PADERBORN UNIVERSITY

Faculty for
Computer Science,
Electrical Engineering
and Mathematics

Matrix and Tensor Decomposition Methods for Joint Blind Source Separation: Theory and Application to Functional Imaging Data

Von der Fakultät für Elektrotechnik, Informatik und Mathematik
der Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

M. Sc. Isabell Lehmann

Erster Gutachter: Prof. Dr. Peter Schreier
Zweite Gutachterin: Prof. Dr. Tülay Adali

Tag der mündlichen Prüfung: 05. November 2025

Paderborn 2025

Diss. EIM-E/391

Abstract

The problem of Joint Blind Source Separation (JBSS), i.e., the joint estimation of latent *sources* from multiple observed datasets, occurs in many areas like speech processing and biomedical signal processing, among others. For example, the latent sources found in biomedical datasets may provide biomarkers for the detection of disorders like schizophrenia.

Several matrix and tensor decomposition methods have been developed to achieve JBSS, i.e., assuming a generative model, to recover the true underlying sources up to permutation and scaling ambiguities and to correctly align the sources across datasets. Independent Vector Analysis (IVA), a multiset extension of the well-known Independent Component Analysis (ICA), achieves JBSS by maximizing independence of sources within a dataset and maximizing dependence of sources across datasets. There also exists multiset Canonical Correlation Analysis (mCCA), which was originally developed to identify variables that are maximally correlated across multiple datasets, called *canonical variables*. As correlation between more than two canonical variables can be quantified in different ways, different mCCA objective functions have been proposed, with the commonly known ones being sumcor, maxvar, minvar, genvar, and ssqcor. While mCCA does not specify a generative model in its formulation, by assuming the JBSS model, we can study under which conditions the canonical variables estimated by mCCA recover the true sources.

The conditions under which a method is able to achieve JBSS are more generally known as *source identification conditions*. Assuming the JBSS generative model matches the underlying physics of the data, the source identification conditions of a JBSS method must be satisfied to interpret the estimated sources, i.e., to attach physical meaning to them. Therefore, it is important to be aware of these conditions. The source identification conditions of IVA and mCCA-maxvar for JBSS have been derived in prior work. In this thesis, we extend the literature by deriving and proving the source identification conditions of mCCA-sumcor and discussing theoretical considerations for those of mCCA-minvar, mCCA-genvar, and mCCA-ssqcor. We substantiate the proposed theoretical conditions with numerical results. We also establish the connections between

an important set of matrix and tensor decomposition methods for JBSS. Understanding how different methods are connected and knowing their source identification conditions helps clarify which methods can be useful in which applications.

In some applications, for example, precision medicine, more important than identifying latent sources is understanding how different (matrix) datasets are related, which is challenging when datasets are high-dimensional and noisy. We propose a novel method for identifying the *relationship structure* among multiple datasets using JBSS, which describes which subsets of datasets have similar latent sources and how these subsets of datasets are related, and verify the success of the proposed method in simulations.

Finally, we apply JBSS methods to real-world functional Magnetic Resonance Imaging (fMRI) and functional Ultrasound (fUS) datasets. We show that the JBSS methods can perform data fusion of multi-task fMRI data and identify potential biomarkers for schizophrenia, and that the proposed method identifies a meaningful relationship structure among the multi-task fMRI datasets. Lastly, we show that JBSS methods estimate active brain networks in fUS data that are similar to those found by the gold standard analysis method, under less strict assumptions.

Zusammenfassung

Das Problem der gemeinsamen blinden Quellentrennung (Joint Blind Source Separation, JBSS), d. h. der gemeinsamen Schätzung verborgener *Quellen* aus mehreren beobachteten Datensätzen, ist von hoher Relevanz in vielen Fachdisziplinen, unter anderem in der Sprachverarbeitung und der biomedizinischen Signalverarbeitung. Beispielsweise können die in biomedizinischen Datensätzen gefundenen Quellen Biomarker für die Erkennung von Erkrankungen wie Schizophrenie liefern.

Verschiedene Matrix- und Tensorzerlegungsmethoden sind entwickelt worden, um JBSS zu erzielen, d. h. unter der Annahme eines generativen Modells die wahren zugrunde liegenden Quellen bis auf Permutations- und Skalierungsmehrdeutigkeiten wiederherzustellen und die Quellen über Datensätze hinweg korrekt anzuordnen. Die Analyse unabhängiger Vektoren (Independent Vector Analysis, IVA), eine Erweiterung der sehr bekannten Analyse unabhängiger Komponenten (Independent Component Analysis, ICA) für mehrere Datensätze, erreicht JBSS durch Maximierung der Unabhängigkeit von Quellen innerhalb eines Datensatzes und Maximierung der Abhängigkeit von Quellen über Datensätze hinweg. Des Weiteren gibt es die Kanonische Korrelationsanalyse für mehrere Datensätze (multiset Canonical Correlation Analysis, mCCA), welche ursprünglich entwickelt wurde, um Variablen zu identifizieren, die über mehrere Datensätze hinweg maximal korreliert sind, sogenannte *kanonische Variablen*. Da die Korrelation von mehr als zwei kanonischen Variablen auf unterschiedliche Weise quantifiziert werden kann, wurden verschiedene mCCA-Optimierungsfunktionen vorgeschlagen. Die fünf bekanntesten sind sumcor, maxvar, minvar, genvar und ssqcor. Obwohl mCCA in seiner Formulierung kein generatives Modell festlegt, können wir durch die Annahme des JBSS-Modells untersuchen, unter welchen Bedingungen die von mCCA geschätzten kanonischen Variablen den wahren Quellen entsprechen.

Die Bedingungen, unter denen eine Methode JBSS erzielen kann, werden allgemein als *Quellenidentifikationsbedingungen* bezeichnet. Unter der Annahme, dass das JBSS-Modell die zugrunde liegende Physik der Daten beschreiben kann, müssen die Quellenidentifikationsbedingungen einer JBSS-Methode erfüllt sein, um die geschätzten Quellen interpretieren zu können, d. h. ihnen eine physikalische Bedeutung zuordnen

zu können. Daher ist es wichtig, diese Bedingungen zu kennen. Die Quellenidentifikationsbedingungen von IVA und mCCA-maxvar für JBSS wurden in der Literatur bereits hergeleitet. In dieser Arbeit erweitern wir den Stand der Literatur, indem wir die Quellenidentifikationsbedingungen von mCCA-sumcor herleiten und beweisen und theoretische Überlegungen zu den Bedingungen von mCCA-minvar, mCCA-genvar und mCCA-ssqcor diskutieren. Wir unterstützen die vorgestellten theoretischen Bedingungen mit numerischen Ergebnissen. Wir stellen außerdem die Zusammenhänge zwischen wichtigen Matrix- und Tensorzerlegungsmethoden für JBSS her. Die Kenntnis über die Zusammenhänge verschiedener Methoden und über ihre Quellenidentifikationsbedingungen trägt dazu bei, Klarheit darüber zu schaffen, welche Methoden für welche Anwendungen nützlich sein können.

Noch wichtiger als die verborgenen Quellen zu identifizieren ist es in einigen Anwendungen, beispielsweise der Präzisionsmedizin, die Beziehung zwischen verschiedenen (Matrix-)Datensätzen zu verstehen. Dies ist bei hochdimensionalen und verrauschten Datensätzen eine Herausforderung. Wir schlagen eine neuartige Methode zur Identifizierung der *Beziehungsstruktur* zwischen mehreren Datensätzen mithilfe von JBSS vor, die beschreibt, welche Untergruppen von Datensätzen ähnliche zugrunde liegende Quellen haben und wie diese Untergruppen zusammenhängen. Anschließend demonstrieren wir den Erfolg der vorgeschlagenen Methode in Simulationen.

Abschließend verwenden wir JBSS-Methoden für die Analyse von realen funktionellen Magnetresonanztomographie (fMRT)-Daten und von funktionellen Ultraschall (fUS)-Daten. Wir zeigen, dass die JBSS-Methoden eine Datenfusion der fMRT-Daten durchführen und potenzielle Biomarker für Schizophrenie identifizieren können. Zudem identifiziert die vorgeschlagene Methode eine sinnvolle Beziehungsstruktur zwischen den fMRT-Datensätzen. Schließlich zeigen wir, dass JBSS-Methoden aktive Gehirnregionen in fUS-Daten identifizieren können, die ähnlich zu denen sind, die mit der als Goldstandard geltenden Analysemethode gefunden wurden, allerdings unter weniger strengen Annahmen.

Acknowledgments

First of all, I would like to thank you, Prof. Dr. Peter Schreier, for your support during my whole Ph.D. time. You made me feel comfortable in talking to you about any struggles that could occur during my Ph.D. time, which I believe is key for creating a group environment as pleasant, respectful, and fun as yours. You are a very thoughtful and reliable person, you always keep your promises, and your feedback has helped me think more critically. I am impressed by how easily you grasp new concepts and how well you explain complex topics in a way that makes them easy to understand. I also enjoyed our non-work-related chats.

Secondly, but no less importantly, I would like to thank you, Prof. Dr. Tülay Adali, for adopting me as your abroad Ph.D. student and making me feel like part of your group, including inviting me for two research visits at UMBC. I'm very grateful for the experiences I made during these times. I enjoyed our research discussions a lot, and your feedback has helped me write better and more convincing papers. When I think of you, I think of making connections—not only between methods but also between people. I truly appreciate the importance you put on creating a strong network and how you are striving to support young researchers, including me, in their academic careers.

I would like to thank my colleagues, Dr. Tanuj Hasija, soon-to-be-Dr. Maurice Kuschel, and Dr. Mohammad Soleymani, for the nice and friendly work atmosphere. Tanuj, I have learned so many things from you, technical as well as personal. You have been my psychologist at work, and there are so many situations in which I don't know what I would have done without you. Maurice, it was very nice to have you as a very motivated Ph.D. companion along the way until we both graduate now. Mohammad, I enjoyed your peaceful company.

I would like to thank Associate Prof. Dr. Bori Hunyadi and the Signal Processing Systems Group from TU Delft. My research visit at SPS had a significant impact on my future plans and how I grew as a person. Bori, your idea of leadership—to create a space where people feel safe and free to open up—has inspired me a lot, and I hope I will adopt this mindset in the future.

Thank you, Dr. Ben Gabrielson, Dr. Cathy Yang, Dr. Trung Vu, and Dr. Fran Lopez,

for making my #phdlife in the US so much fun. I miss our True Grit's visits and the amazing Asilomar experience, which made the conference feel more like a vacation with friends than like a typical work event. Fran, you made my days yellow, and I hope you always keep your happiness. Trung, my bro, I like that we laughed so much about everything and nothing. Cathy, you are such a kind person, and being around you made me feel calm and optimistic. Ben, you are one of the smartest people I know, and I am very happy that we collaborated on my last papers. Thank you for your patience in answering all my questions. I am not sure if you are aware of how much of a relief it has been for me to know that I can always count on your advice and help when I encounter difficult problems. I hope that we will continue to have research discussions every now and then. I like how we all became friends over time, and I enjoyed our joint music sessions so much.

Thank you, Prof. Dr. Sofia Olhede, for being my official mentor. I am impressed by what a kind, understanding, and humble person you are, and I hope to become a bit more like this. Thank you for your motivational words and for asking me always the right questions during our mentoring time and beyond. I feel very valued by you.

I would like to thank my collaborators, Dr. Evrim Acar, Dr. Siddique Akhonda, Dr. Pieter Kruizinga, and Dr. Vince Calhoun. Evrim, I enjoyed our research discussions so much and hope that one day we will finally meet in person.

I would also like to thank the Ph.D. students in our writing club for the super productive, appreciative, and cozy atmosphere. I was looking forward to every Friday, and for sure would not have been able to finish writing my thesis in time without you.

Thank you, Olli Warneke, for your support during almost all of my studies and my Ph.D. Thank you for having believed in me and having made me feel home. I'll always be grateful that I had you at my side.

I would like to thank my roommates, Jan Meyerhans and Ben Reitz, for providing me with a new home after I moved to Paderborn. I would also like to thank Jan for spending weekends and public holidays in the uni with me at the end of my thesis time, which helped me a lot in keeping motivated and finishing writing on time.

Thank you, Ed Focante, for supporting me in everything I do. Thank you for writing me the cutest letters and for making me feel safe when I was feeling insecure. Thank you for being the funniest person ever and making me laugh so many times. Thank you for proofreading countless emails and giving me company during so many VWS (Virtual Working Sessions). Thank you for your help during my final stage of writing. Thank you for the joint piano and singing sessions and your motivational words. Thank you for making my board game hobby bigger than ever. Thank you for making it so easy to talk to you about everything.

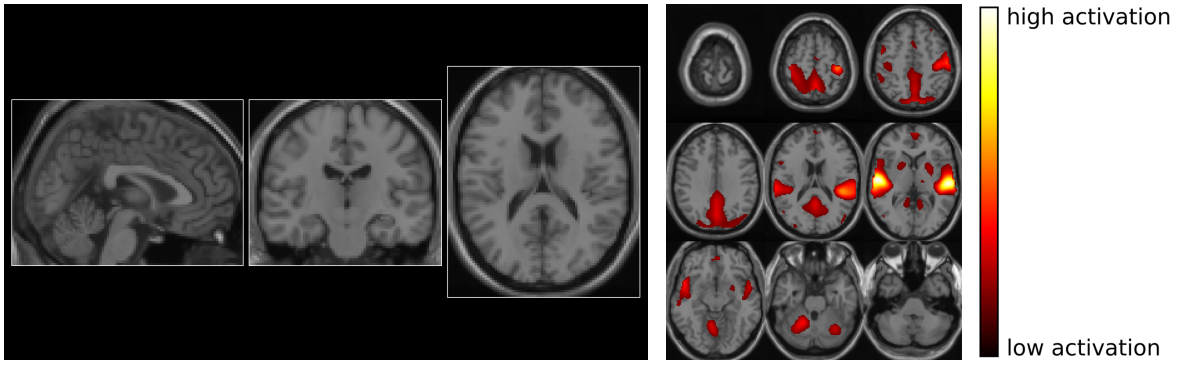
Finally, I would like to thank my family, Horst Lehmann, Iris Lehmann, and Sebastian Lehmann, for always believing in me.

Contents

1 Introduction

1.1 Motivation

A core problem in the field of neuroscience is to localize the sets of brain regions, in the following called *brain networks*, associated with various mental processes or behavior [Pes14], or, more broadly speaking, to understand how the functioning of the brain relates to the mind [Bas11]. The brain can be seen as a complex system consisting of multiple subnetworks [Pes14] that control all our actions and thoughts, and decomposing the brain into these subnetworks helps to understand brain organization [Pes14; Bas11]. In order to gain knowledge about the brain, brain data can be collected using different techniques. Structural Magnetic Resonance Imaging (sMRI), shown in Figure 1.1(a), captures the anatomy of the brain, which is composed of white matter and gray matter and surrounded by Cerebrospinal Fluid (CSF) [Cer01; Gie04]. However, sMRI does not provide insights into the function of brain networks. Instead, functional imaging techniques can infer brain activity indirectly through cerebral blood flow changes [Ror04; Kim23]. For example, functional Magnetic Resonance Imaging (fMRI) measures the



(a) (b)

Figure 1.1: (a) Structural Magnetic Resonance Imaging (sMRI) data of the human brain, visualized in three planes. (b) Functional Magnetic Resonance Imaging (fMRI) data of the human brain (red/yellow color shows active brain networks), overlaid over sMRI data (anatomical gray image).

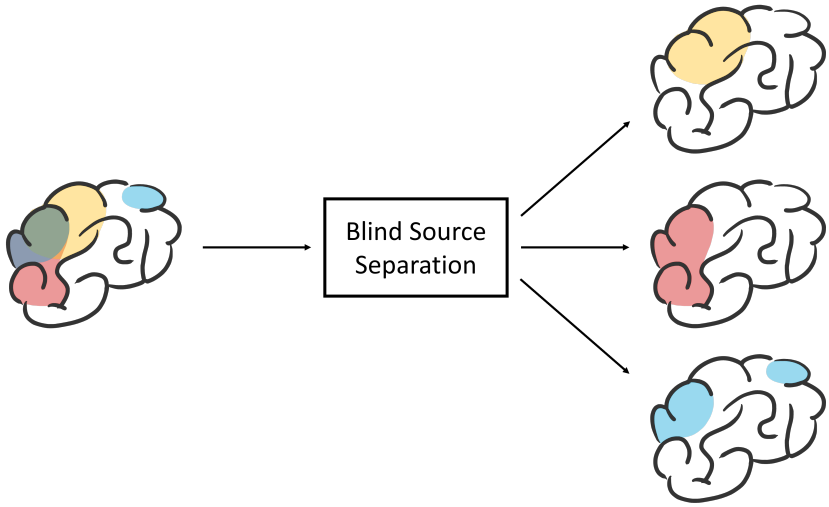


Figure 1.2: Idea of blind source separation visualized for the application in fMRI data. On the left, the observed dataset is shown, which contains a mixture of the sources. On the right, the three separated sources (yellow, red, and blue active brain networks) are shown.

Blood Oxygenation Level Dependent (BOLD) signal, which is higher when brain activity increases [Mum06]. For visualization, fMRI data is typically overlaid over an sMRI brain atlas to see where the active brain networks are located, as shown in Figure 1.1(b).

When functional imaging data of the brain is acquired, a mixture of all active brain networks is observed, as shown symbolically in Figure 1.2 on the left. Therefore, the observed data must be unmixed to reveal the separate active brain networks, as shown in Figure 1.2 on the right. Note that the active brain networks are not necessarily connected, as we see for the blue one. The problem of separating the active brain networks, more generally called latent *sources*, only based on observed mixtures of these sources is called *Blind Source Separation (BSS)* [Jut91]. Several methods have been developed to solve the BSS problem, which primarily work by assuming that the true underlying sources have specific statistical properties and then estimate sources having these assumed properties.

Finding the latent sources in fMRI data with BSS not only helps to understand how the healthy brain functions, but also how neurological disorders affect the brain [Smi04]. To allow generalization of drawn conclusions, fMRI experiments are typically conducted on multiple subjects or multiple times on the same subject [Smi04]. For example, in a study with multiple subjects performing the same tasks, similar brain networks are expected to be active. Therefore, to have a more reliable estimate of the active brain networks, they can be estimated jointly, which is referred to as *Joint Blind Source Separation (JBSS)* [Li09]. Also in JBSS, specific statistical properties are assumed for

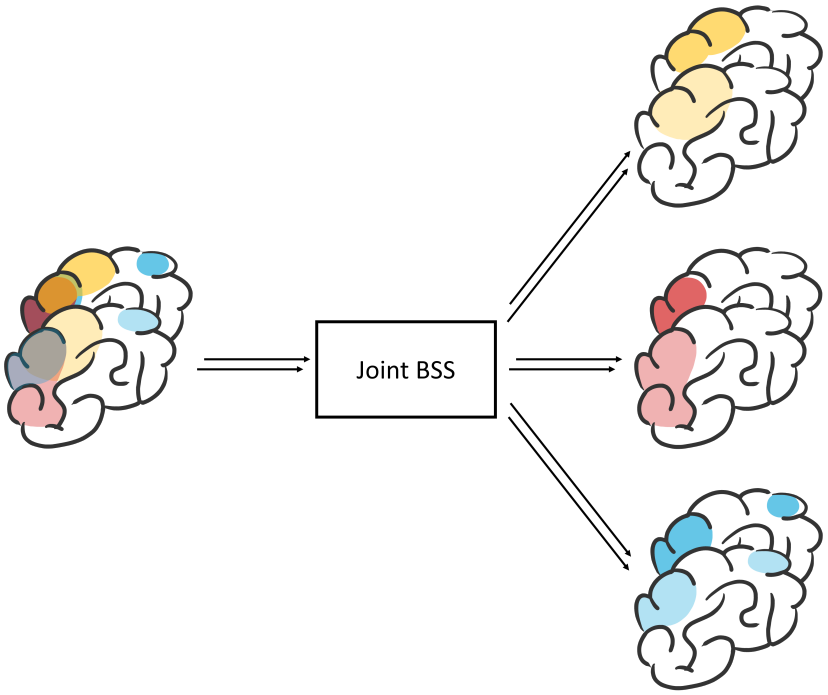


Figure 1.3: Idea of joint blind source separation visualized for the application in fMRI data. On the left, two observed datasets are shown (e.g., from two different subjects), where each observed dataset contains a mixture of the sources within that dataset. On the right, the three separated sources (yellow, red, and blue active brain networks) are shown for each dataset. Each source is dependent across datasets, visualized by the same color in different transparencies.

the sources within a dataset, and additionally, dependence of sources across datasets is exploited. This is visualized in Figure 1.3. On the left, there are two observed datasets containing mixtures of the sources, and on the right, the separated sources are shown in yellow, red, and blue for each dataset. Dependence of sources across datasets is represented by the same color in different transparencies. Note that the sources across datasets are dependent but not necessarily exactly the same; for example, the brain networks can be slightly different in their shape and activation.

1.2 State of the art

Since the introduction of BSS in the mid 1980s [Her86], many solutions for this problem have been proposed. In this thesis, we focus on JBSS methods that assume a linear mixing of the sources. Here, one of the oldest and probably most well-known methods for achieving BSS, i.e., recovering the true underlying sources up to permutation and scaling ambiguities, for one dataset is Independent Component Analysis (ICA) [Jut91; Com94; Hyv00; Ada14]. Based on the assumption that the true sources are statistically

independent, ICA can recover the sources from the observed mixtures up to permutation and scaling ambiguities. ICA has been extended to multiple datasets using either the joint ICA [Cal06] or the group ICA [Cal01] models, which are limited in performance, though, as they assume a common mixing matrix or a group-level source vector for all datasets. More recently, Independent Vector Analysis (IVA) [Kim06] has been proposed, a more flexible extension of ICA to multiple datasets that additionally leverages dependence of sources across datasets. In IVA, a Source Component Vector (SCV) is defined as a vector containing the dependent sources across datasets, e.g., an SCV contains all yellow, all red, or all blue sources in Figure 1.3, and then IVA achieves JBSS, i.e., recovers the true underlying sources up to permutation and scaling ambiguities and correctly aligns the sources across datasets, by minimizing the mutual information across all SCVs [Ada14].

Besides these methods, which are specifically developed for performing (joint) BSS, there exist methods that were initially introduced for jointly analyzing multiple datasets, but have later been shown to also be able to achieve JBSS when assuming a generative model. One of the earliest joint decomposition approaches is Canonical Correlation Analysis (CCA) [Hot36], which transforms two datasets into a new space where the correlation of the transformed variables, called *canonical variables*, is maximized. The extension of CCA to more than two datasets is called multiset Canonical Correlation Analysis (mCCA) [Ket71] or Generalized Canonical Correlation Analysis (GCCA) [Sor21]. While for CCA, maximization of pairwise correlations can be achieved using the singular value decomposition of the data's whitened cross-covariance matrix [Sch08], maximizing correlations in mCCA is not as straightforward because mCCA deals with more than two datasets at a time. As there are multiple pairwise correlations to consider, there are several different ways of quantifying correlation within a covariance matrix (such as the sum of the matrix's entries and the sum of the squared entries, among others), each leading to a different mCCA objective function. Kettenring summarized the five mCCA objective functions sumcor, maxvar, minvar, genvar, and ssqcor along with their solutions in 1971 [Ket71]. Both CCA and mCCA are frequently used methods because they are well-understood and easy to interpret, compared with neural networks.

The conditions under which methods like CCA and mCCA are able to achieve JBSS are known as *source identification conditions*. There are references attempting to derive these conditions, but they are limited in scope. The strongest possible source identification conditions for correlation-based JBSS methods (including mCCA) are outlined in [Vía11], but they are not derived for specific methods such as mCCA. The source identification conditions of CCA are derived and proven in [Li08], and of mCCA-maxvar and a modified sumcor objective function in [Li09]. The source identification conditions of the other mCCA methods remained unknown.

The so far mentioned JBSS methods—IVA, CCA, and mCCA—are all joint matrix decompositions, i.e., each dataset is represented as a matrix that is factorized into a set of matrices. The joint decomposition of multiple datasets is of interest when the datasets are related, e.g., fMRI datasets of multiple subjects or tasks. Besides decomposing multiple matrix datasets using joint matrix decompositions, they can be concatenated along a third dimension such that they form a *tensor*, i.e., a multi-dimensional (here 3D) array, and one can also take advantage of the inherent relationship among the datasets using tensor decompositions, which decompose the tensor into a core tensor multiplied with a matrix along each dimension [Kol09]. Compared with matrix decompositions, which are highly non-unique without imposing additional constraints (such as independence), tensor decompositions are unique under mild conditions [Kol09]. Since the connection between the well-known CANDECOMP/PARAFAC (CP) tensor decomposition [Har70; Car70] and joint (simultaneous) matrix diagonalization [Cha14] has been established in 2006 [DL06], several tensor decomposition methods for (joint) BSS have been developed [Le24]. Such a variety of JBSS methods—matrix and tensor decomposition-based—makes it hard for researchers to pick one for their specific application, and therefore, it is of interest to understand how methods are connected and which methods are suitable for which applications.

The suitability of a JBSS method for a specific application depends on two key factors. First, the JBSS model needs to match the underlying physics of the application, i.e., the way the observed data was generated [Ada22]. Second, the source identification conditions of a specific JBSS method need to be satisfied. Only if these two key factors are given, it makes sense to interpret the estimated sources, i.e., to attach physical meaning to them. Of course, for real-world data without knowledge on the ground-truth sources, one often cannot verify if these conditions are satisfied or not, but at least one should be able to check if it is reasonable to assume that the conditions of a method are satisfied for a specific application, or if another method should be used instead.

BSS and JBSS methods have been used in many applications. For example, brain networks found using ICA show developmental differences in autism spectrum disorder [Bos14] or may provide biomarkers for neurological diseases like Alzheimer's disease [Gre04; de 18] or schizophrenia [Cal04; Koc12]. Furthermore, ICA has been successfully used to evaluate the effect of epilepsy on language functions [Kar11] and for seizure onset localization [Boe24]. Applications of mCCA include identifying abnormal brain networks in patients with schizophrenia [Sui13] and improving the understanding of how the dysregulation of specific microRNAs contributes to the loss of brain areas in major depressive disorders [Qi18]. Besides interpreting the sources directly, they can be combined with machine learning methods to solve classification or prediction tasks.

For example, the combination of ICA with neural networks can help to classify beats in electrocardigram (ECG) data [Yu08], e.g., for arrhythmia classification [Sar14], to separate fetal ECG from the mother's ECG [Zia23], or to classify and remove artifacts in magnetoencephalography (MEG) data [Tre21], and the combination of ICA with a Gaussian Mixture Model has been used for tumor classification in MRI data [Che18].

Furthermore, BSS is not limited to the medical field. ICA is also highly relevant in speech processing [Kim07], where each source corresponds to a speaker and a set of microphones observes mixtures of the speakers, and in wireless communications [Luo18], among others. For example, neural networks in combination with ICA have been used for the prediction of the future value of exchange rates in currency markets [Hen19] and for the prediction of large-scale ground deformation [Pen24], and ICA with Support Vector Machine (SVM) for hyperspectral image classification [Fal14]. IVA has been used for molecular data fusion [Bou21], the combination of IVA with k-means clustering for abandoned object detection [Bhi17], and IVA with SVM for misinformation detection [Dam24]. Extensions of mCCA have been proposed to address new challenges, e.g., scalable mCCA for large-scale word embedding [Fu17], mCCA utilizing structural regularization [Kan18] and subspace intersection approaches [Sor21] for cross-language information retrieval, and supervised mCCA for audio- and video-based emotion recognition [Gao17]. Further advancements include deep-learning-based extensions of mCCA, e.g., Deep GCCA (DGCCA) for phonetic transcription of acoustic and articulatory measurements and for friend recommendation [Ben17], and DGCCA with an attention mechanism for emotion recognition [Lan20]. Friend recommendation has also been enhanced with graph mCCA [Che19] and with a graph autoencoder for mCCA [Kal21].

In some applications, however, more important than identifying latent sources is understanding how different (matrix) datasets are related. For example, the idea of precision medicine is that treatment should be tailored to individual patient characteristics instead of treating every patient in the same way [Kos19a]. By viewing each patient as a dataset and grouping similar datasets together, the treatment of a new patient can be adjusted based on what is known from other patients with similar disease characteristics. Another application where it is of interest to identify the relationship among multiple datasets is in the detection and prediction of epileptic seizures. It has been found that the correlations between measures of the peripheral Autonomic Nervous System (ANS) increased before Generalized Tonic-Clonic Seizures (GTCSs) and decreased after the seizures [Vie21] and that the changes in these correlations therefore might contain information for predicting seizures [Vie23]. Therefore, identifying the changes in the correlations of the ANS measures, i.e., identifying the relationship among multiple ANS datasets at several time instances, can provide novel biomarkers for seizure detection or

even seizure prediction.

However, identifying the relationship among multiple datasets is challenging because the observed datasets are often high-dimensional and noisy. Therefore, using sources estimated by JBSS, where each source has a one-to-one relationship among all datasets, can help tackle this problem. Research in this direction already exists, but we are not aware of a method with the goal to identify the *relationship structure* among multiple datasets using estimated sources from JBSS, which describes which subsets of datasets have similar latent sources and how these subsets of datasets are related. Some of the related methods in the literature identify the relationship among the individual sources within each SCV instead of the relationship structure among the datasets. This is, for example, done in [Akh21], where the dependence structure of latent sources, i.e., datasets across which the sources are dependent, is identified using IVA. Also in [Yan22], an approach based on IVA and Gershgorin discs is proposed for identifying groups of sources that are related with each other. However, for both methods, a postprocessing would be necessary to identify the relationship structure among the datasets from the method's output, which highlights the need to develop a method with this goal directly.

In this thesis, we address the following two important challenges in the field of JBSS:

- Choice of JBSS method for a specific application:

  The large number of JBSS methods makes it hard for researchers to choose one for their application. Different methods are able to achieve JBSS, i.e., to recover the true sources up to permutation and scaling ambiguities and to correctly align sources across datasets, under different assumptions, and the results will depend on whether a method's assumptions are likely to be satisfied. In addition, understanding how methods are connected helps to figure out their advantages and disadvantages, and this information, together with understanding which method has which assumptions, clarifies which method is suitable for which type of application.

- Identification of the relationship among multiple datasets using JBSS:

  Understanding how datasets are related is challenging when datasets are high-dimensional and noisy. One way to simplify the problem is to look at the latent sources estimated by JBSS and make use of how these sources of different datasets are related to each other. However, one source can be related among all datasets, another one only among a subset of datasets, another one among a different subset, and so on. Thus, it is an open question how the relationship among multiple datasets can be summarized and quantified by fusing this information, i.e., how

the sources estimated by JBSS can be used to identify the relationship structure among multiple datasets, which describes which subsets of datasets have similar latent sources and how these subsets of datasets are related.

1.3 Contributions

This thesis aims at addressing the challenges in the field of JBSS described in the previous section. The contributions of this thesis can be divided into theoretical and application-specific contributions. In the theoretical contributions, we establish the connections between different JBSS methods, propose their source identification conditions, and present a robust method for identifying the relationship among multiple datasets. Then, we apply JBSS methods to real-world biomedical applications. In the following, we first summarize the theoretical and then the application-specific contributions in more detail.
Theoretical contributions:

(i) Establish the connections between matrix and tensor decomposition methods for JBSS. *(Chapter 4)*

We first provide a concise overview of the matrix decomposition methods ICA, IVA, CCA, and mCCA, and the tensor decomposition method PARAFAC2, a more flexible alternative to the well-known CP tensor decomposition. Then, we derive the connections between the two popular mCCA methods sumcor and maxvar and prove that under specific constraints, the canonical variables estimated by these two methods are scaled versions of each other. Lastly, we show that the PARAFAC2 and IVA models are intimately related: both can be formulated as the multiplication of a mixing matrix with a source matrix for each dataset, with the difference that PARAFAC2 imposes stronger constraints on the mixing matrices and IVA imposes stronger constraints on the source matrices. In simulations, we demonstrate the differences between PARAFAC2 and IVA in two different scenarios and show that also PARAFAC2 is able to achieve JBSS.

(ii) Propose the source identification conditions of mCCA methods for JBSS. *(Chapter 5)*

After summarizing the source identification conditions of ICA, IVA, CCA, and mCCA-maxvar, we derive and prove the source identification conditions of mCCA-sumcor, conjecture those of mCCA-minvar, and discuss theoretical considerations for those of mCCA-genvar and mCCA-ssqcor. We substantiate the proposed

theoretical conditions with simulations and test the statistical efficiency of the five mCCA methods for finite samples in experiments that are close to (but not exactly) violating their conditions. Based on our experiments, genvar seems to have the least restrictive source identification conditions among all mCCA methods and to be also more statistically efficient than the other methods. Therefore, we generally recommend genvar as the preferred mCCA method for JBSS (over sumcor and maxvar, despite being the perhaps most commonly used methods) and recommend maxvar only in specific scenarios based on our experiments. With these findings, we partly answer the question of which JBSS method should be used in which application.

(iii) Identify the relationship structure among multiple datasets using JBSS methods. *(Chapter 6)*

As the last theoretical contribution, we propose a novel method for identifying the relationship structure among multiple datasets using JBSS, being the first to address this challenge to the best of our knowledge. In the first step, latent sources are estimated using IVA. In step 2, with the help of bootstrap-based hypothesis testing, the SCVs are identified as "common", i.e., all sources in an SCV are correlated with each other, or "structured", i.e., only subsets of sources are correlated. In the third step, hierarchical clustering is applied to features extracted from the structured SCVs, and the relationship structure among the datasets is revealed in the hierarchical clustering. In simulations, we first demonstrate that the proposed method outperforms competing techniques in step 2 and then verify the success of the complete method for identifying the relationship structure among multiple datasets.

Application-specific contributions:

(i) Find biomarkers for schizophrenia. *(Chapter 7)*

We show that IVA and PARAFAC2 can perform data fusion, i.e., joint analysis of multiple related datasets, of fMRI data collected from patients with schizophrenia and healthy controls who performed three different tasks in which they needed to press a button when different auditory or visual stimuli occurred. IVA and PARAFAC2 estimate almost the same sources, which correspond to brain networks that are expected to be active during these tasks. This indicates the success of both methods for achieving JBSS on real-world fMRI data. Furthermore, those brain networks whose activity is significantly different in patients and controls are promising candidates for brain-based biomarkers in the detection of schizophrenia.

(ii) Identify the relationship structure among multiple task fMRI datasets. *(Chapter 7)*

We apply the proposed method to the previously introduced fMRI dataset to identify the relationship structure among the task datasets. The proposed method groups task datasets from similar tasks together, which verifies the method's success on real-world fMRI data and paves the way for future applications like identifying patients with similar disease characteristics for precision medicine.

(iii) Map the 3-dimensional (3D) pathway of visual information processing. *(Chapter 8)*

As the last contribution, we study the application of ICA and IVA to understand where visual information is processed in the 3D mouse brain, based on only 2-dimensional (2D) observations collected with functional Ultrasound (fUS). FUS is a relatively new 2D functional imaging technique, introduced in 2011 [Mac11], with a higher spatiotemporal resolution than fMRI. The gold standard for analyzing fUS data collected during a task is using correlation images. These have the disadvantage that they assume the same brain response strength for all stimuli (which has been shown to vary across stimuli) and that they do not take information across the 2D observations into account. Therefore, we propose analyzing fUS data using ICA and IVA, which do not make any assumptions on the stimulus responses. The sources estimated by both methods match the active brain networks found by correlation images, and thus, both methods can be used to analyze fUS task data. While ICA might split or merge brain networks and requires (time-consuming) manual alignment, IVA might miss brain networks but automatically aligns them by leveraging the information across slices.

In the following, we provide the reader with a mathematical definition of BSS and JBSS and present the matrix and tensor decomposition methods that we study in this thesis. We continue with the theoretical and application-specific contributions as described above and complete the thesis with a summary of the most important findings and a discussion about interesting future directions.

2 Joint Blind Source Separation

In this chapter, after introducing the notation used in this thesis, we mathematically frame the problems of Blind Source Separation (BSS) and Joint Blind Source Separation (JBSS) and define what is meant by source identification conditions.

2.1 Notation

In the remainder of this thesis, scalars are denoted by italicized small letters, a, vectors by bold small letters, $\mathbf{a}$, and matrices by bold capital letters, $\mathbf{A}$. If not stated differently, the n^{th} column of $\mathbf{A} \in \mathbb{R}^{M \times N}$ is denoted as $\mathbf{A}(:, n) \in \mathbb{R}^{M}$ or $\mathbf{a}_n \in \mathbb{R}^{M}$ as a more compact notation of $\mathbf{a}_{:n}$, the m^{th} row of $\mathbf{A}$ is denoted as $\mathbf{A}(m, :) \in \mathbb{R}^{1 \times N}$ or $\mathbf{a}_{m:} \in \mathbb{R}^{1 \times N}$, and the $(m, n)^{\text{th}}$ element of $\mathbf{A}$ is denoted as $\mathbf{A}(m, n)$ or a_{mn}. The superscript $(\cdot)^{\mathsf{T}}$ denotes the transpose of a matrix, $\|\cdot\|_F^2$ denotes the squared Frobenius norm of a matrix, $\text{diag}(\mathbf{a})$ denotes a diagonal matrix with the vector $\mathbf{a}$ on the main diagonal, and $\text{blkdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_N)$ denotes a block-diagonal matrix with the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$ on the main diagonal blocks. The matrix $\mathbf{I}_{N \times N}$ denotes the $(N \times N)$-identity matrix, and $\mathbf{0}_{N \times N}$ denotes a $(N \times N)$-matrix full of zeros. The expected value operator is denoted by $\mathbb{E}\{\cdot\}$, the differential entropy of a continuous random variable a by $\mathcal{H}(a)$, and the mutual information of two random variables a_m and a_n by $\mathcal{I}(a_m; a_n)$. The covariance matrix of a random vector $\mathbf{a}_n$ is denoted as $\mathbf{C}_{\mathbf{a}_n} = \mathbb{E}\left\{\mathbf{a}_n \mathbf{a}_n^{\mathsf{T}}\right\}$, and the covariance matrix of a random vector $\mathbf{a}^{[k]}$ is denoted as $\mathbf{C}_{\mathbf{a}}^{[k]}$ for better readability (instead of $\mathbf{C}_{\mathbf{a}^{[k]}}$). The cross covariance matrix of two random vectors $\mathbf{a}_m$ and $\mathbf{a}_n$ is denoted as $\mathbf{C}_{\mathbf{a}}^{[m,n]}$. When an eigenvalue decomposition (EVD) or Singular Value Decomposition (SVD) is performed, we assume without loss of generality (w.l.o.g.) that the eigenvalues or singular values $a_1, \ldots, a_N$ are sorted in descending order, i.e., $a_1 \geq \cdots \geq a_N$. The big-$\mathcal{O}$ complexity of an algorithm is denoted by $\mathcal{O}(\cdot)$. The most commonly used symbols are summarized in the List of Symbols at the end of the thesis.

2.2 Blind Source Separation for $K = 1$ dataset

Let $\mathbf{x} \in \mathbb{R}^N$ denote a dataset, represented as a random vector, which is assumed to be generated according to the model [Hyv00]

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is an unknown invertible mixing matrix, and

$$\mathbf{s} = \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} \in \mathbb{R}^N$$

is an unknown source vector containing the source components s_n. W.l.o.g., the source vector is assumed to have zero mean, i.e., $\mathbb{E}\{\mathbf{s}\} = \mathbf{0}_N$. The goal of BSS is to estimate the source vectors [Hyv00]

$$\widehat{\mathbf{s}} = \mathbf{W}\mathbf{x},$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is an invertible demixing matrix. BSS is successfully achieved if the estimated source vector $\widehat{\mathbf{s}}$ recover the true source vector $\mathbf{s}$ up to unavoidable permutation and scaling ambiguities [Hyv00], i.e., if

$$\widehat{\mathbf{s}} = \mathbf{P}\mathbf{\Gamma}\mathbf{s},$$

where $\mathbf{P} \in \mathbb{R}^{N \times N}$ is an arbitrary permutation matrix, and $\mathbf{\Gamma} \in \mathbb{R}^{N \times N}$ is a diagonal scale matrix. Then, the desired $\mathbf{W}$ is $\mathbf{W} = \mathbf{P}\mathbf{\Gamma}\mathbf{A}^{-1}$. If the true source components are assumed to have unit variance and unit-variance source components are estimated, then the scale ambiguity reduces to a sign ambiguity, i.e., $\mathbf{W} = \mathbf{P}\widetilde{\mathbf{\Gamma}}\mathbf{A}^{-1}$, where $\widetilde{\mathbf{\Gamma}} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with diagonal elements equal to either -1 or 1.

Since matrix factorizations are generally non-unique, constraints have to be added to get a unique solution for $\mathbf{W}$ and $\widehat{\mathbf{s}}$. These constraints typically enforce the assumptions made in a specific method. In Section 3.1, the ICA method [Hyv00] is presented, where the assumed constraint is independence of the source components in $\mathbf{s}$.

2.3 Joint Blind Source Separation for $K > 1$ datasets

Now let there be K datasets $\mathbf{x}^{[k]}$, $k = 1, \ldots, K$, which are assumed to be generated by the model [Li09]

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K. \tag{2.2}$$

Here, $\mathbf{A}^{[k]} \in \mathbb{R}^{N \times N}$, $k = 1, \ldots, K$, are again unknown invertible mixing matrices, and

$$\mathbf{s}^{[k]} = \begin{bmatrix} s_1^{[k]} \\ \vdots \\ s_N^{[k]} \end{bmatrix} \in \mathbb{R}^N, \quad k = 1, \ldots, K,$$

are unknown zero-mean source vectors containing the source components $s_n^{[k]}$. The goal of JBSS is to estimate the source vectors [Li09]

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}.$$

JBSS is successfully achieved if the estimated source vectors $\widehat{\mathbf{s}}^{[k]}$ recover the true source vectors $\mathbf{s}^{[k]}$ up to unavoidable permutation and scaling ambiguities [And12], i.e., if

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{P}\boldsymbol{\Gamma}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K.$$

While the diagonal scale matrix $\boldsymbol{\Gamma}^{[k]} \in \mathbb{R}^{N \times N}$ can be different for each dataset, the permutation matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ must be the same for all datasets [And12]. This means that the source components can be permuted arbitrarily within a dataset, but the permutation must be the same for each dataset so that the source components are aligned across datasets. Also here, the scale ambiguity can be reduced to a sign ambiguity by assuming and estimating unit-variance source components.

As in ICA, also in IVA, different methods have different assumptions, which are enforced through the corresponding constraints. A common and practical assumption in JBSS is that the source components $s_n^{[k]}$ within a dataset are independent [Li09; Has20], and when additionally dependence of source components across datasets is leveraged, this leads to the IVA method presented in Section 3.2. However, another possible constraint is that source components $s_n^{[k]}$ across datasets are maximally correlated, i.e., linear dependent, leading to the CCA and mCCA methods presented in Sections 3.3 and 3.4.

2.4 Source identification conditions for BSS and JBSS

In this section, we introduce the definition of source identification conditions for BSS and JBSS.

Definition 2.1 (Source identification conditions for BSS)**.** Let $\mathbf{s} \in \mathbb{R}^N$ be the true source vector, and let $\widehat{\mathbf{s}} \in \mathbb{R}^N$ be the estimated source vector of a specific BSS method. We define the *source identification conditions* of the method as the conditions on the true source components in $\mathbf{s}$ under which the method achieves BSS, i.e., under which the estimated source vector $\widehat{\mathbf{s}}$ recovers the true source vector $\mathbf{s}$ up to scaling and permutation:

$$\widehat{\mathbf{s}} = \mathbf{P}\mathbf{\Gamma}\mathbf{s},$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a permutation matrix, and $\mathbf{\Gamma} \in \mathbb{R}^{N\times N}$ is a diagonal scale matrix.

This definition can be easily extended to JBSS:

Definition 2.2 (Source identification conditions for JBSS)**.** Let $\mathbf{s}^{[k]} \in \mathbb{R}^N$, $k = 1, \ldots, K$, be the true source vectors, and let $\widehat{\mathbf{s}}^{[k]} \in \mathbb{R}^N$, $k = 1, \ldots, K$, be the estimated source vectors of a specific JBSS method. We define the *source identification conditions* of the method as the conditions on the true source components in $\mathbf{s}^{[k]}$ under which the method achieves JBSS, i.e., under which the estimated source vectors $\widehat{\mathbf{s}}^{[k]}$ recover the true source vectors $\mathbf{s}^{[k]}$ up to a scaling and a common permutation:

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{P}\mathbf{\Gamma}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a permutation matrix, which is common for all datasets, and $\mathbf{\Gamma}^{[k]} \in \mathbb{R}^{N\times N}$ are diagonal scale matrices.

2.5 A small note on the use of samples

Our notation up to this point represents datasets $\mathbf{x}^{[k]} \in \mathbb{R}^N$ and source vectors $\mathbf{s}^{[k]} \in \mathbb{R}^N$ as random vectors. However, in practice, we handle observed datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N\times V}$, where V is the number of samples, and $\mathbf{X}^{[k]}(:, v) \in \mathbb{R}^N$ denotes the v^{th} sample of $\mathbf{X}^{[k]}$. Then, source vectors are denoted by $\mathbf{S}^{[k]} \in \mathbb{R}^{N\times V}$, where $\mathbf{s}^{[k]}_{r:} \in \mathbb{R}^{1\times V}$ denotes the r^{th} row, i.e., source component, of $\mathbf{S}^{[k]}$, and covariance matrices, e.g., $\mathbf{C}^{[k]}_{\mathbf{x}} = \mathbb{E}\left\{\mathbf{x}\left(\mathbf{x}^{[k]}\right)^{\mathsf{T}}\right\}$, are replaced by their estimates, e.g., $\widehat{\mathbf{C}}^{[k]}_{\mathbf{x}} = \frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}$, and so on. In the remainder

of the thesis, we keep the random vector notation for deriving theoretical concepts and use the sample notation when we discuss practical implementation, e.g., in the form of pseudocode.

