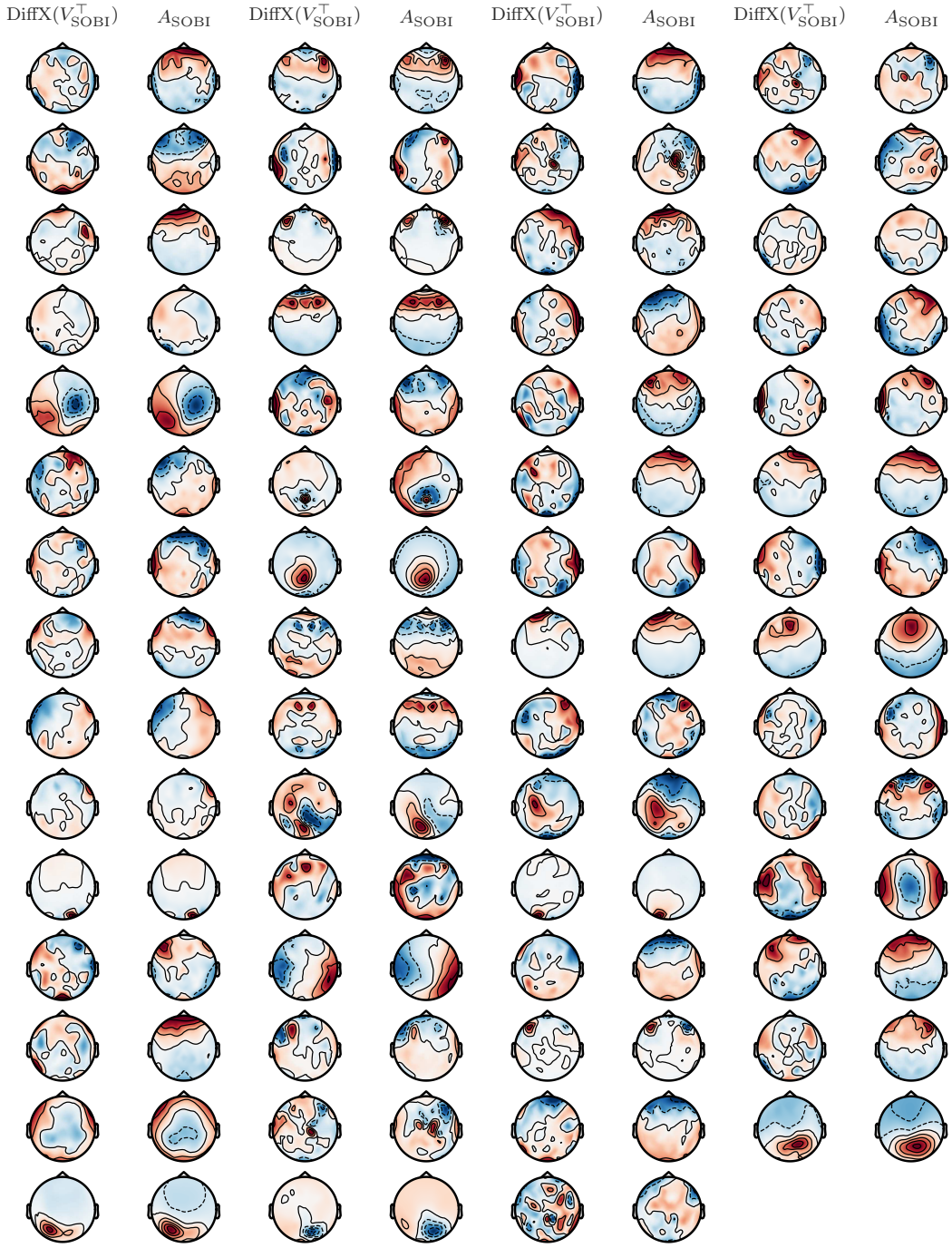


Figure 17. Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by SOBI on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 2.4 and 4.3.3).

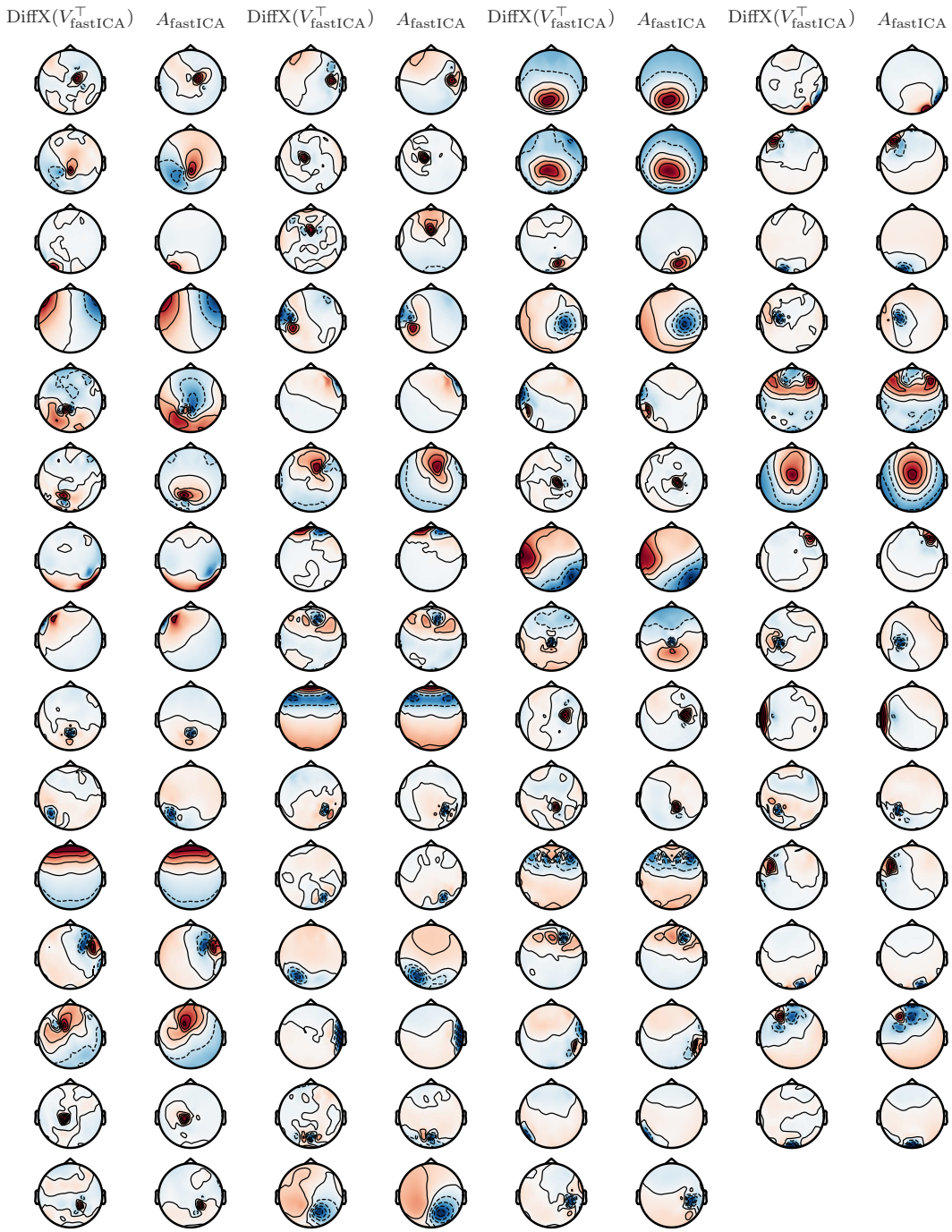


Figure 18. Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by fastICA on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 2.4 and 4.3.3).