3 Matrix and tensor decomposition methods for BSS and JBSS

This chapter provides the reader with an overview of the matrix and tensor decomposition methods studied in this thesis. We explain ICA, IVA, CCA, mCCA, and PARAFAC2 all with a common notation for all methods and provide pseudocodes for an easier understanding of the methods.[1]

3.1 Independent Component Analysis

Independent Component Analysis (ICA) is one of the oldest and most well-known methods for achieving BSS. The generative model of ICA is given by (2.1) and repeated here for better readability:

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

and ICA estimates $\widehat{\mathbf{s}} = \mathbf{W}\mathbf{x}$, with the goal that $\widehat{\mathbf{s}}$ equals $\mathbf{s}$ up to scaling and permutation. By defining the probability density function (pdf) of the n^{th} source component as

$$p_{s_n}(s_n), \quad n = 1, \ldots, N,$$

we can formulate the assumptions of ICA [Hyv00]:

- W.l.o.g., source components have zero mean, i.e.,

$$\mathbb{E}\{\mathbf{s}\} = \mathbf{0}_N. \tag{ICA.1}$$

[1] Section 3.4 of this chapter is based on the paper: "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, *submitted for review*, 2025. I specifically contributed to proving the source identification conditions of sumcor, deriving the connection between maxvar [Ket71] and sumcor [Nie95], implementing and running the experiments, creating all figures, and writing the paper. B. Gabrielson has provided theoretical insights in the connection between maxvar [Car68] and sumcor [Nie95] and the source identification conditions of genvar. Discussions and feedback along the way until the final paper have contributions from all authors.

- W.l.o.g., source components have unit variance, i.e.,

$$\mathbb{E}\left\{s_n^2\right\} = 1, \quad n = 1, \ldots, N. \tag{ICA.2}$$

- Source components are independent, i.e.,

$$p_{s_1,\ldots,s_N}(s_1, \ldots, s_N) = p_{s_1}(s_1) \cdots p_{s_N}(s_N). \tag{ICA.3}$$

Several algorithms for ICA have been developed. The popular Infomax ICA [Bel95] uses a neural network approach with a given non-linearity function, where the observed data is the input and the estimated source components are the output of the neural network. Then, maximizing statistical independence between the estimated source components is achieved by maximizing the mutual information between the input and output of the neural network. FastICA [Hyv00] achieves independence of the estimated source components by making them as non-Gaussian as possible, where non-Gaussianity is measured by an approximation of negentropy. More recently, ICA by Entropy Bound Minimization (ICA-EBM) [Li10] has been proposed, where independence of the estimated source components is achieved by minimizing their entropy, which is approximated by the maximum entropy bound. In this thesis, we use FastICA, which is described in more detail below.

FastICA

The idea of FastICA is based on the central limit theorem, which states that the sum of an infinite number of random variables has a Gaussian distribution under certain conditions [Hyv00]. Therefore, if source components are assumed to be non-Gaussian, the distribution of the sum (mixture) of several source components is more Gaussian than any of the individual components, and maximizing non-Gaussianity of an estimated source component

$$\widehat{s}_n = \mathbf{w}_n^{\mathsf{T}}\mathbf{x},$$

where $\mathbf{w}_n^{\mathsf{T}} \in \mathbb{R}^{1\times N}$ denotes the n^{th} row of the demixing matrix $\mathbf{W}$, gives one of the true source components [Hyv00]. Note that we denote the n^{th} row of $\mathbf{W}$ by $\mathbf{w}_n^{\mathsf{T}}$ instead of $\mathbf{w}_{n:}$ in order to be consistent with the literature. An additional constraint needs to be imposed on the next source component so that it does not estimate the same one as before. Since source components are assumed to be independent, and independence implies uncorrelatedness, Gram-Schmidt-like decorrelation can be used to estimate the remaining source components in a deflationary way.

The question is, therefore, how non-Gaussianity of the random variable $\widehat{s}_n$ can be measured. Two classical measures are Kurtosis and negentropy [Hyv00]. The Kurtosis, defined as

$$\operatorname{kurt}(\widehat{s}_n) = \mathbb{E}\left\{\widehat{s}_n^4\right\} - 3\left(\mathbb{E}\left\{\widehat{s}_n^2\right\}\right)^2,$$

is zero for Gaussian random variables, positive for super-Gaussian random variables (e.g., Laplacian), and negative for sub-Gaussian random variables (e.g., uniform) [Hyv00]. The squared or absolute value of the Kurtosis can be used as a measure for non-Gaussianity, but it is very sensitive to outliers [Hub85]. The alternative measure is negentropy, defined as [Hyv00]

$$\mathcal{J}(\widehat{s}_n) = \mathcal{H}(\nu) - \mathcal{H}(\widehat{s}_n),$$

where

$$\mathcal{H}(\widehat{s}_n) = -\int p_{\widehat{s}_n}(\widehat{s}_n) \ln\left(p_{\widehat{s}_n}(\widehat{s}_n)\right) \mathrm{d}\widehat{s}_n$$

is the differential entropy of $\widehat{s}_n$, and $\mathcal{H}(\nu)$ is the differential entropy of a Gaussian random variable with the same mean and variance as $\widehat{s}_n$. As a Gaussian random variable has the largest differential entropy among all random variables that have the same variance, the negentropy is always non-negative and is zero if $\widehat{s}_n$ is Gaussian [Hyv00]. In practice, the negentropy is hard to compute, as an estimate of the pdf of $\widehat{s}_n$ would be necessary. Therefore, the FastICA algorithm uses the following approximation of negentropy to measure non-Gaussianity:

$$\mathcal{J}(\widehat{s}_n) \approx \left(\mathbb{E}\{g(\widehat{s}_n)\} - \mathbb{E}\{g(\nu)\}\right)^2,$$

where $g(\cdot)$ is a non-quadratic function, for which very useful options are [Hyv00]

$$g(u) = \ln(\cosh(u)) \quad \text{or} \quad g(u) = -e^{-\frac{u^2}{2}}.$$

For the FastICA algorithm using Newton's method, the first and second derivatives of $g(\cdot)$ are needed, which are

$$g'(u) = \tanh(u) \quad \text{or} \quad g'(u) = ue^{-\frac{u^2}{2}}$$

and

$$g''(u) = 1 - \tanh^2(u) \quad \text{or} \quad g'(u) = \left(1 - u^2\right) e^{-\frac{u^2}{2}}.$$

The resulting FastICA algorithm for finding $\mathbf{W} = \left[\mathbf{w}_1, \ldots, \mathbf{w}_N\right]^{\mathsf{T}}$ is derived in [Hyv00] and turns out to be incredibly simple. The pseudocode for FastICA is given in Algorithm

1, where a typical value for ϵ is 0.0001. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(IN^2V)$, where I is the number of iterations. It is dominated by line 8 in Algorithm 1.

Algorithm 1 FastICA [Hyv00]

Input: $\mathbf{X} \in \mathbb{R}^{N \times V}$

1: $\mathbf{X} \leftarrow \left(\frac{1}{V}\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1/2}\mathbf{X}$ ▷ whiten data
2: $\mathbf{W} \leftarrow \text{randn}(N, N)$ ▷ initialize demixing matrix $\mathbf{W}$
3: $\mathbf{W} \leftarrow \left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)^{-1/2}\mathbf{W}$ ▷ make $\mathbf{W}$ orthogonal
4: **for** $i = 1, \ldots, \text{maxIter}$ **do**
5: $\quad\mathbf{W}^{(\text{old})} \leftarrow \mathbf{W}$
6: $\quad$**for** $n = 1, \ldots, N$ **do**
7: $\quad\quad\mathbf{w}_n \leftarrow (\mathbf{W}(n,:))^\mathsf{T}$
8: $\quad\quad\mathbf{W}(n,:) \leftarrow \left(\frac{1}{V}\sum_{v=1}^{V}\mathbf{X}(:,v)g'\left(\mathbf{w}_n^\mathsf{T}\mathbf{X}(:,v)\right) - \frac{1}{V}\sum_{v=1}^{V}g''\left(\mathbf{w}_n^\mathsf{T}\mathbf{X}(:,v)\right)\mathbf{w}_n\right)^\mathsf{T}$
9: $\quad\mathbf{W} \leftarrow \left(\mathbf{W}\mathbf{W}^\mathsf{T}\right)^{-1/2}\mathbf{W}$
10: $\quad$**if** $\max\left(1 - \text{diag}\left(\left|\mathbf{W}^{(\text{old})}\mathbf{W}^\mathsf{T}\right|\right)\right) < \epsilon$ **then** ▷ FastICA converged
11: $\quad\quad$**break**
12: $\widehat{\mathbf{S}} \leftarrow \mathbf{W}\mathbf{X}$

Output: $\mathbf{W} \in \mathbb{R}^{N \times N}$, $\widehat{\mathbf{S}} \in \mathbb{R}^{N \times V}$

When ICA is applied separately to multiple datasets, then $\widehat{\mathbf{s}}^{[k]} = \mathbf{P}^{[k]}\mathbf{\Gamma}^{[k]}\mathbf{s}^{[k]}$, $k = 1, \ldots, K$, are estimated. As the permutation matrix $\mathbf{P}^{[k]}$ can be different for each dataset, the source components may not be aligned across datasets. To overcome this problem and automatically align the independent source components across datasets, IVA has been proposed, which is explained in the next section.

3.2 Independent Vector Analysis

Independent Vector Analysis (IVA) [Kim06] is an extension of ICA to multiple datasets. The generative model for IVA is given by (2.2) and repeated here for better readability:

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

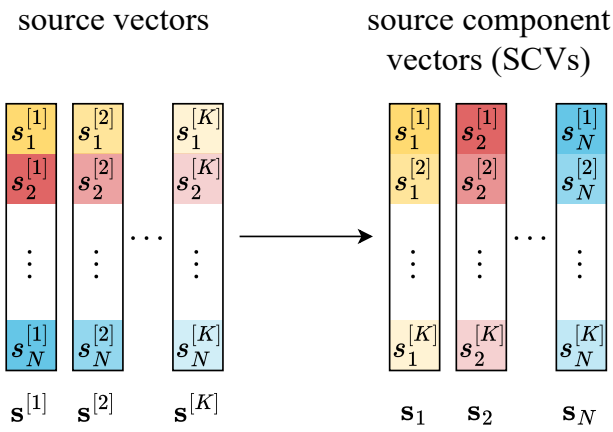


Figure 3.1: The n^{th} element of all source vectors is concatenated to form the n^{th} SCV.

and IVA estimates the source vectors $\widehat{\mathbf{s}}^{[k]} = \mathbf{W}^{[k]}\mathbf{x}^{[k]}$, where

$$\mathbf{W}^{[k]} = \begin{bmatrix} \left(\mathbf{w}_1^{[k]}\right)^{\mathsf{T}} \\ \vdots \\ \left(\mathbf{w}_N^{[k]}\right)^{\mathsf{T}} \end{bmatrix},$$

with the goal that $\widehat{\mathbf{s}}^{[k]}$ equals $\mathbf{s}^{[k]}$ up to scaling and permutation. Note that also here, we denote the n^{th} row of $\mathbf{W}^{[k]}$ by $\left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}} \in \mathbb{R}^{1\times N}$ instead of $\mathbf{w}_{n:}^{[k]}$ in order to be consistent with the literature. In order to easier state the assumptions of IVA, let the n^{th} Source Component Vector (SCV) be defined by stacking the n^{th} source of each dataset as

$$\mathbf{s}_n = \begin{bmatrix} s_n^{[1]} \\ \vdots \\ s_n^{[K]} \end{bmatrix} \in \mathbb{R}^K, \quad n = 1, \ldots, N,$$

which is visualized in Figure 3.1. Then, by defining the joint pdf of the n^{th} SCV as

$$p_{\mathbf{s}_n}(\mathbf{s}_n) = p_{s_n^{[1]},\ldots,s_n^{[K]}}\left(s_n^{[1]}, \ldots, s_n^{[K]}\right), \quad n = 1, \ldots, N,$$

we can formulate the assumptions of IVA [And14]:

- W.l.o.g., source components have zero mean, i.e.,

$$\mathbb{E}\left\{\mathbf{s}^{[k]}\right\} = \mathbf{0}_N, \quad k = 1, \ldots, K. \tag{IVA.1}$$

- W.l.o.g., source components have unit variance, i.e.,

$$\mathbb{E}\left\{\left(s_n^{[k]}\right)^2\right\} = 1, \quad n = 1, \ldots, N, \quad k = 1, \ldots, K. \tag{IVA.2}$$

- Source components are independent across SCVs (but can be dependent within an SCV), i.e.,

$$p_{\mathbf{s}_1,\ldots,\mathbf{s}_N}(\mathbf{s}_1, \ldots, \mathbf{s}_N) = p_{\mathbf{s}_1}(\mathbf{s}_1) \cdots p_{\mathbf{s}_N}(\mathbf{s}_N). \tag{IVA.3}$$

  This can be seen in Figure 3.1, where the independent SCVs have different colors, and the dependent source components within an SCV have the same color in different transparencies. Assumption (IVA.3) is not w.l.o.g. and may not necessarily hold for any given dataset. However, it is a common and practical assumption for deriving solutions and studying their source identification conditions [Li09; Has20].

IVA identifies maximally independent SCVs by minimizing the mutual information among the estimated SCVs, which is defined as [And10]

$$\mathcal{I}_{\text{IVA}}\left(\mathbf{W}^{[1]}; \ldots; \mathbf{W}^{[K]}\right) = \sum_{n=1}^{N} \mathcal{H}\left(\widehat{\mathbf{s}}_n\right) - \sum_{k=1}^{K} \ln\left(\left|\det\left(\mathbf{W}^{[k]}\right)\right|\right) - C, \tag{3.1}$$

where

$$\mathcal{H}\left(\widehat{\mathbf{s}}_n\right) = \mathbb{E}\left\{-\ln\left(p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n)\right)\right\} = -\int_{\mathbb{R}^K} p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n) \ln\left(p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n)\right) \mathrm{d}\widehat{\mathbf{s}}_n$$

is the differential entropy of the n^{th} estimated SCV $\widehat{\mathbf{s}}_n$, and C is a constant term. Since

$$\mathcal{H}\left(\widehat{\mathbf{s}}_n\right) = \sum_{k=1}^{K} \mathcal{H}\left(\widehat{s}_n^{[k]}\right) - \mathcal{I}\left(\widehat{\mathbf{s}}_n\right),$$

where

$$\mathcal{H}\left(\widehat{s}_n^{[k]}\right) = \mathbb{E}\left\{-\ln\left(p_{\widehat{s}_n^{[k]}}\left(\widehat{s}_n^{[k]}\right)\right)\right\} = \int p_{\widehat{s}_n^{[k]}}\left(\widehat{s}_n^{[k]}\right) \ln\left(p_{\widehat{s}_n^{[k]}}\left(\widehat{s}_n^{[k]}\right)\right) \mathrm{d}\widehat{s}_n^{[k]}$$

is the differential entropy of $s_n^{[k]}$ and $\mathcal{I}\left(\widehat{\mathbf{s}}_n\right)$ is the mutual information of the n^{th} SCV, we see that by minimizing the mutual information among the estimated SCVs, IVA simultaneously maximizes the mutual information within each SCV [And10], i.e., maximizes the dependence of the source components within an SCV. Depending on the selection of the multivariate pdf $p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n)$ for the estimated SCVs, IVA can take either second-order or all-order statistics into account [Ada14]. In the two following sections,

we present two IVA methods, of which one takes second-order statistics into account and one takes all-order statistics into account.

Four algorithms for minimizing the IVA optimization function in (3.1) are described in [And12], of which the gradient-based algorithm is the easiest to understand, and the Newton-based algorithm is reported to be most computationally efficient. Therefore, in this thesis, we explain IVA using the gradient update and the Newton SCV update.

The gradient of (3.1) with respect to (w.r.t.) $\mathbf{w}_n^{[k]} \in \mathbb{R}^N$ is [And12]

$$\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}} = \mathbb{E}\left\{\mathbf{x}^{[k]}\phi_n^{[k]}\right\} - \left(\mathbf{W}^{[k]}\right)^{-1}(:,n), \tag{3.2}$$

where $\left(\mathbf{W}^{[k]}\right)^{-1}(:,n)$ is the n^{th} column of $\left(\mathbf{W}^{[k]}\right)^{-1}$, and the score function

$$\phi_n^{[k]} = -\frac{\partial \ln\left(p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n)\right)}{\partial \widehat{s}_n^{[k]}} \tag{3.3}$$

denotes the negative derivative of the n^{th} SCV's pdf by the k^{th} element in the n^{th} SCV. The computation of the expected value depends on $p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n)$ and is described later for specific pdfs. Then, the IVA gradient update is just

$$\mathbf{w}_n^{[k]} \leftarrow \mathbf{w}_n^{[k]} - \mu \frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}},$$

where $0 < \mu \leq 1$ is the step size. Finally, each demixing vector $\mathbf{w}_n^{[k]}$ must be made unit-norm by

$$\mathbf{w}_n^{[k]} \leftarrow \frac{\mathbf{w}_n^{[k]}}{\left\|\mathbf{w}_n^{[k]}\right\|}, \quad k = 1, \ldots, K. \tag{3.4}$$

For the Newton SCV update, the gradient and the Hessian matrix of the IVA optimization function in (3.1) need to be known. Let

$$\mathbf{w}_n = \begin{bmatrix} \mathbf{w}_n^{[1]} \\ \vdots \\ \mathbf{w}_n^{[K]} \end{bmatrix} \in \mathbb{R}^{NK}$$

contain the concatenated demixing vectors of the n^{th} estimated SCV

$$\widehat{\mathbf{s}}_n = \begin{bmatrix} \left(\mathbf{w}_n^{[1]}\right)^{\mathsf{T}} \mathbf{x}^{[1]} \\ \vdots \\ \left(\mathbf{w}_n^{[K]}\right)^{\mathsf{T}} \mathbf{x}^{[K]} \end{bmatrix}$$

The gradient of (3.1) w.r.t. $\mathbf{w}_n$ is given as [And12]

$$\mathbf{g}_n = \frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n} = \begin{bmatrix} \frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[1]}} \\ \vdots \\ \frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[K]}} \end{bmatrix},$$

where $\frac{\partial \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]}}$ is given in (3.2). The Hessian matrix of (3.1) is [And12]

$$\mathbf{H}_n = \frac{\partial^2 \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}} = \begin{bmatrix} \mathbf{H}_n^{[1,1]} & \cdots & \mathbf{H}_n^{[1,K]} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_n^{[K,1]} & \cdots & \mathbf{H}_n^{[K,K]} \end{bmatrix} \in \mathbb{R}^{NK \times NK},$$

where the $(k,l)^{\text{th}}$ matrix block $\mathbf{H}_n^{[k,l]} \in \mathbb{R}^{N \times N}$ is

$$\begin{aligned} \mathbf{H}_n^{[k,l]} &= \frac{\partial^2 \mathcal{I}_{\text{IVA}}}{\partial \mathbf{w}_n^{[k]} \partial \left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}} \\ &= \mathbb{E}\left\{ \mathbf{x}^{[k]} \frac{\partial \phi_n^{[k]}}{\partial \left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}} \right\} + \begin{cases} \left(\mathbf{W}^{[k]}\right)^{-1}(:,n)\left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}}(n,:), & k = l \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{3.5}$$

Then, the Newton SCV update of $\mathbf{w}_n$ is [And12]

$$\mathbf{w}_n \leftarrow \mathbf{w}_n - \mu \mathbf{H}_n^{-1} \mathbf{g}_n,$$

and then $\mathbf{w}_n^{[k]}$ are again made unit-norm as described in (3.4).

3.2.1 IVA-G

IVA with a Gaussian model (IVA-G) [And10] assumes that the SCVs follow the multivariate Gaussian distribution, i.e., that the pdf of the n^{th} estimated SCV is given

by

$$p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n) = \frac{1}{\sqrt{(2\pi)^K \det(\mathbf{C}_{\mathbf{s}_n})}} \exp\left(-\frac{1}{2}\widehat{\mathbf{s}}_n^{\mathsf{T}} \mathbf{C}_{\mathbf{s}_n}^{-1} \widehat{\mathbf{s}}_n\right), \quad n = 1, \ldots, N, \tag{3.6}$$

where

$$\mathbf{C}_{\mathbf{s}_n} = \mathbb{E}\left\{\mathbf{s}_n \mathbf{s}_n^{\mathsf{T}}\right\} \in \mathbb{R}^{K \times K}$$

is the (unknown) covariance matrix of the n^{th} SCV. Then, the differential entropy of the n^{th} estimated SCV is [And12]

$$\mathcal{H}(\widehat{\mathbf{s}}_n) = \frac{K \ln(2\pi e)}{2} + \frac{1}{2} \ln\left(\det(\mathbf{C}_{\mathbf{s}_n})\right). \tag{3.7}$$

Inserting (3.7) in (3.1) gives the IVA-G optimization function that is to be minimized [And10],

$$\begin{aligned} \mathcal{I}_{\text{IVA-G}}\left(\mathbf{W}^{[1]}; \ldots; \mathbf{W}^{[K]}\right) = \frac{NK \ln(2\pi e)}{2} + \frac{1}{2} \sum_{n=1}^{N} \ln\left(\det\left(\mathbf{C}_{\mathbf{s}_n}\right)\right) \\ - \sum_{k=1}^{K} \ln\left(\left|\det\left(\mathbf{W}^{[k]}\right)\right|\right) - C. \end{aligned} \tag{3.8}$$

In the following, we describe the IVA-G Newton SCV update. The score function defined in (3.3) for the Gaussian pdf in (3.6) is [And12]

$$\phi_n^{[k]} = \widehat{\mathbf{s}}_n^{\mathsf{T}} \mathbf{C}_{\mathbf{s}_n}^{-1}(:, k) = \mathbf{C}_{\mathbf{s}_n}^{-1}(k, :)\widehat{\mathbf{s}}_n, \tag{3.9}$$

where $\mathbf{C}_{\mathbf{s}_n}^{-1}(:, k)$ is the k^{th} column of $\mathbf{C}_{\mathbf{s}_n}^{-1}$, and $\mathbf{C}_{\mathbf{s}_n}^{-1}(k, :)$ is the k^{th} row of $\mathbf{C}_{\mathbf{s}_n}^{-1}$. Inserting (3.9) in the expected value in (3.2) gives [And12]

$$\begin{aligned} \mathbb{E}\left\{\mathbf{x}^{[k]} \phi_n^{[k]}\right\} &= \mathbb{E}\left\{\mathbf{x}^{[k]} \widehat{\mathbf{s}}_n^{\mathsf{T}} \mathbf{C}_{\mathbf{s}_n}^{-1}(:, k)\right\} = \mathbb{E}\left\{\mathbf{x}^{[k]} \begin{bmatrix} \widehat{s}_n^{[1]} & \cdots & \widehat{s}_n^{[K]} \end{bmatrix}\right\} \mathbf{C}_{\mathbf{s}_n}^{-1}(:, k) \\ &= \begin{bmatrix} \mathbf{C}_{\mathbf{x}}^{[k,1]} \mathbf{w}_n^{[1]} & \cdots & \mathbf{C}_{\mathbf{x}}^{[k,K]} \mathbf{w}_n^{[K]} \end{bmatrix} \mathbf{C}_{\mathbf{s}_n}^{-1}(:, k), \end{aligned}$$

where we have used that $\widehat{s}_n^{[k]} = \left(\mathbf{x}^{[k]}\right)^{\mathsf{T}} \mathbf{w}_n^{[k]}$, and

$$\mathbf{C}_{\mathbf{x}}^{[k,l]} = \mathbb{E}\left\{\mathbf{x}^{[k]} \left(\mathbf{x}^{[l]}\right)^{\mathsf{T}}\right\}$$

is the cross-covariance matrix of the k^{th} and l^{th} dataset. Then, the IVA-G gradient

w.r.t. $\mathbf{w}_n^{[k]}$ is [And12]

$$\frac{\partial \mathcal{I}_{\text{IVA-G}}}{\partial \mathbf{w}_n^{[k]}} = \left[\mathbf{C}_{\mathbf{x}}^{[k,1]}\mathbf{w}_n^{[1]} \quad \cdots \quad \mathbf{C}_{\mathbf{x}}^{[k,K]}\mathbf{w}_n^{[K]}\right] \mathbf{C}_{\mathbf{s}_n}^{-1}(:,k) - \left(\mathbf{W}^{[k]}\right)^{-1}(:,n).$$

Likewise, the expected value in (3.5) can be evaluated as [And12]

$$\begin{aligned}
\mathbb{E}\left\{\mathbf{x}^{[k]}\frac{\partial \phi_n^{[k]}}{\partial\left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}}\right\} &= \mathbb{E}\left\{\mathbf{x}^{[k]}\mathbf{C}_{\mathbf{s}_n}^{-1}(k,:)\frac{\partial \widehat{\mathbf{s}}n}{\partial\left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}}\right\} \\
&= \mathbb{E}\left\{\mathbf{x}^{[k]}\left[\mathbf{C}_{\mathbf{s}_n}^{-1}(k,1) \quad \cdots \quad \mathbf{C}_{\mathbf{s}_n}^{-1}(k,K)\right]\begin{bmatrix}\mathbf{0}^{\mathsf{T}}\\ \left(\mathbf{x}^{[l]}\right)^{\mathsf{T}}\\ \mathbf{0}^{\mathsf{T}}\end{bmatrix}\right\} \\
&= \mathbb{E}\left\{\mathbf{x}^{[k]}\mathbf{C}_{\mathbf{s}_n}^{-1}(k,l)\left(\mathbf{x}^{[l]}\right)^{\mathsf{T}}\right\} \\
&= \mathbf{C}_{\mathbf{s}_n}^{-1}(k,l)\mathbf{C}_{\mathbf{x}}^{[k,l]},
\end{aligned}$$

and the $(k,l)^{\text{th}}$ matrix block of the IVA-G Hessian matrix is

$$\mathbf{H}_n^{[k,l]} = \mathbf{C}_{\mathbf{s}_n}^{-1}(k,l)\mathbf{C}_{\mathbf{x}}^{[k,l]} + \begin{cases}\left(\mathbf{W}^{[k]}\right)^{-1}(:,n)\left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}}(n,:), & k=l\\ 0, & \text{else.}\end{cases}$$

For observed data, $\mathbf{C}_{\mathbf{x}}^{[k,l]}$ is estimated as

$$\widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]} = \frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[l]}\right)^{\mathsf{T}},$$

and $\mathbf{C}_{\mathbf{s}_n}$ is estimated as

$$\widehat{\mathbf{C}}_{\mathbf{s}_n} = \frac{1}{V}\widehat{\mathbf{S}}_n\widehat{\mathbf{S}}_n^{\mathsf{T}},$$

and the $(k,l)^{\text{th}}$ element of $\widehat{\mathbf{C}}_{\mathbf{s}_n}$ can be efficiently estimated as

$$\widehat{\mathbf{C}}_{\mathbf{s}_n}(k,l) = \left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}}\widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]}\mathbf{w}_n^{[l]}.$$

In practice, it is recommended to whiten the observed data $\mathbf{X}^{[k]} \in \mathbb{R}^{N\times V}$ before applying IVA-G. The pseudocode for IVA-G using the Newton SCV update is given in Algorithms 2–4, where a typical value for ϵ is $1e-06$. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2K^2V + IN^4K^4)$, where I is the number of iterations. It is dominated

Algorithm 2 IVA-G using Newton SCV Update [And12]

Input: $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \ldots, K$

1: **for** $k = 1, \ldots, K$ **do** ▷ recommended: whiten observed datasets

2: $\quad \mathbf{X}^{[k]} \leftarrow \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-1/2} \mathbf{X}^{[k]}$

3: **for** $k = 1, \ldots, K$ **do** ▷ cross-covariance matrix of k^{th} and l^{th} observed dataset

4: $\quad$ **for** $l = 1, \ldots, K$ **do**

5: $\quad\quad \widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]} \leftarrow \frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[l]} \right)^{\mathsf{T}}$

6: **for** $k = 1, \ldots, K$ **do** ▷ initialize an orthogonal demixing matrix

7: $\quad \mathbf{W}^{[k]} \leftarrow \text{randn}(N, N)$

8: $\quad \mathbf{W}^{[k]} \leftarrow \left(\mathbf{W}^{[k]} \left(\mathbf{W}^{[k]} \right)^{\mathsf{T}} \right)^{-1/2} \mathbf{W}^{[k]}$

9: **for** $i = 1, \ldots, \text{maxIter}$ **do**

10: $\quad \mathbf{W}_{\text{old}}^{[k]} \leftarrow \mathbf{W}^{[k]}$

11: $\quad$ **for** $n = 1, \ldots, N$ **do**

12: $\quad\quad$ **for** $k = 1, \ldots, K$ **do** ▷ covariance matrix of the n^{th} estimated SCV

13: $\quad\quad\quad$ **for** $l = 1, \ldots, K$ **do**

14: $\quad\quad\quad\quad \widehat{\mathbf{C}}_{\mathbf{s}_n}(k, l) \leftarrow \left(\mathbf{w}_n^{[k]} \right)^{\mathsf{T}} \widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]} \mathbf{w}_n^{[l]}$ ▷ $\left(\mathbf{w}_n^{[k]} \right)^{\mathsf{T}}$ is the n^{th} row of $\mathbf{W}^{[k]}$

15: $\quad\quad \mathbf{g}_n \leftarrow \text{calculate\_gradient}(\widehat{\mathbf{C}}_{\mathbf{s}_n}, \widehat{\mathbf{C}}_{\mathbf{x}}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]})$

16: $\quad\quad \mathbf{H}_n \leftarrow \text{calculate\_hessian}(\widehat{\mathbf{C}}_{\mathbf{s}_n}, \widehat{\mathbf{C}}_{\mathbf{x}}, \mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]})$

17: $\quad\quad \mathbf{w}_n \leftarrow \begin{bmatrix} \mathbf{w}_n^{[1]} \\ \vdots \\ \mathbf{w}_n^{[K]} \end{bmatrix}$

18: $\quad\quad \mathbf{w}_n \leftarrow \mathbf{w}_n - \mu \mathbf{H}_n^{-1} \mathbf{g}_n$ ▷ update the demixing vectors for the n^{th} SCV

19: $\quad\quad$ **for** $k = 1, \ldots, K$ **do**

20: $\quad\quad\quad \mathbf{w}_n^{[k]} \leftarrow \frac{\mathbf{w}_n^{[k]}}{\left\| \mathbf{w}_n^{[k]} \right\|}$ ▷ normalize $\mathbf{w}_n^{[k]}$

21: $\quad$ **for** $n = 1, \ldots, N$ **do**

22: $\quad\quad$ **for** $k = 1, \ldots, K$ **do**

23: $\quad\quad\quad \psi_n^{[k]} \leftarrow 1 - \left| \left(\mathbf{w}_{\text{old},n}^{[k]} \right)^{\mathsf{T}} \mathbf{w}_n^{[k]} \right|$

24: $\quad$ **if** $\max_{n,k} \psi_n^{[k]} < \epsilon$ **then** ▷ IVA-G converged

25: $\quad\quad$ **break**

26: **for** $k = 1, \ldots, K$ **do** ▷ estimate sources

27: $\quad \widehat{\mathbf{S}}^{[k]} \leftarrow \mathbf{W}^{[k]} \mathbf{X}^{[k]}$

Output: $\mathbf{W}^{[k]} \in \mathbb{R}^{N \times N}, \widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{N \times V}, k = 1, \ldots, K$

by line 18 in Algorithm 2. In order to achieve a more efficient implementation, the

inversion of $\mathbf{W}^{[k]}$ in line 2 in Algorithm 3 and line 5 in Algorithm 4 can be replaced using a decoupling trick as described in [And10].

Algorithm 3 calculate_gradient

Input: $\widehat{\mathbf{C}}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$, $\widehat{\mathbf{C}}_{\mathbf{x}} \in \mathbb{R}^{NK\times NK}$, $\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]} \in \mathbb{R}^{N\times N}$

1: **for** $k = 1, \ldots, K$ **do**

2: $\quad \mathbf{g}_n^{[k]} \leftarrow \sum_{l=1}^{K} \widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]} \mathbf{w}_n^{[l]} \widehat{\mathbf{C}}_{\mathbf{s}_n}^{-1}(l,k) - \left(\mathbf{W}^{[k]}\right)^{-1}(:,n)$ ▷ $\frac{\partial \mathcal{I}_{\text{IVA-G}}}{\partial \mathbf{w}_n^{[k]}}$

3: $\mathbf{g}_n \leftarrow \begin{bmatrix} \mathbf{g}_n^{[1]} \\ \vdots \\ \mathbf{g}_n^{[K]} \end{bmatrix}$ ▷ $\frac{\partial \mathcal{I}_{\text{IVA-G}}}{\partial \mathbf{w}_n}$

Output: $\mathbf{g}_n \in \mathbb{R}^{NK}$

Algorithm 4 calculate_hessian

Input: $\widehat{\mathbf{C}}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$, $\widehat{\mathbf{C}}_{\mathbf{x}} \in \mathbb{R}^{NK\times NK}$, $\mathbf{W}^{[1]}, \ldots, \mathbf{W}^{[K]} \in \mathbb{R}^{N\times N}$

1: **for** $k = 1, \ldots, K$ **do**

2: $\quad$ **for** $l = 1, \ldots, K$ **do**

3: $\quad\quad \mathbf{H}_n^{[k,l]} \leftarrow \widehat{\mathbf{C}}_{\mathbf{s}_n}^{-1}(k,l)\widehat{\mathbf{C}}_{\mathbf{x}}^{[k,l]}$ ▷ $\frac{\partial^2 \mathcal{I}_{\text{IVA-G}}}{\partial \mathbf{w}_n^{[k]} \partial \left(\mathbf{w}_n^{[l]}\right)^{\mathsf{T}}}$

4: $\quad\quad$ **if** $k == l$ **then**

5: $\quad\quad\quad \mathbf{H}_n^{[k,k]} \leftarrow \mathbf{H}_n^{[k,k]} + \left(\mathbf{W}^{[k]}\right)^{-1}(:,n)\left(\mathbf{W}^{[k]}\right)^{-\mathsf{T}}(n,:)$

6: $\mathbf{H}_n \leftarrow \begin{bmatrix} \mathbf{H}_n^{[1,1]} & \cdots & \mathbf{H}_n^{[1,K]} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_n^{[K,1]} & \cdots & \mathbf{H}_n^{[K,K]} \end{bmatrix}$ ▷ $\frac{\partial^2 \mathcal{I}_{\text{IVA-G}}}{\partial \mathbf{w}_n \partial \mathbf{w}_n^{\mathsf{T}}}$

Output: $\mathbf{H}_n \in \mathbb{R}^{NK\times NK}$

As a small note, we would like to add that even though this method is called IVA-G in the literature, a more suitable name would be Uncorrelated Vector Analysis (UVA), as in fact it just makes the estimated SCVs maximally uncorrelated instead of maximally independent.

3.2.2 IVA-L-SOS

IVA with a Laplacian model and second-order statistics (IVA-L-SOS) [Bhi19] assumes that the SCVs follow a multivariate Laplacian distribution that allows for a non-identity covariance matrix, and therefore, dependence is measured by second-order statistics (correlations) and higher-order statistics. The multivariate Laplacian distribution can

be expressed by the Multivariate Generalized Gaussian Distribution (MGGD) [Góm98] with shape parameter $\beta = 0.5$, and the pdf of the n^{th} estimated SCV is

$$p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n) = \frac{\Gamma\left(\frac{K}{2}\right)\left(\det(\boldsymbol{\Sigma}_n)\right)^{-\frac{1}{2}}}{2^{K+1}\pi^{K/2}\Gamma(K)} \exp\left(-\frac{1}{2}\sqrt{\widehat{\mathbf{s}}_n^{\mathsf{T}}\boldsymbol{\Sigma}_n^{-1}\widehat{\mathbf{s}}_n}\right), \tag{3.10}$$

where $\Gamma(\cdot)$ is the Gamma function, and $\boldsymbol{\Sigma}_n \in \mathbb{R}^{K\times K}$ is a positive definite scatter matrix that accounts for second-order statistics [Bhi19]. The relationship of $\boldsymbol{\Sigma}_n$ and the n^{th} SCV's covariance matrix $\mathbf{C}_{\mathbf{s}_n}$ is [Góm98]

$$\mathbf{C}_{\mathbf{s}_n} = \frac{4\Gamma(K+2)}{K\Gamma(K)}\boldsymbol{\Sigma}_n = \frac{4(K+1)\Gamma(K+1)}{K\Gamma(K)}\boldsymbol{\Sigma}_n = \frac{4(K+1)K\Gamma(K)}{K\Gamma(K)}\boldsymbol{\Sigma}_n = 4(K+1)\boldsymbol{\Sigma}_n.$$

Therefore, the score function of (3.10) is [And13a]

$$\phi_n^{[k]} = \frac{1}{2}\frac{\boldsymbol{\Sigma}_n^{-1}(k,:)}{\sqrt{\widehat{\mathbf{s}}_n^{\mathsf{T}}\boldsymbol{\Sigma}_n^{-1}\widehat{\mathbf{s}}_n}} = \frac{\sqrt{K+1}\mathbf{C}_{\mathbf{s}_n}^{-1}(k,:)\widehat{\mathbf{s}}_n}{\sqrt{\widehat{\mathbf{s}}_n^{\mathsf{T}}\mathbf{C}_{\mathbf{s}_n}^{-1}\widehat{\mathbf{s}}_n}}, \tag{3.11}$$

and inserting (3.11) in (3.2) gives the IVA-L-SOS gradient w.r.t. $\mathbf{w}_n^{[k]}$:

$$\frac{\partial \mathcal{I}_{\text{IVA-L-SOS}}}{\partial \mathbf{w}_n^{[k]}} = \mathbb{E}\left\{\mathbf{x}^{[k]}\frac{\sqrt{K+1}\mathbf{C}_{\mathbf{s}_n}^{-1}(k,:)\widehat{\mathbf{s}}_n}{\sqrt{\widehat{\mathbf{s}}_n^{\mathsf{T}}\mathbf{C}_{\mathbf{s}_n}^{-1}\widehat{\mathbf{s}}_n}}\right\} - \left(\mathbf{W}^{[k]}\right)^{-1}(:,n).$$

IVA-L-SOS is optimized using the gradient update, i.e,.

$$\mathbf{w}_n^{[k]} \leftarrow \mathbf{w}_n^{[k]} - \mu\frac{\partial \mathcal{I}_{\text{IVA-L-SOS}}}{\partial \mathbf{w}_n^{[k]}},$$

and then demixing vectors are again made unit-norm as in (3.4).

For observed data, $\mathbf{C}_{\mathbf{s}_n}$ is estimated as

$$\widehat{\mathbf{C}}_{\mathbf{s}_n} = \frac{1}{V}\widehat{\mathbf{S}}_n\widehat{\mathbf{S}}_n^{\mathsf{T}}, \tag{3.12}$$

and the gradient is

$$\frac{\partial \mathcal{I}_{\text{IVA-L-SOS}}}{\partial \mathbf{w}_n^{[k]}} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{X}^{[k]}(:,v)\frac{\sqrt{K+1}\;\widehat{\mathbf{C}}_{\mathbf{s}_n}^{-1}(k,:)\widehat{\mathbf{S}}_n(:,v)}{\sqrt{\left(\widehat{\mathbf{S}}_n(:,v)\right)^{\mathsf{T}}\widehat{\mathbf{C}}_{\mathbf{s}_n}^{-1}\widehat{\mathbf{S}}_n(:,v)}} - \left(\mathbf{W}^{[k]}\right)^{-1}(:,n).$$

3.3 Canonical Correlation Analysis

Compared with ICA and IVA, Canonical Correlation Analysis (CCA) does not assume a generative model on the datasets. Instead, CCA linearly transforms two datasets $\mathbf{x}^{[1]} \in \mathbb{R}^N$ and $\mathbf{x}^{[2]} \in \mathbb{R}^N$ into a new space where the correlation of the transformed variables, called *canonical variables*, is maximized. Interestingly, it has been shown that under specific conditions, which are presented in Section 5.3, CCA is able to achieve JBSS.

The n^{th} canonical variables are defined as

$$u_n^{[1]} = \left(\mathbf{t}_n^{[1]}\right)^{\mathsf{T}} \mathbf{x}^{[1]} \quad \text{and} \quad u_n^{[2]} = \left(\mathbf{t}_n^{[2]}\right)^{\mathsf{T}} \mathbf{x}^{[2]}, \quad n = 1, \ldots, N,$$

where $\mathbf{t}_n^{[1]} \in \mathbb{R}^N$ and $\mathbf{t}_n^{[2]} \in \mathbb{R}^N$ are unknown transformation vectors. W.l.o.g., the datasets are assumed to have zero mean, i.e.,

$$\mathbb{E}\left\{\mathbf{x}^{[1]}\right\} = \mathbb{E}\left\{\mathbf{x}^{[2]}\right\} = \mathbf{0}_N,$$

which then also holds for the canonical variables, i.e.,

$$\mathbb{E}\left\{u_n^{[1]}\right\} = \mathbb{E}\left\{u_n^{[2]}\right\} = 0, \quad n = 1, \ldots, N. \tag{3.13}$$

As correlation and covariance of the n^{th} canonical variables coincide due to (3.13), the optimization function for finding the n^{th} canonical variables $u_1^{[1]}$ and $u_1^{[2]}$ therefore is

$$\max_{\mathbf{t}_n^{[1]}, \mathbf{t}_n^{[2]}} \mathbb{E}\left\{u_n^{[1]} u_n^{[2]}\right\}, \tag{3.14}$$

where

$$\kappa_n = \mathbb{E}\left\{u_n^{[1]} u_n^{[2]}\right\} = \left(\mathbf{t}_n^{[1]}\right)^{\mathsf{T}} \mathbb{E}\left\{\mathbf{x}^{[1]} \left(\mathbf{x}^{[2]}\right)^{\mathsf{T}}\right\} \mathbf{t}_n^{[2]} = \left(\mathbf{t}_n^{[1]}\right)^{\mathsf{T}} \mathbf{C}_{\mathbf{x}}^{[1,2]} \mathbf{t}_n^{[2]}, \quad n = 1, \ldots, N,$$

is called the canonical correlation of the n^{th} canonical variables. To avoid maximizing (3.14) simply by scaling the transformation vectors, the canonical variables are constrained to have unit variance, i.e.,

$$\mathbb{E}\left\{\left(u_n^{[1]}\right)^2\right\} = \mathbb{E}\left\{\left(u_n^{[2]}\right)^2\right\} = 1, \quad n = 1, \ldots, N, \tag{3.15}$$

where

$$\mathbb{E}\left\{\left(u_n^{[k]}\right)^2\right\} = \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbb{E}\left\{\mathbf{x}^{[k]}\left(\mathbf{x}^{[k]}\right)^{\mathsf{T}}\right\} \mathbf{t}_n^{[k]} = \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{C}_{\mathbf{x}}^{[k]} \mathbf{t}_n^{[k]}, \quad k = 1, 2.$$

After the first canonical variables are found, an additional constraint needs to be imposed on the next canonical variables so that they do not estimate the same ones as before. Thus, canonical variables are constrained to be uncorrelated within a dataset [Hot36], i.e.,

$$\mathbb{E}\left\{u_n^{[1]} u_m^{[1]}\right\} = \mathbb{E}\left\{u_n^{[2]} u_m^{[2]}\right\} = 0, \quad n \neq m, \quad n, m = 1, \ldots, N, \tag{3.16}$$

where

$$\mathbb{E}\left\{u_n^{[k]} u_m^{[k]}\right\} = \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{C}_{\mathbf{x}}^{[k]} \mathbf{t}_m^{[k]}, \quad k = 1, 2.$$

Finally, the N canonical variables of each dataset form the canonical vectors

$$\mathbf{u}^{[k]} = \begin{bmatrix} u_1^{[k]} \\ \vdots \\ u_N^{[k]} \end{bmatrix}, \quad k = 1, 2.$$

They are found as

$$\mathbf{u}^{[k]} = \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}, \quad k = 1, 2.$$

where

$$\mathbf{T}^{[k]} = \begin{bmatrix} \mathbf{t}_1^{[k]} & \cdots & \mathbf{t}_N^{[k]} \end{bmatrix}, \quad k = 1, 2,$$

is the transformation matrix for the k^{th} dataset, $k = 1, 2$. By performing an SVD on the coherence matrix [Sch08]

$$\widetilde{\mathbf{C}}_{\mathbf{x}}^{[1,2]} = \left(\mathbf{C}_{\mathbf{x}}^{[1,2]}\right)^{-\frac{1}{2}} \mathbf{C}_{\mathbf{x}}^{[1,2]} \left(\mathbf{C}_{\mathbf{x}}^{[2,2]}\right)^{-\frac{1}{2}} = \mathbf{F}^{[1]} \mathbf{K} \left(\mathbf{F}^{[2]}\right)^{\mathsf{T}},$$

where the diagonal elements of $\mathbf{K} = \operatorname{diag}(\kappa_1, \ldots, \kappa_N)$ turn out to be the canonical correlations sorted that $\kappa_1 \geq \cdots \geq \kappa_N$, the transformation matrices $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}$ are found as [Sch08]

$$\mathbf{T}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \mathbf{F}^{[k]}, \quad k = 1, 2,$$

and the canonical vectors are

$$\mathbf{u}^{[k]} = \left(\mathbf{F}^{[k]}\right)^{\mathsf{T}} \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \mathbf{x}^{[k]}, \quad k = 1, 2.$$

It is interesting to see that this solution can be interpreted as first whitening the datasets

using Mahalanobis whitening [Kes18], i.e, $\mathbf{y}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \mathbf{x}^{[k]}$, $k = 1, 2$, and then finding the transformation vectors as the singular vectors of the cross-covariance matrix of the whitened datasets, $\mathbf{C}_{\mathbf{y}}^{[1,2]} = \mathbb{E}\left\{\mathbf{y}^{[1]} \left(\mathbf{y}^{[2]}\right)^{\mathsf{T}}\right\}$.

The pseudocode for CCA using the SVD is given in Algorithm 5. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2 V)$. It is dominated by lines 2 and 3.

Algorithm 5 CCA [Hot36]

Input: $\mathbf{X}^{[1]}, \mathbf{X}^{[2]} \in \mathbb{R}^{N \times V}$

1: **for** $k = 1, 2$ **do**
2: $\quad \mathbf{C}_{\mathbf{x}}^{[k]} \leftarrow \frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}$
3: $\mathbf{C}_{\mathbf{x}}^{[1,2]} \leftarrow \frac{1}{V} \mathbf{X}^{[1]} \left(\mathbf{X}^{[2]}\right)^{\mathsf{T}}$
4: $\widetilde{\mathbf{C}}_{\mathbf{x}}^{[1,2]} \leftarrow \left(\mathbf{C}_{\mathbf{x}}^{[1]}\right)^{-1/2} \mathbf{C}_{\mathbf{x}}^{[1,2]} \left(\mathbf{C}_{\mathbf{x}}^{[2]}\right)^{-1/2}$ ▷ coherence matrix
5: $\mathbf{F}^{[1]}, \mathbf{K}, \mathbf{F}^{[2]} \leftarrow \mathrm{SVD}(\mathbf{C})$
6: **for** $k = 1, 2$ **do**
7: $\quad \mathbf{T}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-1/2} \mathbf{F}^{[k]}$ ▷ transformation matrix
8: **for** $k = 1, 2$ **do**
9: $\quad \mathbf{U}^{[k]} \leftarrow \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]}$ ▷ canonical vectors

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}$, $\mathbf{U}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, 2$

3.4 Multiset Canonical Correlation Analysis

Besides the wide use of multiset Canonical Correlation Analysis (mCCA), there is still a lack of comprehensive understanding of its theory and implementation with different objective functions all under one umbrella. Therefore, in this section, we present the optimization problems of the five mCCA methods sumcor, maxvar, minvar, ssqcor, and genvar and provide a clear and concise overview of their solutions and underlying factor models.[2]

[2]This section is based on the paper: "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, *submitted for review*, 2025. I specifically contributed to proving the source identification conditions of sumcor, deriving the connection between maxvar [Ket71] and sumcor [Nie95], implementing and running the experiments, creating all figures, and writing the paper. B. Gabrielson has provided theoretical insights in the connection between maxvar [Car68] and sumcor [Nie95] and the source identification conditions of genvar. Discussions and feedback along the way until the final paper have contributions from all authors.

MCCA extends CCA to $K \geq 2$ datasets and aims to estimate canonical variables [Ket71]

$$u_n^{[k]} = \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}, \quad n = 1, \ldots, N, \quad k = 1, \ldots, K,$$

such that the n^{th} canonical variables $u_n^{[1]}, \ldots, u_n^{[K]}$ are maximally correlated for each $n = 1, \ldots, N$. Here, $u_n^{[k]}$ denotes the n^{th} canonical variable in the k^{th} dataset, and $\mathbf{t}_n^{[k]} \in \mathbb{R}^N$ is the corresponding transformation vector. Also mCCA is able to achieve JBSS under specific conditions, which we will see later in Section 5.4.

The datasets are again assumed to have zero mean, i.e.,

$$\mathbb{E}\left\{\mathbf{x}^{[k]}\right\} = \mathbf{0}_N, \quad k = 1, \ldots, K,$$

and therefore also the canonical variables, i.e.,

$$\mathbb{E}\left\{u_n^{[k]}\right\} = 0, \quad n = 1, \ldots, N, \quad k = 1, \ldots, K. \tag{3.17}$$

Maximizing correlation across multiple random variables requires a measure of this correlation. Such a measure is not uniquely defined; however, there are different ways of empirically measuring the correlation of more than two random variables. One intuitive measure is the sum of the pairwise covariances $\mathbb{E}\left\{u_n^{[k]} u_n^{[l]}\right\}$ among the n^{th} canonical variables from all K datasets, as covariances and correlations coincide due to the canonical variables being zero-mean. This leads to the following optimization problem to find the first canonical variable of all datasets, $u_1^{[k]}$, $k = 1, \ldots, K$:

$$\max_{\mathbf{t}_1^{[1]}, \ldots, \mathbf{t}_1^{[K]}} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbb{E}\left\{u_1^{[k]} u_1^{[l]}\right\}. \tag{3.18}$$

Later we discuss an additional constraint needed on the $u_n^{[k]}$ that, like in CCA, is necessary to avoid maximizing the objective function simply by scaling the transformation vectors $\mathbf{t}_n^{[k]}$, which we call the "trivial solution". To simplify the notation in the following, let all K canonical variables with index n be stacked in the n^{th} Canonical Component Vector (CCV)

$$\mathbf{u}_n = \begin{bmatrix} u_n^{[1]} \\ \vdots \\ u_n^{[K]} \end{bmatrix} \in \mathbb{R}^K, \quad n = 1, \ldots, N.$$

This is similar to the definition of SCVs visualized in Figure 3.1, where now the same color in different transparencies represents correlation, i.e., linear dependence. Now, for the next CCV, the sum of correlations between the elements in $\mathbf{u}_2$ is maximized. Also here an additional constraint is needed in the optimization problem, as otherwise the same CCV would be estimated as before. We will later introduce possible options for such a constraint. Finally, after estimating the N CCVs, the N canonical variables of the k^{th} dataset form the canonical vector

$$\mathbf{u}^{[k]} = \begin{bmatrix} u_1^{[k]} \\ \vdots \\ u_N^{[k]} \end{bmatrix} \in \mathbb{R}^N,$$

which is found as

$$\mathbf{u}^{[k]} = \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{x}^{[k]}, \quad k = 1, \ldots, K,$$

with transformation matrix

$$\mathbf{T}^{[k]} = \begin{bmatrix} \mathbf{t}_1^{[k]} & \cdots & \mathbf{t}_N^{[k]} \end{bmatrix} \in \mathbb{R}^{N \times N}.$$

As noted before, there is more than one way to measure correlation across multiple random variables, leading to several mCCA optimization problems. Let

$$\mathbf{C}_{\mathbf{u}_n} = \mathbb{E}\left\{\mathbf{u}_n \mathbf{u}_n^{\mathsf{T}}\right\} \in \mathbb{R}^{K \times K}, \quad n = 1, \ldots, N,$$

be the covariance matrix of the n^{th} CCV, which consists of the pairwise covariances $\mathbb{E}\left\{u_n^{[k]} u_n^{[l]}\right\}$, $k, l = 1, \ldots, K$, and let

$$\boldsymbol{\ell}_n = \begin{bmatrix} \ell_n^{[1]} \\ \vdots \\ \ell_n^{[K]} \end{bmatrix}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{u}_n}$. W.l.o.g., we assume in this thesis that eigenvalues are sorted in descending order, i.e., $\ell_n^{[1]} \geq \cdots \geq \ell_n^{[K]}$. Kettering [Ket71] summarizes five mCCA objective functions, which all reduce to the CCA objective function (3.14) for $K = 2$ datasets, as a function of $\mathbf{C}_{\mathbf{u}_n}$ for the n^{th} CCV. If all canonical variables have unit variance and are uncorrelated, then $\mathbf{C}_{\mathbf{u}_n} = \mathbf{I}_{K \times K}$. Ultimately, the goal of all of the following mCCA objective functions boils down to making $\mathbf{C}_{\mathbf{u}_n}$ as far away as possible from $\mathbf{I}_{K \times K}$, i.e., increasing correlation within $\mathbf{C}_{\mathbf{u}_n}$.

- mCCA-sumcor [Hor61]:

  Maximize the sum of covariances,

$$\mathcal{J}_{\text{sumcor}}(\mathbf{C}_{\mathbf{u}_n}) = \sum_{k=1}^{K}\sum_{l=1}^{K} c_{u_n}^{[k,l]}, \tag{F1}$$

  where $c_{u_n}^{[k,l]}$ denotes the $(k,l)^{\text{th}}$ entry in $\mathbf{C}_{\mathbf{u}_n}$. This is the expression from (3.18) written in terms of $\mathbf{C}_{\mathbf{u}_n}$ and is the natural extension of CCA to more than two datasets [Ase15].

- mCCA-ssqcor [Ket71]:

  Maximize the sum of squared covariances,

$$\mathcal{J}_{\text{ssqcor}}(\mathbf{C}_{\mathbf{u}_n}) = \sum_{k=1}^{K}\sum_{l=1}^{K} \left(c_{u_n}^{[k,l]}\right)^2 = \sum_{k=1}^{K} \left(\ell_n^{[k]}\right)^2. \tag{F2}$$

  This removes the effect of the sign of the covariances compared with sumcor, such that negative covariances do not cancel out positive ones. The fact that the sum of squared elements in $\mathbf{C}_{\mathbf{u}_n}$ can be expressed in terms of its eigenvalues $\ell_n^{[k]}$ makes it reasonable that also other functions of these eigenvalues can be used in the optimization problem, which are stated in the following.

- mCCA-maxvar [Hor61]:

  Maximize the largest eigenvalue of $\mathbf{C}_{\mathbf{u}_n}$,

$$\mathcal{J}_{\text{maxvar}}(\mathbf{C}_{\mathbf{u}_n}) = \ell_n^{[1]}, \tag{F3}$$

  where $\ell_n^{[1]}$ is the largest eigenvalue of $\mathbf{C}_{\mathbf{u}_n}$. This $\mathbf{C}_{\mathbf{u}_n}$ is the best approximation to a rank-1 matrix, i.e., a matrix in which the covariances between all variables in the CCV $\mathbf{u}_n$ are as close as possible to one.

- mCCA-minvar [Ket71]:

  Minimize the smallest eigenvalue of $\mathbf{C}_{\mathbf{u}_n}$,

$$\mathcal{J}_{\text{minvar}}(\mathbf{C}_{\mathbf{u}_n}) = \ell_n^{[K]}, \tag{F4}$$

  where $\ell_n^{[K]}$ is the smallest eigenvalue of $\mathbf{C}_{\mathbf{u}_n}$. This idea can be viewed as the reverse of maxvar.

- mCCA-genvar [Ste51]:

  Minimize the determinant of $\mathbf{C}_{\mathbf{u}_n}$,

$$\mathcal{J}_{\text{genvar}}(\mathbf{C}_{\mathbf{u}_n}) = \det(\mathbf{C}_{\mathbf{u}_n}) = \prod_{k=1}^{K} \ell_n^{[k]}. \tag{F5}$$

  The determinant of a covariance matrix is referred to as its generalized variance [Wil32], thus the name. This last objective function is similar to ssqcor in being defined over all eigenvalues, but while the sum of squared eigenvalues is heavily influenced by large eigenvalues, the product of eigenvalues places greater emphasis on the small eigenvalues.

As mentioned before, constraints must be added to all the presented optimization problems to avoid the trivial solution. In this thesis, we consider the following two constraints on the canonical variables within the n^{th} CCV $\mathbf{u}_n$:

- $\mathbb{E}\left\{\left(u_n^{[k]}\right)^2\right\} = \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{C}_{\mathbf{x}}^{[k]} \mathbf{t}_n^{[k]} = 1, \quad k = 1, \ldots, K,$ (3.19)

- $\sum_{k=1}^{K} \mathbb{E}\left\{\left(u_n^{[k]}\right)^2\right\} = \sum_{k=1}^{K} \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{C}_{\mathbf{x}}^{[k]} \mathbf{t}_n^{[k]} = 1,$ (3.20)

where

$$\mathbf{C}_{\mathbf{x}}^{[k]} = \mathbb{E}\left\{\mathbf{x}^{[k]} \left(\mathbf{x}^{[k]}\right)^{\mathsf{T}}\right\} \in \mathbb{R}^{N \times N}$$

is the covariance matrix of the k^{th} dataset. We know the first constraint (3.19), which means that each canonical variable has unit variance, already from CCA. Constraint (3.19) is also used by Kettenring [Ket71] and corresponds to the third constraint in Nielsen [Nie95]. The second constraint (3.20) means that the sum over the variances of all K canonical variables within a CCV is one and corresponds to the fourth constraint in [Nie95].

Furthermore, we consider the following two constraints on the canonical variables across CCVs $\mathbf{u}_n$ and $\mathbf{u}_m$:

- $\mathbb{E}\left\{u_n^{[k]} u_m^{[k]}\right\} = 0, \quad n \neq m, \quad n, m = 1, \ldots, N, \quad k = 1, \ldots, K,$ (3.21)

- $\sum_{k=1}^{K} \mathbb{E}\left\{u_n^{[k]} u_m^{[k]}\right\} = 0, \quad n \neq m, \quad n, m = 1, \ldots, N.$ (3.22)

Constraint (3.21), which we again know from CCA, denotes that canonical variables within the same dataset are uncorrelated [Ket71, eq. (9.5)], while (3.22) is a weaker constraint [Nie95, eq. (3.63)].

In the following subsections, we present the solutions of the five mCCA methods.

3.4.1 mCCA-sumcor

For the n^{th} CCV, sumcor aims to find the transformation vectors $\mathbf{t}_n^{[k]}$ that maximize the objective function given in (F1), i.e., the sum of the elements of the n^{th} CCV's covariance matrix $\mathbf{C}_{\mathbf{u}_n}$ [Ket71]. Kettenring proposed a *deflationary* numerical solution for sumcor under (3.19) and (3.21) [Ket71], i.e., a solution where the $\mathbf{t}_n^{[1]}, \ldots, \mathbf{t}_n^{[K]}$ are estimated sequentially for $n = 1, \ldots, N$, using a numerical method for each CCV $\mathbf{u}_n$. Under (3.20) and (3.22), however, sumcor also enjoys an *all-at-once* analytical solution, i.e., a solution where all CCVs are found simultaneously using an analytical method. This method is considerably more straightforward and more efficient regarding runtime and JBSS performance than the deflationary solution. For the remainder of this thesis, when discussing sumcor, we focus on this all-at-once analytical solution, which we present in the following.

All-at-once analytical solution

Nielsen [Nie95] derived the following all-at-once analytical solution for (F1) under (3.20) and (3.22). First, the datasets are concatenated as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{[1]} \\ \vdots \\ \mathbf{x}^{[k]} \end{bmatrix} \in \mathbb{R}^{NK},$$

and the covariance matrix of the concatenated datasets is

$$\mathbf{C}_{\mathbf{x}} = \mathbb{E}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} \in \mathbb{R}^{NK \times NK}.$$

Defining $\mathbf{D}_{\mathbf{x}}$ as the block-diagonal matrix consisting of the covariance matrices of the datasets,

$$\mathbf{D}_{\mathbf{x}} = \begin{bmatrix} \mathbf{C}_{\mathbf{x}}^{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{C}_{\mathbf{x}}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times NK},$$

the transformation vectors $\mathbf{t}_n^{[k]}$ can be found by solving the Generalized EVD (GEVD) problem [Nie95]

$$\mathbf{C}_{\mathbf{x}}\mathbf{T} = \mathbf{D}_{\mathbf{x}}\mathbf{T}\mathbf{\Phi}, \tag{3.23}$$

where $\mathbf{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_{NK}) \in \mathbb{R}^{NK \times NK}$ is a diagonal matrix consisting of the eigenvalues of $\mathbf{C_x}$ w.r.t. $\mathbf{D_x}$, and $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{NK}] \in \mathbb{R}^{NK \times NK}$ is an orthogonal matrix with the corresponding eigenvectors as columns. Let

$$\widetilde{\mathbf{T}} = [\mathbf{t}_1, \ldots, \mathbf{t}_N] \in \mathbb{R}^{NK \times N}$$

be the matrix consisting of the N principal eigenvectors (corresponding to the N largest eigenvalues), then the transformation vectors $\mathbf{t}_n^{[k]} \in \mathbb{R}^N$, $k = 1, \ldots, K$, for the n^{th} CCV are found by partitioning $\mathbf{t}_n$ as

$$\mathbf{t}_n = \begin{bmatrix} \mathbf{t}_n^{[1]} \\ \vdots \\ \mathbf{t}_n^{[K]} \end{bmatrix} \in \mathbb{R}^{NK},$$

and

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{T}^{[1]} \\ \vdots \\ \mathbf{T}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times N} \tag{3.24}$$

contains the transformation matrices for the k^{th} dataset, $\mathbf{T}^{[k]}$. In the end, we normalize the canonical variables to unit-variance and multiply $\mathbf{t}_n^{[k]}$ by the variance of $u_n^{[k]}$, $n = 1, \ldots, N$, $k = 1, \ldots, K$, for comparison with the other methods. The pseudocode for the all-at-once analytical sumcor method is given in Algorithm 6. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2 K^2 V)$, where we have assumed that $V > NK$. It is dominated by line 2.

Underlying factor model

The CCVs estimated by sumcor underlie the model [Ket71]

$$\mathbf{u}_n = \mathbf{1}_K f_n + \boldsymbol{\nu}_n, \quad n = 1, \ldots, N, \tag{3.25}$$

where f_n is the best fitting zero-mean unit-variance common factor present in all $u_n^{[k]}$, $k = 1, \ldots, K$, $\mathbf{1}_K$ is a vector full of ones, and $\boldsymbol{\nu}_n \in \mathbb{R}^K$ is a noise vector. In the JBSS setting, this means that sumcor assumes an *effective rank* of one for the SCVs, i.e., there is only one common factor for all source components $s_n^{[k]}$ within the n^{th} SCV, which is represented by f_n [Gab24].

Algorithm 6 All-at-once analytical sumcor [Nie95]

Input: $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \ldots, K$

$\mathbf{X} \leftarrow \begin{bmatrix} \mathbf{X}^{[1]} \\ \vdots \\ \mathbf{X}^{[K]} \end{bmatrix}$

$\widehat{\mathbf{C}}_{\mathbf{x}} \leftarrow \frac{1}{V}\mathbf{X}\mathbf{X}^{\mathsf{T}}$

$\widehat{\mathbf{D}}_{\mathbf{x}} \leftarrow \text{blkdiag}\left(\frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)$

$\boldsymbol{\phi}, \mathbf{T} \leftarrow \text{GEVD}(\widehat{\mathbf{C}}_{\mathbf{x}}, \widehat{\mathbf{D}}_{\mathbf{x}})$

$\widetilde{\mathbf{T}} \leftarrow \mathbf{T}(:, 1:N)$

for $k = 1, \ldots, K$ **do**

  $\mathbf{T}^{[k]} = \widetilde{\mathbf{T}}\left((k-1)N + 1 : kN, :\right)$

  $\mathbf{U}^{[k]} \leftarrow \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]}$

for $k = 1, \ldots, K$ **do**

  $\mathbf{T}^{[k]}, \mathbf{U}^{[k]} \leftarrow \text{normalize}\left(\mathbf{T}^{[k]}, \mathbf{U}^{[k]}\right)$ ▷ normalize canonical variables

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}, \mathbf{U}^{[k]} \in \mathbb{R}^{N \times V}, k = 1, \ldots, K$

3.4.2 mCCA-maxvar

For the n^{th} CCV, maxvar aims to find the transformation vectors $\mathbf{t}_n^{[k]}$ that maximize the largest eigenvalue of its covariance matrix $\mathbf{C}_{\mathbf{u}_n}$ as denoted in (F3) [Ket71].

Deflationary analytical solution

Kettenring proposes in [Ket71, section 10] the following deflationary solution for maxvar using constraints (3.19) and (3.21), where each CCV $\mathbf{u}_n$ is found analytically as follows. First, each dataset $\mathbf{x}^{[k]} \in \mathbb{R}^N$ is whitened using Mahalanobis whitening [Kes18] to obtain

$$\mathbf{y}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \mathbf{x}^{[k]} \in \mathbb{R}^N, \quad k = 1, \ldots, K. \tag{3.26}$$

The whitened datasets are concatenated into the vector

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}^{[1]} \\ \vdots \\ \mathbf{y}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK},$$

and the covariance matrix of the whitened datasets is

$$\mathbf{C}_{\mathbf{y}} = \mathbb{E}\left\{\mathbf{y}\mathbf{y}^{\mathsf{T}}\right\} \in \mathbb{R}^{NK \times NK}. \tag{3.27}$$

Now, the following EVD is performed sequentially for $n = 1, \ldots, N$:

$$\mathbf{H}_n \mathbf{C}_\mathbf{y} \mathbf{H}_n \mathbf{v}_n = \phi_n \mathbf{v}_n, \tag{3.28}$$

where ϕ_n denotes the largest eigenvalue of $\mathbf{H}_n \mathbf{C}_\mathbf{y} \mathbf{H}_n \in \mathbb{R}^{NK \times NK}$, $\mathbf{v}_n \in \mathbb{R}^{NK}$ the corresponding eigenvector, and the matrix $\mathbf{H}_n$ is described in the following. For estimating the first CCV, $\mathbf{H}_1 = \mathbf{I}_{NK \times NK}$. For the next CCVs, $\mathbf{H}_n \in \mathbb{R}^{NK \times NK}$ is calculated using the eigenvectors from the previous $n-1$ EVDs as follows. First, the eigenvectors are partitioned as

$$\mathbf{v}_i = \begin{bmatrix} \mathbf{v}_i^{[1]} \\ \vdots \\ \mathbf{v}_i^{[K]} \end{bmatrix},$$

and then the normalized eigenvectors of the k^{th} dataset,

$$\widetilde{\mathbf{v}}_i^{[k]} = \frac{\mathbf{v}_i^{[k]}}{\left\| \mathbf{v}_i^{[k]} \right\|}, \quad i = 1, \ldots, n-1,$$

are concatenated in

$$\widetilde{\mathbf{V}}_{(n-1)}^{[k]} = \begin{bmatrix} \widetilde{\mathbf{v}}_1^{[k]} & \cdots & \widetilde{\mathbf{v}}_{n-1}^{[k]} \end{bmatrix} \in \mathbb{R}^{N \times (n-1)}$$

to form the block-diagonal matrix

$$\begin{aligned} \widetilde{\mathbf{V}}_{(n-1)} &= \begin{bmatrix} \widetilde{\mathbf{V}}_{(n-1)}^{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \widetilde{\mathbf{V}}_{(n-1)}^{[K]} \end{bmatrix} \\ &= \begin{bmatrix} \widetilde{\mathbf{v}}_1^{[1]} & \cdots & \widetilde{\mathbf{v}}_{n-1}^{[1]} & & & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \ddots & & \vdots & \vdots & \\ \vdots & & \vdots & & \ddots & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & & & \widetilde{\mathbf{v}}_1^{[K]} & \cdots & \widetilde{\mathbf{v}}_{n-1}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times (n-1)K}. \end{aligned}$$

Using this matrix, $\mathbf{H}_n$ is found as the projection matrix onto the subspace that is orthogonal to $\widetilde{\mathbf{V}}_{(n-1)}$:

$$\mathbf{H}_n = \mathbf{I} - \widetilde{\mathbf{V}}_{(n-1)} \left(\widetilde{\mathbf{V}}_{(n-1)}^\mathsf{T} \widetilde{\mathbf{V}}_{(n-1)} \right)^{-1} \widetilde{\mathbf{V}}_{(n-1)}^\mathsf{T}.$$

This is repeated until the eigenvectors of all N EVDs are found. Finally, they are concatenated in

$$\widetilde{\mathbf{V}}^{[k]} = \begin{bmatrix} \widetilde{\mathbf{v}}_1^{[k]} & \dots & \widetilde{\mathbf{v}}_N^{[k]} \end{bmatrix} \in \mathbb{R}^{N \times N} \in \mathbb{R}^{N \times N}$$

to find the transformation matrix for the k^{th} dataset

$$\mathbf{T}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \widetilde{\mathbf{V}}^{[k]}, \quad k = 1, \dots, K. \tag{3.29}$$

The multiplication by $\left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}}$ is necessary because $\widetilde{\mathbf{V}}^{[k]}$ is for the whitened datasets.

The pseudocode for the deflationary analytical maxvar method is given in Algorithm 7. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2K^2V + N^4K^3)$. It is dominated by lines 4 and 11.

Algorithm 7 Deflationary analytical maxvar [Ket71]

Input: $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \dots, K$

1: **for** $k = 1, \dots, K$ **do** ▷ whiten observed datasets
2: $\quad \mathbf{Y}^{[k]} \leftarrow \left(\frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1/2} \mathbf{X}^{[k]}$
3: $\mathbf{Y} \leftarrow \begin{bmatrix} \mathbf{Y}^{[1]} \\ \vdots \\ \mathbf{Y}^{[K]} \end{bmatrix}$
4: $\widehat{\mathbf{C}}_{\mathbf{y}} \leftarrow \frac{1}{V}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}$
5: **for** $n = 1, \dots, N$ **do**
6: $\quad$ **if** $n = 1$ **then**
7: $\quad\quad \mathbf{H}_n \leftarrow \mathbf{I}_{NK \times NK}$
8: $\quad$ **else**
9: $\quad\quad \widetilde{\mathbf{V}}_{(n-1)} \leftarrow \text{blkdiag}\left(\begin{bmatrix} \widetilde{\mathbf{v}}_1^{[1]} & \dots & \widetilde{\mathbf{v}}_{n-1}^{[1]} \end{bmatrix}, \dots, \begin{bmatrix} \widetilde{\mathbf{v}}_1^{[K]} & \dots & \widetilde{\mathbf{v}}_{n-1}^{[K]} \end{bmatrix}\right)$
10: $\quad\quad \mathbf{H}_n \leftarrow \mathbf{I} - \widetilde{\mathbf{V}}_{(n-1)}\left(\widetilde{\mathbf{V}}_{(n-1)}^{\mathsf{T}}\widetilde{\mathbf{V}}_{(n-1)}\right)^{-1}\widetilde{\mathbf{V}}_{(n-1)}^{\mathsf{T}}$
11: $\quad \phi_n, \mathbf{v}_n \leftarrow \text{EVD}(\mathbf{H}_n\widehat{\mathbf{C}}_{\mathbf{y}}\mathbf{H}_n, 1)$ ▷ largest eigenvalue and eigenvector
12: $\quad$ **for** $k = 1, \dots, K$ **do**
13: $\quad\quad \mathbf{v}_n^{[k]} \leftarrow \mathbf{v}_n\left((k-1)N + 1 : kN\right)$
14: $\quad\quad \widetilde{\mathbf{v}}_n^{[k]} \leftarrow \frac{\mathbf{v}_n^{[k]}}{\left\|\mathbf{v}_n^{[k]}\right\|}$ ▷ $\widetilde{\mathbf{v}}_n^{[k]}$ is the n^{th} column of $\widetilde{\mathbf{V}}^{[k]}$
15: **for** $k = 1, \dots, K$ **do**
16: $\quad \mathbf{T}^{[k]} \leftarrow \left(\frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1/2} \widetilde{\mathbf{V}}^{[k]}$
17: $\quad \mathbf{U}^{[k]} \leftarrow \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]}$

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}, \mathbf{U}^{[k]} \in \mathbb{R}^{N \times V}, k = 1, \dots, K$

All-at-once analytical solution

Contrasting this deflationary analytical approach, Carroll proposed an all-at-once analytical solution for maxvar [Car68]. Therefore, he introduces an auxiliary variable z_n and seeks $u_n^{[k]}$ of all datasets that are maximally correlated to z_n, $n = 1, \ldots, N$. Instead of imposing constraints (3.19) or (3.20), [Car68] imposes

$$\mathbb{E}\{z_n^2\} = 1,$$

and also makes the $z_1, \ldots, z_N$ uncorrelated instead of the canonical variables across CCVs, i.e., imposes the constraint

$$\mathbb{E}\{z_n z_m\} = 0, \quad n \neq m, \quad n, m = 1, \ldots, N,$$

instead of (3.21) or (3.22). Carroll's solution directly considers observed sample datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, where N is the dimension of the data and V is the number of samples, instead of random variables. Carroll allows for a different scalar weighting factor for each dataset, which we choose in this thesis to be equal to one, i.e., all datasets have an equal weight. The choice of this value will become clear once we derive the connection to sumcor. Carroll defines

$$\mathbf{Q} = \sum_{k=1}^{K} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]} \in \mathbb{R}^{V \times V}, \tag{3.30}$$

which can be interpreted as the summation of the K datasets' projection matrices, representing the row space of the datasets. Then, by performing an EVD on $\mathbf{Q}$, we have

$$\mathbf{Q}\mathbf{Z} = \mathbf{Z}\mathbf{\Phi}, \tag{3.31}$$

with $\mathbf{Z} \in \mathbb{R}^{V \times V}$ and $\mathbf{\Phi} \in \mathbb{R}^{V \times V}$. We denote $\widetilde{\mathbf{Z}} = [\mathbf{z}_1, \ldots, \mathbf{z}_N] \in \mathbb{R}^{V \times N}$, as the matrix consisting of the N principal eigenvectors of $\mathbf{Q}$. Then, the the transformation matrix for the k^{th} dataset $\mathbf{T}^{[k]}$ can be found as [Car68]

$$\mathbf{T}^{[k]} = \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]} \widetilde{\mathbf{Z}}. \tag{3.32}$$

In the end, we normalize the canonical variables to unit-variance and multiply $\mathbf{t}_n^{[k]}$ by the variance of the n^{th} row of $\mathbf{U}^{[k]}$, $n = 1, \ldots, N$, $k = 1, \ldots, K$, for comparison with the other methods. The pseudocode for the all-at-once analytical maxvar method is

given in Algorithm 8. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(V^3)$. It is dominated by line 2 under the assumption that $V > NK$.

Algorithm 8 All-at-once analytical maxvar [Car68]

Input: $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \ldots, K$

1: $\mathbf{Q} \leftarrow \sum_{k=1}^{K} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]}$
2: $\boldsymbol{\phi}, \mathbf{Z} \leftarrow \mathrm{EVD}(\mathbf{Q})$
3: $\widetilde{\mathbf{Z}} \leftarrow \mathbf{Z}(:, 1:N)$
4: **for** $k = 1, \ldots, K$ **do**
5: $\quad \mathbf{T}^{[k]} \leftarrow \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]} \widetilde{\mathbf{Z}}$
6: $\quad \mathbf{U}^{[k]} \leftarrow \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]}$
7: **for** $k = 1, \ldots, K$ **do**
8: $\quad \mathbf{T}^{[k]}, \mathbf{U}^{[k]} \leftarrow \mathrm{normalize}\left(\mathbf{T}^{[k]}, \mathbf{U}^{[k]}\right)$ ▷ normalize canonical variables

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}, \mathbf{U}^{[k]} \in \mathbb{R}^{N \times V}, k = 1, \ldots, K$

Underlying factor model

The CCVs estimated by maxvar underlie the model [Ket71]

$$\mathbf{u}_n = \mathbf{m}_n f_n + \boldsymbol{\nu}_n, \quad n = 1, \ldots, N, \tag{3.33}$$

where f_n is the best fitting zero-mean unit-variance common factor, the elements of $\mathbf{m}_n \in \mathbb{R}^K$ weight the contribution of this factor to each $u_n^{[k]}$, and $\boldsymbol{\nu}_n \in \mathbb{R}^K$ is again a noise vector. Thus, also maxvar assumes the SCVs to have an effective rank of one in the JBSS setting [Gab24], but compared with sumcor, here the common factor has a different contribution to each source component $s_n^{[k]}$.

Given this factor model, another interpretation of (3.32) is to consider the overdetermined system

$$\left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]} = \mathbf{z}_n,$$

and then

$$\mathbf{t}_n^{[k]} = \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]} \mathbf{z}_n$$

is just the optimal solution of the Ordinary Least Squares (OLS) problem

$$\min_{\mathbf{t}_n^{[k]}} \left\| \mathbf{z}_n^{\mathsf{T}} - \left(\mathbf{t}_n^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]} \right\|^2,$$

and therefore the $\mathbf{t}_n^{[k]}$ can be interpreted as regression coefficients that extract $\mathbf{z}_n$ from each dataset $\mathbf{X}^{[k]}$, where the $\mathbf{z}_n \in \mathbb{R}^V$ can be interpreted as sample versions of the common factors in (3.33).

3.4.3 mCCA-minvar

The goal of minvar is very similar to maxvar, with the difference that now for the n^{th} CCV, minvar aims to find transformation vectors $\mathbf{t}_n^{[k]}$ that minimize the smallest eigenvalue of each $\mathbf{C}_{\mathbf{u}_n}$ as denoted in (F4) [Ket71].

Deflationary analytical solution

Kettenring [Ket71] proposed, using constraints (3.19) and (3.21), the same deflationary analytical solution as for maxvar, i.e., performing the EVD

$$\mathbf{H}_n \mathbf{C}_{\mathbf{y}} \mathbf{H}_n \mathbf{v}_n = \phi_n \mathbf{v}_n$$

sequentially for $n = 1, \ldots, N$, but while in maxvar ϕ_n denoted the largest eigenvalue, here in minvar ϕ_n denotes the smallest non-zero eigenvalue, and $\mathbf{v}_n \in \mathbb{R}^{NK}$ is the corresponding eigenvector. The pseudocode for the deflationary analytical minvar method is the same as that for the deflationary analytical maxvar method given in Algorithm 7 on page 41, but line 11 is replaced as shown in Algorithm 9, where $NK - (n-1)K$ is the position index of the smallest non-zero eigenvalue for the n^{th} EVD. The big-$\mathcal{O}$ complexity of this method is the same as that of the deflationary analytical maxvar method, i.e., $\mathcal{O}(N^2K^2V + N^4K^3)$.

Algorithm 9 Deflationary analytical minvar [Ket71]

11: $\phi_n, \mathbf{v}_n \leftarrow \text{EVD}(\mathbf{H}_n \widehat{\mathbf{C}}_{\mathbf{y}} \mathbf{H}_n, NK - (n-1)K)$ ▷ smallest non-zero eigenvalue and eigenvector

Underlying factor model

The CCVs estimated by minvar underlie the model [Ket71]

$$\mathbf{u}_n = \sum_{r=1}^{K-1} \mathbf{m}_n^{(r)} f_n^{(r)} + \boldsymbol{\nu}_n, \tag{3.34}$$

where $f_n^{(r)}$, $r = 1, \ldots, K-1$, are the $K-1$ zero-mean unit-variance uncorrelated common factors, the elements of $\mathbf{m}_n^{(r)} \in \mathbb{R}^K$ weight the contribution of each $f_n^{(r)}$ to

each $u_n^{[k]}$, and $\boldsymbol{\nu}_n \in \mathbb{R}^K$ is again a noise vector. Whereas maxvar assumes a CCV to have an effective rank-1 model (one common factor), minvar assumes a CCV to have an effective rank-$(K-1)$ model.

3.4.4 mCCA-genvar

For the n^{th} CCV, genvar aims to find the transformation vectors $\mathbf{t}_n^{[k]}$ that minimize the determinant of its covariance matrix $\mathbf{C}_{\mathbf{u}_n}$ as denoted in (F5) [Ket71].

Deflationary numerical solution

Kettenring proposed in [Ket71] the following deflationary solution for genvar using constraints (3.19) and (3.21), where each CCV is found using the following numerical method. First, the datasets are whitened using Mahalanobis whitening as in (3.26), and the joint covariance matrix $\mathbf{C}_{\mathbf{y}}$ of the whitened datasets $\mathbf{y}^{[k]}$ is calculated as in (3.27). Then, the transformation vectors $\mathbf{v}_n^{[k]}$ for the whitened datasets are initialized as unit-norm vectors with all elements equal to $\frac{1}{\sqrt{N}}$. As for the deflationary analytical maxvar method in (3.28), N EVDs are performed sequentially for $n = 1, \ldots, N$, but now separately for each dataset, i.e.,

$$\mathbf{H}_n^{[k]} \widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \mathbf{H}_n^{[k]} \mathbf{v}_n^{[k]} = \phi_n^{[k]} \mathbf{v}_n^{[k]}, \quad k = 1, \ldots, K, \tag{3.35}$$

where $\phi_n^{[k]}$ denotes the largest eigenvalue, $\mathbf{v}_n^{[k]} \in \mathbb{R}^N$ the corresponding eigenvector, and we explain in the following how to form the k^{th} datasets' projection matrices $\mathbf{H}_n^{[k]} \in \mathbb{R}^{N \times N}$ and the matrices $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \in \mathbb{R}^{N \times N}$. For estimating the first CCV, $\mathbf{H}_1^{[k]} = \mathbf{I}_{N \times N}$. For the other CCVs,

$$\mathbf{H}_n^{[k]} = \mathbf{I}_{N \times N} - \mathbf{V}_{(n-1)}^{[k]} \left(\left(\mathbf{V}_{(n-1)}^{[k]} \right)^{\mathsf{T}} \mathbf{V}_{(n-1)}^{[k]} \right)^{-1} \left(\mathbf{V}_{(n-1)}^{[k]} \right)^{\mathsf{T}},$$

where

$$\mathbf{V}_{(n-1)}^{[k]} = \begin{bmatrix} \mathbf{v}_1^{[k]} & \ldots & \mathbf{v}_{n-1}^{[k]} \end{bmatrix} \in \mathbb{R}^{N \times (n-1)}$$

contains the concatenated transformation vectors for the k^{th} whitened dataset from the $n-1$ previous EVDs. While $\mathbf{H}_n^{[k]}$ is only dependent on the $\mathbf{v}_n^{[k]}$ of the previous stages, the matrix $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$ is updated in every iteration. The matrix $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$ is defined as

$$\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} = \mathbf{N}_n^{[k]} \left(\mathbf{C}_{\mathbf{u}_n}^{[-k]} \right)^{-1} \left(\mathbf{N}_n^{[k]} \right)^{\mathsf{T}} \in \mathbb{R}^{N \times N}, \tag{3.36}$$

where

$$\mathbf{N}_n^{[k]} = \left[\mathbf{C}_{\mathbf{y}}^{[k,1]}\mathbf{v}_n^{[1]} \quad \dots \quad \mathbf{C}_{\mathbf{y}}^{[k,k-1]}\mathbf{v}_n^{[k-1]} \quad \mathbf{C}_{\mathbf{y}}^{[k,k+1]}\mathbf{v}_n^{[k+1]} \quad \dots \quad \mathbf{C}_{\mathbf{y}}^{[k,K]}\mathbf{v}_n^{[K]}\right] \in \mathbb{R}^{N\times(K-1)}$$

consists of the covariances of the whitened dataset $\mathbf{y}^{[k]}$ with the canonical variables $u_n^{[l]}$, $l \neq k$, $l, k = 1, \dots, K$, i.e.,

$$\mathbf{C}_{\mathbf{y}}^{[k,l]}\mathbf{v}_n^{[l]} = \mathbb{E}\left\{\mathbf{y}^{[k]}\left(\mathbf{y}^{[l]}\right)^{\mathsf{T}}\mathbf{v}_n^{[l]}\right\} = \mathbb{E}\left\{\mathbf{y}^{[k]}u_n^{[l]}\right\} \in \mathbb{R}^N, \quad l \neq k,$$

and $\mathbf{C}_{\mathbf{u}_n}^{[-k]} \in \mathbb{R}^{(K-1)\times(K-1)}$ is the matrix obtained by removing the k^{th} row and column of the n^{th} CCV's covariance matrix $\mathbf{C}_{\mathbf{u}_n}$, which is calculated as

$$\mathbf{C}_{\mathbf{u}_n} = \text{blkdiag}\left(\mathbf{v}_n^{[1]}, \dots, \mathbf{v}_n^{[K]}\right)^{\mathsf{T}}\mathbf{C}_{\mathbf{y}}\text{blkdiag}\left(\mathbf{v}_n^{[1]}, \dots, \mathbf{v}_n^{[K]}\right) \in \mathbb{R}^{K\times K}.$$

Using the updated $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$, the EVD in (3.35) is performed to update $\mathbf{v}_n^{[k]}$, and $\mathbf{C}_{\mathbf{u}_n}$, and then $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$, $\mathbf{v}_n^{[k]}$ and $\mathbf{C}_{\mathbf{u}_n}$ are updated again in the next iteration until either convergence or a maximum number of iterations has been reached. Convergence is achieved when the sum of the differences between the largest eigenvalues of the current iteration i and the previous iteration $i-1$ is smaller than a user-defined threshold $\epsilon > 0$, i.e., when

$$\sum_{k=1}^{K}\left|\psi_n^{[k]}(i) - \psi_n^{[k]}(i-1)\right| < \epsilon,$$

where $\psi_n^{[k]}(i) = \phi_n^{[k]}(i)$ is the largest eigenvalue of the i^{th} iteration [Ket71]. After the iterative process is completed, the transformation matrices $\mathbf{T}^{[k]}$ are found as

$$\mathbf{T}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}}\mathbf{V}^{[k]}.$$

The multiplication by $\left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}}$ is necessary because $\mathbf{V}^{[k]}$ is found for the whitened datasets. The pseudocode for the deflationary numerical genvar method is given in Algorithms 10 and 11. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2K^2V + IN^3K^4)$, where I is the number of iterations. It is dominated by line 5 in Algorithm 10 and line 3 in Algorithm 11.

Algorithm 10 Deflationary numerical genvar [Ket71]

Input: $\mathbf{X}^{[k]} \in \mathbb{R}^{N\times V}$, $k = 1, \ldots, K$

for $k = 1, \ldots, K$ **do** ▷ whiten observed datasets
  $\mathbf{Y}^{[k]} \leftarrow \left(\frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1/2}\mathbf{X}^{[k]}$
for $k = 1, \ldots, K$ **do** ▷ cross-covariance matrix of k^{th} and l^{th} whitened dataset
  for $l = 1, \ldots, K$ **do**
    $\widehat{\mathbf{C}}_{\mathbf{y}}^{[k,l]} \leftarrow \frac{1}{V}\mathbf{Y}^{[k]}\left(\mathbf{Y}^{[l]}\right)^{\mathsf{T}}$
for $k = 1, \ldots, K$ **do** ▷ initialize transformation matrix for whitened data
  $\mathbf{V}^{[k]} \leftarrow \frac{1}{\sqrt{N}} I_{K\times K}$
for $n = 1, \ldots, N$ **do**
  for $k = 1, \ldots, K$ **do**
    if $n = 1$ **then**
      $\mathbf{H}_n^{[k]} \leftarrow \mathbf{I}_{N\times N}$
    else
      $\mathbf{V}_{(n-1)}^{[k]} \leftarrow \begin{bmatrix}\mathbf{v}_1^{[k]} & \cdots & \mathbf{v}_{n-1}^{[k]}\end{bmatrix}$ ▷ $\mathbf{v}_n^{[k]}$ is the n^{th} column of $\mathbf{V}^{[k]}$
      $\mathbf{H}_n^{[k]} \leftarrow \mathbf{I}_{N\times N} - \mathbf{V}_{(n-1)}^{[k]}\left(\left(\mathbf{V}_{(n-1)}^{[k]}\right)^{\mathsf{T}}\mathbf{V}_{(n-1)}^{[k]}\right)^{-1}\left(\mathbf{V}_{(n-1)}^{[k]}\right)^{\mathsf{T}}$
  $\boldsymbol{\psi}_n \leftarrow \mathbf{0}$
  for $i = 2, \ldots,$ maxIter **do**
    for $k = 1, \ldots, K$ **do**
      $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \leftarrow$ calculate_$\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}(k, \widehat{\mathbf{C}}_{\mathbf{y}}, \mathbf{v}_n^{[1]}, \ldots, \mathbf{v}_n^{[K]})$
      $\phi_n^{[k]}, \mathbf{v}_n^{[k]} \leftarrow \text{EVD}(\mathbf{H}_n^{[k]}\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}\mathbf{H}_n^{[k]}, 1)$ ▷ largest eigenvalue and eigenvector
      $\psi_n^{[k]}(i) \leftarrow \phi_n^{[k]}$
    if $\text{sum}(|\boldsymbol{\psi}_n(i) - \boldsymbol{\psi}_n(i-1)|) < \epsilon$ **then** ▷ genvar converged
      break
for $k = 1, \ldots, K$ **do**
  $\mathbf{T}^{[k]} \leftarrow \left(\frac{1}{V}\mathbf{X}^{[k]}\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1/2}\mathbf{V}^{[k]}$
  $\mathbf{U}^{[k]} \leftarrow \left(\mathbf{T}^{[k]}\right)^{\mathsf{T}}\mathbf{X}^{[k]}$

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N\times N}, \mathbf{U}^{[k]} \in \mathbb{R}^{N\times V}, k = 1, \ldots, K$

Underlying factor model

The CCVs estimated by genvar underlie the model [Ket71]

$$\mathbf{u}_n = \sum_{r=1}^{K} \mathbf{m}_n^{(r)} f_n^{(r)} + \boldsymbol{\nu}_n = \mathbf{M}_n \mathbf{f}_n + \boldsymbol{\nu}_n, \tag{3.37}$$

Algorithm 11 calculate_$\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$ [Ket71]

Input: k, $\widehat{\mathbf{C}}_{\mathbf{y}} \in \mathbb{R}^{NK \times NK}$, $\mathbf{v}_n^{[1]}, \ldots, \mathbf{v}_n^{[K]} \in \mathbb{R}^N$

1: $\mathbf{N}_n^{[k]} \leftarrow \left[\widehat{\mathbf{C}}_{\mathbf{y}}^{[k,1]}\mathbf{v}_n^{[1]} \quad \ldots \quad \widehat{\mathbf{C}}_{\mathbf{y}}^{[k,k-1]}\mathbf{v}_n^{[k-1]} \quad \widehat{\mathbf{C}}_{\mathbf{y}}^{[k,k+1]}\mathbf{v}_n^{[k+1]} \quad \ldots \quad \widehat{\mathbf{C}}_{\mathbf{y}}^{[k,K]}\mathbf{v}_n^{[K]}\right]$

2: $\mathbf{V}_n \leftarrow \text{blkdiag}\left(\mathbf{v}_n^{[1]}, \ldots, \mathbf{v}_n^{[K]}\right)$

3: $\widehat{\mathbf{C}}_{\mathbf{u}_n} \leftarrow \mathbf{V}_n^{\mathsf{T}}\widehat{\mathbf{C}}_{\mathbf{y}}\mathbf{V}_n$ ▷ covariance matrix of the n^{th} estimated CCV

4: $\widehat{\mathbf{C}}_{\mathbf{u}_n}^{[-k]} \leftarrow$ delete k^{th} row and column of $\widehat{\mathbf{C}}_{\mathbf{u}_n}$

5: $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \leftarrow \mathbf{N}_n^{[k]}\left(\widehat{\mathbf{C}}_{\mathbf{u}_n}^{[-k]}\right)^{-1}\left(\mathbf{N}_n^{[k]}\right)^{\mathsf{T}}$

Output: $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \in \mathbb{R}^{N \times N}$

with $\mathbf{M}_n = \left[\mathbf{m}_n^{(1)} \quad \cdots \quad \mathbf{m}_n^{(K)}\right] \in \mathbb{R}^{K \times K}$, $\mathbf{f}_n = \left[f_n^{(1)} \quad \cdots \quad f_n^{(K)}\right]^{\mathsf{T}} \in \mathbb{R}^K$, and $\boldsymbol{\nu}_n \in \mathbb{R}^K$. The model (3.37) is similar to the maxvar / minvar models in (3.33) and (3.34) but now assumes an effective rank-K model with K different $\mathbf{m}_n^{(r)}$ and $\mathbf{f}_n^{(r)}$. It was noted in [Ket71] that genvar can be seen as more "unpredictable" as the determinant cost is more sensitive to the smaller eigenvalues. While this may sound like a disadvantage, we later demonstrate that genvar is consistently among the best-performing mCCA methods for JBSS.