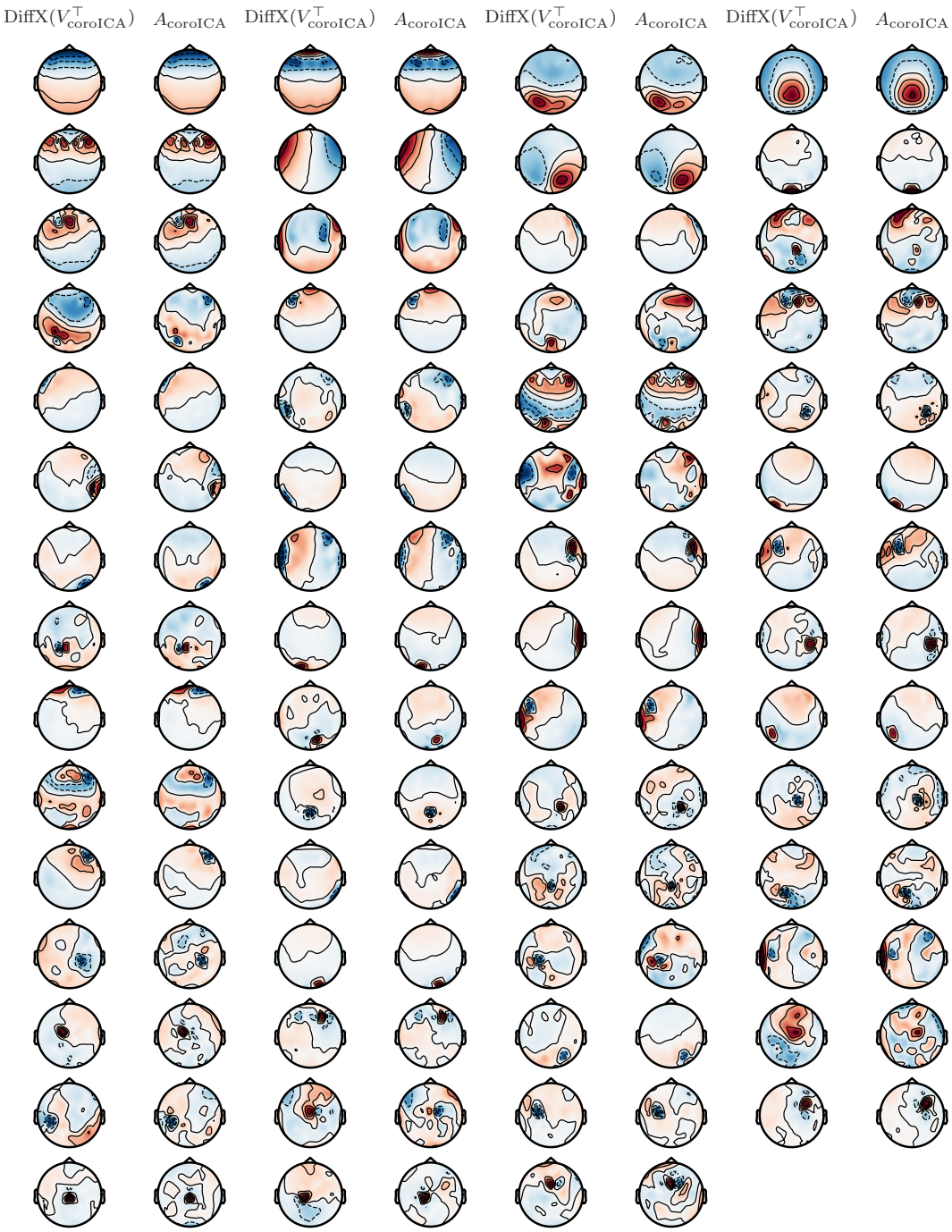


Figure 19. Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by coroICA (var) on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 2.4 and 4.3.3).

References

P. Ablin, J.-F. Cardoso, and A. Gramfort. Beyond Pham's algorithm for joint diagonalization. *arXiv preprint arXiv:1811.11433v1*, 2018.

S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, pages 757–763. NIPS Foundation, 1995.

A. Back and A. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(4):473–484, 1997.

A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

C. Beckmann and S. M. Smith. Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage*, 25(1):294–311, 2005.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2): 434–444, 1997.

B. Bereiter, S. Eggleston, J. Schmitt, C. Nehrbass-Ahles, T. F. Stocker, H. Fischer, S. Kipfstuhl, and J. Chappellaz. Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophysical Research Letters*, 42(2):542–549, 2015.

V. D. Calhoun, T. Adalı, L. K. Hansen, J. Larsen, and J. J. Pekar. ICA of functional MRI data: An overview. In *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 281–288, 2003.

J.-F. Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2109–2112 vol.4, May 1989a.

J.-F. Cardoso. Blind identification of independent components with higher-order statistics. In *Workshop on Higher-Order Spectral Analysis*, pages 157–162, 06 1989b.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *IEE Proceedings F - Radar and Signal Processing*, 140(6):362–370, 1993.

J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

S. Choi and A. Cichocki. Blind separation of nonstationary and temporally correlated sources from noisy mixtures. In *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, pages 405–414. IEEE, 2000a.

S. Choi and A. Cichocki. Blind separation of nonstationary sources in noisy mixtures. *Electronics Letters*, 36(9):848–849, 2000b.

S. Choi and A. Cichocki. Algebraic differential decorrelation for nonstationary source separation. *Electronics Letters*, 37(23):1414–1415, 2001.

S. Choi, A. Cichocki, and A. Belouchrani. Blind separation of second-order nonstationary and temporally colored sources. In *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 444–447. IEEE, 2001.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994.

S. Dähne, F. Meinecke, S. Haufe, J. Höhne, M. Tangermann, K.-R. Müller, and V. Nikulin. SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, 86:111–122, 2014.

A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

A. Delorme, T. Sejnowski, and S. Makeig. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, 34(4): 1443–1449, 2007.

A. Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig. Independent EEG Sources Are Dipolar. *PLOS One*, 7:1–14, 02 2012.