3.4.5 mCCA-ssqcor

For the n^{th} CCV, ssqcor aims to find the transformation vectors $\mathbf{t}_n^{[k]}$ that maximize the sum of squared correlations within each $\mathbf{C}_{\mathbf{u}_n}$ as denoted in (F2) [Ket71].

Deflationary numerical solution

Kettenring proposed in [Ket71], using constraints (3.19) and (3.21), a solution very similar to that of genvar, with the difference that $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]}$ in (3.36) is now defined as

$$\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} = \mathbf{N}_n^{[k]}\left(\mathbf{N}_n^{[k]}\right)^{\mathsf{T}}, \tag{3.38}$$

The pseudocode for the deflationary numerical ssqcor method is the same as that for the deflationary numerical genvar method given in Algorithms 10 and 11 on pages 47 and 48, but line 5 in Algorithm 11 is replaced as shown in Algorithm 12, so that lines 2, 3, and 4 in Algorithm 11 are no longer necessary. The big-$\mathcal{O}$ complexity of this method is $\mathcal{O}(N^2K^2V + IN^3K(K+N))$, where I is the number of iterations. It is dominated by lines 5 and 19 in Algorithm 10, line 1 in Algorithm 11, and modified

Table 3.1: Overview of the computational complexity of the mCCA methods, categorized by the nature of their solutions.

	all-at-once	deflationary
analytical	sumcor [Nie95]: $\mathcal{O}(N^2K^2V)$ maxvar [Car68]: $\mathcal{O}(V^3)$	maxvar [Ket71]: $\mathcal{O}(N^2K^2V + N^4K^3)$ minvar [Ket71]: $\mathcal{O}(N^2K^2V + N^4K^3)$
numerical	–	sumcor [Ket71]: (not further explained) genvar [Ket71]: $\mathcal{O}(N^2K^2V + IN^3K^4)$ ssqcor [Ket71]: $\mathcal{O}(N^2K^2V + IN^3K(K+N))$

line 5 in Algorithm 12.

Algorithm 12 Deflationary analytical ssqcor [Ket71]

5: $\widetilde{\mathbf{C}}_{\mathbf{u}_n}^{[k]} \leftarrow \mathbf{N}_n^{[k]} \left(\mathbf{N}_n^{[k]}\right)^{\mathsf{T}}$

Underlying factor model

The CCVs estimated by ssqcor also underlie the model in (3.37) [Ket71]. By comparing (3.37) to (3.33), we observe that if only one single factor heavily influences the CCVs, i.e., all except for one $\mathbf{m}_n^{(l)}$ are nearly zero, ssqcor will estimate CCVs that are very similar to those of maxvar. However, in the JBSS setting, ssqcor is more flexible than maxvar in that ssqcor can estimate SCVs with *any* effective rank (from 1 to K).

3.4.6 Comment on different solutions

We have seen in the previous subsections that the different optimization problems of mCCA are solved by various methods, which we have summarized in Table 3.1. The mCCA methods presented in [Ket71] all require estimating the transformation vectors in a deflationary way, i.e., sequentially estimating each CCV, where the solution for each CCV can be found either analytically (typically manifesting as EVD) or numerically, e.g., using gradient descent. On the other hand, the methods from [Car68] and [Nie95] estimate the transformation vectors all at once together for all CCVs with an analytical solution.

3.5 PARAFAC2

After having introduced the matrix decomposition methods ICA, IVA, CCA, and IVA, in this section, we present the tensor decomposition method PARAFAC2, which is closely related to IVA-G, as we will show in Section 4.3. The PARAFAC2 [Har72] tensor model has proven useful for jointly analyzing datasets [Che07; Mad17; Roa20] as it does not impose strong constraints on them, in contrast to the well-known CP [Har70; Car70] tensor decomposition method. As PARAFAC2 is a more flexible version of the well-known CP tensor decomposition, we briefly present this method before explaining PARAFAC2.

Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N \times V \times K}$ be the tensor consisting of the K observed datasets as frontal slices, i.e.,

$$\boldsymbol{\mathcal{X}}(:,:,k) = \mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}, \quad k = 1, \ldots, K.$$

The well-known CP tensor model [Har70; Car70] represents a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N \times V \times K}$ consisting of R components as follows:

$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{3.39}$$

with the factor matrices

$$\begin{aligned} \mathbf{A} &= [\mathbf{a}_1 \ \ldots \ \mathbf{a}_R] \in \mathbb{R}^{N \times R}, \\ \mathbf{B} &= [\mathbf{b}_1 \ \ldots \ \mathbf{b}_R] \in \mathbb{R}^{V \times R}, \\ \mathbf{C} &= [\mathbf{c}_1 \ \ldots \ \mathbf{c}_R] \in \mathbb{R}^{K \times R}, \end{aligned}$$

and the outer product denoted by $\circ$. The solution of (3.39) is found by solving the following optimization problem [Kol09]:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \boldsymbol{\mathcal{X}} - \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \right\|_F^2. \tag{3.40}$$

When fixing all but one of the factor matrices, (3.40) reduces to a linear least-squares problem, and therefore, the solution for $\mathbf{A}, \mathbf{B}, \mathbf{C}$ can be found using Alternating Least Squares (ALS) [Kol09].

With $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$ being the k^{th} frontal slice of $\boldsymbol{\mathcal{X}}$, (3.39) can be rewritten as [Kol09]

$$\mathbf{X}^{[k]} \approx \mathbf{A} \operatorname{diag}(\mathbf{c}_{k:}) \mathbf{B}^{\mathsf{T}}, \quad k = 1, \ldots, K, \tag{3.41}$$

where $\mathbf{c}_{k:}$ is the k^{th} row of

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{1:} \\ \vdots \\ \mathbf{c}_{K:} \end{bmatrix} \in \mathbb{R}^{K \times R},$$

and we see that the CP model assumes common factor matrices $\mathbf{A}$ and $\mathbf{B}$ for all K frontal slices, i.e., observed datasets. In contrast to this, PARAFAC2 [Har72; Kie99] is a more flexible model, as it allows for changes of the $\mathbf{B}$ matrix, meaning that the factor matrix $\mathbf{B}^{[k]} \in \mathbb{R}^{V \times R}$ can be different in each frontal slice (observed dataset):

$$\begin{aligned} \mathbf{X}^{[k]} &\approx \mathbf{A} \operatorname{diag}(\mathbf{c}_{k:}) \left(\mathbf{B}^{[k]}\right)^{\mathsf{T}} \\ \text{s.t. } &\left(\mathbf{B}^{[k]}\right)^{\mathsf{T}} \mathbf{B}^{[k]} = \mathbf{M}, \quad k = 1, \ldots, K, \end{aligned} \tag{3.42}$$

where $\mathbf{M} \in \mathbb{R}^{R \times R}$ is an arbitrary matrix, and with the constraint introduced to preserve uniqueness of the components up to permutation and scaling ambiguities [Har72; Kie99]. The solution of (3.42) is found by introducing the matrices $\mathbf{J}^{[k]} \in \mathbb{R}^{V \times R}$ and $\mathbf{B} \in \mathbb{R}^{R \times R}$ such that

$$\mathbf{B}^{[k]} = \mathbf{J}^{[k]} \mathbf{B}, \quad k = 1, \ldots, K,$$

and solving the following optimization problem using an ALS-based algorithm [Kie99]:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{J}^{[1]},\ldots,\mathbf{J}^{[K]}} & \sum_{k=1}^{K} \left\| \mathbf{X}^{[k]} - \mathbf{A} \operatorname{diag}(\mathbf{c}_{k:}) \mathbf{B}^{\mathsf{T}} \left(\mathbf{J}^{[k]}\right)^{\mathsf{T}} \right\|_F^2 \\ \text{s.t. } &\left(\mathbf{J}^{[k]}\right)^{\mathsf{T}} \mathbf{J}^{[k]} = \mathbf{I}, \quad k = 1, \ldots, K. \end{aligned} \tag{3.43}$$

Therefore, after initializing $\mathbf{A}, \mathbf{B}, \mathbf{C}$, the solution of (3.43) for $\mathbf{J}^{[k]}$ is found as [Kie99]

$$\mathbf{J}^{[k]} = \mathbf{F}^{[k]} \left(\mathbf{G}^{[k]}\right)^{\mathsf{T}}, \quad k = 1, \ldots, K,$$

where $\mathbf{F}^{[k]} \in \mathbb{R}^{V \times R}$ and $\mathbf{G}^{[k]} \in \mathbb{R}^{R \times R}$ are found using the SVD of

$$\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \mathbf{A} \operatorname{diag}(\mathbf{c}_{k:}) \mathbf{B}^{\mathsf{T}} = \mathbf{F}^{[k]} \mathbf{K}^{[k]} \left(\mathbf{G}^{[k]}\right)^{\mathsf{T}}.$$

Then, the optimization problem can be written equivalently as

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{k=1}^{K} \left\| \mathbf{X}^{[k]} \mathbf{J}^{[k]} - \mathbf{A} \operatorname{diag}(\mathbf{c}_{k:}) \mathbf{B}^{\mathsf{T}} \right\|_F^2, \tag{3.44}$$

and we see that the solution of (3.44) for $\mathbf{A}, \mathbf{B}, \mathbf{C}$ just corresponds to the CP decomposition of a tensor $\widetilde{\boldsymbol{\mathcal{X}}}$ with $\mathbf{X}^{[k]}\mathbf{J}^{[k]}$ as frontal slices, i.e.,

$$\widetilde{\boldsymbol{\mathcal{X}}}(:,:,k) = \mathbf{X}^{[k]}\mathbf{J}^{[k]}.$$

Here, one update for $\mathbf{A}, \mathbf{B}, \mathbf{C}$ in the CP-ALS algorithm is sufficient for the PARAFAC2-ALS algorithm to converge. Then, using the updated $\mathbf{A}, \mathbf{B}, \mathbf{C}$, the $\mathbf{J}^{[1]}, \ldots, \mathbf{J}^{[K]}$ are updated, until convergence is reached. For better convergence, $\mathbf{A}$ is initialized with the R principal eigenvectors of $\sum_{k=1}^{K} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}$, while $\mathbf{B}^{[1]}, \ldots, \mathbf{B}^{[K]}$ and $\mathbf{C}$ are initialized as identity matrices [Kie99]. The pseudocode for PARAFAC2 is given in Algorithm 13, where CP_update($\widetilde{\boldsymbol{\mathcal{X}}}, \mathbf{A}, \mathbf{B}, \mathbf{C}$) denotes one CP-ALS update of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ given the current tensor $\widetilde{\boldsymbol{\mathcal{X}}}$, and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ from the previous iteration.

Algorithm 13 PARAFAC2 [Kie99]

Input: R, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N \times V \times K}$

$\boldsymbol{\phi}, \mathbf{V} \leftarrow \text{EVD}\left(\sum_{k=1}^{K} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)$
$\mathbf{A} \leftarrow \mathbf{V}(:, 1:R)$
$\mathbf{B} \leftarrow \mathbf{I}_{R \times R}$
$\mathbf{C} \leftarrow \mathbf{I}_{K \times R}$
$\psi(1) \leftarrow 0$
for $i = 2, \ldots,$ maxIter **do**
  for $k = 1, \ldots, K$ **do**
    $\mathbf{F}^{[k]}, \mathbf{K}^{[k]}, \mathbf{G}^{[k]} \leftarrow \text{SVD}\left(\left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \mathbf{A}\, \text{diag}\left(\mathbf{c}_{k:}\right) \mathbf{B}^{\mathsf{T}}\right)$
    $\mathbf{J}^{[k]} = \mathbf{F}^{[k]} \left(\mathbf{G}^{[k]}\right)^{\mathsf{T}}$
  for $k = 1, \ldots, K$ **do**
    $\widetilde{\boldsymbol{\mathcal{X}}}(:,:,k) = \mathbf{X}^{[k]}\mathbf{J}^{[k]}$
  $\mathbf{A}, \mathbf{B}, \mathbf{C} \leftarrow$ CP_update($\widetilde{\boldsymbol{\mathcal{X}}}$, **A, B, C**)
  $\psi(i) \leftarrow \left\| \mathbf{X}^{[k]} - \mathbf{A}\, \text{diag}\left(\mathbf{c}_{k:}\right) \mathbf{B}^{\mathsf{T}} \left(\mathbf{J}^{[k]}\right)^{\mathsf{T}} \right\|$
  if $\psi(i-1) - \psi(i) < \epsilon\psi(i-1)$ **then** ▷ PARAFAC2 converged
    break

Output: $\mathbf{T}^{[k]} \in \mathbb{R}^{N \times N}, \mathbf{U}^{[k]} \in \mathbb{R}^{N \times V}, k = 1, \ldots, K$

4 Connections between JBSS methods

This chapter is dedicated to establishing connections between an important set of JBSS methods. We start by comparing the two most closely related, and perhaps the most commonly used [Sor21], mCCA methods: the sumcor method and the maxvar method. We prove that the canonical variables estimated by the all-at-once analytical maxvar method [Car68] are scaled versions of those from the all-at-once analytical sumcor method [Nie95], and thus, the canonical variables of both methods are the same if they are normalized to unit variance. This is important to understand because the latter is computationally much faster. Furthermore, we show that the canonical vectors from the deflationary analytical maxvar method [Ket71] and those of the all-at-once analytical sumcor method [Nie95] are closely related but different, with the difference that the canonical variables estimated by the latter are not constrained to be uncorrelated within a dataset. Thereafter, we address the connection of IVA-G [And12] and PARAFAC2 [Har72; Kie99]. We show that the PARAFAC2 model and the model of IVA-G with dimension reduction are intimately related and can both be formulated as the multiplication of a mixing matrix with a source matrix for each dataset, where the latter is more constrained in terms of the sources and the first is more constrained in terms of the mixing matrices, and demonstrate these differences in simulations.[1]

[1]Sections 4.1 and 4.2 of this chapter are based on the paper: "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, *submitted for review*, 2025. I specifically contributed to proving the source identification conditions of sumcor, deriving the connection between maxvar [Ket71] and sumcor [Nie95], implementing and running the experiments, creating all figures, and writing the paper. B. Gabrielson has provided theoretical insights in the connection between maxvar [Car68] and sumcor [Nie95] and the source identification conditions of genvar. Discussions and feedback along the way until the final paper have contributions from all authors. Section 4.3 of this chapter is based on the paper: "Multi-task fMRI Data Fusion Using IVA and PARAFAC2," I. Lehmann, E. Acar, T. Hasija, M. A. B. S. Akhonda, V. D. Calhoun, P. Schreier, and T. Adali, *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1466–1470.

4.1 Connection between all-at-once analytical sumcor [Nie95] and all-at-once analytical maxvar [Car68]

Comparing the sumcor GEVD in (3.23) and the maxvar EVD in (3.31), we see that we can write both EVDs in terms of the whitened datasets. By doing so, we find the connection between the EVDs and, thereby, also the connection between the methods. As Carroll formulated the maxvar method [Car68] using samples, we also use sample notation for the sumcor method [Nie95] to state their connection in Theorem 4.1.

Theorem 4.1. *Let the concatenated whitened datasets be*

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}^{[1]} \\ \vdots \\ \mathbf{Y}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times V},$$

where

$$\mathbf{Y}^{[k]} = \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}, \quad k = 1, \ldots, K.$$

The canonical vectors estimated by the all-at-once analytical maxvar method [Car68] are scaled versions of those estimated by the all-at-once analytical sumcor method [Nie95] (before normalization), i.e.,

$$\mathbf{U}_{\text{maxvar}}^{[k]} = \frac{1}{V} \widetilde{\mathbf{\Omega}} \, \mathbf{U}_{\text{sumcor}}^{[k]}, \quad k = 1, \ldots, K,$$

or in CCV notation,

$$\mathbf{U}_{\text{maxvar},n} = \frac{\omega_n}{V} \mathbf{U}_{\text{sumcor},n}, \quad n = 1, \ldots, N,$$

where V is the number of samples, and $\widetilde{\mathbf{\Omega}} = \operatorname{diag}(\omega_1, \ldots, \omega_N) \in \mathbb{R}^{N \times N}$ is a diagonal matrix consisting of the N largest singular values of $\mathbf{Y}$.

The proof of Theorem 4.1 consists of three steps.

1. Write the sumcor GEVD in (3.23) and the maxvar EVD in (3.31) in terms of the concatenated whitened datasets $\mathbf{Y}$.
2. Derive the relation of the eigenvectors in these two EVDs.
3. Find the connection of the canonical vectors.

Proof. 1) The covariance matrix of the concatenated whitened datasets is

$$\widehat{\mathbf{C}}_{\mathbf{y}} = \frac{1}{V}\mathbf{Y}\mathbf{Y}^{\mathsf{T}} \in \mathbb{R}^{NK\times NK},$$

and

$$\widehat{\mathbf{D}}_{\mathbf{x}} = \begin{bmatrix} \frac{1}{V}\mathbf{X}^{[1]}\left(\mathbf{X}^{[1]}\right)^{\mathsf{T}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{V}\mathbf{X}^{[K]}\left(\mathbf{X}^{[K]}\right)^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{NK\times NK} \tag{4.1}$$

The sumcor GEVD in (3.23), given by

$$\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{T} = \widehat{\mathbf{D}}_{\mathbf{x}}\mathbf{T}\mathbf{\Phi},$$

can be written in terms of $\widehat{\mathbf{C}}_{\mathbf{y}}$ by multiplying it from the left with $\widehat{\mathbf{D}}_{\mathbf{x}}^{-\frac{1}{2}}$ [Nie95]:

$$\begin{aligned} & \underbrace{\widehat{\mathbf{D}}_{\mathbf{x}}^{-\frac{1}{2}}\widehat{\mathbf{C}}_{\mathbf{x}}\widehat{\mathbf{D}}_{\mathbf{x}}^{-\frac{1}{2}}}_{\widehat{\mathbf{C}}_{\mathbf{y}}}\underbrace{\widehat{\mathbf{D}}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}}_{\mathbf{B}} = \underbrace{\widehat{\mathbf{D}}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T}}_{\mathbf{B}}\mathbf{\Phi} \\ \Leftrightarrow \quad & \widehat{\mathbf{C}}_{\mathbf{y}}\mathbf{B} = \mathbf{B}\mathbf{\Phi}, \end{aligned} \tag{4.2}$$

where the matrix

$$\mathbf{B} = \widehat{\mathbf{D}}_{\mathbf{x}}^{\frac{1}{2}}\mathbf{T} \in \mathbb{R}^{NK\times NK}$$

contains the eigenvectors of $\widehat{\mathbf{C}}_{\mathbf{y}}$, and $\mathbf{\Phi} \in \mathbb{R}^{NK\times NK}$ contains its eigenvalues on the main diagonal. Lastly, let

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{t}_1 & \cdots & \mathbf{t}_N \end{bmatrix} \in \mathbb{R}^{NK\times N}$$

consist of the N principal generalized eigenvectors of $\widehat{\mathbf{C}}_{\mathbf{x}}$ w.r.t. $\widehat{\mathbf{D}}_{\mathbf{x}}$, and let

$$\widetilde{\mathbf{B}} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_N \end{bmatrix} \in \mathbb{R}^{NK\times N}$$

consist of the N principal eigenvectors of $\widehat{\mathbf{C}}_{\mathbf{y}}$, with $\widetilde{\mathbf{B}}^{\mathsf{T}}\widetilde{\mathbf{B}} = \mathbf{I}_{N\times N}$.

In the maxvar EVD in (3.31), given by

$$\mathbf{Q}\mathbf{Z} = \mathbf{Z}\mathbf{\Phi},$$

the matrix $\mathbf{Q}$ can be written in terms of $\mathbf{Y}$ as

$$\begin{aligned}
\mathbf{Q} &= \sum_{k=1}^{K} \frac{1}{V} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \left(\frac{1}{V}\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-1} \mathbf{X}^{[k]} \\
&= \sum_{k=1}^{K} \frac{1}{V} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \left(\frac{1}{V}\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-\frac{\mathsf{T}}{2}} \left(\frac{1}{V}\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}}\right)^{-\frac{1}{2}} \mathbf{X}^{[k]} \\
&= \frac{1}{V} \sum_{k=1}^{K} \left(\mathbf{Y}^{[k]}\right)^{\mathsf{T}} \mathbf{Y}^{[k]} \\
&= \frac{1}{V} \left[\left(\mathbf{Y}^{[1]}\right)^{\mathsf{T}} \dots \left(\mathbf{Y}^{[K]}\right)^{\mathsf{T}}\right] \begin{bmatrix} \mathbf{Y}^{[1]} \\ \vdots \\ \mathbf{Y}^{[K]} \end{bmatrix} \\
&= \frac{1}{V} \mathbf{Y}^{\mathsf{T}} \mathbf{Y} \in \mathbb{R}^{V \times V},
\end{aligned}$$

where we have chosen the weighting factor to be equal to one. Let

$$\widetilde{\mathbf{Z}} = \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_N \end{bmatrix} \in \mathbb{R}^{V \times N}$$

consist of the N principal eigenvectors of $\mathbf{Q}$, with $\widetilde{\mathbf{Z}}^{\mathsf{T}} \widetilde{\mathbf{Z}} = \mathbf{I}_{N \times N}$.

2) Now, let the SVD of $\mathbf{Y}$ be

$$\mathbf{Y} = \mathbf{B} \mathbf{\Omega} \mathbf{Z}^{\mathsf{T}},$$

where $\mathbf{B} \in \mathbb{R}^{NK \times NK}$, $\mathbf{\Omega} \in \mathbb{R}^{NK \times NK}$, and $\mathbf{Z} \in \mathbb{R}^{V \times NK}$, and let the rank-$N$ approximation of $\mathbf{Y}$ be

$$\mathbf{Y} = \widetilde{\mathbf{B}} \widetilde{\mathbf{\Omega}} \widetilde{\mathbf{Z}}^{\mathsf{T}},$$

with $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{Z}}$ as defined above, and $\widetilde{\mathbf{\Omega}} = \operatorname{diag}(\omega_1, \dots, \omega_N) \in \mathbb{R}^{N \times N}$. With the partitions of $\mathbf{Y}$ and $\widetilde{\mathbf{B}}$, we have

$$\begin{bmatrix} \mathbf{Y}^{[1]} \\ \vdots \\ \mathbf{Y}^{[K]} \end{bmatrix} = \begin{bmatrix} \mathbf{B}^{[1]} \\ \vdots \\ \mathbf{B}^{[K]} \end{bmatrix} \widetilde{\mathbf{\Omega}} \widetilde{\mathbf{Z}}^{\mathsf{T}},$$

and thus, for each dataset

$$\mathbf{Y}^{[k]} \widetilde{\mathbf{Z}} = \mathbf{B}^{[k]} \widetilde{\mathbf{\Omega}}.$$

3) We know that the sumcor transformation matrices $\mathbf{T}^{[k]}$ are the blocks of

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{T}^{[1]} \\ \vdots \\ \mathbf{T}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times N}.$$

Thus, by inserting (4.1) in $\widetilde{\mathbf{T}} = \widehat{\mathbf{D}}_{\mathbf{x}}^{-\frac{1}{2}} \widetilde{\mathbf{B}}$, we find

$$\mathbf{T}_{\text{sumcor}}^{[k]} = \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \mathbf{B}^{[k]}. \tag{4.3}$$

Furthermore, we know from (3.32) that the maxvar transformation matrices are

$$\mathbf{T}^{[k]} = \left(\mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-1} \mathbf{X}^{[k]} \widetilde{\mathbf{Z}},$$

and rewriting gives

$$\begin{aligned} \mathbf{T}_{\text{maxvar}}^{[k]} &= \frac{1}{V} \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \mathbf{X}^{[k]} \widetilde{\mathbf{Z}} \\ &= \frac{1}{V} \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \mathbf{Y}^{[k]} \widetilde{\mathbf{Z}} \\ &= \frac{1}{V} \left(\frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]} \right)^{\mathsf{T}} \right)^{-\frac{1}{2}} \mathbf{B}^{[k]} \widetilde{\mathbf{\Omega}} \\ &= \frac{1}{V} \mathbf{T}_{\text{sumcor}} \widetilde{\mathbf{\Omega}}. \end{aligned}$$

For the canonical vectors, this means

$$\mathbf{U}_{\text{maxvar}}^{[k]} = \left(\mathbf{T}_{\text{maxvar}}^{[k]} \right)^{\mathsf{T}} \mathbf{X}^{[k]} = \frac{1}{V} \widetilde{\mathbf{\Omega}} \left(\mathbf{T}_{\text{sumcor}}^{[k]} \right)^{\mathsf{T}} \mathbf{X}^{[k]} = \frac{1}{V} \widetilde{\mathbf{\Omega}} \, \mathbf{U}_{\text{sumcor}}^{[k]},$$

i.e., the canonical variables of the all-at-once analytical maxvar [Car68] and the all-at-once analytical sumcor [Nie95] are scaled versions of each other.

□

Note that the canonical variables estimated by these two methods are the same when they are normalized to unit variance. As the all-at-once analytical maxvar method is computationally more expensive than the all-at-once analytical sumcor method (because the EVD is performed on a $V \times V$ matrix instead of an $N \times N$ matrix), we do not report the all-at-once analytical maxvar results but only the sumcor results in the remainder

of this thesis.

4.2 Connection between all-at-once analytical sumcor [Nie95] and deflationary analytical maxvar [Ket71]

After having seen how the sumcor EVD in (3.23) can be written in terms of the whitened datasets in (4.2), we can write this EVD in a deflationary way and compare it with the deflationary maxvar EVD in (3.28). By doing so, we find the connection between the EVDs and, thereby, also the connection between the methods. As the maxvar method [Ket71] and the sumcor method [Nie95] are formulated using random vector notation, we also use this notation to state their connection in Theorem 4.2.

Theorem 4.2. *The transformation matrices, and therefore the canonical variables, of the all-at-once analytical sumcor method [Nie95] and the deflationary analytical maxvar method [Ket71] are different but closely related. More precisely, let*

$$\mathbf{V}^{[k]} = \begin{bmatrix} \mathbf{v}_1^{[k]} & \cdots & \mathbf{v}_N^{[k]} \end{bmatrix} \in \mathbb{R}^{N \times N}$$

be the transformation matrix for the whitened datasets

$$\mathbf{y}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]} \right)^{-\frac{1}{2}} \mathbf{x}^{[k]} \in \mathbb{R}^N$$

such that

$$\mathbf{u}^{[k]} = \left(\mathbf{V}^{[k]} \right)^{\mathsf{T}} \mathbf{y}^{[k]}, \quad k = 1, \ldots, K.$$

Then, for maxvar [Ket71], the transformation vectors $\mathbf{v}_n^{[k]}$ *for all datasets are orthogonal, i.e.,*

$$\left(\mathbf{v}_n^{[k]} \right)^{\mathsf{T}} \mathbf{v}_m^{[k]} = 0, \quad m \neq n, \quad m = 1, \ldots, N, \quad k = 1, \ldots, K,$$

while for sumcor [Nie95], the concatenated transformation vectors

$$\mathbf{v}_n = \begin{bmatrix} \mathbf{v}_n^{[1]} \\ \vdots \\ \mathbf{v}_n^{[K]} \end{bmatrix} \in \mathbb{R}^{NK}$$

are orthogonal, i.e.,

$$\mathbf{v}_n^{\mathsf{T}}\mathbf{v}_m = \sum_{k=1}^{K}\left(\mathbf{v}_n^{[k]}\right)^{\mathsf{T}}\mathbf{v}_m^{[k]} = 0, \quad m \neq n, \quad n, m = 1, \ldots, N.$$

The proof of Theorem 4.2 consists of two steps:

1. Write the sumcor EVD in (3.23) depending on the whitened datasets $\mathbf{y}^{[k]}$ in a deflationary way.

2. Compare the deflationary sumcor EVD with the deflationary maxvar EVD in (3.28).

Proof. 1) Let

$$\mathbf{y}^{[k]} = \left(\mathbf{C}_{\mathbf{x}}^{[k]}\right)^{-\frac{1}{2}} \in \mathbb{R}^N$$

be the k^{th} whitened dataset, and let

$$\mathbf{C}_{\mathbf{y}} = \mathbb{E}\left\{\mathbf{y}\mathbf{y}^{\mathsf{T}}\right\} \in \mathbb{R}^{NK\times NK}$$

be the covariance matrix of the concatenated whitened datasets

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}^{[1]}\\ \vdots \\ \mathbf{y}^{[K]}\end{bmatrix} \in \mathrm{R}^{NK}.$$

As in (4.2), the sumcor EVD can be written as

$$\mathbf{C}_{\mathbf{y}}\mathbf{B} = \mathbf{\Phi}\mathbf{B},$$

where

$$\mathbf{B} = \begin{bmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_{NK}\end{bmatrix} \in \mathbb{R}^{NK\times NK}$$

contains the eigenvectors of $\mathbf{C}_{\mathbf{y}}$, and

$$\mathbf{b}_n = \begin{bmatrix}\mathbf{b}_n^{[1]}\\ \vdots \\ \mathbf{b}_n^{[K]}\end{bmatrix} \in \mathbb{R}^{NK}, \quad n = 1, \ldots, N$$

contains the transformation vectors $\mathbf{b}_n^{[k]}$ for the k^{th} whitened dataset. Now we define

$$\mathbf{B}_{(n-1)} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_{n-1} \end{bmatrix} \in \mathbb{R}^{NK \times (n-1)}$$

and $\mathbf{H}_n$ as [Sta09]

$$\mathbf{H}_n = \mathbf{I} - \mathbf{B}_{(n-1)} \left(\mathbf{B}_{(n-1)}^{\mathsf{T}} \mathbf{B}_{(n-1)} \right)^{-1} \mathbf{B}_{(n-1)}^{\mathsf{T}} \in \mathbb{R}^{NK \times NK}.$$

Then, $\mathbf{H}_n \mathbf{C}_{\mathbf{y}}$ is the projection of $\mathbf{C}_{\mathbf{y}}$ on the subspace that is orthogonal to $\mathbf{b}_1, \ldots, \mathbf{b}_{n-1}$ [Sta09]. Let us denote the principal eigenvector of $\mathbf{H}_n \mathbf{C}_{\mathbf{y}}$ by $\mathbf{v}_n$. Now, it is obvious to see that $\mathbf{v}_n = \mathbf{b}_n$, i.e., that the principal eigenvector of $\mathbf{H}_n \mathbf{C}_{\mathbf{y}}$ equals the n^{th} eigenvector of $\mathbf{C}_{\mathbf{y}}$. As $\mathbf{H}_n$ is symmetric and idempotent, the eigenvectors of $\mathbf{H}_n \mathbf{C}_{\mathbf{y}}$ are equal to those of $\mathbf{H}_n \mathbf{C}_{\mathbf{y}} \mathbf{H}_n$ [Ket71]. Thus, the all-at-once sumcor EVD in (4.2) can also be performed in a deflationary way by performing the following EVD sequentially for $n = 1, \ldots, N$:

$$\mathbf{H}_n \mathbf{C}_{\mathbf{y}} \mathbf{H}_n \mathbf{v}_n = \phi_n \mathbf{v}_n,$$

where ϕ_n denotes the largest eigenvalue of $\mathbf{H}_n \mathbf{C}_{\mathbf{y}} \mathbf{H}_n$,

$$\mathbf{v}_n = \begin{bmatrix} \mathbf{v}_n^{[1]} \\ \vdots \\ \mathbf{v}_n^{[K]} \end{bmatrix} \in \mathbb{R}^{NK}$$

denotes the corresponding eigenvector, and $\mathbf{H}_n$ is calculated as follows. In the first EVD, $\mathbf{H}_1 = \mathbf{I}_{NK \times NK}$, and then,

$$\mathbf{H}_n = \mathbf{I} - \mathbf{V}_{(n-1)} \left(\mathbf{V}_{(n-1)}^{\mathsf{T}} \mathbf{V}_{(n-1)} \right)^{-1} \mathbf{V}_{(n-1)}^{\mathsf{T}}, \quad n = 2, \ldots, N,$$

where

$$\mathbf{V}_{(n-1)} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^{[1]} & \cdots & \mathbf{v}_{n-1}^{[1]} \\ \vdots & & \vdots \\ \mathbf{v}_1^{[K]} & \cdots & \mathbf{v}_{n-1}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times (n-1)}$$

consists of the eigenvectors of the $n-1$ previous EVDs.

2) In the deflationary maxvar EVD in (3.28), also given by $\mathbf{H}_n \mathbf{C}_{\mathbf{y}} \mathbf{H}_n \mathbf{v}_n = \phi_n \mathbf{v}_n$, the definition of $\mathbf{H}_n$ is different:

$$\mathbf{H}_n = \mathbf{I} - \widetilde{\mathbf{V}}_{(n-1)} \left(\widetilde{\mathbf{V}}_{(n-1)}^{\mathsf{T}} \widetilde{\mathbf{V}}_{(n-1)} \right)^{-1} \widetilde{\mathbf{V}}_{(n-1)}^{\mathsf{T}},$$

where

$$\widetilde{\mathbf{V}}_{(n-1)} = \begin{bmatrix} \widetilde{\mathbf{v}}_1^{[1]} & \cdots & \widetilde{\mathbf{v}}_{n-1}^{[1]} & & & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \ddots & & \vdots & & \vdots \\ \vdots & & \vdots & & \ddots & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & & & \widetilde{\mathbf{v}}_1^{[K]} & \cdots & \widetilde{\mathbf{v}}_{n-1}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK \times (n-1)K}$$

consists of the partitioned and normalized eigenvectors

$$\widetilde{\mathbf{v}}_n^{[k]} = \frac{\mathbf{v}_n^{[k]}}{\left\| \mathbf{v}_n^{[k]} \right\|} \in \mathbb{R}^N$$

of the $n-1$ previous EVDs. While the first eigenvectors $\mathbf{v}_1^{[k]}$ from Kettenring are the same as $\mathbf{v}_1^{[k]}$ from Nielsen, $k = 1, \ldots, K$, all following eigenvectors are different because the matrix $\widetilde{\mathbf{V}}_{(n-1)}$ is built differently than $\mathbf{V}_{(n-1)}$: Instead of concatenating the $n-1$ previous eigenvectors horizontally, Kettenring partitions the eigenvectors according to the datasets they belong to, normalizes these partitions, and then stacks all normalized partitions into a block-diagonal matrix. As a result, in sumcor [Nie95], the $\mathbf{v}_n$, $\mathbf{v}_m$ are orthogonal, i.e., $\mathbf{v}_n^{\mathsf{T}} \mathbf{v}_m = \sum_{k=1}^{K} \left(\mathbf{v}_n^{[k]} \right)^{\mathsf{T}} \mathbf{v}_m^{[k]} = 0$, while in maxvar [Ket71], the $\widetilde{\mathbf{v}}_n^{[k]}$, $\widetilde{\mathbf{v}}_m^{[k]}$ are orthogonal, i.e., $\left(\widetilde{\mathbf{v}}_n^{[k]} \right)^{\mathsf{T}} \widetilde{\mathbf{v}}_m^{[k]} = 0$, $k = 1, \ldots, K$. □

4.3 Connection between PARAFAC2 and IVA-G

In this section, we establish the connection between the PARAFAC2 and IVA-G models. Consider the PARAFAC2 model, shown in Figure 4.1, which we repeat here for better readability:

$$\begin{aligned} \mathbf{X}^{[k]} &\approx \mathbf{A} \operatorname{diag}\left(\mathbf{c}_{k:}\right) \left(\mathbf{B}^{[k]} \right)^{\mathsf{T}} \\ &\text{s.t. } \left(\mathbf{B}^{[k]} \right)^{\mathsf{T}} \mathbf{B}^{[k]} = \mathbf{M}, \quad k = 1, \ldots, K, \end{aligned} \tag{4.4}$$

with $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $\mathbf{A} \in \mathbb{R}^{N \times R}$, $\mathbf{c}_{k:} \in \mathbb{R}^{1 \times R}$, $\left(\mathbf{B}^{[k]} \right)^{\mathsf{T}} = \begin{bmatrix} \left(\mathbf{b}_1^{[k]} \right)^{\mathsf{T}} \\ \vdots \\ \left(\mathbf{b}_R^{[k]} \right)^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{R \times V}$, and $\mathbf{M} \in \mathbb{R}^{R \times R}$ is an arbitrary matrix. We see that PARAFAC2 represents the datasets through R components, where R is typically smaller than N, and therefore indirectly performs a dimension reduction on the datasets. Furthermore, the PARAFAC2

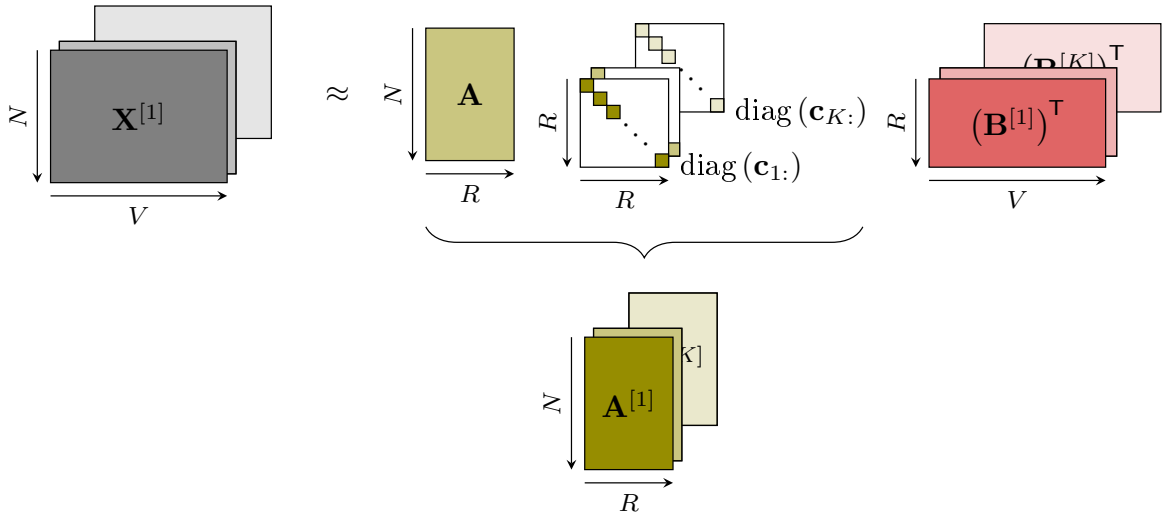


Figure 4.1: Illustration of an R-component PARAFAC2 model. The $\mathbf{A}^{[k]}$ are more constrained than in IVA-G; however, the $\mathbf{B}^{[k]}$ are less constrained by only fulfilling $\left(\mathbf{B}^{[k]}\right)^{\mathsf{T}} \mathbf{B}^{[k]} = \mathbf{M}$.

formulation considers observed datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \ldots, K$, where V is the number of samples, instead of random vectors $\mathbf{x}^{[k]} \in \mathbb{R}^{N}$. Therefore, we compare the PARAFAC2 model with the IVA-G model for dimension-reduced observed datasets, which we derive in the following.

First, dimension reduction via Principal Component Analysis (PCA) is performed on the observed datasets $\mathbf{X}^{[k]}$ to get the dimension-reduced observed datasets

$$\widetilde{\mathbf{X}}^{[k]} = \left(\mathbf{\Psi}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]} \in \mathbb{R}^{R \times V}, \tag{4.5}$$

where $\mathbf{\Psi}^{[k]} \in \mathbb{R}^{N \times R}$ is the matrix consisting of the $R < N$ principal eigenvectors of the estimated covariance matrix of the k^{th} observed dataset,

$$\widehat{\mathbf{C}}_{\mathbf{x}}^{[k]} = \frac{1}{V} \mathbf{X}^{[k]} \left(\mathbf{X}^{[k]}\right)^{\mathsf{T}} \in \mathbb{R}^{N \times N}.$$

Now, $\widetilde{\mathbf{X}}^{[k]}$ is assumed to be generated according to the model

$$\widetilde{\mathbf{X}}^{[k]} = \widetilde{\mathbf{A}}^{[k]} \mathbf{S}^{[k]}, \tag{4.6}$$

where $\mathbf{S}^{[k]} \in \mathbb{R}^{R \times V}$ consists of the sample versions of the R source components, and

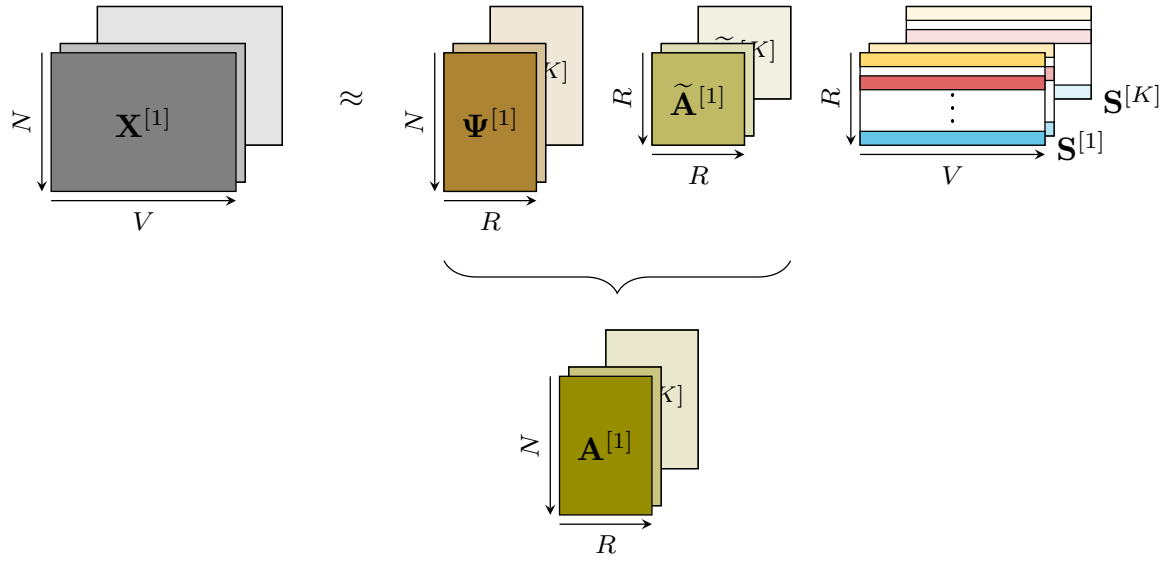


Figure 4.2: Illustration of the PCA-IVA-G model with reduced dimension R.

$\widetilde{\mathbf{A}}^{[k]} \in \mathbb{R}^{R\times R}$ is the mixing matrix. Then,

$$\widehat{\mathbf{S}}^{[k]} = \widetilde{\mathbf{W}}^{[k]}\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R\times V},$$

where $\widetilde{\mathbf{W}}^{[k]} \in \mathbb{R}^{R\times R}$ is estimated by applying IVA-G to $\widetilde{\mathbf{X}}^{[k]}$. Finally, by approximating the original data from the dimension-reduced datasets as

$$\mathbf{X}^{[k]} \approx \mathbf{\Psi}^{[k]}\widetilde{\mathbf{X}}^{[k]}, \tag{4.7}$$

where we misused notation and denoted the approximated dataset the same as the original dataset, and inserting (4.6) in (4.7), the PCA-IVA-G model is found as

$$\mathbf{X}^{[k]} \approx \underbrace{\mathbf{\Psi}^{[k]}\widetilde{\mathbf{A}}^{[k]}}_{\mathbf{A}^{[k]}}\mathbf{S}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}, \tag{4.8}$$

where $\mathbf{A}^{[k]} \in \mathbb{R}^{N\times R}$ is the non-square mixing matrix for the k^{th} dataset. The PCA-IVA-G model is shown in Figure 4.2.

Comparing (4.4) and (4.8), we see that the PCA-IVA-G and PARAFAC2 models are equivalent when $\mathbf{A}^{[k]} = \mathbf{A}\text{diag}(\mathbf{c}_{k:})$ and $\mathbf{S}^{[k]} = \left(\mathbf{B}^{[k]}\right)^{\mathsf{T}}$. Similar to the r^{th} SCV $\mathbf{S}_r$ in

PCA-IVA-G, which consists of the r^{th} row $\mathbf{s}_{r:}^{[k]}$ of all $\mathbf{S}^{[k]}$, i.e.,

$$\mathbf{S}_r = \begin{bmatrix} \mathbf{s}_{r:}^{[1]} \\ \vdots \\ \mathbf{s}_{r:}^{[K]} \end{bmatrix} \in \mathbb{R}^{K \times V},$$

we can define the r^{th} SCV in PARAFAC2 by concatenating the r^{th} row of all $\left(\mathbf{B}^{[k]}\right)^{\mathsf{T}}$ as

$$\mathbf{B}_r = \begin{bmatrix} \left(\mathbf{b}_r^{[1]}\right)^{\mathsf{T}} \\ \vdots \\ \left(\mathbf{b}_r^{[K]}\right)^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{K \times V}.$$

However, each method has different cost functions and hence different assumptions, i.e., PARAFAC2 is more constrained in terms of $\mathbf{A}^{[k]}$, while PCA-IVA-G is more constrained in terms of $\mathbf{S}^{[k]}$. If $\mathbf{M} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, the assumptions of the two models become more similar as both assume uncorrelatedness of the sources. While PARAFAC2 may have an upper limit on the number of components to uniquely estimate, which is $K \geq R(R+1)(R+2)(R+3)/24$ [Kie99], typical values for R in PCA-IVA-G are significantly higher. This higher value of R allows for a more detailed decomposition using PCA-IVA-G that is easier to interpret, as order selection is highly linked to the interpretability of the source components [Ada22]. However, as PARAFAC2 has a smaller set of parameters to estimate than PCA-IVA-G, it is likely to be more robust against noise. In the following subsections, we demonstrate the differences between PCA-IVA-G and PARAFAC2 in simulations.

4.3.1 Simulation setup

We simulate a tensor of dimension $N = 300$ subjects, $V = 5000$ voxels, and $K = 12$ datasets, i.e., $\mathbf{X}^{[k]} \in \mathbb{R}^{300 \times 5000}$, $k = 1, \ldots, 12$. The number and dimension of the observed datasets are similar to those of the real-world fMRI datasets described in Section 7.1. A typical assumption in fMRI analysis is that the observed datasets can be modeled with a lower-dimensional set of latent sources [Ada22], and therefore, we model only $R = 4$ underlying source components, i.e., $\mathbf{S}^{[k]} = \left(\mathbf{B}^{[k]}\right)^{\mathsf{T}} \in \mathbb{R}^{4 \times 5000}$. Then, the observed data can be generated as

$$\mathbf{X}^{[k]} = \mathbf{A}^{[k]} \mathbf{S}^{[k]},$$

where $\mathbf{A}^{[k]} \in \mathbb{R}^{300\times 5}$ is the non-square mixing matrix.[2]

We compare PCA-IVA-G and PARAFAC2 in two different scenarios. In the first, $\mathbf{A}^{[k]}$ is modeled as in (4.4), i.e.,

$$\mathbf{A}^{[k]} = \mathbf{A}\,\mathrm{diag}\left(\mathbf{c}_{k:}\right), \quad k = 1, \ldots, K, \tag{4.9}$$

but in the second, (4.9) is violated. For the *subjects mode* in scenario 1, the elements of the first and third column of $\mathbf{A} \in \mathbb{R}^{300\times 4}$ are drawn from the normal distribution $\mathcal{N}(0, 1)$, and then 0.5 is added to the last 150 elements of these columns, i.e., $\mathbf{a}_1(151 : 300)$ and $\mathbf{a}_3(151 : 300)$, to simulate a difference between patients and controls. The second and fourth column are also distributed normally, with a mean of zero and standard deviation as the average of the standard deviations of $\mathbf{a}_1$ and $\mathbf{a}_3$, i.e., $\sigma_2 = \sigma_4 = \frac{\sigma_1+\sigma_3}{2}$. Then, the $\mathbf{A}^{[k]}$ are calculated as in (4.9). In scenario 2, the $\mathbf{A}^{[k]}$ are calculated as before, but then, we subtract 0.5 from $\mathbf{a}_3(151 : 300)$, and the third column of $\mathbf{A}^{[k]}$ is recalculated for the last eight datasets, i.e., $\mathbf{a}^{[k]} = c_{k3}\mathbf{a}_3^{(\text{new})}$, $k = 5, \ldots, 8$. This way, $\mathbf{A}^{[k]} \neq \mathbf{A}\mathrm{diag}(\mathbf{c}_{k:})$ for the third component. In both scenarios, for the *voxels mode*, the sources $\mathbf{S}^{[k]} = \left(\mathbf{B}^{[k]}\right)^{\mathsf{T}} \in \mathbb{R}^{4\times 5000}$ are generated using a multivariate Gaussian distribution with the covariance matrix $\mathbf{C}_{\mathbf{s}_n}$ for each SCV shown in Figure 4.3. Note that $\mathbf{C}_{\mathbf{s}_n}$ denotes the covariance matrices and $\mathbf{C}$ the factor matrix. These covariance matrices are designed to have similar values for the correlations between sources like those of the fMRI datasets in Section 7.2. We see that in $\mathbf{C}_{\mathbf{s}_4}$, all source components are uncorrelated; thus, there is no dependence that IVA-G can leverage for this SCV. For the *task/datasets mode*, the factor matrix $\mathbf{C} \in \mathbb{R}^{12\times 4}$ is also generated equally for both scenarios. The elements of the first three columns $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ are drawn from $\mathcal{N}(1.5, 0.01)$. Then, from the elements corresponding to the first four datasets in $\mathbf{c}_2$, i.e., $\mathbf{c}_2(1 : 4)$, and the last eight datasets in $\mathbf{c}_3$, i.e., $\mathbf{c}_3(5 : 12)$, we subtract one to simulate that the component is not present in these datasets. The elements of $\mathbf{c}_4$ are drawn from $\mathcal{N}(1.5, 0.25)$ to make sure that the columns of $\mathbf{C}$ are not too similar.

4.3.2 Performance evaluation

The SCVs in IVA-G are sorted according to their determinant to match the order of the true SCVs. For PARAFAC2, non-negativity constraints are imposed on the *task/datasets mode* to solve the potential sign indeterminacy per component [Har72]. Then, the correct permutation of the components is found by maximizing the cosine similarity of the concatenated true and estimated $\mathbf{B}^{[k]}$ matrices. To test for statistically

[2]The Python code for the simulations is available at: https://github.com/SSTGroup/ICASSP2022.

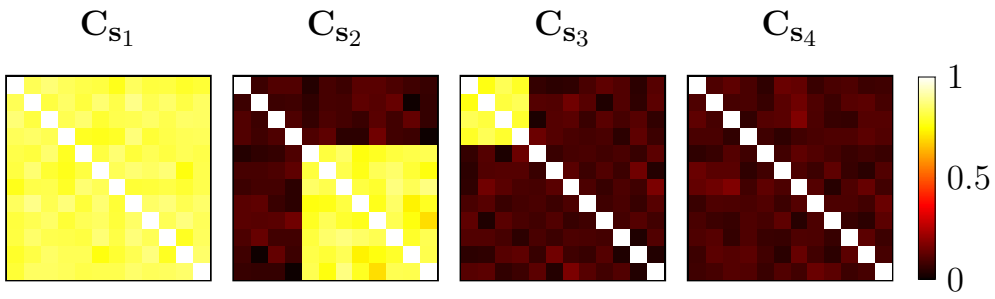


Figure 4.3: Covariance matrices (12×12) of simulated sources for each SCV, which are estimated reliably by both models (not shown).

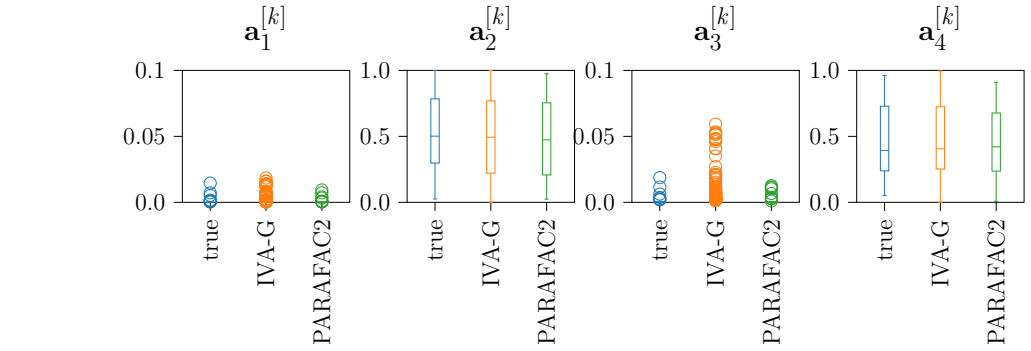
significant difference ($p < 0.05$) between patients and controls (first vs. second half of each column in $\mathbf{A}^{[k]}$), a two-sample t-test (unequal variances) is applied on each column of $\mathbf{A}^{[k]}$, and the corresponding p-values are calculated. The distributions of the true and estimated p-values are compared in boxplots. Furthermore, the average correlation of the true and estimated sources across datasets for each SCV is calculated.

4.3.3 Results of the simulated data analysis

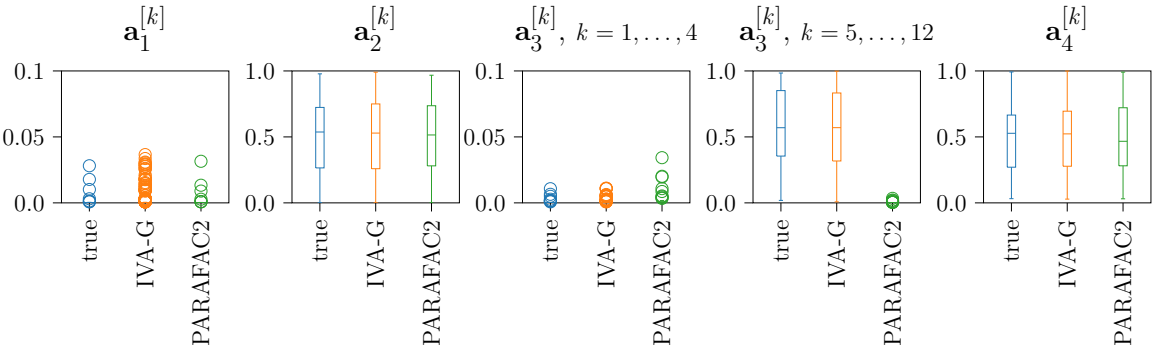
For PCA-IVA-G, the dimension of the simulated datasets is reduced to $R = 4$, as we have simulated $R = 4$ underlying source components, and IVA-G is applied to the dimension-reduced datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$. PARAFAC2 is applied to the simulated datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$ to estimate $R = 4$ components, with non-negativity constraints imposed on $\mathbf{C}$. Both IVA-G and PARAFAC2 are run for multiple initializations, and the results for the most consistent run [Lon18] in IVA-G[3] and for the smallest reconstruction error in PARAFAC2[4] are reported. The results for the simulated data analysis, averaged across 50 independent Monte-Carlo runs, show that IVA-G and PARAFAC2 both achieve an almost perfect correlation (> 0.97) between the true and estimated sources. Thus, PARAFAC2 is able to estimate $\mathbf{B}^{[k]}$ correctly, i.e., to achieve JBSS, despite the assumption on $\mathbf{A}^{[k]}$ being violated. Figure 4.4 shows the distributions of the true and estimated p-values for both scenarios. As $\mathbf{A}$ is the same for all datasets in scenario 1, the true p-value for all datasets is also the same, while in scenario 2, $\mathbf{A}$ is different for the first four and last eight datasets in component 3 and therefore, we show the distribution of the p-value for the first four and last eight datasets of $\mathbf{a}_3^{[k]}$ separately. We see in Figure 4.4(a) that for scenario 1, component 3, IVA-G overestimates the p-value for some datasets, while PARAFAC2 estimates the p-value small enough matching the

[3]The Python implementation is available at: https://github.com/SSTGroup/independent_vector_analysis.

[4]The Python implementation is available at: https://github.com/tensorly/tensorly [Roa20; Kos19b].



(a) Scenario 1



(b) Scenario 2

Figure 4.4: Distribution of the true and estimated p-values in $\mathbf{A}^{[k]}$.

ground truth. However, for scenario 2, we see in Figure 4.4(b) that IVA-G can capture the different p-values in component 3 due to its more flexible model design in terms of $\mathbf{A}^{[k]}$, while PARAFAC2 cannot capture the differences and finds the same p-value for all datasets because it assumes the same mixing matrix.

4.4 Summary

In this chapter, we have established the connections between an important set of JBSS methods. First, we have derived the connection between the two perhaps most commonly used mCCA methods sumcor and maxvar, proving that if the canonical variables are normalized to unit variance, those estimated by the all-at-once analytical sumcor method [Nie95] are the same as those estimated by the all-at-once analytical maxvar method [Car68]. Furthermore, we have shown that the canonical variables estimated by the all-at-once analytical sumcor method [Nie95] and by the deflationary analytical maxvar method [Ket71] are closely related, with their main difference being that the canonical variables within a dataset are constrained to be uncorrelated in maxvar but

not in sumcor. Finally, we have shown that the PCA-IVA-G and PARAFAC2 models are intimately related. We show the differences between the two methods in simulations inspired by real-world fMRI data, which reveal that while both methods can accurately capture the underlying source components, i.e., achieve JBSS, PARAFAC2 captures group differences more reliably, providing a compact representation when subject scores differ only up to a scaling across different datasets, and IVA-G performs better when different subject scores are expected in different datasets.

To be able to also connect the mCCA methods with IVA and PARAFAC2, a next step is to formally derive the connection between mCCA-genvar and a deflationary version of IVA-G with orthogonality constraint, which have similar cost functions [And10]. A more detailed comparison of IVA using its different versions (not only IVA-G but also IVA-L-SOS) and PARAFAC2 using different constraints could further complete the comparison between JBSS methods. Having a complete picture of the connections between these methods and their performance in JBSS in different scenarios would help to ultimately solve the challenge of choosing the right JBSS method for a specific application. In order to move another step forward towards this goal, in the next chapter, we discuss the conditions under which the methods are able to achieve JBSS.

5 Source identification conditions for BSS and JBSS

This chapter is dedicated to summarizing the source identification conditions of matrix decomposition methods for BSS and JBSS as defined in Definitions 2.1 and 2.2, i.e., the conditions on the true source components under which a method is able to achieve BSS or JBSS. Knowing these conditions is important because without having guarantees to identify the true sources, it would not make sense to interpret them, i.e., to attach physical meaning to them [Ada22]. Of course, for real-world data without knowledge of the ground-truth sources, one often cannot verify if the conditions of a specific method are satisfied, but at least one should be able to check if it is reasonable to assume this, or if another method should be chosen instead. This knowledge on the source identification conditions, together with our insights from Chapter 4 on the connections of the methods, will ultimately help to solve the challenge of choosing a JBSS method for a specific application.

To simplify the formulations of the source identification conditions, we make the following assumption throughout the whole chapter:

- All source components within an SCV are sufficiently correlated such that its covariance matrix cannot be partitioned into a (permuted) block-diagonal matrix. (JBSS)

If Assumption (JBSS) does not hold, the conditions to identify the true source components can be extended to account for block-diagonal covariance matrices using so-called α-Gaussian SCVs, similar to the development in [And14]. However, for block-diagonal matrices, methods may not be able to correctly align sources across datasets.[1]

[1] Section 5.4 of this chapter is based on the paper: "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, *submitted for review*, 2025. I specifically contributed to proving the source identification conditions of sumcor, deriving the connection between maxvar [Ket71] and sumcor [Nie95], implementing and running the experiments, creating all figures, and writing the paper. B. Gabrielson has provided theoretical insights in the connection between maxvar [Car68] and sumcor [Nie95] and the source identification conditions of genvar. Discussions and feedback along the way until the final paper have contributions from all authors.

5.1 Source identification conditions of ICA for BSS ($K = 1$ dataset)

The source identification conditions of ICA for BSS depend on the ICA method. In Theorem 5.1, we present the source identification conditions of FastICA, which are stated in [Hyv00].

Theorem 5.1 (Source identification conditions of FastICA from [Hyv00])**.** *We assume Assumptions* (ICA.1)*,* (ICA.2)*, and* (ICA.3) *from page 17. Let*

$$\mathbf{s} = \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix}$$

be the true source vector. If at most one of the source components s_n has a Gaussian distribution, then FastICA [Hyv00] achieves BSS, i.e., the estimated source vector $\widehat{\mathbf{s}}$ recovers the true source vector $\mathbf{s}$ up to sign and permutation:

$$\widehat{\mathbf{s}} = \mathbf{P}\widetilde{\mathbf{\Gamma}}\mathbf{s},$$

where $\mathbf{P} \in \mathbb{R}^{N \times N}$ is a permutation matrix and $\widetilde{\mathbf{\Gamma}} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with diagonal elements equal to either -1 or 1.

5.2 Source identification conditions of IVA for JBSS ($K \geq 2$ datasets)

The conditions under which IVA can identify all sources up to a scaling and an *individual* permutation for each dataset are stated in [And14; Ada14]. These conditions do not imply that the source components are also correctly aligned across datasets, i.e., that the permutation matrix is the same for all datasets. In order for IVA to also align the identified sources correctly, i.e., to recover the true sources up to a scaling and a *common* permutation for all datasets, there must not exist two SCVs that both contain a subset of source components that is independent of the other source components in that same SCV [And14]. This condition is included in Assumption (JBSS) from page 69, as the covariance matrix of such SCVs would be a (permuted) block-diagonal matrix. In the next two subsections, we present the source identification conditions of IVA-G and IVA-L-SOS for JBSS.

5.2.1 Source identification conditions of IVA-G

In IVA-G, the SCVs are assumed to have a multivariate Gaussian distribution, and JBSS is therefore achieved by maximizing uncorrelatedness of the SCVs instead of statistical independence. Therefore, Assumption (IVA.3) on page 20 reduces for IVA-G to:

- Source components are uncorrelated across SCVs (but can be correlated within an SCV), i.e.,

$$\mathbb{E}\left\{\mathbf{s}_n\mathbf{s}_m^{\mathsf{T}}\right\} = \mathbf{0}_{K\times K} \quad m \neq n, \quad n, m = 1, \ldots, N. \tag{IVA.3b}$$

The source identification conditions of IVA-G are derived and proven in [And12] and presented in Theorem 5.2.

Theorem 5.2 (Source identification conditions of IVA-G [And12])**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b)*, and Assumption* (JBSS) *from page 69. Let* $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$ *be the covariance matrix of the* n^{th} *SCV* $\mathbf{s}_n \in \mathbb{R}^K$*. If the following condition holds for this SCV* $\mathbf{s}_n$*:*
$\mathbf{C}_{\mathbf{s}_n}$ *is distinct from all other* $\mathbf{C}_{\mathbf{s}_m}$ *by more than a sign, i.e.,*

$$\mathbf{C}_{\mathbf{s}_n} \neq \mathbf{D}_m\mathbf{C}_{\mathbf{s}_m}\mathbf{D}_m, \quad m \neq n, \quad m = 1, \ldots, N,$$

where $\mathbf{D}_m \in \mathbb{R}^{K\times K}$ *is a diagonal matrix with diagonal elements equal to either* -1 *or* 1*, then this SCV* $\mathbf{s}_n$ *can be identified by IVA-G [And12]. If this condition holds for all* $n = 1, \ldots, N$*, then IVA-G achieves JBSS, i.e., the estimated source vectors* $\widehat{\mathbf{s}}^{[k]}$ *recover the true source vectors* $\mathbf{s}^{[k]}$ *up to a sign and a common permutation:*

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{P}\widetilde{\mathbf{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ *is a common permutation matrix and* $\widetilde{\mathbf{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ *are diagonal matrices with diagonal elements equal to either* -1 *or* 1*.*

Note that if only a few elements of $\mathbf{C}_{\mathbf{s}_n}$ and $\mathbf{D}_m\mathbf{C}_{\mathbf{s}_m}\mathbf{D}_m$ are different, the condition $\mathbf{C}_{\mathbf{s}_n} \neq \mathbf{D}_m\mathbf{C}_{\mathbf{s}_m}\mathbf{D}_m$ is almost violated. When more elements are different, the condition is better satisfied, and IVA-G can better achieve JBSS.

5.2.2 Source identification conditions of IVA-L-SOS

In IVA-L-SOS, the assumptions on the SCVs are as stated in Section 3.2, as IVA-L-SOS also considers higher-order statistics. The source identification conditions of IVA-L-SOS are derived and proven in [And14] and presented in Theorem 5.3.

Theorem 5.3 (Source identification conditions of IVA-L-SOS [And14])**.** *We assume Assumptions* (IVA.1)*,* (IVA.2)*, and* (IVA.3) *from page 21, and Assumption* (JBSS) *from page 69. Let* $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$ *be the covariance matrix of the* n^{th} *SCV* $\mathbf{s}_n \in \mathbb{R}^K$*. If one of the following two conditions holds for this SCV* $\mathbf{s}_n$*:*

1. *the assumed distribution of the estimated SCV* $\widehat{\mathbf{s}}_n$ *matches the distribution of the true SCV* $\mathbf{s}_n$*, i.e.,*

$$p_{\widehat{\mathbf{s}}_n}(\widehat{\mathbf{s}}_n) = p_{\mathbf{s}_n}(\mathbf{s}_n) = \frac{\Gamma\left(\frac{K}{2}\right)\left(\det(\boldsymbol{\Sigma}_n)\right)^{-\frac{1}{2}}}{2^{K+1}\pi^{K/2}\Gamma(K)} \exp\left(-\frac{1}{2}\sqrt{\mathbf{s}_n^{\mathsf{T}}\boldsymbol{\Sigma}_n^{-1}\mathbf{s}_n}\right),$$

   where $\Gamma(\cdot)$ *is the Gamma function and* $\boldsymbol{\Sigma}_n = \frac{1}{4(K+1)}\mathbf{C}_{\mathbf{s}_n}$*,*

2. *or* $\mathbf{C}_{\mathbf{s}_n}$ *is distinct from all other* $\mathbf{C}_{\mathbf{s}_m}$ *by more than a sign, i.e.,*

$$\mathbf{C}_{\mathbf{s}_n} \neq \mathbf{D}_m\mathbf{C}_{\mathbf{s}_m}\mathbf{D}_m, \quad m \neq n, \quad m = 1,\ldots,N,$$

   where $\mathbf{D}_m \in \mathbb{R}^{K\times K}$ *is a diagonal matrix with diagonal elements equal to either* -1 *or* 1*,*

then this SCV $\mathbf{s}_n$ *can be identified by IVA-L-SOS [Bhi19]. If one of these conditions holds for all* $n = 1,\ldots,N$*, then IVA-L-SOS achieves JBSS, i.e., the estimated source vectors* $\widehat{\mathbf{s}}^{[k]}$ *recover the true source vectors* $\mathbf{s}^{[k]}$ *up to a sign and a common permutation:*

$$\widehat{\mathbf{s}}^{[k]} = \mathbf{P}\widetilde{\boldsymbol{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1,\ldots,K,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ *is a common permutation matrix and* $\widetilde{\boldsymbol{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ *are diagonal matrices with diagonal elements equal to either* -1 *or* 1*.*

5.3 Source identification conditions of CCA for JBSS ($K = 2$ datasets)

By its formulation, CCA does not have a generative model. However, to investigate its ability to achieve JBSS, a generative model needs to be assumed for CCA. It is therefore

assumed in [Li08] that datasets $\mathbf{x}^{[k]} \in \mathbb{R}^N$ are generated according to the JBSS model in (2.2), given as $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}$, $k = 1, 2$. Since CCA achieves JBSS by maximizing correlations of canonical variables (not their higher-order statistics), the assumptions on the sources for CCA are the same as those for IVA-G, which minimizes correlations of the SCVs, i.e., considers only second-order statistics. The source identification conditions of CCA, which have been proven in [Li08], are presented in Theorem 5.4.

Theorem 5.4 (Source identification conditions of CCA from [Li08])**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b) *from page 71, and Assumption* (JBSS) *from page 69. Let*

$$\mathbf{C}_{\mathbf{s}_n} = \begin{bmatrix} 1 & \rho_n \\ \rho_n & 1 \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

be the covariance matrix of the n^{th} SCV $\mathbf{s}_n \in \mathbb{R}^2$, where ρ_n is the correlation of the n^{th} source component across the 2 datasets. If the following condition holds for this SCV $\mathbf{s}_n$:
the magnitude of the correlation $|\rho_n|$ is distinct from all other $|\rho_m|$, i.e.,

$$|\rho_n| \neq |\rho_m|, \quad m \neq n, \quad m = 1, \ldots, N,$$

then, this SCV $\mathbf{s}_n$ can be identified by CCA [Hot36]. If this condition holds for all $n = 1, \ldots, N$, then CCA achieves JBSS for $K = 2$ datasets, i.e., the canonical vectors $\mathbf{u}^{[k]}$ recover the true source vectors $\mathbf{s}^{[k]}$ up to a sign and a common permutation:

$$\mathbf{u}^{[k]} = \mathbf{P}\widetilde{\mathbf{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, 2,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a common permutation matrix and $\widetilde{\mathbf{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ are diagonal matrices with diagonal elements equal to either -1 or 1.

We can rewrite Theorem 5.4 in terms of the eigenvalues of $\mathbf{C}_{\mathbf{s}_n}$, as we know that

$$\lambda_n^{[1]} = 1 + \rho_n, \quad \lambda_n^{[2]} = 1 - \rho_n, \quad n = 1, \ldots, N.$$

Thus, if $|\rho_n| \neq |\rho_m|$, then $\lambda_n^{[1]} \neq \lambda_m^{[1]}$ and $\lambda_n^{[2]} \neq \lambda_m^{[2]}$, and we find Theorem 5.5 for the source identification conditions of CCA depending on the eigenvalues of $\mathbf{C}_{\mathbf{s}_n}$.

Theorem 5.5 (Source identification conditions of CCA from [Li08])**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b) *from page 71,*

and Assumption (JBSS) *from page 69. Let*

$$\boldsymbol{\lambda}_n = \begin{bmatrix} \lambda_n^{[1]} \\ \lambda_n^{[2]} \end{bmatrix}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{2\times 2}$*, the covariance matrix of the* n^{th} *SCV* $\mathbf{s}_n \in \mathbb{R}^2$*, sorted such that* $\lambda_n^{[1]} \geq \lambda_n^{[2]}$*. If the following condition holds for this SCV* $\mathbf{s}_n$*:*
the largest eigenvalue of $\mathbf{C}_{\mathbf{s}_n}$ *is distinct from the largest eigenvalue of all other* $\mathbf{C}_{\mathbf{s}_m}$*, i.e.,*

$$\lambda_n^{[1]} \neq \lambda_m^{[1]}, \quad m \neq n, \quad m = 1, \ldots, N,$$

then, this SCV $\mathbf{s}_n$ *can be identified by CCA [Hot36]. If this condition holds for all* $n = 1, \ldots, N$*, then CCA achieves JBSS for* $K = 2$ *datasets, i.e., the canonical vectors* $\mathbf{u}^{[k]}$ *recover the true source vectors* $\mathbf{s}^{[k]}$ *up to a sign and a common permutation:*

$$\mathbf{u}^{[k]} = \mathbf{P}\widetilde{\mathbf{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, 2,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ *is a common permutation matrix and* $\widetilde{\mathbf{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ *are diagonal matrices with diagonal elements equal to either* -1 *or* 1*.*

5.4 Source identification conditions of mCCA for JBSS ($K \geq 2$ datasets)

In this section, the source identification conditions of the different mCCA methods are presented. The source identification conditions of maxvar have been derived in [Li09]. We derive and prove the source identification conditions of sumcor, conjecture those of minvar, and discuss theoretical considerations for those of genvar and ssqcor. We substantiate the proposed theoretical conditions with simulations, test the statistical efficiency of the five mCCA methods for finite samples in experiments that are close to (but not exactly) violating their conditions, and investigate what influence the effective rank of the SCVs has on the methods' ability to achieve JBSS.

Like CCA, also mCCA does not have a generative model. To investigate its ability to achieve JBSS, we therefore assume again that datasets $\mathbf{x}^{[k]} \in \mathbb{R}^N$ are generated according to the JBSS model in (2.2), given as $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}$, $k = 1, \ldots, K$. Following the same logic as in CCA, the assumptions on the sources for mCCA are the same as

those for IVA-G.[2]

5.4.1 Source identification conditions of mCCA-sumcor

The sumcor transformation vectors are the partitions of the principal eigenvectors in the GEVD in (3.23). We can write this GEVD in terms of the concatenated SCVs' covariance matrix to find the relationship between the sumcor transformation vectors and the eigenvectors of the concatenated SCVs' covariance matrix. By doing so, we find the conditions that the eigenvalues of the SCV covariance matrices must satisfy such that the canonical vectors estimated by sumcor recover the true source vectors (up to permutation and scaling). These conditions are the source identification conditions of sumcor, which we propose in Theorem 5.6.

Theorem 5.6 (Source identification conditions of sumcor)**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b) *from page 71, and Assumption* (JBSS) *from page 69. Let*

$$\boldsymbol{\lambda}_n = \begin{bmatrix} \lambda_n^{[1]} \\ \vdots \\ \lambda_n^{[K]} \end{bmatrix}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$*, the covariance matrix of the* n^{th} *SCV* $\mathbf{s}_n \in \mathbb{R}^K$*, sorted such that* $\lambda_n^{[1]} \geq \cdots \geq \lambda_n^{[K]}$*. If the following two conditions hold for this SCV* $\mathbf{s}_n$*:*

1. *the largest eigenvalue of* $\mathbf{C}_{\mathbf{s}_n}$ *is distinct from the largest eigenvalue of all other* $\mathbf{C}_{\mathbf{s}_m}$*, i.e.,*
$$\lambda_n^{[1]} \neq \lambda_m^{[1]}, \quad m \neq n, \quad m = 1, \ldots, N,$$

2. *and the largest eigenvalue of this* $\mathbf{C}_{\mathbf{s}_n}$ *is greater than the second-largest (and all following) eigenvalues of all other* $\mathbf{C}_{\mathbf{s}_m}$*, i.e.,*
$$\lambda_n^{[1]} > \lambda_m^{[k]}, \quad k = 2, \ldots, K, \quad m \neq n, \quad m = 1, \ldots, N,$$

then this SCV $\mathbf{s}_n$ *can be identified by sumcor [Nie95]. If these conditions hold for all*

[2]This section is based on the paper: "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, *submitted for review*, 2025. I specifically contributed to proving the source identification conditions of sumcor, deriving the connection between maxvar [Ket71] and sumcor [Nie95], implementing and running the experiments, creating all figures, and writing the paper. B. Gabrielson has provided theoretical insights in the connection between maxvar [Car68] and sumcor [Nie95] and the source identification conditions of genvar. Discussions and feedback along the way until the final paper have contributions from all authors.

$n = 1, \ldots, N$, then sumcor achieves JBSS, i.e., the canonical vectors $\mathbf{u}^{[k]}$ recover the source vectors $\mathbf{s}^{[k]}$ up to a scaling and a common permutation:

$$\mathbf{u}^{[k]} = \mathbf{P}\mathbf{\Gamma}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

where $\mathbf{P} \in \mathbb{R}^{N \times N}$ is a common permutation matrix and $\mathbf{\Gamma}^{[k]} \in \mathbb{R}^{N \times N}$ are diagonal scale matrices.

The proof is given in Appendix 5.6. Note that the ambiguity here is up to a scaling and not only up to a sign, as the canonical variables estimated by sumcor [Nie95] are not constrained to have unit variance. However, if they are post-normalized to have unit variance, which is what we do in this thesis, they can recover the true sources up to a sign and a common permutation. The conditions in Theorem 5.6 also hold for the all-at-once analytical maxvar [Car68] since the estimated canonical variables are the same up to a scaling.

5.4.2 Source identification conditions of mCCA-maxvar

In Theorem 5.7, we present the source identification conditions of the deflationary analytical maxvar [Ket71], which are derived and proven in [Li09].

Theorem 5.7 (Source identification conditions of maxvar from [Li09])**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b) *from page 71, and Assumption* (JBSS) *from page 69. Let*

$$\boldsymbol{\lambda}_n = \begin{bmatrix} \lambda_n^{[1]} \\ \vdots \\ \lambda_n^{[K]} \end{bmatrix}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K \times K}$, the covariance matrix of the n^{th} SCV $\mathbf{s}_n \in \mathbb{R}^K$, sorted such that $\lambda_n^{[1]} \geq \cdots \geq \lambda_n^{[K]}$. If the following condition holds for this SCV $\mathbf{s}_n$: the largest eigenvalue of $\mathbf{C}_{\mathbf{s}_n}$ is distinct from the largest eigenvalue of all other $\mathbf{C}_{\mathbf{s}_m}$, i.e.,

$$\lambda_n^{[1]} \neq \lambda_m^{[1]}, \quad m \neq n, \quad m = 1, \ldots, N,$$

then this SCV $\mathbf{s}_n$ can be identified by maxvar [Ket71]. If this condition holds for all $n = 1, \ldots, N$, then maxvar achieves JBSS, i.e., the canonical vectors $\mathbf{u}^{[k]}$ recover the source vectors $\mathbf{s}^{[k]}$ up to a sign and a common permutation:

$$\mathbf{u}^{[k]} = \mathbf{P}\widetilde{\mathbf{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ *is a common permutation matrix and* $\widetilde{\mathbf{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ *are diagonal matrices with diagonal elements equal to either* -1 *or* 1*.*

These conditions are less restrictive than the source identification conditions of sumcor, which also depend on the second-largest eigenvalues.

5.4.3 Source identification conditions of mCCA-minvar

As minvar minimizes the smallest eigenvalue, while maxvar maximizes the largest eigenvalue, we conjecture that the source identification conditions of minvar are identical to those of maxvar but depend on the *smallest* eigenvalue instead. We present the conjectured source identification conditions of minvar in 5.1.

Conjecture 5.1 (Source identification conditions of minvar)**.** *We assume Assumptions* (IVA.1) *and* (IVA.2) *from page 21, Assumption* (IVA.3b) *from page 71, and Assumption* (JBSS) *from page 69. Let*

$$\boldsymbol{\lambda}_n = \begin{bmatrix} \lambda_n^{[1]} \\ \vdots \\ \lambda_n^{[K]} \end{bmatrix}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$*, the covariance matrix of the* n^{th} *SCV* $\mathbf{s}_n \in \mathbb{R}^K$*, sorted such that* $\lambda_n^{[1]} \geq \cdots \geq \lambda_n^{[K]}$*. If the following condition holds for this SCV* $\mathbf{s}_n$*: the smallest eigenvalue of* $\mathbf{C}_{\mathbf{s}_n}$ *is distinct from the smallest eigenvalue of all other* $\mathbf{C}_{\mathbf{s}_m}$*, i.e.,*

$$\lambda_n^{[K]} \neq \lambda_m^{[K]}, \quad m \neq n, \quad m = 1, \ldots, N,$$

then this SCV $\mathbf{s}_n$ *can be identified by minvar [Ket71]. If this condition holds for all* $n = 1, \ldots, N$*, then minvar achieves JBSS, i.e., the canonical vectors* $\mathbf{u}^{[k]}$ *recover the source vectors* $\mathbf{s}^{[k]}$ *up to a sign and a common permutation:*

$$\mathbf{u}^{[k]} = \mathbf{P}\widetilde{\mathbf{\Gamma}}^{[k]}\mathbf{s}^{[k]}, \quad k = 1, \ldots, K,$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ *is a common permutation matrix and* $\widetilde{\mathbf{\Gamma}}^{[k]} \in \mathbb{R}^{N\times N}$ *are diagonal matrices with diagonal elements equal to either* -1 *or* 1*.*

Note that these conditions are unlikely to be satisfied in real-world datasets, as the smallest eigenvalues in practice are typically very similar to each other.

5.4.4 Source identification conditions of mCCA-genvar

We discuss theoretical considerations for the source identification conditions of genvar using its connection to IVA-G. More precisely, suppose that IVA-G does not estimate the demixing vectors $\mathbf{w}_n^{[1]}, \ldots, \mathbf{w}_n^{[K]}$ for all $n = 1, \ldots, N$ simultaneously, but instead in a deflationary way. Furthermore, suppose that the demixing vectors $\mathbf{w}_n^{[k]}$ and $\mathbf{w}_m^{[k]}$ for different SCVs $n \neq m$ are constrained to be orthogonal for all $k = 1, \ldots, K$. Under these assumptions, $\ln\left(\left|\det\left(\mathbf{W}^{[k]}\right)\right|\right)$ is equal to zero, and the IVA-G optimization function in (3.8) would result in the following deflationary orthogonal IVA-G optimization function for the n^{th} SCV:

$$\begin{aligned} \mathcal{I}_{\text{defl. orth. IVA-G}}\left(\mathbf{w}_n^{[1]}; \ldots; \mathbf{w}_n^{[K]}\right) &= \ln\left(\det\left(\mathbf{C}_{\mathbf{s}_n}\right)\right) - \widetilde{C} \\ \text{s.t. } \left(\mathbf{w}_n^{[k]}\right)^{\mathsf{T}} \mathbf{w}_i^{[k]} &= 0, \quad i = 1, \ldots, n-1, \end{aligned} \tag{5.1}$$

where $\widetilde{C}$ is a constant. For unit-variance canonical variables and sources, it holds that $\det\left(\mathbf{C}_{\mathbf{u}_n}\right) = \det\left(\mathbf{C}_{\mathbf{s}_n}\right)$ if genvar achieves JBSS. Then, (5.1) is equal to the genvar optimization function in (F5) on page 36, given by $\det\left(\mathbf{C}_{\mathbf{u}_n}\right)$, with constraint (3.21) up to the natural logarithm and a constant. Since $\ln(\cdot)$ is a monotonic function, maximizing or minimizing the $\ln(\cdot)$ of some function $f(x)$ is respectively equivalent to maximizing or minimizing $f(x)$, and vice versa. Thus, since $\ln\left(\det\left(\mathbf{C}_{\mathbf{s}_n}\right)\right)$ has the same minima as $\det\left(\mathbf{C}_{\mathbf{s}_n}\right)$, it seems plausible to assume that the source identification conditions of genvar are similar to those of deflationary orthogonal IVA-G. The source identification conditions of deflationary orthogonal IVA-G are currently unknown and an interesting question for future work. As the source identification conditions of unconstrained IVA-G, given in Theorem 5.2, depend on the SCV covariance matrices and therefore both on their eigenvalues and eigenvectors, we expect that the source identification conditions of genvar also depend on the eigenvalues and eigenvectors of the SCV covariance matrices. This would mean that genvar can achieve JBSS even for equal eigenvalues across SCVs as long as the eigenvectors are different, and thereby seems to have the least restrictive source identification conditions of all mCCA methods.

5.4.5 Source identification conditions of mCCA-ssqcor

As the genvar and ssqcor methods are the only mCCA methods with objective functions defined over all eigenvalues of the covariance matrices $\mathbf{C}_{\mathbf{u}_n}$, we expect the source identification conditions of ssqcor to be similar to those of genvar, and therefore to also depend on the eigenvalues and eigenvectors of the SCV covariance matrices.

5.4.6 Simulation study of source identification conditions of mCCA

In this section, we substantiate the source identification conditions of the different mCCA methods with simulations. We start with an experiment where we violate the conditions of all methods and continue with experiments that substantiate the conditions of specific mCCA methods. As the methods have different underlying models and therefore assume different *effective ranks* for the SCVs, we end this simulation study with an experiment where we investigate the influence of the effective rank on the methods' ability to achieve JBSS.

In all experiments, we model the n^{th} SCVs' covariance matrix $\mathbf{C}_{\mathbf{s}_n} \in \mathbb{R}^{K\times K}$ as in [Gab24],

$$\mathbf{C}_{\mathbf{s}_n} = \alpha_n \mathbf{L}_n \mathbf{L}_n^{\mathsf{T}} + (1-\alpha_n)\mathbf{I}_{K\times K}, \tag{5.2}$$

where the elements of $\mathbf{L}_n \in \mathbb{R}^{K\times R}$ are drawn from $\mathcal{N}(0,1)$ and each row of $\mathbf{L}_n$ is normalized to unit-norm so that the diagonal elements of $\mathbf{C}_{\mathbf{s}_n}$ are equal to one, and $\frac{-1}{K-1} < \alpha_n \leq 1$. The matrix $\mathbf{L}_n\mathbf{L}_n^{\mathsf{T}}$ thus has rank R, the addition of the identity matrix is necessary to make $\mathbf{C}_{\mathbf{s}_n}$ have full rank, and $(1-\alpha_n)$ determines the proportion of the full-rank part of $\mathbf{C}_{\mathbf{s}_n}$. The model (5.2) corresponds to the SCV model [Gab24]

$$\mathbf{s}_n = \mathbf{M}_n \mathbf{f}_n + \boldsymbol{\nu}_n,$$

where $\mathbf{f}_n \in \mathbb{R}^R$ consists of the R common factors in the n^{th} SCV, $\mathbf{M}_n \in \mathbb{R}^{K\times R}$ weights the contribution of each factor (element of $\mathbf{f}_n$) to each source component (element of $\mathbf{s}_n$), and $\boldsymbol{\nu}_n \in \mathbb{R}^K$ is a noise vector with diagonal covariance matrix. This SCV model generalizes all mCCA factor models. If we choose $R = 1$, the covariance matrix $\mathbf{C}_{\mathbf{s}_n}$ in (5.2) has $K-1$ identical eigenvalues with the value $\lambda_n^{[k]} = 1-\alpha_n$, $k = 1,\ldots,K-1$, and one different eigenvalue with the value $\lambda_n^{[K]} = K-(K-1)(1-\alpha_n)$. Thus, to make the different eigenvalue $\lambda_n^{[K]} > 0$ have a specific value, we can choose

$$\alpha_n = 1 - \frac{K-\lambda_n^{[K]}}{K-1}. \tag{5.3}$$

Now, when testing the ability of the mCCA methods to achieve JBSS, it is important to note that estimation performance is determined not only by whether the sources satisfy the *source identification conditions* of the method but also by the method's *statistical efficiency*, i.e., how efficient the method makes use of observed data (with finite samples) when estimating the sources. A more statistically efficient estimator needs fewer samples to adequately estimate the sources. Since in this chapter we focus on

theoretical conditions, we need to decouple testing *source identification conditions* from testing *statistical efficiency* when possible. Therefore, the simulations are performed twice, first representing an *infinite* number of samples (testing *only* source identification conditions) and second for a *finite* number of samples (testing both statistical efficiency and source identification conditions).

To represent infinite samples, the true covariance matrices, instead of their estimates, are used by the methods. Therefore, starting from the true joint covariance matrix $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})} = \mathrm{blkdiag}\left(\mathbf{C}_{\mathbf{s}_1}, \ldots, \mathbf{C}_{\mathbf{s}_N}\right) \in \mathbb{R}^{NK \times NK}$, the covariance matrix of the concatenated source vectors $\mathbf{s} \in \mathbb{R}^{NK}$ is found as $\mathbf{C}_{\mathbf{s}} = \mathbf{\Pi}\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\mathbf{\Pi}^{\mathsf{T}}$, where $\mathbf{\Pi} \in \mathbb{R}^{NK \times NK}$ is a specific permutation matrix to sort the elements of $\mathbf{C}_{\mathbf{s}}$ in source vector ordering instead of SCV ordering. Then, the joint mixing matrix $\mathbf{A} = \mathrm{blkdiag}\left(\mathbf{A}^{[1]}, \ldots, \mathbf{A}^{[K]}\right)$ is generated, where $\mathbf{A}^{[k]}$ is an invertible mixing matrix with elements drawn from $\mathcal{N}(0, 1)$. Finally, the covariance matrix of the concatenated datasets $\mathbf{x} \in \mathbb{R}^{NK}$ is $\mathbf{C}_{\mathbf{x}} = \mathbf{A}\mathbf{C}_{\mathbf{s}}\mathbf{A}^{\mathsf{T}}$, and the covariance matrix of the concatenated whitened datasets is $\mathbf{C}_{\mathbf{y}} = \mathbf{D}_{\mathbf{x}}^{-\frac{1}{2}}\mathbf{C}_{\mathbf{x}}\mathbf{D}_{\mathbf{x}}^{-\frac{1}{2}}$, where $\mathbf{D}_{\mathbf{x}}$ is a block-diagonal matrix consisting of the diagonal blocks of $\mathbf{C}_{\mathbf{x}}$.

To simulate with finite samples, $V = 10000$ samples are drawn for the SCVs $\mathbf{S}_n \in \mathbb{R}^{K \times V}$ from a multivariate Gaussian distribution with zero mean and the previously specified $\mathbf{C}_{\mathbf{s}_n}$, and the source components are normalized to have zero mean and unit variance. Then, the source components of the SCVs $\mathbf{S}_n$ are distributed to each $\mathbf{S}^{[k]}$, and the observed datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$ are generated by mixing the source matrices as $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}$, where $\mathbf{A}^{[k]} \in \mathbb{R}^{N \times N}$ is generated as described previously. Then, the covariance matrices $\widehat{\mathbf{C}}_{\mathbf{x}}$ and $\widehat{\mathbf{C}}_{\mathbf{y}}$ are estimated using the observed datasets $\mathbf{X}^{[k]}$.

We simulate $N = 5$ SCVs and $K = 100$ datasets. In the following paragraphs, we describe how we choose the values of R and α_n for experiments A–E.[3] Then, after introducing the performance metric, we report the results observed in 50 runs of the experiments.

A. Violating the source identification conditions of all methods

We start with an experiment where we violate the source identification conditions of all methods. Consider the EVD of the SCV covariance matrices $\mathbf{C}_{\mathbf{s}_n} = \mathbf{\Theta}_n\mathbf{\Lambda}_n\mathbf{\Theta}_n^{\mathsf{T}}$, where $\mathbf{\Lambda}_n = \mathrm{diag}\left(\boldsymbol{\lambda}_n\right) \in \mathbb{R}^{K \times K}$ contains the eigenvalues of $\mathbf{C}_{\mathbf{s}_n}$ on its main diagonal, and $\mathbf{\Theta}_n \in \mathbb{R}^{K \times K}$ contains the corresponding eigenvectors. This experiment is designed so that all $\mathbf{C}_{\mathbf{s}_n}$ have the same eigenvalues $\boldsymbol{\lambda}_n$ and the same eigenvectors $\mathbf{\Theta}_n$ up to a sign (i.e., for some n^{th} and m^{th} SCVs, $\mathbf{\Theta}_n = \mathbf{D}_n\mathbf{\Theta}_m$ for any diagonal matrix $\mathbf{D}_n$

[3]The Python implementation of the mCCA methods, as well as the code for the experiments, is available at: https://github.com/SSTGroup/multiset_canonical_correlation_analysis.

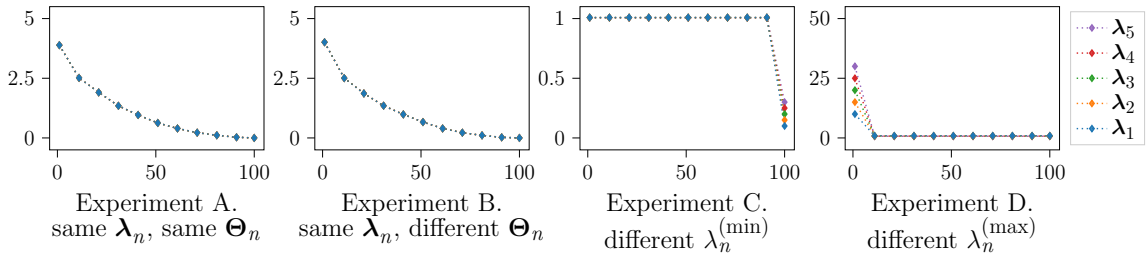


Figure 5.1: Example eigenvalues for experiments A–D; every tenth value is plotted. A. The source identification conditions of all methods are violated. B. The experiment substantiates the conditions of genvar and ssqcor. C. The experiment substantiates the conditions of minvar. D. The experiment substantiates the conditions of sumcor and maxvar.

with diagonal elements equal to either -1 or 1). To achieve this, we first generate $\mathbf{C}_{\mathbf{s}_1} \in \mathbb{R}^{K\times K}$ with $R = K = 100$ and $\alpha_1 = 1$ and perform the EVD $\mathbf{C}_{\mathbf{s}_1} = \boldsymbol{\Theta}_1\boldsymbol{\Lambda}_1\boldsymbol{\Theta}_1^\mathsf{T}$. We then generate $\boldsymbol{\Theta}_n = \mathbf{D}_n\boldsymbol{\Theta}_1$, $n = 2,\dots,N$, where $\mathbf{D}_n$ is a diagonal matrix with randomly generated ± 1 elements on its diagonal. Finally, we generate the remaining SCV covariance matrices as $\mathbf{C}_{\mathbf{s}_n} = \boldsymbol{\Theta}_n\boldsymbol{\Lambda}_1\boldsymbol{\Theta}_n^\mathsf{T}$, $n = 2,\dots,N$. The eigenvalues for one run of the experiment are shown in Figure 5.1(a). As the eigenvalues are the same for each SCV covariance matrix and the eigenvectors are also the same (up to a sign), the source identification conditions of all methods are violated.