D. Ghahremani, S. Makeig, T. Jung, A. Bell, and T. Sejnowski. Independent component analysis of simulated EEG using a three-shell spherical head model. Technical report, Naval Health Research Center San Diego CA, 1996.

I. Goncharova, D. McFarland, T. Vaughan, and J. Wolpaw. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114(9):1580–1593, 2003.

T. Haavelmo. The Probability Approach in Econometrics. *Econometrica*, 12:iii–115, 1944.

R. Hardstone, S.-S. Poil, G. Schiavone, R. Jansen, V. Nikulin, H. Mansvelder, and K. Linkenkaer-Hansen. Detrended fluctuation analysis: a scale-free view on neuronal oscillations. *Frontiers in Physiology*, 3:450, 2012.

R. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis, 1970.

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

P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

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

A. Hyvärinen. Fast ICA for noisy data using Gaussian moments. In *ISCAS'99. Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI*, pages 57–61. IEEE, 1999.

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

A. Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1984):20110534, 2013.

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

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2002.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (5):1709–1731, 2010.

P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila. A new performance index for ICA: Properties, computation and asymptotic analysis. In *Latent Variable Analysis and Signal Separation*, pages 229–236. Springer, 2010.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Intergovernmental Panel on Climate Change. *Climate Change 2013 – The Physical Science Basis*. Cambridge University Press, 2014.

J. Jouzel, V. Masson-Delmotte, O. Cattani, G. Dreyfus, S. Falourd, G. Hoffmann, B. Minster, J. Nouet, J.-M. Barnola, J. Chappellaz, H. Fischer, J. C. Gallet, S. Johnsen, M. Leuenberger, L. Loulergue, D. Lüthi, H. Oerter, F. Parrenin, Raisbeck G., D. Raynaud, J. Schilt, A. Schwander, E. Selmo, R. Souchez, R. Spahni, B. Stauffer, J. P. Steffensen, B. Stenni, T. Stocker, J. L. Tison, M. Werner, and E. W. Wolff. Orbital and millennial antarctic climate variability over the past 800,000 years. *Science*, 317(5839): 793–796, 2007.

T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. Mckeown, V. Iragui, and T. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

M. Kleinsteuber and H. Shen. Uniqueness analysis of non-unitary matrix joint diagonalization. *IEEE Transactions on Signal Processing*, 61(7):1786–1796, 2013.

Z. J. Koles, M. S. Lazar, and S. Z. Zhou. Spatial patterns underlying population differences in the background EEG. *Brain Topography*, 2(4):275–284, 1990.

F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.

S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems 8 (NIPS 1995)*, pages 145–151. NIPS Foundation, 1995.

S. Makeig, T.-P. Jung, A. Bell, D. Ghahremani, and T. Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, 94(20):10979–10984, 1997.

S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694, 2002.

K. Matsuoka, M. Ohoya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

M. McKeown, T.-P. Jung, S. Makeig, G. Brown, S. Kindermann, T.-W. Lee, and T. Sejnowski. Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. *Proceedings of the National Academy of Sciences*, 95(3):803–810, 1998a.

M. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. Kindermann, A. Bell, and T.-J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998b.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Statistical properties of a blind source separation estimator for stationary time series. *Statistics & Probability Letters*, 82 (11):1865–1873, 2012.

E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3617–3620. IEEE, 1997.

K. Nordhausen. On robustifying some second order blind source separation methods for nonstationary time series. *Statistical Papers*, 55(1):141–156, Feb 2014.

P.L. Nunez and R. Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2006.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(10):2825–2830, 2011.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

Pexels GmbH. Pexels. https://www.pexels.com/, 2018. Accessed: 2018-10-30.

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, 49(9):1837–1848, 2001.

D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1513–1521. NIPS Foundation, 2015.

S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10):2003–2030, 2006.

M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz. Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

P. Tichavsky and A. Yeredor. Fast approximate joint diagonalization incorporating weight matrices. *IEEE Transactions on Signal Processing*, 57(3):878–891, 2009.

L. Tong, V. C. Soon, Y. F. Huang, and R. Liu. AMUSE: A new blind identification algorithm. In *IEEE International Symposium on Circuits and Systems*, pages 1784–1787. IEEE, 1990.

M. Treder, A. Bahramisharif, N. Schmidt, M. Van Gerven, and B. Blankertz. Brain-computer interfacing using modulations of alpha activity induced by covert shifts of attention. *Journal of NeuroEngineering and Rehabilitation*, 8(1):24, 2011.

L. Zhukov, D. Weinstein, and C. Johnson. Independent component analysis for EEG source localization. *IEEE Engineering in Medicine and Biology Magazine*, 19(3):87–96, 2000.

A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research*, 5(7):777–800, 2004.