B. Substantiating the source identification conditions of genvar and ssqcor

In this experiment, we substantiate our theoretical considerations for the source identification conditions of genvar and ssqcor from Subsections 5.4.4 and 5.4.5. Therefore, we generate the SCV covariance matrices $\mathbf{C}_{\mathbf{s}_n}$ such that they still have the same eigenvalues but different eigenvectors. First, we generate $\mathbf{C}_{\mathbf{s}_1}$ as described in Subsection 5.4.6, perform the EVD, and use $\boldsymbol{\lambda}_1 \in \mathbb{R}^K$ as the same eigenvalues for all N SCV covariance matrices, shown in Figure 5.1(b) for one run. However, now we use the scipy function `random_correlation` [Vir20] to generate $\mathbf{C}_{\mathbf{s}_2},\dots,\mathbf{C}_{\mathbf{s}_N}$ having these same eigenvalues $\boldsymbol{\lambda}_1$ but different eigenvectors. Thus, we expect genvar and ssqcor to achieve JBSS but sumcor/maxvar and minvar not, as their proposed source identification conditions are still violated since all SCV covariance matrices have the same largest and smallest eigenvalue.

C. Substantiating the source identification conditions of minvar

In this experiment, we generate $\mathbf{C}_{\mathbf{s}_n}$ such that the proposed source identification conditions of minvar are satisfied. Therefore, we choose $R = 1$ and specify the val-

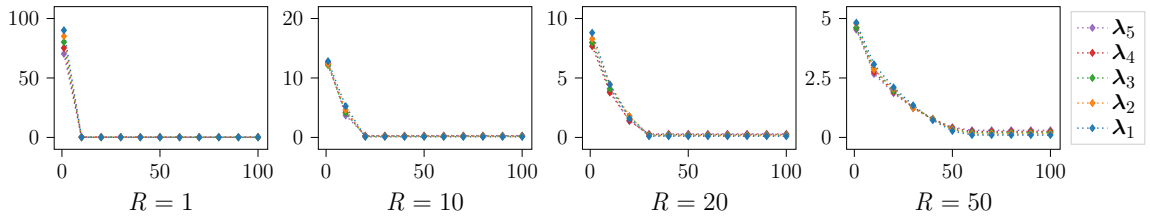


Figure 5.2: Example eigenvalues for experiment E for different values of the effective rank R; every tenth value is plotted.

ues of α_n according to (5.3) such that the smallest eigenvalues $\lambda_n^{(\min)}$ are $\lambda_n^{[K]} = \{0.1, 0.15, 0.2, 0.25, 0.3\}$ for $n = 1, \ldots, 5$. The eigenvalues for one run are shown in Figure 5.1(d). Note that this is a very artificial scenario, as the resulting covariance matrices are very close to identity matrices, and thus the correlations between the source components are almost zero. The source identification conditions of maxvar are almost violated, as the largest eigenvalues are very close to each other, and those of sumcor are violated, as the largest eigenvalue of each SCV covariance matrix is not greater than all second-largest eigenvalues. We expect genvar to achieve JBSS here, and we expect ssqcor to perform similarly to maxvar as they estimate similar SCVs if $R = 1$.

D. Substantiating the source identification conditions of sumcor and maxvar

In this experiment, we generate $\mathbf{C}_{\mathbf{s}_n}$ such that the source identification conditions of sumcor and maxvar are satisfied. Therefore, we choose $R = 1$ and specify the values of α_n according to (5.3) such that the largest eigenvalues $\lambda_n^{(\max)}$ are $\lambda_n^{[K]} = \{10, 15, 20, 25, 30\}$ for $n = 1, \ldots, 5$. Note that here we call the largest eigenvalue $\lambda_n^{[K]}$ instead of $\lambda_n^{[1]}$ to keep the notation as in (5.3). The eigenvalues for one run are shown in Figure 5.1(c). The source identification conditions of minvar are almost violated, as the smallest eigenvalues are very close to each other. We expect genvar and ssqcor to achieve JBSS here based on our theoretical considerations for their source identification conditions.

E. Changing the effective rank R of the SCVs

In this experiment, we generate $\mathbf{C}_{\mathbf{s}_n}$ such that we can investigate the influence of the effective rank R on the methods' ability to achieve JBSS. Therefore, we vary the value of $R = 1, 2, 5, 10, 20, 50$, and for each value of R we choose $\alpha_n = \{0.9, 0.85, 0.8, 0.75, 0.7\}$ for $n = 1, \ldots, 5$. To satisfy the source identification conditions of sumcor, we make sure that $\max_n \left(\lambda_n^{[2]}\right) < \min_n \left(\lambda_n^{[1]}\right) - \epsilon$, where $\epsilon = \frac{1}{K}$, i.e., that the second-largest eigenvalue of all SCVs is smaller than the largest eigenvalue of all SCVs. The eigenvalues

for one run for $R = 1, 10, 20, 50$ are shown in Figure 5.2. As we see, the eigenvalues of all SCV covariance matrices are distinct, but as the smallest eigenvalues are very close to each other, the conditions of minvar are almost violated, similar to the case in experiment D.

Performance metric

To measure if JBSS has been successful, we use the joint Inter-Symbol-Interference (jISI) from [And12], which is defined as

$$\text{jISI} = \frac{1}{2N(N-1)} \left[\sum_{n=1}^{N} \left(\frac{\sum_{m=1}^{N} g_{nm}}{\max_m g_{nm}} - 1 \right) + \sum_{m=1}^{N} \left(\frac{\sum_{n=1}^{N} g_{nm}}{\max_n g_{nm}} - 1 \right) \right],$$

where g_{nm} is the $(n,m)^{\text{th}}$ element of

$$\mathbf{G} = \sum_{k=1}^{K} \left| \mathbf{W}^{[k]} \mathbf{A}^{[k]} \right| \in \mathbb{R}^{N \times N}.$$

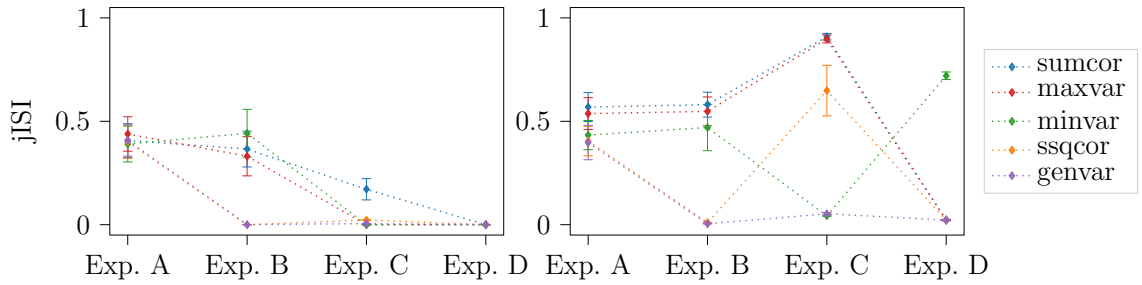
The jISI is normalized between zero and one, where values close to zero indicate successful JBSS, i.e., source components are separated correctly within a dataset and also aligned correctly across datasets. To get more insight into what the joint ISI measures, let us look at the formula in more detail. If JBSS is achieved, i.e., if $\mathbf{W}^{[k]} = \mathbf{P}\mathbf{\Gamma}^{[k]} \left(\mathbf{A}^{[k]} \right)^{-1}$, then,

$$\mathbf{G} = \sum_{k=1}^{K} \left| \mathbf{W}^{[k]} \mathbf{A}^{[k]} \right| = \sum_{k=1}^{K} \left| \mathbf{P}\mathbf{\Gamma}^{[k]} \right| = \sum_{k=1}^{K} \mathbf{P} = K\mathbf{P},$$

where we have used that the diagonal elements of $\mathbf{\Gamma}^{[k]}$ can only be ± 1 because we assume sources and estimated sources to have unit variance. As for any permutation matrix it holds that there is one element in each row and column that is equal to one and the other elements are equal to zero, $\sum_{m=1}^{N} g_{nm} = \max_m g_{mn} = \sum_{n=1}^{N} g_{nm} = \max_n g_{mn} = 1$, and thus jISI $= 0$.

Results

Figures 5.3 and 5.4 show how successful the five mCCA methods (sumcor from [Nie95], and maxvar, minvar, ssqcor and genvar from [Ket71]) are in achieving JBSS in experiments A–D and E, respectively, where the results representing infinite samples are shown in Figures 5.3(a) and 5.4(a) for finite samples in Figures 5.3(b) and 5.4(b). The diamonds and lines indicate the mean and standard deviation of the jISI across 50 runs



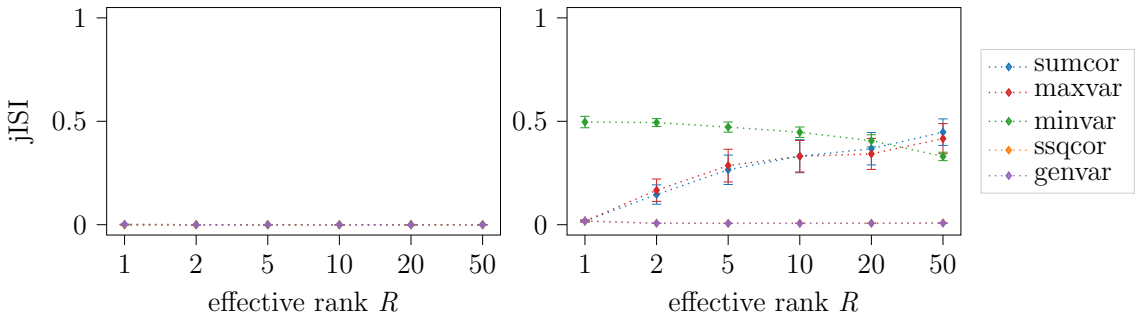
(a) Infinite samples (b) $V = 10000$ samples

Figure 5.3: Joint ISI value (lower is better) for experiments A–D, averaged across 50 runs. The diamonds show the mean value and the end lines plus or minus one standard deviation. The sumcor method is according to Nielsen [Nie02], and the other methods are according to Kettenring [Ket71]. (a) For the simulation representing infinite samples, the methods perform as expected. (b) With finite samples, we see that in experiment C also maxvar and ssqcor fail, and in experiment D minvar fails. This is because their source identification conditions are almost violated in these experiments.

for each experiment.

In Figure 5.3(a) in experiment A, where all SCV covariance matrices have the same eigenvalues and the same eigenvectors (up to a sign), the source identification conditions of all methods are violated, and we see that, indeed, no method is able to achieve JBSS. In experiment B, where the eigenvalues are still the same but the eigenvectors are different across SCVs, ssqcor and genvar achieve JBSS, while sumcor, maxvar, and minvar fail. This supports our theoretical considerations in Subsections 5.4.4 and 5.4.5 that genvar could achieve JBSS even if the eigenvalues are the same across SCVs as long as the eigenvectors are different, and that the source identification conditions of ssqcor could be similar to those of genvar. In experiment C, where all SCVs have a different smallest eigenvalue, minvar achieves JBSS, supporting Conjecture 5.1. Maxvar also achieves JBSS, since the largest eigenvalue is also different for each SCV (even though the difference is very small), but sumcor fails as the largest eigenvalues are equal to the second-largest, substantiating Theorems 5.6 and 5.7. Finally, in experiment D, where the largest eigenvalue is different for each SCV (and the smallest eigenvalue is also different with a very small difference), all methods are successful. To conclude these experiments, our simulations support the proposed theoretical source identification conditions.

In Figure 5.3(b), we see how using finite samples, i.e., estimating the covariance matrices, affects the methods' ability to achieve JBSS. The trends for experiments A and B are the same as those in the simulation representing infinite samples. In



(a) Infinite samples (b) $V = 10000$ samples

Figure 5.4: Joint ISI value (lower is better) for experiment E for different values of R, averaged across 50 runs. (a) For the simulation representing infinite samples, all methods are able to achieve JBSS. (b) With finite samples, minvar fails to achieve JBSS as its source identification conditions are almost violated. With an increasing value of $R > 1$, sumcor and maxvar fail to achieve JBSS because both of them assume that SCVs have an effective rank of $R = 1$. Lastly, genvar and ssqcor achieve JBSS for all values of R.

experiment C, minvar and genvar again successfully achieve JBSS, but sumcor and maxvar fail now because the largest eigenvalues of the estimated covariance matrices are too similar, which violates their source identification conditions. Also ssqcor fails to achieve JBSS, substantiating that it performs similarly to maxvar for $R = 1$. In experiment D, minvar is the only method that fails to achieve JBSS because the smallest eigenvalues of the estimated covariance matrices are too similar. To conclude these experiments, it is observed that genvar is more statistically efficient than the other methods, as its performance is not so strongly affected by using finite samples compared with the other methods.

In Figure 5.4(a), we see that all methods are able to achieve JBSS in experiment E for all values of the effective rank R, which is expected as the eigenvalues of all SCV covariance matrices are distinct. In Figure 5.4(b), where the covariance matrices are estimated from finite samples, we see that minvar fails to achieve JBSS, which is also expected as the smallest eigenvalues are very similar to each other. For $R = 1$, all other methods achieve JBSS, but for an increasing value of R, sumcor and maxvar fail. This can be explained by the fact that both methods assume that the SCVs have an effective rank of $R = 1$ because of their underlying factor models in (3.25) and (3.33). In contrast, genvar and ssqcor achieve JBSS for all values of R in this experiment, as they can estimate SCVs with any effective rank from 1 to K according to their underlying factor model in (3.37).

5.5 Summary

In this chapter, we have discussed the source identification conditions of several matrix decomposition methods, i.e., the conditions on the true underlying source components under which the methods are able to achieve BSS or JBSS. After summarizing the source identification conditions of ICA, IVA, CCA, and mCCA-maxvar, we have derived and proven the source identification conditions of mCCA-sumcor, conjectured those of mCCA-minvar, and discussed theoretical considerations for those of mCCA-genvar and mCCA-ssqcor. We have substantiated the proposed source identification conditions of mCCA with simulations, where we have observed that the ability of the methods to achieve JBSS is related to their underlying factor models. Some of them have similar underlying factor models, and as such, we discuss them together. First, maxvar [Ket71] and sumcor [Nie95] both are effective in estimating SCVs with an effective rank of one, meaning that the sources of all datasets within each SCV have the same underlying factor. Their main difference is that the sources within a dataset are constrained to be uncorrelated in maxvar, but not in sumcor, which leads to better interpretability of the maxvar sources and makes maxvar the preferred method of these two. Then, genvar [Ket71] and ssqcor [Ket71] can both effectively estimate SCVs with an arbitrary effective rank, while it is worth mentioning that ssqcor performs similarly to maxvar if the SCVs have an effective rank of one. Lastly, minvar [Ket71] is only effective for estimating SCVs in a very artificial scenario. Based on our experiments, genvar seems to have the least restrictive source identification conditions among all mCCA methods and thus can identify sources that the other methods may not recover. This is especially notable for a finite number of samples, where we observed in our experiments that genvar seems to be the most statistically efficient method. As the source identification conditions of ICA, IVA, and CCA were already known, they were not included in these simulations that focused on demonstrating the novel findings of this chapter. Building on our findings, we summarize our recommended choices of mCCA methods as follows:

- In general, we recommend choosing genvar [Ket71] for achieving JBSS based on our experiments, as it seems to have the least restrictive source identification conditions and to be more statistically efficient than the other methods. This is a significant finding, as sumcor and maxvar, despite being perhaps the most commonly used mCCA methods [Sor21], are outperformed by genvar in our simulation study.

- However, the computational complexity of genvar is the highest among all mCCA methods, with differences becoming stronger with an increasing number of datasets

or dimension of data. Therefore, in specific scenarios, it makes sense to use a faster mCCA method. As the ability of the methods to achieve JBSS depends on their underlying factor models, we recommend the use of maxvar [Ket71] if it is reasonable to assume that the true SCVs have an effective rank of one.

Based on our comparison of IVA-G and PARAFAC2 in the simulations in the previous chapter, where both methods have been able to achieve JBSS, we recommend the use of PARAFAC2 when the columns of the mixing matrices across different datasets are expected to differ only up to a scaling and of IVA-G when different mixing matrices are expected in different datasets. These findings partly answer the question of which JBSS method (more specifically, which mCCA method) should be used in which application and form the basis for further research. The next logical steps are to formally prove the conjecture for minvar and to derive and prove the source identification conditions of genvar, ssqcor, and PARAFAC2. After filling these theoretical gaps, more extensive simulations testing the JBSS performance of all mCCA methods, IVA-G, IVA-L-SOS, and PARAFAC2 together would help to ultimately solve the challenge of choosing a JBSS method for a specific application. In the next chapter, we address the second challenge of this thesis, identifying the relationship among multiple datasets using JBSS methods.

5.6 Appendix: Proof of source identification conditions of sumcor

In order to prove Theorem 5.6, we prove that if the source identification conditions of sumcor are satisfied, then the sumcor transformation matrices $\mathbf{T}^{[k]}$ and the mixing matrices $\mathbf{A}^{[k]}$ are related as follows:

$$\left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Gamma}^{[k]}\left(\mathbf{A}^{[k]}\right)^{-1},$$

which is equivalent to $\mathbf{u}^{[k]} = \mathbf{P}\boldsymbol{\Gamma}^{[k]}\mathbf{s}^{[k]}$ but slightly shorter to prove. The proof of Theorem 5.6 consists of three steps.

1. Derive the relationship of $\widetilde{\mathbf{T}}$, the concatenated sumcor transformation matrices in (3.29), and $\widetilde{\boldsymbol{\Delta}}$, the eigenvectors of the concatenated SCVs' covariance matrix $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$: $\widetilde{\mathbf{T}} = \mathbf{A}^{-\mathsf{T}}\boldsymbol{\Pi}\widetilde{\boldsymbol{\Delta}}$, where $\boldsymbol{\Pi}$ is a specific permutation matrix.

2. Show that the structure of $\widetilde{\boldsymbol{\Delta}}$ is $\widetilde{\boldsymbol{\Delta}} = \operatorname{diag}\left(\boldsymbol{\theta}_1^{[1]}, \ldots, \boldsymbol{\theta}_N^{[1]}\right)\mathbf{P}^{\mathsf{T}}$ under the conditions in Theorem 5.6.

3. Use 1 and 2 to show that $\left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Gamma}^{[k]}\left(\mathbf{A}^{[k]}\right)^{-1}$.

Proof. 1) First, we express $\mathbf{C_x}$ and $\mathbf{D_x}$ in the GEVD in (3.23), given by $\mathbf{C_x T} = \mathbf{D_x T \Phi}$, in terms of $\mathbf{C}_\mathbf{s}^{(\mathrm{SCV})}$, the covariance matrix of the concatenated SCVs. Therefore, we define the concatenated source vector

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}^{[1]} \\ \vdots \\ \mathbf{s}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK}$$

with covariance matrix

$$\mathbf{C_s} = \mathbb{E}\left\{\mathbf{s}\mathbf{s}^{\mathsf{T}}\right\} = \begin{bmatrix} \mathbf{I}_{N\times N} & \mathbf{C}_\mathbf{s}^{[1,2]} & \cdots & \mathbf{C}_\mathbf{s}^{[1,K]} \\ \mathbf{C}_\mathbf{s}^{[2,1]} & \mathbf{I}_{N\times N} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{C}_\mathbf{s}^{[K-1,K]} \\ \mathbf{C}_\mathbf{s}^{[K,1]} & \cdots & \mathbf{C}_\mathbf{s}^{[K,K-1]} & \mathbf{I}_{N\times N} \end{bmatrix},$$

where we used that $\mathbf{C}_\mathbf{s}^{[k]} = \mathbf{I}_{N\times N}$, $k = 1, \ldots, K$, according to Assumptions (IVA.2) from page 22 and (IVA.3b) from page 71, we define the joint mixing matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{A}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK\times NK},$$

and we define the concatenated datasets

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{[1]} \\ \vdots \\ \mathbf{x}^{[K]} \end{bmatrix} \in \mathbb{R}^{NK}.$$

Now, we can write the JBSS model in (2.2), given by $\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]}$, $k = 1, \ldots, K$, for all datasets in one equation:

$$\mathbf{x} = \mathbf{A}\mathbf{s}.$$

Then, the covariance matrix of the concatenated datasets is

$$\mathbf{C_x} = \mathbb{E}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} = \mathbf{A}\mathbf{C_s}\mathbf{A}^{\mathsf{T}} \in \mathbb{R}^{NK\times NK},$$

and

$$\mathbf{D}_{\mathbf{x}} = \mathbf{A} \begin{bmatrix} \mathbf{I}_{N\times N} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{I}_{N\times N} \end{bmatrix} \mathbf{A}^{\mathsf{T}} = \mathbf{A}\mathbf{A}^{\mathsf{T}} \in \mathbb{R}^{NK\times NK}.$$

Now, we define the covariance matrix of the concatenated SCVs as the block-diagonal matrix

$$\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})} = \mathbb{E}\left\{ \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_N \end{bmatrix} \begin{bmatrix} \mathbf{s}_1^{\mathsf{T}} & \cdots & \mathbf{s}_N^{\mathsf{T}} \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{C}_{\mathbf{s}_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{C}_{\mathbf{s}_N} \end{bmatrix} \in \mathbb{R}^{NK\times NK},$$

where the off-diagonal blocks are equal to zero because the SCVs are assumed to be uncorrelated by Assumption (IVA.3b) from page 71. Then, we can write $\mathbf{C}_{\mathbf{s}}$, the covariance matrix of the concatenated source vectors, as a permuted version of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$:

$$\mathbf{C}_{\mathbf{s}} = \boldsymbol{\Pi}\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\boldsymbol{\Pi}^{\mathsf{T}} \in \mathbb{R}^{N\times N},$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{NK\times NK}$ is a specific permutation matrix. This gives $\mathbf{C}_{\mathbf{x}} = \mathbf{A}\boldsymbol{\Pi}\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$, and the GEVD in (3.23) becomes

$$\mathbf{A}\boldsymbol{\Pi}\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{T} = \mathbf{A}\mathbf{A}^{\mathsf{T}}\mathbf{T}\boldsymbol{\Phi}. \tag{5.4}$$

Multiplying (5.4) from the left with $\boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{-1}$ results in $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{T} = \boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{T}\boldsymbol{\Phi}$, and defining $\boldsymbol{\Delta} = \boldsymbol{\Pi}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{T}$ yields the EVD

$$\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}\boldsymbol{\Delta} = \boldsymbol{\Delta}\boldsymbol{\Phi}.$$

Let $\widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{t}_1 & \cdots & \mathbf{t}_N \end{bmatrix} \in \mathbb{R}^{NK\times N}$ consist of the N principal generalized eigenvectors of $\mathbf{C}_{\mathbf{x}}$ w.r.t. $\mathbf{D}_{\mathbf{x}}$, and let $\widetilde{\boldsymbol{\Delta}} = \begin{bmatrix} \boldsymbol{\delta}_1 & \cdots & \boldsymbol{\delta}_N \end{bmatrix} \in \mathbb{R}^{NK\times N}$ consist of the N principal eigenvectors of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$. We know that $\widetilde{\mathbf{T}}$ consists of the concatenated sumcor transformation matrices $\mathbf{T}^{[k]}$, i.e.,

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \left(\mathbf{T}^{[1]}\right)^{\mathsf{T}} \\ \vdots \\ \left(\mathbf{T}^{[K]}\right)^{\mathsf{T}} \end{bmatrix}, \tag{5.5}$$

and their relationship with the eigenvectors of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$ is

$$\widetilde{\mathbf{T}} = \mathbf{A}^{-\mathsf{T}} \boldsymbol{\Pi} \widetilde{\boldsymbol{\Delta}}. \tag{5.6}$$

2) We now determine the structure of $\widetilde{\boldsymbol{\Delta}}$. Let

$$\boldsymbol{\lambda}_n = \begin{bmatrix} \lambda_n^{[1]} \\ \vdots \\ \lambda_n^{[K]} \end{bmatrix} \in \mathbb{R}^K$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}_n}$ such that $\lambda_n^{[1]} \geq \cdots \geq \lambda_n^{[K]}$, and let

$$\boldsymbol{\Theta}_n = \begin{bmatrix} \boldsymbol{\theta}_n^{[1]} & \cdots & \boldsymbol{\theta}_n^{[K]} \end{bmatrix} \in \mathbb{R}^{K \times K}$$

be the corresponding eigenvectors. Let

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{NK} \end{bmatrix} \in \mathbb{R}^{NK}$$

be the eigenvalues of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$ such that $\phi_1 \geq \cdots \geq \phi_{NK}$. Under the conditions in Theorem 5.6, the N largest eigenvalues of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$ are equal to the set consisting of the largest eigenvalue of each $\mathbf{C}_{\mathbf{s}_n}$, $n = 1, \ldots, N$, i.e.,

$$\left\{ \phi_1, \ldots, \phi_N \right\} = \left\{ \lambda_1^{[1]}, \ldots, \lambda_N^{[1]} \right\},$$

and the non-zero chunks of the N principal eigenvectors of $\mathbf{C}_{\mathbf{s}}^{(\mathrm{SCV})}$ thus consist of the principal eigenvector $\boldsymbol{\theta}_n^{[1]}$ of each $\mathbf{C}_{\mathbf{s}_n}, n = 1, \ldots, N$, i.e.,

$$\widetilde{\boldsymbol{\Delta}} = \begin{bmatrix} \boldsymbol{\theta}_1^{[1]} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\theta}_N^{[1]} \end{bmatrix} \mathbf{P}^{\mathsf{T}} = \begin{bmatrix} \theta_1^{[1]}(1) & 0 & 0 \\ \vdots & \vdots & \vdots \\ \theta_1^{[1]}(K) & 0 & \vdots \\ 0 & \ddots & 0 \\ \vdots & 0 & \theta_N^{[1]}(1) \\ \vdots & \vdots & \vdots \\ 0 & 0 & \theta_N^{[1]}(K) \end{bmatrix} \mathbf{P}^{\mathsf{T}}, \tag{5.7}$$

Note that we have introduced a permutation matrix $\mathbf{P} \in \mathbb{R}^{N\times N}$ as the columns of $\widetilde{\mathbf{\Delta}}$ can be in arbitrary ordering depending on the ordering of the largest eigenvalue of each $\mathbf{C}_{\mathbf{s}_n}$.

3) Now, inserting (5.7) in (5.6), the permutation matrix $\mathbf{\Pi}$ resorts the rows of $\widetilde{\mathbf{\Delta}}$, and we have

$$\widetilde{\mathbf{T}} = \mathbf{A}^{-\mathsf{T}}\mathbf{\Pi}\widetilde{\mathbf{\Delta}} = \mathbf{A}^{-\mathsf{T}} \begin{bmatrix} \theta_1^{[1]}(1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \theta_N^{[1]}(1) \\ \vdots & \vdots & \vdots & \vdots \\ \theta_1^{[1]}(K) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \theta_N^{[1]}(K) \end{bmatrix} \mathbf{P}^{\mathsf{T}}.$$

By defining the diagonal matrix

$$\mathbf{\Gamma}^{[k]} = \begin{bmatrix} \theta_1^{[1]}(k) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \theta_N^{[1]}(k) \end{bmatrix} \in \mathbb{R}^{N\times N}$$

and using the partitioning of $\widetilde{\mathbf{T}}$ in (5.5), we see that

$$\begin{aligned} \begin{bmatrix} \mathbf{T}^{[1]} \\ \vdots \\ \mathbf{T}^{[K]} \end{bmatrix} &= \begin{bmatrix} \left(\mathbf{A}^{[1]}\right)^{-\mathsf{T}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \left(\mathbf{A}^{[K]}\right)^{-\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}^{[1]} \\ \vdots \\ \mathbf{\Gamma}^{[K]} \end{bmatrix} \mathbf{P}^{\mathsf{T}} \\ &= \begin{bmatrix} \left(\mathbf{A}^{[1]}\right)^{-\mathsf{T}} \mathbf{\Gamma}^{[1]}\mathbf{P}^{\mathsf{T}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \left(\mathbf{A}^{[K]}\right)^{-\mathsf{T}} \mathbf{\Gamma}^{[K]}\mathbf{P}^{\mathsf{T}} \end{bmatrix}, \end{aligned}$$

and thus,

$$\left(\mathbf{T}^{[k]}\right)^{\mathsf{T}} = \mathbf{P}\mathbf{\Gamma}^{[k]} \left(\mathbf{A}^{[k]}\right)^{-1}, \quad k = 1, \ldots, K,$$

i.e., $\left(\mathbf{T}^{[k]}\right)^{\mathsf{T}}$ correctly estimates $\left(\mathbf{A}^{[k]}\right)^{-1}$ up to a scaling and a common permutation.

□

6 Identifying the relationship structure among multiple datasets using JBSS

In recent years, identifying relationships among multiple datasets has received growing attention in medical applications, such as for making group inferences when estimating brain activations [Cor10] or for better localization of brain activity [Cal12]. One potential application for identifying the relationship among multiple datasets is subgroup identification [Lip17]. In this context, each dataset corresponds to a subject, and a group of related subject datasets is called a subject group or *subgroup*. The identified subgroups can be used in multiple applications. For example, in Electronic Health Records (EHR) data, identifying subgroups can help uncover previously unknown connections between illnesses [Wan20b], while in precision medicine, estimating the dose of medication for a patient can be achieved based on the known dose of other patients in the same subgroup [Loh19; Nat11]. Another application where it is of interest to identify the relationship among multiple datasets over time is in the detection and prediction of epileptic seizures, as it has been found that the correlations between measures of the peripheral Autonomic Nervous System (ANS) increased before Generalized Tonic-Clonic Seizure (GTCS) and decreased after the seizures [Vie21] and that the changes in these correlations therefore might contain information for predicting seizures [Vie23].

However, identifying the relationship among multiple datasets is challenging because datasets are often high-dimensional and noisy. JBSS methods simplify the problem by extracting interpretable sources that are linked across datasets. Based on these latent sources, a *group* can be defined as a subset of datasets with similar latent sources. Analyzing how the datasets are distributed in these groups and how these groups are related reveals the *relationship structure* among the datasets. To the best of our knowledge, no method has yet been developed to achieve this goal. Some of the related methods in the literature identify the relationship among the individual sources within each SCV instead of the relationship structure among the whole datasets. This

is, for example, done in [Akh21], where the dependence structure of latent sources, i.e., datasets across which they are dependent, is identified using IVA. That work extends [Has20], where the correlation structure between latent signal components is identified without estimating the components themselves. Also in [Lon20], IVA has been used to identify groups of sources that are related with each other, but this method is based on user-defined thresholds. To overcome the need for subject-specific thresholds, an approach based on IVA and Gershgorin discs is proposed in [Yan22] for identifying groups of related sources. For all these methods, a postprocessing would be necessary to identify the relationship structure among the datasets from the method's output.

In this chapter, we therefore propose a method for identifying the relationship structure among multiple datasets based on JBSS and thereby address the second challenge of this thesis. The proposed method works as follows. 1) The latent sources in the datasets are estimated using IVA-L-SOS, 2) the SCVs are identified that contain information about the relationship among the datasets, and 3) features are extracted from these SCVs as the input of a hierarchical clustering, which reveals the relationship structure among the datasets in a resulting dendrogram. Unlike related approaches for Step 2, the proposed method can effectively handle non-Gaussian data. It achieves this by incorporating higher-order statistics through IVA-L-SOS and extracting features with an eigenvalue decomposition-based approach without distributional assumptions. In simulations, we experimentally demonstrate the superior performance of the second step of the proposed method over the competing techniques in [Akh21] and [Yan22], and we show that the proposed method reveals the true relationship structure among the datasets, as measured by a perfect Adjusted Mutual Information (AMI) for different values of the correlation between the sources.[1]

6.1 Problem formulation

In the following, we formulate the problem for which we are proposing a solution in this chapter.

Problem 6.1. *Given K datasets, identify their relationship structure, i.e., group datasets based on the similarity (high statistical dependence) between their latent variables. Note that this cannot be achieved through simple clustering approaches, as revealing the relationship structure relies on the dependence of the latent variables across datasets.*

[1] This chapter is based on the paper: "Identifying the Relationship Structure among Multiple Datasets Using Independent Vector Analysis: Application to Multi-task fMRI Data," I. Lehmann, T. Hasija, B. Gabrielson, M. A. B. S. Akhonda, V. D. Calhoun, T. Adali, *IEEE Access*, 2024, pp. 109443–109456.

In [Lon20] and [Leh22], it is shown that information about the relationship among the datasets can be inferred from the linear dependence, i.e., correlation, of the source components across datasets. This information is revealed in the covariance matrices of the SCVs. An SCV covariance matrix with all non-zero off-diagonal values implies that all source components within that SCV are correlated, and hence describes a *common* SCV, i.e., an SCV with source components common across all datasets [Lon20]. Similarly, an SCV covariance matrix exhibiting a block-diagonal structure (with off-block elements equal to zero) implies that source components are correlated within each block but uncorrelated across blocks. We call SCVs with these covariance matrices *structured*. Note that this also holds for SCV covariance matrices $\mathbf{C}_{\mathbf{s}_n}$ that can be transformed into block-diagonal matrices using an orthogonal permutation matrix $\mathbf{P}_n \in \mathbb{R}^{K\times K}$ s.t. $\widetilde{\mathbf{C}}_{\mathbf{s}_n} = \mathbf{P}_n \mathbf{C}_{\mathbf{s}_n} \mathbf{P}_n^\mathsf{T}$ is block-diagonal. While common SCVs only provide limited useful information about the relationship among the datasets, structured SCVs are most informative for this. Consequently, in this chapter, we are primarily interested in identifying the structured SCVs and then using them to identify the relationship structure among the datasets.

6.2 Proposed method for identifying the relationship structure among multiple datasets

In the following, we propose a method for identifying the relationship structure among multiple datasets. Figure 6.1 shows the three steps of the method: 1) estimation of latent sources using a JBSS method, 2) identification of common and structured SCVs, and 3) identification of the relationship structure using structured SCVs. The following sections explain the details of each step.

6.2.1 Step 1: Estimation of SCVs

To be able to infer information about the relationship structure from the SCV covariance matrices, the unknown source components in the SCVs must be estimated from the observed datasets. For this, we use IVA-L-SOS [Bhi19], described in Subsection 3.2.2, as it has been shown that the assumed pdf of the SCVs in IVA-L-SOS is a good match for real-world fMRI data [Bhi19; Lon20], which is the main application field in this thesis. However, if the method is applied to data with different source distributions, Step 1 can be replaced by another JBSS method with appropriate assumptions for that type of data. We assume that the dimension N of the observed datasets $\mathbf{X}^{[k]} \in \mathbb{R}^{N\times V}$, $k = 1, \ldots, K$, is much higher than the number R of latent source components of interest—a typical

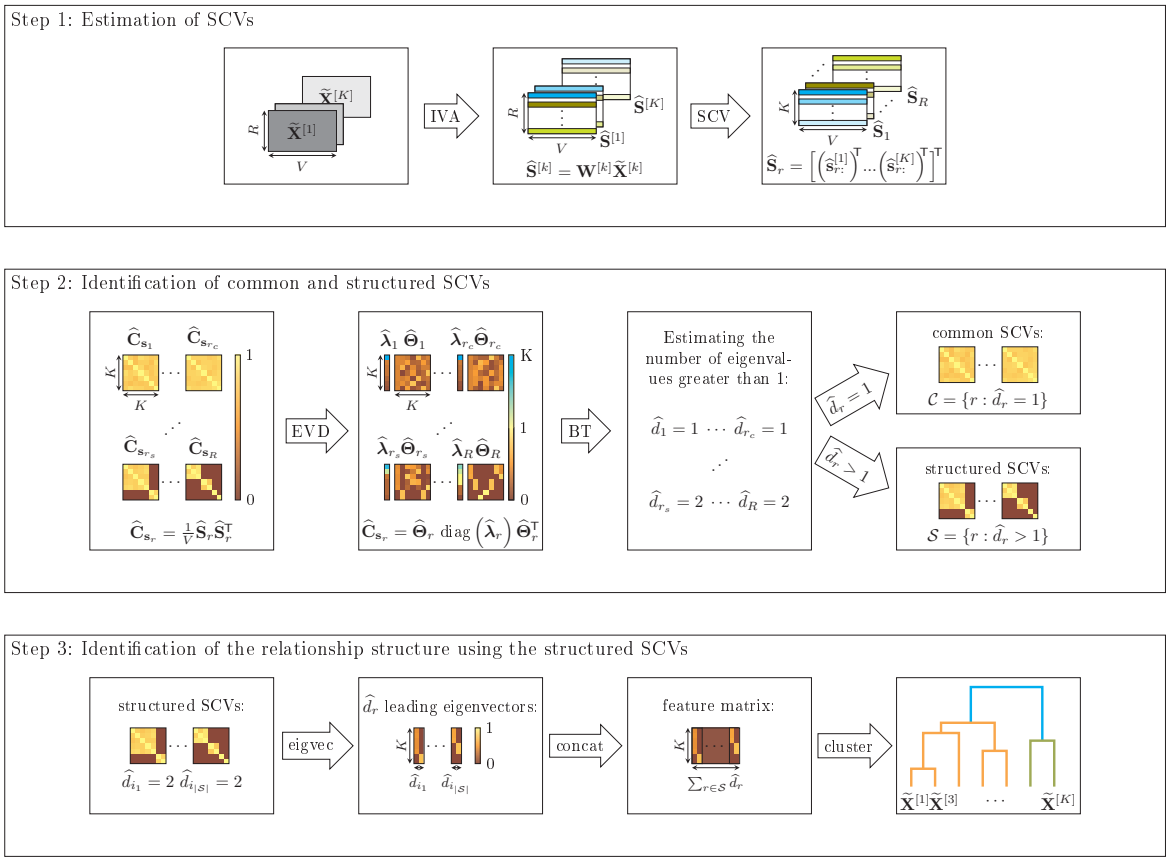


Figure 6.1: Visualization of the proposed method for identifying the relationship structure among multiple datasets. We have K (dimension-reduced) observed datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$, where V is the number of voxels and R the (reduced) dimension of the data. In step 1, the latent sources in the SCVs $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K \times V}$ are estimated by applying IVA-L-SOS on the observed datasets. In step 2, the SCVs are identified as structured or common by applying an eigenvalue decomposition (EVD) on their covariance matrices and using the proposed bootstrap technique (BT) to find $\widehat{d}_r$, the estimated number of eigenvalues greater than one in the r^{th} SCV. If $\widehat{d}_r = 1$, the r^{th} SCV is identified as common, and if $\widehat{d}_r > 1$, as structured. We denote the indices of the structured SCVs as $\mathcal{S} = \left\{i_1, \ldots, i_{|\mathcal{S}|}\right\}$. In Step 3, the $\widehat{d}_r$ leading eigenvectors of the structured SCVs are concatenated to form a feature matrix, which is the input of the hierarchical clustering. The resulting clusters are the identified groups (in this example, there are two groups: orange and green), and the dendrogram reveals the relationship structure among the datasets.

assumption in many applications including fMRI analysis [Ada22]—and therefore dimension reduction via PCA has been performed on $\mathbf{X}^{[k]}$ as preprocessing. The estimation of the R SCVs $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K \times V}$ from the dimension-reduced observed datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$ is shown in Step 1 in Figure 6.1.

6.2.2 Step 2: Identification of common and structured SCVs

In Step 2, it must be determined which of the estimated SCVs are common and which are structured. Only the structured SCVs contain information about the relationship among the datasets, and thus, only these structured SCVs are utilized in Step 3 to determine features for identifying the relationship structure.

In theory, the true SCVs can be identified as either common or structured based on the eigenvalues of their covariance matrices. Let the EVD of

$$\mathbf{C}_{\mathbf{s}_r} = \mathbb{E}\left\{\mathbf{s}_r \mathbf{s}_r^{\mathsf{T}}\right\},$$

the true covariance matrix of the r^{th} SCV, be

$$\mathbf{C}_{\mathbf{s}_r} = \boldsymbol{\Theta}_r \operatorname{diag}(\boldsymbol{\lambda}_r) \boldsymbol{\Theta}_r^{\mathsf{T}},$$

where $\boldsymbol{\lambda}_r \in \mathbb{R}^K$ contains the true eigenvalues, and $\boldsymbol{\Theta}_r \in \mathbb{R}^{K \times K}$ contains the corresponding true eigenvectors of $\mathbf{C}_{\mathbf{s}_r}$ as columns. If $\mathbf{C}_{\mathbf{s}_r} = \mathbf{I}_{K \times K}$, all its eigenvalues are equal to one. However, if $\mathbf{C}_{\mathbf{s}_r} \neq \mathbf{I}$, some eigenvalues are different from one. Let d_r be defined as the number of eigenvalues of $\mathbf{C}_{\mathbf{s}_r}$ that are greater than one. We assume that the SCV covariance matrices have only one of the following two structures:

- For *common* SCVs, where all the source components within an SCV are correlated with each other, the covariance matrix $\mathbf{C}_{\mathbf{s}_r}$ has ones on the diagonal, and all off-diagonal values (corresponding to the correlation coefficients) are non-zero. In [Has20], it is shown that, under certain conditions on correlation coefficients, $\mathbf{C}_{\mathbf{s}_r}$ has exactly one eigenvalue greater than one, i.e., $d_r = 1$.

- For *structured* SCVs, the covariance matrix $\mathbf{C}_{\mathbf{s}_r}$ is block-diagonal with ones on the diagonal and zero off-block values (or can be permuted into a block-diagonal matrix $\widetilde{\mathbf{C}}_{\mathbf{s}_r} = \mathbf{P}_r \mathbf{C}_{\mathbf{s}_r} \mathbf{P}_r^{\mathsf{T}}$, where $\mathbf{P}_r \in \mathbb{R}^{K \times K}$ is an orthogonal permutation matrix). We assume that the covariance matrices of structured SCVs have at least two blocks. To determine the number of eigenvalues greater than one for such SCV covariance matrices, we propose the following corollary:

  Corollary 6.1. *Let* $\mathbf{C}_{\mathbf{s}_r} \in \mathbb{R}^{K \times K}$ *be a block-diagonal matrix consisting of* $d_r \geq 2$

blocks $\mathbf{B}_i \in \mathbb{R}^{K_i \times K_i}$ *such that*

$$\mathbf{C}_{\mathbf{s}_r} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{B}_{d_r} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I} \end{bmatrix},$$

and $\mathbf{B}_i$ *contains ones on the diagonal and positive entries on its off-diagonal elements. Then,* $\mathbf{C}_{\mathbf{s}_r}$ *has exactly* d_r *eigenvalues greater than one.*

Since it is proven in [Has20] that $|\{\lambda(\mathbf{B}_i) > 1\}| = 1$ (and the remaining eigenvalues are less than or equal to one), it follows naturally for d_r blocks in $\mathbf{C}_{\mathbf{s}_r}$ that $|\{\lambda(\mathbf{C}_{\mathbf{s}_r}) > 1\}| = d_r$. Note that Corollary 6.1 also holds if $\widetilde{\mathbf{C}}_{\mathbf{s}_r}$ (and not $\mathbf{C}_{\mathbf{s}_r}$) is block-diagonal because $\widetilde{\mathbf{C}}_{\mathbf{s}_r}$ has the same eigenvalues as $\mathbf{C}_{\mathbf{s}_r}$.

As we assume that a structured SCV consists of at least two blocks, we can differentiate between a common SCV ($d_r = 1$) and a structured SCV ($d_r > 1$) by counting the number of eigenvalues greater than one. We define $\mathcal{C} = \{r : d_r = 1\}$ as the set of indices of the common SCVs and $\mathcal{S} = \{r : d_r > 1\}$ as the set of indices of the structured SCVs. In the following, we denote the indices in $\mathcal{S}$ as

$$\mathcal{S} = \{i_1, \dots, i_{|\mathcal{S}|}\}.$$

Note that we do not consider SCVs with identity covariance matrices, which would correspond to completely uncorrelated datasets, because we assume that in real-world data, some correlations typically exist in each SCV.

Now, when covariance matrices are estimated over finite samples V,

$$\widehat{\mathbf{C}}_{\mathbf{s}_r} = \frac{1}{V} \widehat{\mathbf{S}}_r \widehat{\mathbf{S}}_r^{\mathsf{T}},$$

the estimated correlation coefficients corresponding to the uncorrelated datasets will not be exactly zero, and thus, more than d_r eigenvalues for a structured SCV will be greater than one. Thus, by just counting the estimated eigenvalues greater than one, d_r would be overestimated. Consequently, it is necessary to estimate d_r. Estimating the number of significant eigenvalues is commonly addressed as a model-order selection problem in the literature [Wax85]. However, these model-order selection techniques assume certain asymptotic properties (as $V \to \infty$) on the non-significant eigenvalues, for example, assuming they all or a subset of them are equal to each other (in [Has20; Wax85]). These properties are not applicable to the SCV covariance matrices. For example, if

all the source components in an SCV covariance matrix belonged to one of the blocks, and thus there were no uncorrelated source components, then for arbitrary correlation coefficients, none of the non-significant eigenvalues would be equal to each other.

To estimate d_r for the r^{th} SCV covariance matrix $\mathbf{C}_{\mathbf{s}_r}$, we perform a binary hypothesis test for each $k = 1, \ldots, K$ with the null hypothesis

$$H_0 : \quad \lambda_r^{[k]} \leq 1,$$

and the alternative

$$H_1 : \quad \lambda_r^{[k]} > 1,$$

where $\lambda_r^{[k]}$ is the true (unknown) k^{th} eigenvalue of $\mathbf{C}_{\mathbf{s}_r}$. The estimated eigenvalues $\widehat{\lambda}_r^{[k]}$ are used as follows to perform the hypothesis test. A test statistic $\widehat{T} = \widehat{\lambda}_r^{[k]} - 1$ is defined, and to perform the hypothesis test, the distribution of the statistic under H_0 [Kay98] must be known. However, neither the sample nor the asymptotic distribution of the test statistic is known. We, therefore, propose a bootstrap-based hypothesis test [Zou04] to estimate this distribution. Under certain conditions, the distribution estimated by bootstrap converges to the true distribution if the number of samples goes to infinity ($V \to \infty$) [Dav06], and thereby we can estimate d_r for each SCV.

The pseudocode for the proposed bootstrap technique for estimating d_r is given in Algorithm 14. In the following, we describe the steps of the algorithm.

1. The sample covariance matrix $\widehat{\mathbf{C}}_{\mathbf{s}_r} = \left| \frac{1}{V} \widehat{\mathbf{S}}_r \widehat{\mathbf{S}}_r^{\mathsf{T}} \right|$ is calculated (line 1). The absolute value is necessary because of a possible sign ambiguity in the sources. Then, an EVD is applied on $\widehat{\mathbf{C}}_{\mathbf{s}_r}$ to get the eigenvalues $\widehat{\boldsymbol{\lambda}}_r = \left[\widehat{\lambda}_r^{[1]}, \ldots, \widehat{\lambda}_r^{[K]} \right]^{\mathsf{T}}$, sorted in descending order (line 2). Here, $\widehat{\lambda}_r^{[k]}$ denotes the estimate of $\lambda_r^{[k]}$.

2. The SCVs are resampled with replacement for B times, where for each resampling, V indices are drawn from $\mathcal{U}_{\text{int}}(1, V)$ (uniform distribution of integers between 1 and V) (line 4), and the SCVs are resampled on those indices (line 5). The resampled SCVs ${}_b\widehat{\mathbf{S}}_r$ thus also have V samples. The prescript b denotes the b^{th} bootstrap resample.

3. The eigenvalues ${}_b\widehat{\boldsymbol{\lambda}}_r$ from the covariance matrices ${}_b\widehat{\mathbf{C}}_{\mathbf{s}_r} = \left| \frac{1}{V} {}_b\widehat{\mathbf{S}}_r \, {}_b\widehat{\mathbf{S}}_r^{\mathsf{T}} \right|$ are calculated (lines 6–7).

4. The test statistic $\widehat{T} = \widehat{\lambda}_r^{[k]} - 1$ is defined (line 9).

5. The test statistic ${}_b\widehat{T}^* = {}_b\widehat{\lambda}_r^{[k]} - \widehat{\lambda}_r^{[k]}$ is calculated (line 11).

6. The values of ${}_b\widehat{T}^*$ are sorted in ascending order (line 12). Using a given false alarm probability P_{fa}, a threshold ${}_\eta\widehat{T}^*$ is found, where $\eta = \lfloor (B+1)\cdot(1-P_{\text{fa}})\rfloor$ (line 13), and ${}_\eta\widehat{T}^*$ is ${}_b\widehat{T}^*$ for $b=\eta$. If $\widehat{T} > {}_\eta\widehat{T}^*$, then H_0 is rejected, i.e., $\lambda_r^{[k]}$ is greater than one with a significance of $1-P_{\text{fa}}$, and the k^{th} element in a vector $\mathbf{z}$ is set to one (lines 14–17).

7. In the end, the number of ones in $\mathbf{z}$ is counted, which equals to $\widehat{d}_r$, the estimated number of eigenvalues greater than one in the r^{th} SCV (line 18).

Algorithm 14 Bootstrap technique for estimating d_r

Input: $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K\times V}$, B, P_{fa}

1: $\widehat{\mathbf{C}}_{\mathbf{s}_r} \leftarrow \text{abs}\left(\frac{1}{V}\widehat{\mathbf{S}}_r\widehat{\mathbf{S}}_r^{\mathsf{T}}\right)$
2: $\widehat{\boldsymbol{\lambda}}_r, \widehat{\boldsymbol{\Theta}}_r \leftarrow \text{EVD}(\widehat{\mathbf{C}}_{\mathbf{s}_r})$ ▷ s.t. $\widehat{\lambda}_r^{[1]} \geq \cdots \geq \widehat{\lambda}_r^{[K]}$
3: **for** $b = 1, \ldots, B$ **do**
4: ${}_b\mathbf{j} \leftarrow \text{randint}(1, V, V)$ ▷ V integers $\sim \mathcal{U}_{\text{int}}(1, V)$
5: ${}_b\widehat{\mathbf{S}}_r \leftarrow \widehat{\mathbf{S}}_r(:, {}_b\mathbf{j})$
6: ${}_b\widehat{\mathbf{C}}_{\mathbf{s}_r} \leftarrow \text{abs}\left(\frac{1}{V}\,{}_b\widehat{\mathbf{S}}_r\,{}_b\widehat{\mathbf{S}}_r^{\mathsf{T}}\right)$
7: ${}_b\widehat{\boldsymbol{\lambda}}_r,\ {}_b\widehat{\boldsymbol{\Theta}}_r \leftarrow \text{EVD}({}_b\widehat{\mathbf{C}}_{\mathbf{s}_r})$
8: **for** $k = 1, \ldots, K$ **do**
9: $\widehat{T} \leftarrow \widehat{\lambda}_r^{[k]} - 1$
10: **for** $b = 1, \ldots, B$ **do**
11: ${}_b\widehat{T}^* \leftarrow {}_b\widehat{\lambda}_r^{[k]} - \widehat{\lambda}_r^{[k]}$
12: $\widehat{\mathbf{T}}^* \leftarrow \text{sort}(\widehat{\mathbf{T}}^*)$ ▷ s.t. ${}_1\widehat{T}^* \leq \cdots \leq {}_B\widehat{T}^*$
13: $\eta = \lfloor (B+1)\cdot(1-P_{\text{fa}})\rfloor$
14: **if** $\widehat{T} < {}_\eta T^*$ **then**
15: $z^{[k]} \leftarrow 0$ ▷ $\lambda_r^{[k]} \leq 1$
16: **else**
17: $z^{[k]} \leftarrow 1$ ▷ H_0 is rejected. $\lambda_r^{[k]} > 1$
18: $\widehat{d}_r = \text{count\_nonzero}(\mathbf{z})$ ▷ number of elements in $\mathbf{z}$ equal to one

Output: $\widehat{d}_r$ ▷ estimate of d_r

The identification of common and structured SCVs is shown in Step 2 in Figure 6.1.

6.2.3 Step 3: Identification of the relationship structure using the structured SCVs

In Step 3, the eigenvectors of the covariance matrices of the structured SCVs are used as features for the hierarchical clustering. In [Has20], it is shown that the eigenvector

corresponding to an eigenvalue greater than one characterizes the correlated datasets in the corresponding block. More specifically, for each eigenvalue greater than one, all eigenvector elements corresponding to datasets that are not part of the correlated datasets are zero, while the eigenvector elements corresponding to the correlated datasets are greater than zero. This means that the d_r leading eigenvectors contain information about the relationship among the datasets within the r^{th} SCV. As stated in Step 2, the covariance matrices are estimated from finite samples, and d_r is also estimated with the proposed bootstrap-based hypothesis test. We will thus use this estimated $\widehat{d}_r$ from Step 2. As different SCVs provide complementary information about the relationship among the datasets, the $\widehat{d}_r$ leading eigenvectors from all structured SCV covariance matrices ($r \in \mathcal{S}$) are horizontally concatenated to form a feature matrix, $\mathbf{F}$, of $\sum_{r\in\mathcal{S}} \widehat{d}_r$ eigenvector columns, which is then fed into the hierarchical clustering:

$$\mathbf{F} = \text{concat}\left(\widehat{\mathbf{\Theta}}_{i_1}\left(:, 1:\widehat{d}_{i_1}\right), \ldots, \widehat{\mathbf{\Theta}}_{i_{|\mathcal{S}|}}\left(:, 1:\widehat{d}_{i_{|\mathcal{S}|}}\right)\right) \in \mathbb{R}^{K\times\sum_{r\in\mathcal{S}}\widehat{d}_r}.$$

This way, the proposed method leverages the knowledge of all SCVs together instead of performing an analysis separately on each SCV. The advantage of using hierarchical clustering, compared with, e.g., K-means clustering, is that no prior knowledge or estimation of the number of clusters is necessary. Additionally, while K-means clustering would only identify the groups of datasets, which correspond to the resulting clusters, hierarchical clustering also identifies the relationship structure among the datasets, which is revealed in the dendrogram. The identification of the relationship structure is shown in Step 3 in Figure 6.1.

6.2.4 Computational complexity of the proposed method

In the following, we compute the big-$\mathcal{O}$ complexity for each step of the proposed method. In Step 1, the complexity is dominated by the multiplication of $\widehat{\mathbf{C}}_{\mathbf{s}_r}^{-1} \in \mathbb{R}^{K\times K}$ with $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K\times V}$ in the main loop of IVA-L-SOS. In each iteration, $\widehat{\mathbf{C}}_{\mathbf{s}_r}^{-1}$ is updated K times per SCV (and thus this multiplication is performed K times per SCV), which results for all R SCVs in a complexity of $\mathcal{O}(IRK^3V)$, where I is the number of iterations. In Step 2, the dominant cost is the calculation of the covariance matrices of the resampled SCVs ${}_b\widehat{\mathbf{S}}_r \in \mathbb{R}^{K\times V}$, which for all B bootstrap resamples in all R SCVs has complexity $\mathcal{O}(BRK^2V)$. In Step 3, the complexity is dominated by the hierarchical clustering of the K datasets, which is of complexity $\mathcal{O}(K^3)$ [Sas13]. Thus, the method has a big-$\mathcal{O}$ complexity of $\mathcal{O}(RK^2V(IK+B))$.

6.3 Simulations

To demonstrate the performance of the proposed bootstrap technique, we simulate common and structured SCVs.[2] We generate the entries of the true SCV covariance matrices $\mathbf{C}_{\mathbf{s}_r}$ according to the following model:

$$\mathbf{C}_{\mathbf{s}_r}(k,l) = \begin{cases} 1, & k = l \quad \text{(diagonal elements)} \\ 0, & k \neq l, \mathbf{C}_{\mathbf{s}_r}(k,l) \notin \mathbf{B}_i, 1 \leq i \leq d_r \\ \rho + n_r(k,l) & k \neq l, \mathbf{C}_{\mathbf{s}_r}(k,l) \in \mathbf{B}_i, 1 \leq i \leq d_r \end{cases}$$

where $\mathbf{B}_i$ is the i^{th} block, ρ is the correlation coefficient of the correlated sources, and $n_r(k,l) = n_r(l,k) \sim \mathcal{N}(0, 0.0025)$ is added variability so that IVA-L-SOS can identify the sources correctly. Note that this does not mean that the sources are also aligned correctly; in fact, the conditions for correct alignment are not satisfied because of the block structure of the SCVs, which violates Assumption (JBSS) on page 69. However, if the blocks of correlated sources are interchanged between SCVs, this is not a problem for our proposed method as the $\widehat{d}_r$ eigenvectors of all structured SCVs are concatenated in the end. Figure 6.2 shows the structure of the six 10×10 SCV covariance matrices in our simulations for $\rho = 0.2$. $\mathbf{C}_{\mathbf{s}_1}$ corresponds to a common SCV, i.e., $d_1 = 1$, while the other SCVs are structured. $\mathbf{C}_{\mathbf{s}_2}$, $\mathbf{C}_{\mathbf{s}_3}$ and $\mathbf{C}_{\mathbf{s}_4}$ all have two blocks, i.e., $d_2 = d_3 = d_4 = 2$, $\mathbf{C}_{\mathbf{s}_5}$ has three blocks, i.e., $d_5 = 3$, and $\mathbf{C}_{\mathbf{s}_6}$ has four blocks, i.e., $d_6 = 4$. Note that $\mathbf{C}_{\mathbf{s}_3}$ and $\mathbf{C}_{\mathbf{s}_4}$ contain one and four uncorrelated source components, respectively. Using these SCV covariance matrices, we generate $R = 6$ SCVs $\mathbf{S}_r \in \mathbb{R}^{K \times V}$, each with $K = 10$ source components and $V = 1000$ samples. The samples are drawn from a Laplacian distribution as described in [Kot01, section 6.4], with zero mean and with covariance matrices specified as shown in Figure 6.2.

We perform two experiments. In the first experiment in Subsection 6.3.1, we compare the performance of the proposed bootstrap technique and two competing methods for estimating d_r. Here, we estimate d_r from the generated sample SCVs because, as explained before, the identified sources from IVA-L-SOS may not be aligned correctly among SCVs, i.e., the correlated blocks in the covariance matrices may be permuted among SCVs. Due to this, $\widehat{d}_r$ of the output of IVA-L-SOS may not match $\widehat{d}_r$ of the generated SCVs. In the second experiment in Subsection 6.3.2, we apply the complete method, including the IVA step and the hierarchical clustering, on the simulated observed datasets and investigate the robustness of the method for different correlation

[2] The Python code for the simulations is available at: https://github.com/SSTGroup/relationship_structure_identification.

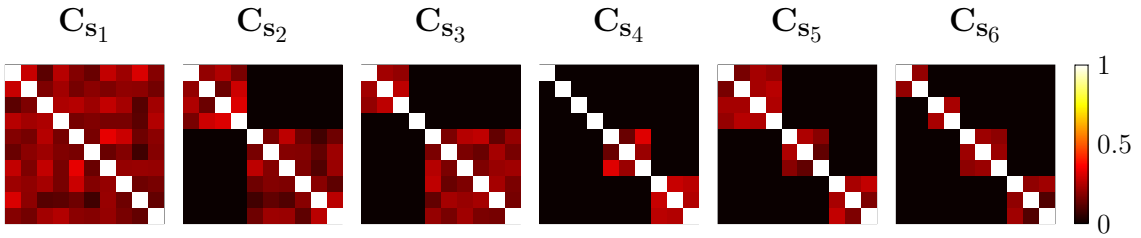


Figure 6.2: $R = 6$ simulated SCV covariance matrices (of dimension 10×10) for $\rho = 0.2$. We have $d_1 = 1$, $d_2 = d_3 = d_4 = 2$, $d_5 = 3$, and $d_6 = 4$. Using these covariance matrices, 6 SCVs with Laplacian-distributed sources are generated.

coefficients ρ. In both of these scenarios, we did not consider noise, as we assumed that for real-world data, noise is effectively removed during the PCA-based dimension reduction step in the preprocessing. This is based on the assumption that the dimension of the observed datasets is higher than the number of latent source components, which is a common assumption in fMRI data analysis [Ada22].

6.3.1 Estimating d_r

We evaluate the performance of three techniques for estimating d_r from the generated sample SCVs. The first technique is our proposed bootstrap (BT) technique, described in Algorithm 14 on page 100, with $B = 1000$ bootstrap resamples and $P_{\text{fa}} = 0.05$. The second technique based on Hard Thresholding (HT) of the eigenvalues [Akh21] directly counts how many eigenvalues are greater than one in the r^{th} SCV's covariance matrix, i.e., $\widehat{d}_r = |\{k : \lambda_r^{[k]} > 1\}|$. The third technique is the Gershgorin Disc (GD) technique from [Yan22].

We simulate 50 Monte-Carlo runs. We investigate the two performance metrics

$$P\left(\widehat{d}_r = d_r\right) = \frac{\#\left(\widehat{d}_r = d_r\right)}{\#\text{runs}},$$

which estimates the probability that d_r is estimated correctly, and

$$\mu_{\widehat{d}_r},$$

which is the average value of the estimated $\widehat{d}_r$.

The boxplot of $P\left(\widehat{d}_r = d_r\right)$ for the $R = 6$ SCVs is shown in Figure 6.3 for different values of ρ, with circles denoting outliers. Notably, the proposed BT technique demonstrates robust performance by accurately estimating d_r even for very small correlation values in the underlying data, showcasing its effectiveness in handling the Laplacian

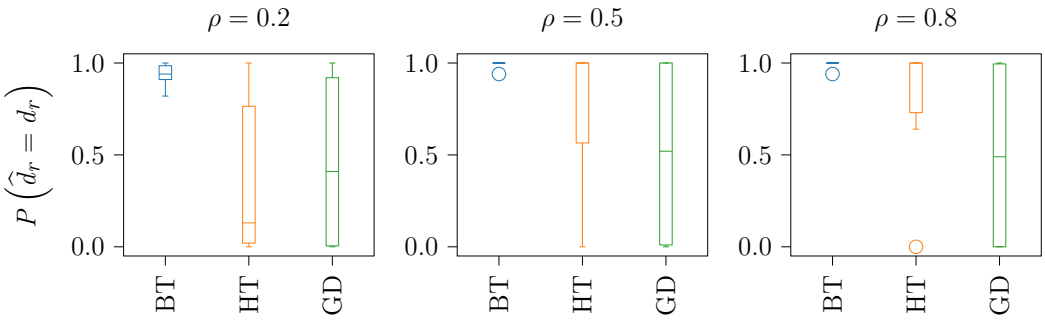


Figure 6.3: Boxplot of $P\left(\widehat{d}_r = d_r\right), r = 1, \ldots, 6$, for different values of ρ for the BT, HT and GD technique. The higher ρ, the more accurate is the estimation of d_r for the BT and HT techniques. The GD technique does not depend on ρ but fails for all values. The proposed BT technique is superior to HT and GD for all values of ρ.

(non-Gaussian) data distribution and varying correlation values ρ. In contrast, the HT technique only achieves high $P\left(\widehat{d}_r = d_r\right)$ values with increasing ρ, while the GD technique does not perform well for all values of ρ.

To investigate the reason for the small values of $P\left(\widehat{d}_r = d_r\right)$ in the HT and GD technique, we show $\mu_{\widehat{d}_4}$ and $\mu_{\widehat{d}_5}$ in Figure 6.4, estimated for the BT, HT and GD techniques. We see that the HT technique (blue triangles) overestimates d_4 (also d_3, not shown here), with an improvement for higher ρ values, while the BT (blue circles) and GD (blue squares) techniques estimate d_4 close to the ground truth for all ρ values. Furthermore, we see that the GD technique (orange squares) strongly underestimates d_5 (also d_2 and d_6, not shown) compared with the BT (orange circles) and HT (orange triangles) techniques for all ρ values. The over- and underestimation of d_r in the HT and GD techniques is the reason for the decreased values of $P\left(\widehat{d}_r = d_r\right)$ compared with the BT technique. Thus, the proposed BT technique is superior to HT and GD in

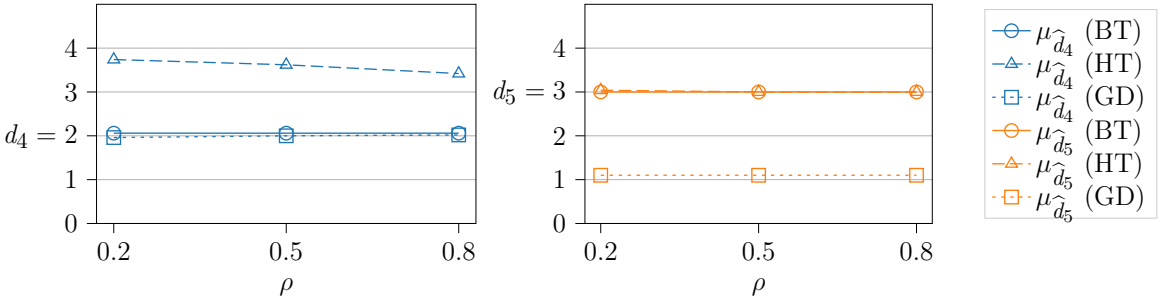


Figure 6.4: Average value of $\widehat{d}_r$ for SCVs 4 and 5 for different values of ρ for the BT, HT and GD technique. The true values are $d_4 = 2$ and $d_5 = 3$. The proposed BT technique estimates both values close to the ground truth, while the HT technique overestimates d_4 and the GD technique underestimates d_5.

terms of estimating d_r.

6.3.2 Identifying the relationship structure among multiple datasets

We have seen in the previous subsection that the BT technique is superior to the HT and GD techniques in terms of estimating d_r. Therefore, we evaluate the complete method for identifying the relationship structure only with the BT technique. We are not aware of other methods that identify the relationship structure—and not only groups of datasets—with which we could compare the complete proposed method.

We generate the SCVs as described in Subsection 6.3.1 and the datasets as $\mathbf{X}^{[k]} = \mathbf{A}^{[k]}\mathbf{S}^{[k]}$, where the elements of $\mathbf{A}^{[k]} \in \mathbb{R}^{6\times 6}$ are drawn from $\mathcal{N}(0,1)$. After performing IVA-L-SOS with 20 random initializations and choosing the estimated sources of the most consistent run[3] as in [Lon18], the BT technique is applied to the estimated SCVs to find $\widehat{d}_r$. The $\widehat{d}_r$ leading eigenvectors of the structured SCVs are concatenated in a feature matrix $\mathbf{F}$, which we feed into the hierarchical clustering. Hierarchical clustering is performed using the `linkage` function from scikit-learn [Ped11] with the 'ward' linkage (minimizing the variance of the clusters) and visualized using the `dendrogram` function. Also, here, we simulate 50 Monte-Carlo runs.

Figure 6.5 shows the dendrogram revealing the true relationship structure among the datasets (for $\rho = 0.2$, dendrograms for the other values of ρ are very similar). For creating this ground truth dendrogram, we set d_r manually and concatenate the d_r eigenvectors of the true (instead of sample) covariance matrices as the input for the hierarchical clustering. The labels of the datasets are denoted on the x-axis of the dendrogram, and the y-axis shows the distance between the clusters. A small distance means that the corresponding clusters are very similar. We see that three groups exist in the datasets, i.e., datasets 5–7 form one group (orange), datasets 8–10 form a second group (green), and datasets 1–4 form a third group (red). We also see that the groups of datasets 5–7 and 8–10 are closer to each other than to those of datasets 1–4. The labels that result from the clustering are the same within each cluster but with arbitrary ordering. We denote the true label for the k^{th} dataset with c_k and choose $c_1 = c_2 = c_3 = c_4 = 1$, $c_5 = c_6 = c_7 = 2$, and $c_8 = c_9 = c_{10} = 3$.

Figure 6.6 shows the SCV covariance matrices estimated by IVA-L-SOS for one of the 50 Monte-Carlo runs. Comparing these estimated covariance matrices with the true covariance matrices in Figure 6.2, we see that some of the correlated blocks are

[3]The Python implementation is available at https://github.com/SSTGroup/independent_vector_analysis.

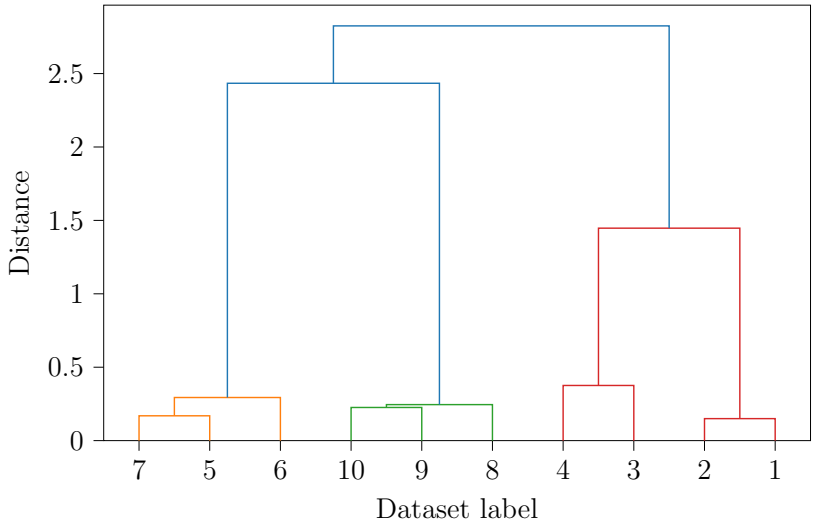


Figure 6.5: The ground truth dendrogram (for $\rho = 0.2$) shows the relationship structure among the 10 datasets. There exist three groups: consisting of datasets 5–7 (orange), 8–10 (green), and 1–4 (red).

permuted among the SCV covariance matrices. For example, the (4×4)-block in $\mathbf{C}_{\mathbf{s}_2}$ has moved to $\widehat{\mathbf{C}}_{\mathbf{s}_4}$, and the (2×2)-block in $\mathbf{C}_{\mathbf{s}_6}$ has moved to $\widehat{\mathbf{C}}_{\mathbf{s}_3}$. Even though the blocks are permuted, the SCV covariance matrices still provide information about the relationship among the datasets, and therefore, the identification of the relationship structure is not affected by this, as we will see in the following.

We are not aware of a performance metric that captures how well the relationship structure among the datasets is estimated; instead, we use the Adjusted Mutual Information (AMI) [Vin09] between the true and estimated clusters, which evaluates if the groups are identified correctly. The AMI is a normalized metric based on the mutual information of the true and estimated clustering, i.e., the AMI is equal to one if the true and the estimated clusters are equal, and equal to zero if the mutual information between the true and estimated clusters equals the expected value of the mutual information between the true and a random clustering. Furthermore, the AMI

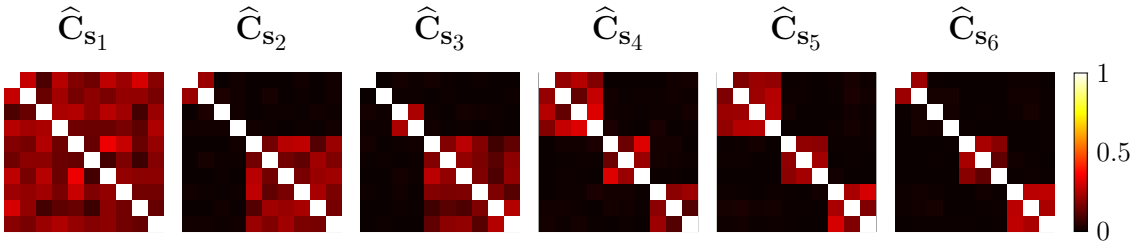


Figure 6.6: $R = 6$ estimated SCV covariance matrices (of dimension 10×10) for $\rho = 0.2$.

corrects for the permutation ambiguity between the true and the estimated clusters. The proposed method achieves an AMI of one for $\rho = 0.2, 0.5, 0.8$, i.e., it correctly identifies the groups in 100% of the runs for all the correlation values.

6.4 Summary

In this chapter, we have proposed a method to identify the relationship structure among multiple datasets using JBSS and thereby addressed the second challenge of this thesis. The proposed method consists of three steps: 1) estimating latent sources from observed datasets using IVA-L-SOS, 2) identifying structured SCVs (SCVs whose covariance matrices have more than one eigenvalue greater than one), and 3) extracting features from the structured SCVs, which are then used by hierarchical clustering to identify the relationship structure among the datasets. In simulations, we have first demonstrated the superior performance of the second step of the method against the techniques in [Akh21] and [Yan22]. Compared with these approaches, the proposed method eliminates the need to assume Gaussianity in the data by 1) including higher-order statistics through the use of IVA-L-SOS for source estimation, which leads to more interpretable components, and 2) not relying on any distributional assumptions, which is achieved by estimating the number of eigenvalues greater than one using a bootstrap-based hypothesis test. Thereafter, we have verified the success of the proposed method for identifying the relationship structure among multiple datasets, where it has achieved perfect AMI for different correlation values between the sources. To the best of our knowledge, no competing technique exists in the literature that identifies the relationship structure of multiple datasets using JBSS. A limitation of the proposed method is that it does not identify SCVs with identity covariance matrices, i.e., SCVs that consist of completely uncorrelated sources, as they typically do not exist in real-world data. However, a possible way to overcome this limitation in the future may be to adapt the bootstrap test. In the next chapter, we apply the proposed method to multi-task fMRI data and demonstrate its success also for real-world data.

7 JBSS for multi-task fMRI data analysis

One of the important applications of JBSS is *data fusion*, i.e., the joint analysis of multiple related datasets such that they can interact and inform each other [Lah15]. JBSS methods perform data fusion by making use of the dependence of the latent sources across datasets [Lah15; Ada15; Hun17; Ada19]. Over the years, there has been an increase in the utilization of IVA for functional Magnetic Resonance Imaging (fMRI) data fusion because IVA 1) naturally aligns the sources across datasets [LS17], 2) effectively retains subject variability in multi-subject fMRI data [Bhi19], and 3) can provide information about the relationship among multiple datasets [Lon20]. The interest in data fusion of *multi-task* fMRI data, where datasets are collected from the same subjects while they are performing different tasks, has grown [Cal06; LS17; San20]. FMRI datasets from multiple tasks provide complementary information about the brain [Ram15; Mij12] because different tasks involve cognitive functions that are either task-specific or common across all tasks [San20]. By performing data fusion, the function of and relationship among brain networks can be identified [San20], which eventually helps to understand the brain organization [Con22].

In this chapter, we analyze fMRI datasets from patients with schizophrenia and healthy controls who perform multiple tasks. Schizophrenia is a neuropsychiatric disorder associated with cognitive deficits [Ram15; Mwa17] and altered connections between brain networks [Mic09], which might not be captured when analyzing only the data from a single task [Ram15]. Performing data fusion on these multi-task datasets, therefore, may help to better understand schizophrenia and to reveal latent neural patterns called *biomarkers*, e.g., a common network among the tasks, that capture differences between patients with schizophrenia and healthy controls [Cal06; Mic09]. Previous studies that concentrated on performing JBSS to distinguish between patients with schizophrenia and healthy controls made use of a single fMRI feature vector per task, e.g., [LS17]. It is also of interest to analyze multiple feature vectors per task, which report on different aspects of the task, to examine the relationship among different brain networks in this

task.

In the following section, we introduce the multi-task fMRI datasets. Thereafter, we apply IVA-G and PARAFAC2 for data fusion of the multi-task fMRI data, where we identify potential biomarkers for schizophrenia and show that the SCV covariance matrices provide information about the relationship among the task datasets. Building on these findings, we apply the proposed method from Chapter 6 for identifying the relationship structure among multiple datasets. We verify the success of the proposed method by identifying a meaningful relationship structure among the multi-task fMRI datasets, as the method groups together the task datasets from similar tasks, and we again reveal brain networks with significantly different activations in patients with schizophrenia and healthy controls that might be potential biomarkers for the disorder.[1]

7.1 fMRI dataset and preprocessing

We analyze fMRI datasets from the MIND Clinical Imaging Consortium (MCIC) collection [Gol13], which are collected from 271 subjects (121 patients with schizophrenia and 150 healthy controls) that perform three different tasks: Auditory Oddball (AOD), Sensory Motor (SM), and Sternberg Item Recognition Paradigm (SIRP).

In the AOD task, three stimuli are played: a frequent standard stimulus (1 kHz tone), an infrequent target stimulus (1.2 kHz tone), and an infrequent novel stimulus (computer-generated, complex sound). Whenever the target stimulus occurs, the subject must press a button with the right index finger [Gol13]. During the SM task, in total 16 different audio tones are played, in increasing order until the highest pitch is reached, then in decreasing order. The subject must press a button with the right thumb every time a new tone occurs [Gol13]. The SIRP task consists of two phases, encoding (SIRP-E) and probe (SIRP-P). In the encoding phase, a set of one, three, or five integer digits (randomly selected from zero to nine) is presented on a screen, and the subject needs to memorize this set. In the following probe (SIRP-P), digits are presented subsequently in a pseudo-random order. The subject must press a button with the right thumb if the digit was in the set and with the left thumb if not [Gol13].

For each of the subjects, multiple three-dimensional brain scans were collected over

[1] Sections 7.1 and 7.3 of this chapter are based on the paper: "Identifying the Relationship Structure among Multiple Datasets Using Independent Vector Analysis: Application to Multi-task fMRI Data," I. Lehmann, T. Hasija, B. Gabrielson, M. A. B. S. Akhonda, V. D. Calhoun, T. Adali, *IEEE Access*, 2024, pp. 109443–109456. Section 7.2 of this chapter is based on the paper: "Multi-task fMRI Data Fusion Using IVA and PARAFAC2," I. Lehmann, E. Acar, T. Hasija, M. A. B. S. Akhonda, V. D. Calhoun, P. Schreier, and T. Adali, *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1466–1470.

time during each task. The recorded scans were then preprocessed as in [Mic09; LS17]: Using the Statistical Parametric Mapping (SPM) MATLAB toolbox [Mem20], a simple voxelwise linear regression was applied to the data to eliminate the temporal dimension. The regressors were created by convolving the default Hemodynamic Response Function (HRF) in SPM with the desired *predictors* for each task, which will be described in the next paragraph. For each subject and task, the resulting regression coefficient maps, also called "contrast images", are flattened and used as one-dimensional feature vectors that capture the variations across subjects. The flattened feature vectors of length $V = 48546$ voxels are concatenated for the $N = 271$ subjects to create the k^{th} task dataset $\mathbf{X}^{[k]} \in \mathbb{R}^{N \times V}$, $k = 1, \ldots, K$.

In the AOD task, the occurrences of the novel stimuli (AOD-N), the novel with standard stimuli (AOD-NS), the target stimuli (AOD-T), and the target with standard stimuli (AOD-TS) are each modeled as delta functions and used as predictors [LS17]. Thus, there are four task datasets for the AOD task. In the SM task, the entire increase and decrease block is used as the predictor [LS17]; thus, there is one task dataset for the SM task. For both phases of SIRP (E and P), also the whole block is used as the predictor. This way, the data allows us to analyze the learning and retrieval phases of this task separately. There is one task dataset for the encoding phase and the probe phase for each one, three, and five digits, and the average data; thus, there are eight task datasets for the SIRP task.

We expect task-specific activations of the auditory brain networks for the auditory tasks (AOD and SM) and of the visual brain networks for the visual tasks (SIRP), along with activations of the DMN for all tasks. As the task datasets are collected from patients and controls, we can use them to evaluate which tasks more clearly discriminate between patients with schizophrenia and healthy controls and thus may be involved in the development of novel biomarkers. Furthermore, our prior knowledge about the relationship among the task datasets provides us with the opportunity to directly assess the success of the proposed method from Chapter 6 with real-world data, which is typically not easily achievable, as for real-world data, one often does not know the true relationship structure.

7.2 IVA-G and PARAFAC2 for multi-task fMRI data fusion

In this section, we study IVA-G and PARAFAC2 for data fusion of the multi-task fMRI data. We use all $K = 13$ task datasets described in Section 7.1, i.e, four AOD datasets,

one SM dataset, and eight SIRP datasets. By making use of these rich multi-task fMRI datasets, we show how the two methods can be used for achieving two important goals at once, namely 1) finding interpretable latent source components that allow to discriminate between patients with schizophrenia and healthy controls and 2) providing information about the relationship among the tasks in the SCVs. To the best of our knowledge, this is the first application of PARAFAC2 to perform JBSS on multiple fMRI task datasets with the goal of obtaining information about the relationship among the task datasets based on the latent sources.[2]

7.2.1 Implementation details

As a common assumption in many applications, including fMRI analysis, is that the datasets can be modeled with a lower-dimensional set of latent sources [Ada22], a dimension reduction is performed via PCA before applying IVA-G, as described in Section 4.3. Like it is mostly the case for real-world data, also in the MCIC data, we do not know the true number R of the latent sources. Since PARAFAC2 may have an upper limit on R as described in Section 4.3, we have studied different values for R up to the limit for PARAFAC2, and we obtained the most interpretable source components for $R = 2$. To have a fair comparison of the two methods, we also have chosen the PCA dimension $R = 2$ and run IVA-G on the dimension-reduced datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$, $k = 1, \ldots, K$. We report the results of the best run of IVA-G and PARAFAC2, which are found as described in Subsection 4.3.3.

The R source components $\widehat{\mathbf{s}}_{r:}^{[k]} \in \mathbb{R}^{1 \times V}$, i.e., the r^{th} rows of $\widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{R \times V}$ estimated by IVA-G and PARAFAC2, are normalized to unit variance. To visualize the active brain networks, we calculate the (scalar) mean $\mu_r^{[k]}$ and standard deviation $\sigma_r^{[k]}$ of each $\widehat{\mathbf{s}}_{r:}^{[k]}$, and calculate the z-score of each voxel in $\widehat{\mathbf{s}}_{r:}^{[k]}$ by subtracting $\mu_r^{[k]}$ and dividing by $\sigma_r^{[k]}$. Then, each z-scored source component is reshaped into a tensor $\boldsymbol{\mathcal{Z}}_r^{[k]} \in \mathbb{R}^{V_x \times V_y \times V_z}$. We show $L = 9$ slices of the resulting tensor, where we visualize only the voxels (i, j) where $\left|\boldsymbol{\mathcal{Z}}_r^{[k]}(i, j, l)\right| > 2$. The visible voxels in $\boldsymbol{\mathcal{Z}}_r^{[k]}(:, :, l)$ form the active brain networks in the l^{th} slice.

The estimated mixing matrices $\widehat{\mathbf{A}}^{[k]} \in \mathbb{R}^{N \times R}$ are reconstructed by multiplying the PCA transformation matrix $\boldsymbol{\Psi}^{[k]}$ with the inverse of the estimated demixing matrix

[2]This section is based on the paper: "Multi-task fMRI Data Fusion Using IVA and PARAFAC2," I. Lehmann, E. Acar, T. Hasija, M. A. B. S. Akhonda, V. D. Calhoun, P. Schreier, and T. Adali, *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1466–1470.

$\widetilde{\mathbf{W}}^{[k]}$:

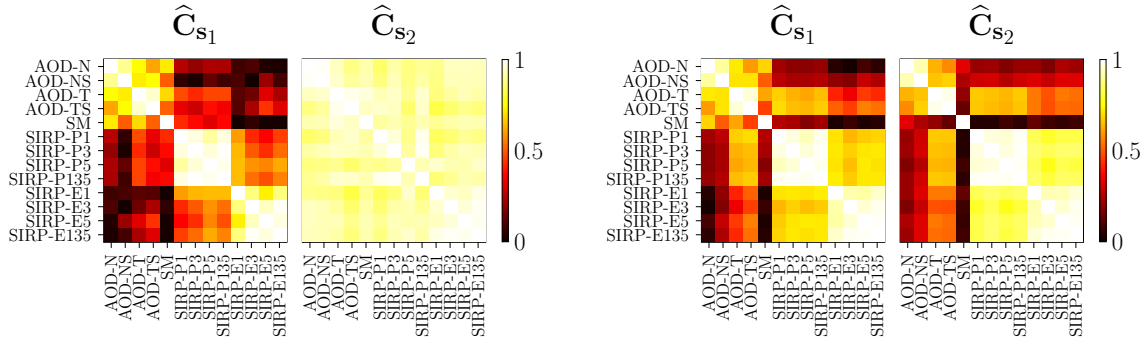
$$\widehat{\mathbf{A}}^{[k]} = \mathbf{\Psi}^{[k]} \left(\widetilde{\mathbf{W}}^{[k]} \right)^{-1} \in \mathbb{R}^{N \times R},$$

and the r^{th} column $\widehat{\mathbf{a}}_r^{[k]}$ is the subject profile corresponding to the r^{th} brain network in the k^{th} dataset. As we are also interested in identifying the source components that discriminate between patients and controls, we apply a two-sample t-test on the first 150 (controls) and the following 121 (patients) values of each column $\widehat{\mathbf{a}}_r^{[k]} \in \mathbb{R}^{271}$. We consider p-values smaller than 0.05 to be significant, indicating that the corresponding active brain networks in $\widehat{\mathbf{s}}_{r:}^{[k]}$ are different in patients with schizophrenia and healthy controls. Using this t-test, we corrected the signs of the estimated sources $\widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{R \times 48546}$ to overcome the sign ambiguity of IVA: We made sure that the t-values of the datasets that show a significant difference between patients and controls ($p < 0.05$) are positive or made positive by multiplying the estimated source components $\widehat{\mathbf{s}}_{r:}^{[k]}$ and corresponding subject profiles $\widehat{\mathbf{a}}_r^{[k]}$ by -1 (if the t-value is negative). This way, positive values of the (zero-mean) sources indicate higher activations in controls, and negative values indicate higher activation in patients.

7.2.2 Results

The SCV covariance matrices (absolute values) estimated by both IVA-G and PARAFAC2 are shown in Figure 7.1. Light values correspond to high correlations. We see that the estimated SCV covariance matrices $\widehat{\mathbf{C}}_{\mathbf{s}_n}$ of both methods provide information about the relationship among the task datasets. In Figure 7.1(a), the first SCV in IVA-G shows high correlations only across subsets of task datasets, while the second SCV shows high correlations across all task datasets. On the other hand, shown in Figure 7.1(b), both SCVs in PARAFAC2 show high correlations only across subsets of task datasets. Thus, although PARAFAC2 does not model the dependence within an SCV, in the final decomposition, the estimated SCVs also contain the information about the relationship among the task datasets. For both methods, the block-structured covariance matrices indicate that correlations across task datasets form two distinct groups, one across the AOD and SM datasets and another across the SIRP datasets. This is because the AOD and SM tasks are more similar to each other compared with the SIRP task [LS17].

A subset of the source components, i.e., fMRI maps, corresponding to the block-structured covariance matrix of SCV 1 in IVA-G and the corresponding SCV 1 in PARAFAC2, are visualized (thresholded at $|z| = 2$) over the anatomical gray images in Figure 7.2(a) and Figure 7.2(b), respectively. Because of the sign correction, red or yellow voxels indicate that a brain network is more active in controls, and blue voxels

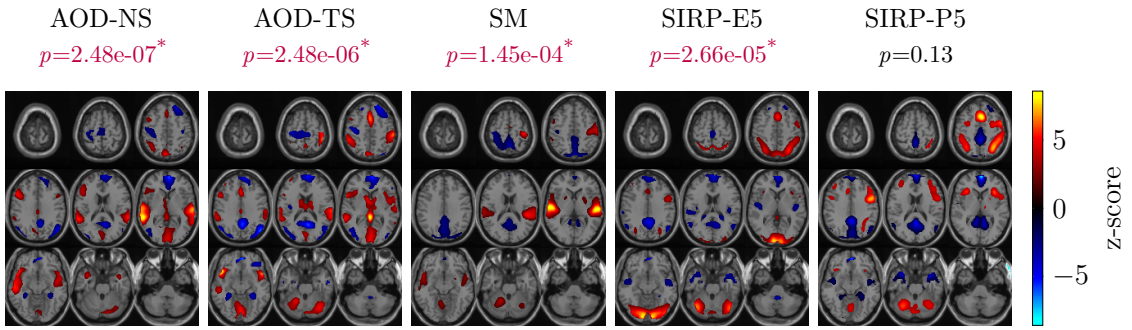


(a) estimated by IVA-G (b) estimated by PARAFAC2

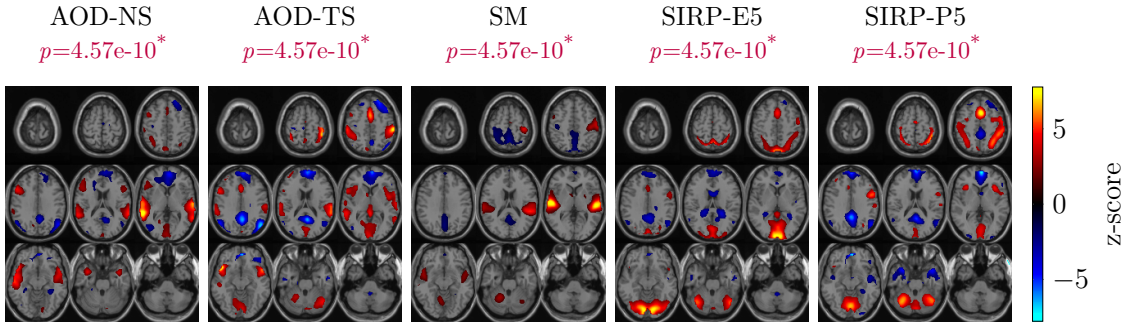
Figure 7.1: Covariance matrices (absolute values) of the $R = 2$ estimated SCVs for the multi-task fMRI data.

indicate higher activation in patients. Then, we adjusted the sign of the non-significant datasets ($p > 0.05$) manually by matching the color (red/blue) of the active brain networks to the significant datasets ($p < 0.05$) of the same task. Significant p-values are displayed in magenta with the superscript $*$.

In both methods, source components estimated for the AOD and SM datasets are showing higher activations in controls in the auditory and motor networks (red/yellow focal areas in slices 5–7 and 2–3), while the source components in the SIRP-E datasets are showing higher activations mostly in the visual network (red/yellow focal areas in slices 6–8), and in the SIRP-P datasets in the motor and in parts of the visual network (red/yellow focal areas in slices 2–3 and 7–8). The p-value is the same for all task datasets in PARAFAC2 and is very small; thus, all task datasets seem to provide a high discrimination between patients and controls. However, IVA-G can identify which of the task datasets are providing significantly different activations between patients and controls, which improves the interpretability, especially if IVA-G estimates more components. The active brain networks in the AOD tasks, which show the largest difference (i.e., smallest p-value) between patients and controls, are known to be affected by schizophrenia from previous studies [Cal04; Du12], and therefore are promising candidates for biomarkers for schizophrenia. The Default Mode Network (DMN) (blue focal areas in slices 4–6), which is present in all but the SM datasets, is more active in patients than in controls. This makes sense as the DMN mostly represents the resting state: the less concentrated one is on a given task, the more dominant are the resting state source components, and typically, controls show a higher task-related suppression of the DMN than patients [Hu17; WG09]. The DMN is more clearly visible in the SIRP-P dataset for PARAFAC2 than for IVA-G. Also the DMN has been shown to be affected by schizophrenia [WG09] and thus may be used as another potential



(a) estimated by IVA-G



(b) estimated by PARAFAC2

Figure 7.2: Subset of estimated fMRI maps corresponding to SCV 1. The auditory network is more active in controls in the AOD and SM datasets (red/yellow areas in slices 5–7) and not active in the SIRP datasets. The visual network is more active in controls in the SIRP-E and SIRP-P datasets (red/yellow areas in slices 6–8) and not active in the AOD and SM datasets. The motor network is more active in controls in the AOD-TS, SM, and SIRP-P datasets (red/yellow areas in slices 2–3) and not active in the AOD-NS and SIRP-E datasets, as no motor action is required in the latter tasks. The default mode network is more active in patients in the AOD and SIRP datasets (blue areas in slices 4–6) and is almost not active in the SM dataset.

biomarker for the disorder, in agreement with [Hu17].

Finally, PARAFAC2 provides us with an additional summary of the task datasets through the factor matrix $\widehat{\mathbf{C}} \in \mathbb{R}^{2\times 13}$. As in the simulations in Section 4.3, $\widehat{\mathbf{C}}$ should not be confused with the covariance matrices $\widehat{\mathbf{C}}_{\mathbf{s}_n}$. The two columns of $\widehat{\mathbf{C}}$, $\widehat{\mathbf{c}}_1$ and $\widehat{\mathbf{c}}_2$, are shown in Figure 7.3, where each element of $\widehat{\mathbf{c}}_r$ is plotted. The x-axis shows to which dataset each element of $\widehat{\mathbf{c}}_r$ corresponds, and the y-axis shows how strong the source component is present in the dataset. More specifically, the k^{th} element in $\widehat{\mathbf{c}}_r$, $\widehat{c}_{kr}$, represents how strong the k^{th} source component of the r^{th} SCV, $\left(\widehat{\mathbf{b}}_r^{[k]}\right)^{\mathsf{T}}$, is present in the k^{th} dataset. We see that the first SCV is present in all task datasets, thus showing an average component, while the second SCV is more present in the SIRP datasets.

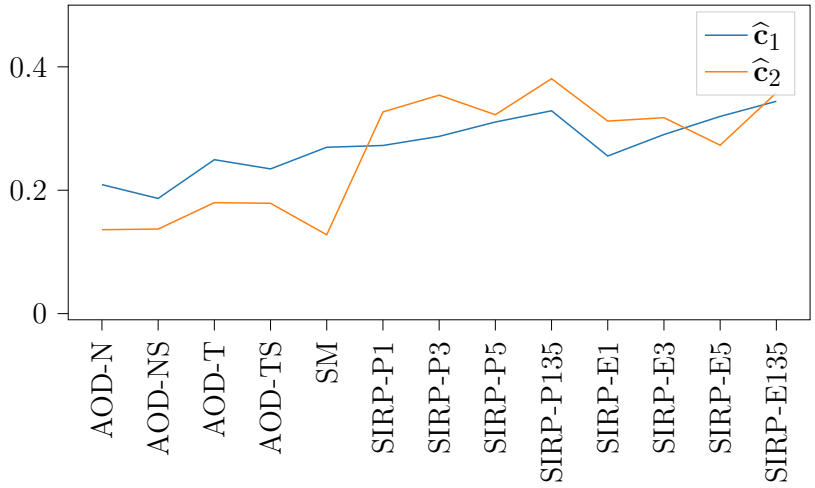


Figure 7.3: Dataset covariations estimated by PARAFAC2. The first component is present in all task datasets, and the second component is mainly present in the SIRP datasets.

To conclude this section, both IVA-G and PARAFAC2 identify source components that discriminate between patients with schizophrenia and healthy controls, with the strongest discrimination found in the AOD task, and provide information about the relationship among task datasets in the SCV covariance matrices.

7.3 Identifying the relationship structure among multi-task fMRI datasets

In the previous section, we have only estimated a small number of sources with IVA-G to have a fair comparison with PARAFAC2, which can only estimate a small number of sources due to its uniqueness conditions. Even for this small number of sources, the SCV covariance matrices provide information about the relationship among the task datasets. Based on these results, in this section, we apply the proposed method from Chapter 6 to identify the relationship structure among the task datasets, which are described in Section 7.1. This time, we have removed redundant datasets and used $K = 10$ task datasets, which are AOD-N, AOD-T, AOD-TS, SM, and SIRP-E and SIRP-P each for one, three, and five digits in the set that needed to be remembered. We now estimate more sources (through a higher PCA dimension R) than in the previous section, which has the advantage that now the active brain networks in one SCV correspond to one task, whereas before, one single SCV captured the active brain networks of all tasks. Furthermore, instead of IVA-G, we now use IVA-L-SOS, as it has been shown that the assumed pdf of the SCVs in IVA-L-SOS is a good match for real-world fMRI

data [Bhi19; Lon20]. We show that the proposed method successfully reveals active brain networks associated with the disorder and identifies the relationship structure among task datasets that matches our prior knowledge of the experiment.[3]

7.3.1 Implementation details

We again perform dimension reduction via PCA to transform each observed dataset in a lower-dimensional subspace of dimension R. Then, IVA-L-SOS is performed on the dimension-reduced datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$ to estimate the source matrices $\widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{R \times V}$. From these, the SCVs $\widehat{\mathbf{S}}_r \in \mathbb{R}^{K \times V}$, $r = 1, \ldots, R$, are formed. Selecting an appropriate value for R is important to get meaningful results. The proposed method achieves robust clustering results for a wide range of values for $R = 20, 21, \ldots, 30$. In this section, we present the results for $R = 25$ because at this order, the estimated source components are 1) stable, i.e., active brain networks are not split (as in higher orders) or merged (as in lower orders), and 2) meaningful, i.e., active brain networks are physically interpretable. The visualization of active brain networks and calculation of p-values is done as described in the previous section.

The proposed method is primarily data-driven, with the exception of two user-selected parameters in Step 2, the probability of false alarm P_{fa} and the number of bootstrap resamples B. With P_{fa}, we can directly control the risk of overestimating d_r, i.e., the number of eigenvalues greater than one in the r^{th} SCV covariance matrix. The higher the value for B is chosen, the better the distribution is estimated, but for a too high value of B, there will not be a better estimate at some point. We choose $B = 1000$ bootstrap resamples and $P_{\text{fa}} = 0.05$ to estimate d_r, as these are typical values for these parameters and achieve good results in general.

7.3.2 Results

Identification of common and structured SCVs

A subset of the covariance matrices of the estimated SCVs (absolute values) is shown in Figure 7.4 along with the estimated values $\widehat{d}_r$. Light values correspond to high correlations. For SCVs 3 and 13, $\widehat{d}_r = 1$, thus, these SCVs are identified as common. SCVs 15–17, 22, and 25 are identified as structured because $\widehat{d}_r > 1$. Like in the previous section, the visible block structure in each structured SCV provides information about

[3] This section is based on the paper: "Identifying the Relationship Structure among Multiple Datasets Using Independent Vector Analysis: Application to Multi-task fMRI Data," I. Lehmann, T. Hasija, B. Gabrielson, M. A. B. S. Akhonda, V. D. Calhoun, T. Adali, *IEEE Access*, 2024, pp. 109443–109456.

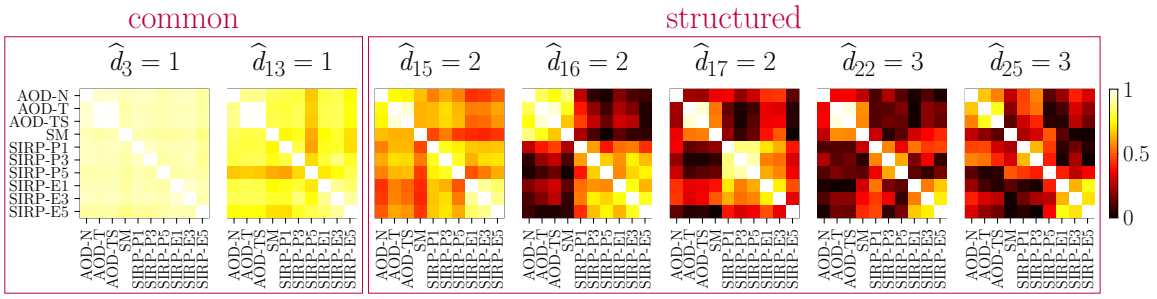


Figure 7.4: Subset of the estimated SCV covariance matrices (absolute values). SCVs 3 and 13 are identified as common, and SCVs 15–17, 22, and 25 are identified as structured.

the relationship among the task datasets. For example, the covariance matrix of SCV16 clearly shows high correlations within the AOD and SM datasets and within the SIRP datasets, but small correlations across these task datasets.

Estimation of sources with IVA-L-SOS

The estimated source components, i.e., fMRI maps, show which brain networks are active in which tasks. In the following, we present SCV3, SCV16, and SCV17 as examples of common and structured SCVs because they either show active brain networks that

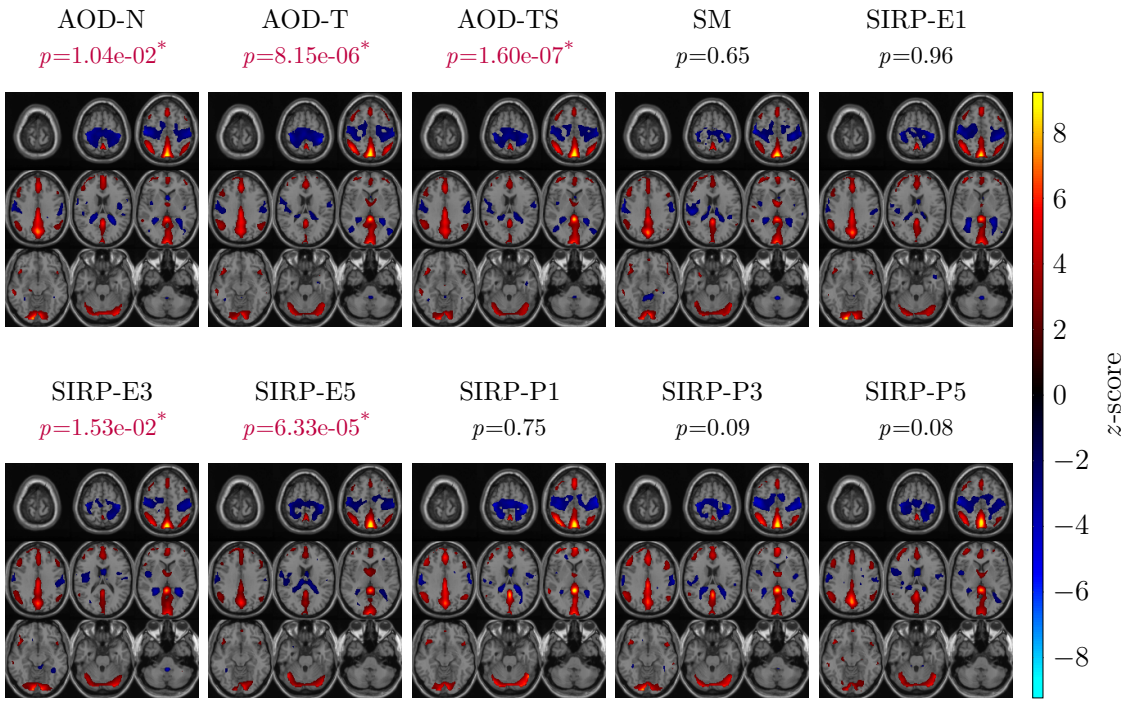


Figure 7.5: Estimated fMRI maps corresponding to SCV3. The default mode network and the visual network are more active in controls (red/yellow areas in slices 4–6 and 7–9), while the sensorimotor areas are more active in patients (blue areas in slices 2–3).

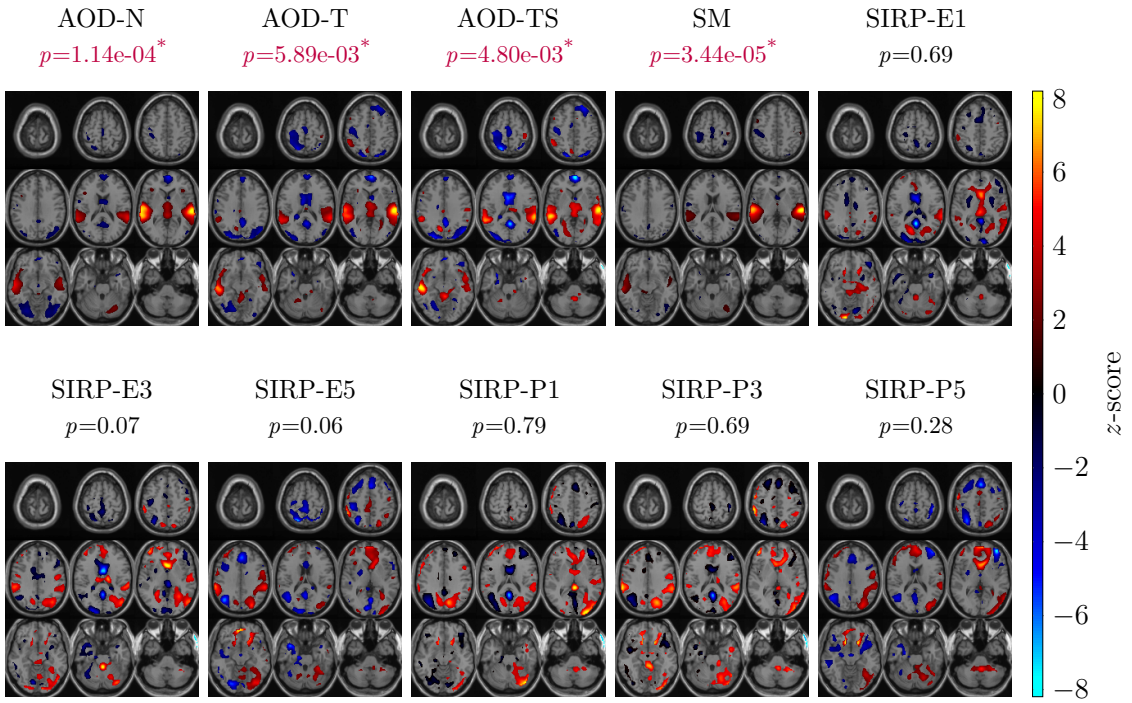


Figure 7.6: Estimated fMRI maps corresponding to SCV16. The auditory network is more active in controls in the AOD and SM datasets (red/yellow areas in slices 5–7) and not active in the SIRP datasets.

are common in all tasks or correspond to specific tasks. All fMRI maps are visualized (thresholded at $|z| = 2$) over the anatomical gray images. Significant p-values are displayed in magenta with the superscript $*$. The sign correction is performed as in the previous section. Therefore, red or yellow voxels again indicate that a brain network is more active in controls, and blue voxels again indicate higher activation in patients.

In SCV 3 in Figure 7.5, we see activations of the DMN (red/yellow focal areas in slices 4–6). The DMN is known to have a decreased activation when a task is performed [San20; Hu17]. The higher activation of the DMN in controls means that the deactivation is stronger in patients. This can be interpreted as patients needing to focus more on a task to perform it well. The sensorimotor network (blue focal areas in slices 2–3) is more active in patients and therefore supports this interpretation. There are minor activations in the visual network (red/yellow focal areas in slices 7–9), which are expected because the subjects had their eyes open during all tasks. The very small p-values in the AOD-T and AOD-TS datasets indicate that especially when the target stimulus occurs, the patients are significantly more engaged with the task, i.e., have significantly smaller activation of the DMN. These lower p-values in the AOD datasets are expected since the AOD task has been shown to be important in discriminating patients with schizophrenia

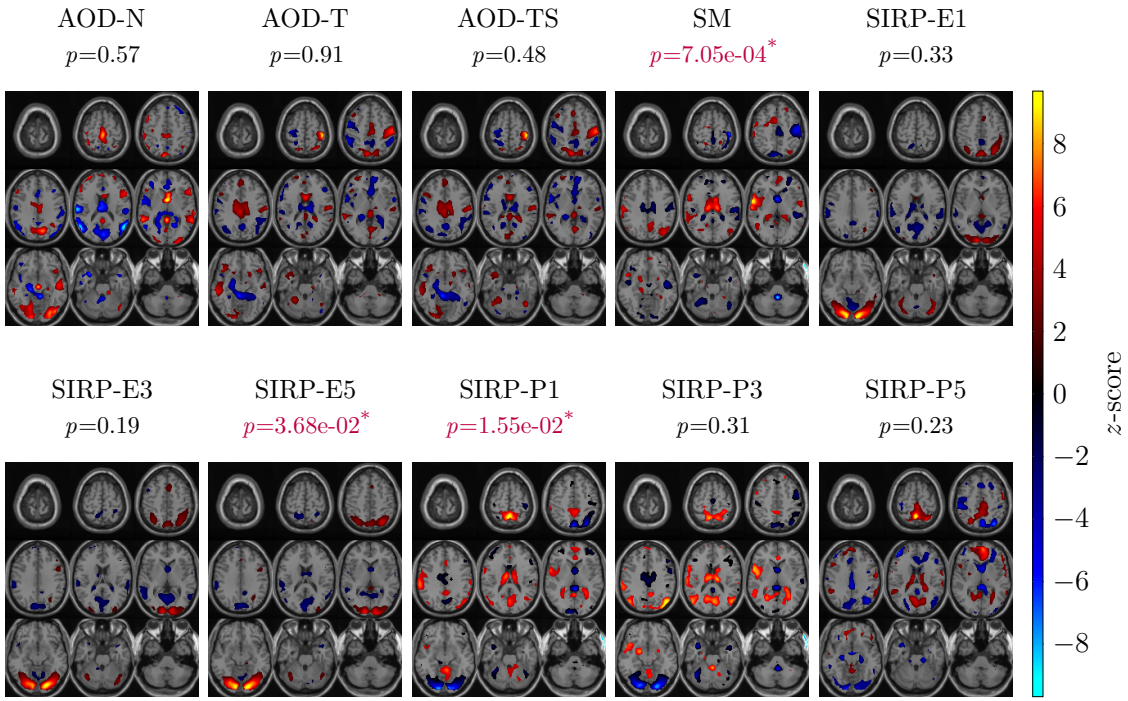


Figure 7.7: Estimated fMRI maps corresponding to SCV17. The visual network is more active in controls in the SIRP-E datasets (red/yellow areas in slices 6–7), more active in patients in the SIRP-P datasets (blue areas in slice 7), and not active in the AOD and SM datasets.

and healthy controls, as patients have a smaller oddball response [Kim09]. In the SIRP-P datasets, there is no significant difference between patients and controls for the DMN. In contrast, in the SIRP-E datasets, the p-values become smaller with increasing task difficulty, i.e., the deactivations of the DMN become stronger for patients. This coincides with the literature, as with an increasing level of difficulty of a task, the deactivation of the DMN becomes stronger [San20].

The fMRI maps corresponding to SCV16 are shown in Figure 7.6. The auditory network (red/yellow focal areas in slices 5–7) is more active in controls in the AOD and SM datasets and not active in the SIRP datasets. The p-values are also significant for the AOD and SM datasets, supporting the literature that activations in the auditory network may be a biomarker for differentiating patients with schizophrenia and healthy controls [Cal04].

The fMRI maps corresponding to SCV17, shown in Figure 7.7, show strong activations in the visual network for the SIRP datasets (red/yellow focal areas in slices 6–7 in the SIRP-E datasets and blue focal areas in slice 7 in the SIRP-P datasets). A significant difference between patients and controls is found for SIRP-E5 (in accordance with SCV3) and SIRP-P1. What is most surprising here is that the visual network is more

active in controls for the SIRP encoding phase, but more active in patients for the SIRP probe phase. An explanation might be that patients need to focus more on a digit to remember it, while the controls just briefly see the digit, and their memory can be accessed faster.

Identification of the relationship structure

As described in Subsection 6.2.3, the $\widehat{d}_r$ leading eigenvectors of the structured SCVs are concatenated to form a feature matrix, which is the input of the hierarchical clustering. The dendrogram for the hierarchical clustering, which reveals the relationship structure among the task datasets, is shown in Figure 7.8. The labels of the datasets are again denoted on the x-axis of the dendrogram, and the y-axis again shows the distance between the clusters. Each cluster refers to a group of task datasets (in contrast to a group of subjects, as in subgroup identification). The SIRP tasks form one group (orange), and the AOD and SM tasks form a second group (green). As the SIRP task involves a visual stimulus and the AOD and SM tasks both involve auditory stimuli, these resulting groups are meaningful. Within the SIRP group, there are two finer groups visible: one consisting of the encoding datasets and the other of the probe

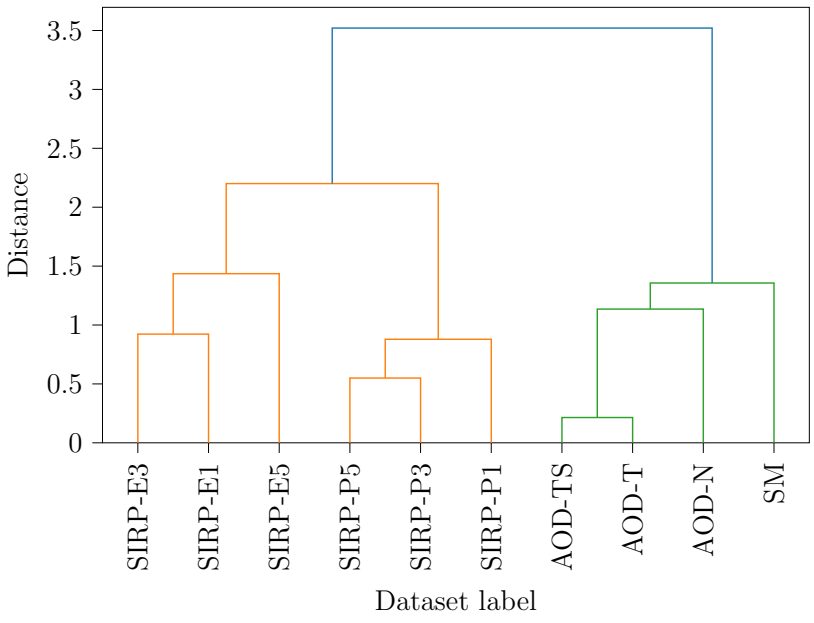


Figure 7.8: The dendrogram for the multi-task fMRI data reveals the relationship structure among the task datasets. Two groups are found, consisting of the visual task datasets (orange), SIRP-E and SIRP-P, and of the auditory task datasets (green), AOD and SM. Within the visual datasets, the SIRP-E and SIRP-P datasets form two groups. Within the auditory datasets, there is one group consisting of the AOD datasets.

datasets. This makes sense because they refer to two different phases of the SIRP task. Within the auditory group, the AOD datasets form a finer group.

To conclude this section, the proposed method reveals active brain networks that allow to discriminate between patients with schizophrenia and healthy controls and successfully identifies the relationship structure among the task datasets, as it groups together datasets from similar tasks in the hierarchical clustering.

7.4 Summary

In this chapter, we have analyzed multi-task fMRI datasets from patients with schizophrenia and healthy controls using JBSS methods. We have started by performing data fusion of these datasets with IVA-G and PARAFAC2, where we have observed that both methods identify source components corresponding to brain networks that discriminate between patients with schizophrenia and healthy controls. The active brain networks in the AOD task show the strongest discrimination between patients and controls and, therefore, are promising candidates for biomarkers for schizophrenia. Besides having found brain networks that are expected to be active during the tasks, which indicates the success of both methods for achieving JBSS on real-world fMRI data, we have seen that the SCV covariance matrices provide information about the relationship among task datasets.

Based on these results, we have applied the proposed method from Chapter 6 to the task datasets. Again, the estimated source components correspond to brain networks that are known to be affected by schizophrenia: We see significantly stronger deactivations of the DMN in patients and significantly stronger activations of the auditory brain network in controls. While we were able to draw conclusions from the here presented estimated fMRI maps, it is important to note that interpretation is not always straightforward. We have carefully selected and presented SCVs that show activations in meaningful brain networks, which we identified according to Brodmann areas [Lac08]. Despite the simplicity of the task design, many of the unrepresented components cannot easily be explained in a straightforward manner. However, interpreting a subset of the most meaningful components can already help in understanding how the brain functions. Thus, a good guideline for using the proposed method for fMRI analysis is to find a range of R values that lead to stable results after the dimension reduction and then compare the estimated components with established brain networks such as the Brodmann areas [Lac08] to facilitate interpretation. It is important to remember that relying solely on comparisons with known brain atlas components might cause us to

overlook brain networks that are not included in the user-defined atlas. Nevertheless, comparing the brain networks with an atlas provides an initial reference point before conducting further investigations.

Finally, we have verified the success of the proposed method for identifying the relationship structure among the multi-task fMRI datasets, as it groups datasets from similar tasks together. Importantly, the proposed method is not limited to task datasets but is applicable to more general problems. For example, by identifying the relationship structure among subject datasets of, e.g., patients with schizophrenia, bipolar disorder, and their subtypes, the proposed method could be used for subgroup identification in precision medicine. Applying the proposed method to ANS measures from several time instances before and after epileptic seizures, and thereby identifying the changes in the relationship among these measures over time, might help to develop novel techniques for seizure detection or prediction. After having demonstrated the suitability of JBSS methods for fMRI data analysis, in the next chapter, we study their application for fUS data, an alternative measure to fMRI for inferring brain activity.

8 Deriving 3D functional brain networks from multi-slice fUS data using JBSS

In addition to the gold-standard method fMRI, also functional Ultrasound (fUS) can be used to infer brain activity. FUS, first introduced in 2011 [Mac11], invasively measures the cerebral blood volume in the brain. Because of neurovascular coupling, an increase in this blood flow and/or volume can be linked to increased neural activity [Mac11]. Compared with fMRI, fUS has a higher spatio-temporal resolution: the spatial resolution of fUS is 50–350 µm compared with 1–3 mm in fMRI, and the temporal resolution is 4–10 Hz for fUS and 0.3–1 Hz for fMRI, respectively [Mar21; Bru21]. Typically, fUS data is collected in a 2-dimensional (2D) imaging plane, which means that in order to collect full-brain information, several 2D images are collected for a fixed position of the ultrasound probe for the duration of the experiment, before the probe is moved to the next position. As a result, a 3-dimensional (3D) functional volume consists of subsequent, time-disjunct 2D datasets. In contrast, in fMRI, the magnetic field is moved after scanning each slice, and after some time correction preprocessing, the resulting tensor naturally is a 3D scan of the brain over time.

The gold standard way to analyze multi-slice fUS datasets is using correlation images or a General Linear Model (GLM)-based approach [Bru21]. However, besides the fact that these slice-by-slice analyses are not making use of the dependence across datasets, it was shown in previous work [Ero22] that the brain response strength varies a lot across consecutive stimuli, so correlation/GLM-based approaches may be too rigid and cannot capture this variability. Furthermore, these approaches can only be applied if the data is collected during a task experiment and the expected stimulus responses for the task are known. Therefore, they are not feasible for analyzing resting state data, which is of great importance to obtain information about the functional organization of the brain [Urb15; PR21]. Instead, the use of data-driven analysis using SVD or ICA for retrieving functional networks [Osm14] or regions of interest [Ero22; Wij23] within a

single slice has become more popular in recent years, as these methods do not require prior information. Very recently, ICA has also been used to map whole-brain mouse functional connectivity from novel 3D recordings [Hik23] and for slice-by-slice evaluation of consecutive 2D recordings [Ber23]. However, in the latter case, ICA does not exploit the information that is common across the slices.

In this chapter, we therefore propose the use of IVA for JBSS of fUS datasets, with the goal to map the mouse visual information processing pathway in 3D, i.e., to track the estimated active brain networks over multiple 2D slices. The preprocessing of the fUS datasets is very important but not straightforward, and is also explained in this chapter. By evaluating ICA and IVA on a multi-slice fUS dataset collected during an experiment with visual stimuli at different locations, we demonstrate the successful application of these methods, as they estimate similar active brain networks as found by the gold standard correlation images, without prior knowledge about the experiment design. Furthermore, we provide guidance as to when which method might be desirable.[1]

8.1 fUS dataset and preprocessing

fUS imaging data is collected from a head-fixed mouse on a wheel, which is looking at two screens. During the experiment, black-and-white flickerings occurred at nine different locations on the screens. The flickering at each location is called a stimulus; thus, there are nine different stimuli in this experiment. The stimuli occur in a pseudo-random order, where each stimulus is presented five times. The stimulus duration is 2.92 seconds, and the duration of the breaks between stimuli is between 7.52 and 10.88 seconds. After collecting the Power Doppler Images (PDIs) over T time points (with a sampling frequency $f_s = 4.65$ Hz) for a fixed probe position, the probe is moved to a different position, and the experiment is repeated. The order of the stimulus occurrences is the same for each probe position. We use data collected from the following probe positions (distance from Bregma): -3.16 mm, -3.36 mm, -4.26 mm, -4.56 mm, -4.76 mm, -4.96 mm. Thus, for each probe position, one slice of the brain is captured, and in total we have $K = 6$ datasets. We denote the data of the k^{th} slice as a tensor $\boldsymbol{\mathcal{X}}^{[k]} \in \mathbb{R}^{V_z \times V_x \times T}$, where $V_z = 150$ and $V_x = 256$ are the numbers of pixels in the z and x direction, respectively, and $T = 2530$ is the number of time points over which data is collected.

This experiment, by virtue of its simplicity, enables us to easily demonstrate and evaluate the performance of ICA and IVA by comparing their results with the correlation

[1] This chapter is based on the paper: "Deriving 3D Functional Brain Regions from Multi-Slice Functional Ultrasound Data Using ICA and IVA," I. Lehmann, P. Kruizinga, B. Hunyadi, *2023 57th Asilomar Conference on Signals, Systems, and Computers*, 2023, pp. 1484–1490.

images and interpreting the identified active brain networks.

The following preprocessing is applied to the PDIs of each slice $\boldsymbol{\mathcal{X}}^{[k]} \in \mathbb{R}^{V_z \times V_x \times T}$, $k = 1, \ldots, K$, separately.

1. (a) For correlation images: image standardization: The image for each time point t, $\boldsymbol{\mathcal{X}}^{[k]}(:, :, t) \in \mathbb{R}^{V_z \times V_x}$, is made zero-mean and unit-variance.

   (b) For ICA/IVA: temporal standardization: Each pixel's time course, $\boldsymbol{\mathcal{X}}^{[k]}(i, j, :) \in \mathbb{R}^{T}$, is made zero-mean and unit-variance.

2. Spatial smoothing: A 2-dimensional Gaussian filter with $\sigma = 1$ is applied to the image collected for each time point t, $\boldsymbol{\mathcal{X}}^{[k]}(:, :, t) \in \mathbb{R}^{V_z \times V_x}$.

3. Temporal smoothing: A fifth-order Butterworth lowpass filter with a cut-off frequency of 0.2 Hz is applied on each pixel's time course $\boldsymbol{\mathcal{X}}^{[k]}(i, j, :)$.

The following two steps are only applied for ICA/IVA.

4. Reshaping: The tensor $\boldsymbol{\mathcal{X}}^{[k]} \in \mathbb{R}^{V_z \times V_x \times T}$ is unfolded in a matrix $\mathbf{X} \in \mathbb{R}^{T \times V}$, where $V = V_z V_x$. Compared with the fMRI datasets, where each dataset $\mathbf{X}^{[k]}$ was of dimension $N \times V$, with N being the number of subjects, now each dataset is of dimension $T \times V$, with T being the number of time points.

5. Dimensionality reduction and whitening: PCA is applied to $\mathbf{X}^{[k]}$ to get the dimension-reduced observed datasets $\widetilde{\mathbf{X}}^{[k]} = \left(\boldsymbol{\Psi}^{[k]}\right)^{\mathsf{T}} \mathbf{X}^{[k]} \in \mathbb{R}^{R \times V}$, where $\boldsymbol{\Psi}^{[k]} \in \mathbb{R}^{T \times R}$, $R < T$, is the PCA transformation matrix. As with the fMRI data, ICA and IVA are then applied to $\widetilde{\mathbf{X}}^{[k]}$.

8.2 Estimation of active brain networks

8.2.1 Using correlation images

The gold standard for identifying active brain networks in fUS task data is to use correlation images [Urb15], which are calculated from the image-standardized, spatially and temporally smoothed tensor $\boldsymbol{\mathcal{X}}^{[k]} \in \mathbb{R}^{V_z \times V_x \times T}$ as follows. We denote the correlation image for the r^{th} stimulus in the k^{th} slice as $\mathbf{G}_r^{[k]} \in \mathbb{R}^{V_z \times V_x}$. The value of each pixel (i, j) in $\mathbf{G}_r^{[k]}$ is the Pearson correlation coefficient of $\boldsymbol{\mathcal{X}}^{[k]}(i, j, :) \in \mathbb{R}^{T}$, the time course of that pixel, with the expected stimulus response $\mathbf{e}_r \in \mathbb{R}^{T}$, which is the convolution of the time course of the r^{th} stimulus $c_r(t)$ with the HRF $h(t)$:

$$e_r(t) = c_r(t) * h(t).$$

We model the HRF by a shifted Dirac impulse:

$$h(t) = \delta(t - t_0),$$

and we choose $t_0 = 7$ samples as time shift, which corresponds to approximately 1.75 seconds. The highly correlated pixels in $\mathbf{G}_r^{[k]}$ form the active brain network in the k^{th} slice corresponding to the r^{th} stimulus.

8.2.2 Using ICA

As the measured data is assumed to be a mixture of the activity fluctuations of multiple spatially independent functio-anatomical brain networks, we apply ICA to each dimension-reduced dataset $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$ separately to get the estimated demixing matrix $\widetilde{\mathbf{W}}^{[k]} \in \mathbb{R}^{R \times R}$ and the estimated source matrices $\widehat{\mathbf{S}}^{[k]} \in \mathbb{R}^{R \times V}$, with $R = 20$ and $V = 38,400$. The r^{th} estimated source component of the k^{th} dataset, denoted by $\widehat{\mathbf{s}}_{r:}^{[k]} \in \mathbb{R}^{1 \times V}$, is reshaped into an image $\mathbf{G}_r^{[k]} \in \mathbb{R}^{V_z \times V_x}$, and the R images for the k^{th} slice are independent of each other. The estimated mixing matrices $\widehat{\mathbf{A}}^{[k]} \in \mathbb{R}^{T \times R}$ are reconstructed by multiplying the PCA transformation matrix $\mathbf{\Psi}^{[k]}$ with the inverse of the estimated demixing matrix $\widetilde{\mathbf{W}}^{[k]}$, as in the fMRI data:

$$\widehat{\mathbf{A}}^{[k]} = \mathbf{\Psi}^{[k]} \left(\widetilde{\mathbf{W}}^{[k]} \right)^{-1} \in \mathbb{R}^{T \times R},$$

and the r^{th} column $\widehat{\mathbf{a}}_r^{[k]}$ is the time course corresponding to the r^{th} active brain network in the k^{th} dataset. To overcome the sign ambiguity of ICA, we made sure that the majority of the pixels in each active brain network is positive or made positive by multiplying the corresponding row of $\widehat{\mathbf{S}}^{[k]}$ and column of $\widehat{\mathbf{A}}^{[k]}$ by -1. To overcome the permutation ambiguity, we manually aligned the 20 components across the six slices, which is prone to subjective biases/errors. In the final step, we removed the components corresponding to noise, and eight components remained.

8.2.3 Using IVA

IVA is jointly applied on the K dimension-reduced datasets $\widetilde{\mathbf{X}}^{[k]} \in \mathbb{R}^{R \times V}$. Reconstructing the mixing matrices, solving the sign ambiguity, and reshaping the source components into images $\mathbf{G}_r^{[k]}$ is performed in the same way as for ICA. As IVA jointly estimates the sources, the permutation ambiguity is automatically resolved by this method. Finally, we match the estimated IVA components to the selected ICA components to

compare the results of the methods.

8.2.4 Visualization of active brain networks

The significant pixels ($p < 0.01$) in each image form an active brain network. To visualize these brain networks, we calculate the (scalar) mean $\mu_r^{[k]}$ and standard deviation $\sigma_r^{[k]}$ of each image $\mathbf{G}_r^{[k]}$, and calculate the z-score of each pixel in $\mathbf{G}_r^{[k]}$ by subtracting $\mu_r^{[k]}$ and dividing by $\sigma_r^{[k]}$. From the resulting z-scored image $\mathbf{Z}_r^{[k]} \in \mathbb{R}^{V_z \times V_x}$, we visualize only the significant pixels, i.e., the pixels (i, j) where $\left|\mathbf{Z}_r^{[k]}(i,j)\right| > 2.58$, which corresponds to a significance level of $p < 0.01$ for a two-sided t-test.

The active brain networks are visualized over the logarithmic mean image $\mathbf{X}_{\text{logmean}}^{[k]}$, which is calculated from the raw PDIs (without any preprocessing) for each slice as:

$$\mathbf{X}_{\text{logmean}}^{[k]} = \log\left(\frac{\mathbf{X}_{\text{mean}}^{[k]}}{\max\left(\mathbf{X}_{\text{mean}}^{[k]}\right)}\right),$$

where $\mathbf{X}_{\text{mean}}^{[k]} = \sum_{t=1}^{T} \boldsymbol{\mathcal{X}}^{[k]}(:,:,t)$. Then, we overlay the Allen brain atlas [Wan20a] for the corresponding slice so that we can interpret the active brain networks in terms of anatomical regions.

8.3 Results

8.3.1 Correlation images

Figure 8.1 shows the correlation images of probe position -3.36 mm for the nine stimuli. In the following, we link the active brain networks to the anatomical regions defined in the Allen brain atlas [Wan20a]. We see positive correlations mainly in V1 on the right side of the brain with stimuli 1–4 and on the left side with stimuli 8–9, which makes sense as V1 is part of the visual pathway with contra-lateral activation. For stimuli 5–7, we see the movement of the V1 and RSG networks from right to left. Thus, the location of the stimulus can be visually tracked in the correlation images for probe positions -3.36 mm (and also for -4.26 mm and -4.56 mm, which are not shown).

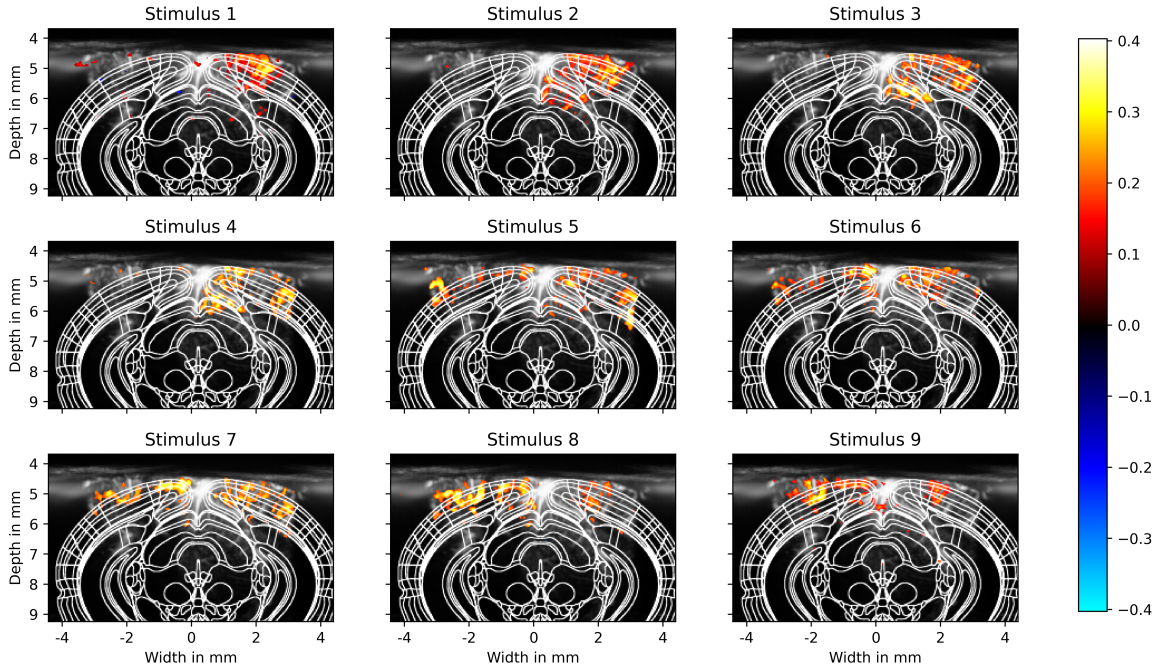
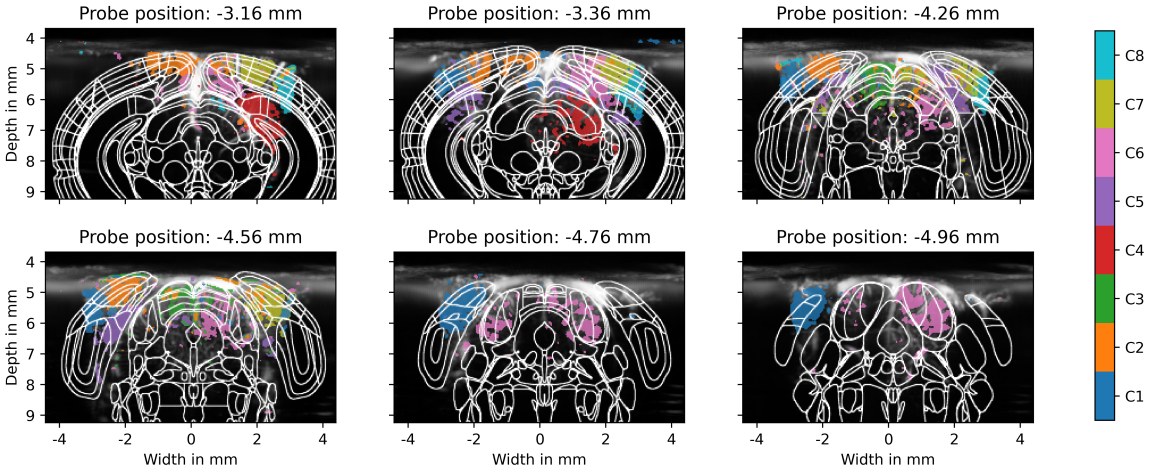


Figure 8.1: For probe position -3.36 mm, the location of the stimuli can be tracked in the correlation images: From stimulus 1 to 9 (location most left to most right), we see a movement of the active brain networks from right to left.

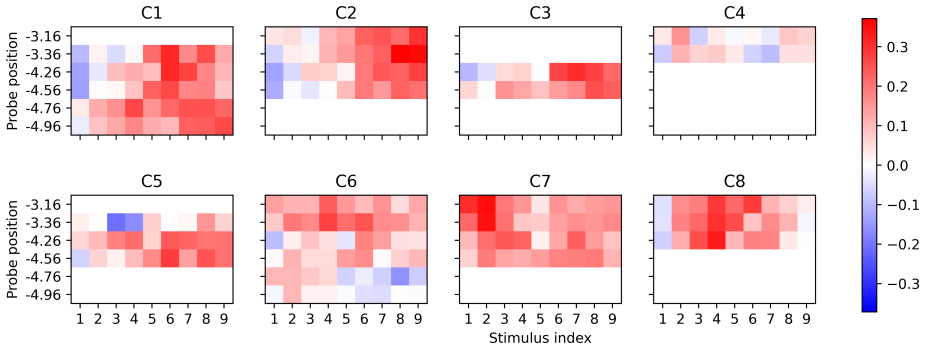
8.3.2 ICA results

We run FastICA[2], described in Section 3.1, 20 times with random initializations and present the results of the most consistent run [Lon18]. Figure 8.2(a) shows the active brain networks estimated by ICA for the six probe positions. All brain networks for one slice are plotted in the same image. To distinguish between them, each brain network is plotted in a different color. Furthermore, a component that occurs in multiple slices (as identified manually) is plotted in the same color for each slice. Figure 8.2(b) shows the correlation of the ICA time courses (corresponding to the active brain networks in Figure 8.2(a)) with the nine expected stimulus responses. If a component is not present in a slice, the corresponding row is empty. The identified active brain networks correspond to meaningful anatomical brain regions. Component 1 (C1) (dark blue) belongs mainly to the left V1 for probe positions -3.36 mm, -4.26 mm, -4.56 mm, -4.76 mm, and -4.96 mm. The corresponding time courses of C1 are more highly correlated with the stimuli located on the right side of the screen (stimuli 6–9), which makes sense as the brain network is also present in the correlation images for those stimuli. C2 (orange) corresponds to left V2ML, V2MM, and RSA/RSG at -3.16 mm, and then evolves to

[2]The Python implementation is available at: https://scikit-learn.org/stable/index.html [Ped11].



(a) Active brain networks estimated by ICA for the six probe positions. Each brain network is visualized in a different color within a slice, and the corresponding brain networks are visualized in the same color across slices.



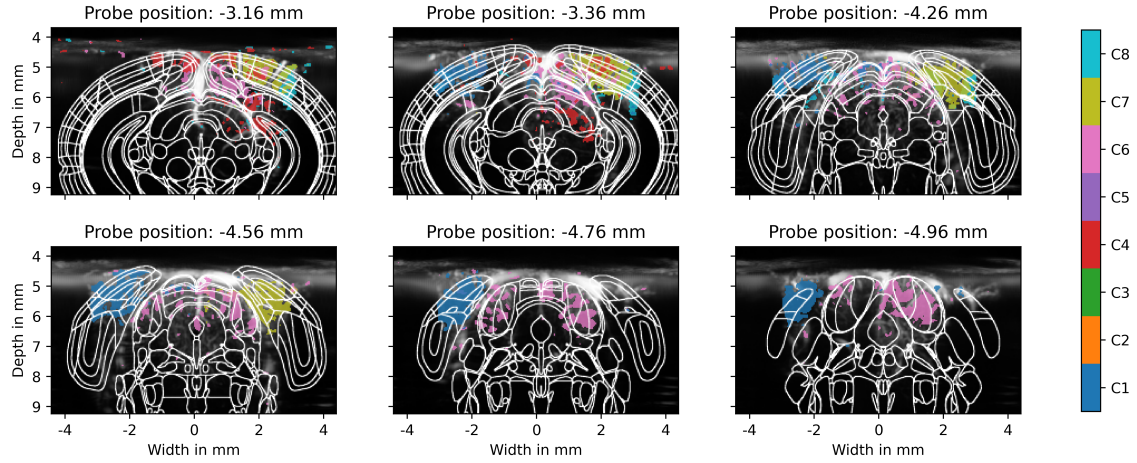
(b) Correlation of ICA time courses and expected stimulus responses.

Figure 8.2: ICA results. The active brain networks of Components 1/2 (C1/C2) and C7 belong to the left and right V1, respectively. The time course of C1/C2 is more highly correlated with the stimuli on the right side, and the time course of C7 is more highly correlated with the stimuli on the left side.

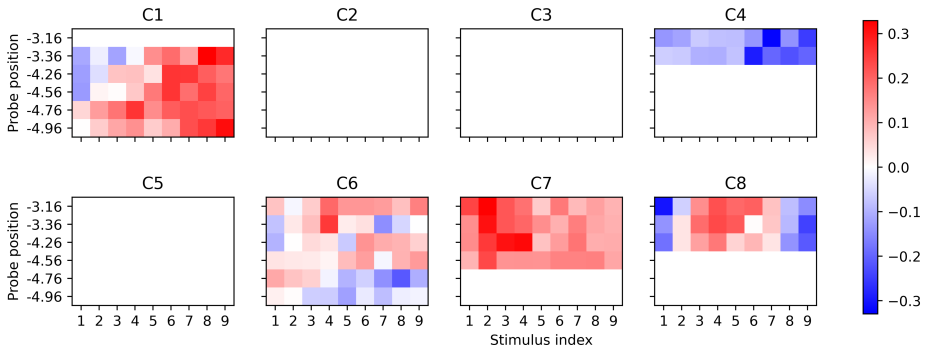
mainly activations of V1 (shared with C1) at -3.36 mm to -4.56 mm. Also for the time courses corresponding to this component, we see higher correlations with the stimuli on the right. C5 (purple) belongs to left S (-3.36 mm), left and right PrS (-4.26 mm), and left PrS (-4.56 mm). C7 (olive) shows activations of right V2ML and right V1 (at -3.16 mm) and then mainly right V1 at -3.52 mm to -4.56 mm. The corresponding time course is more highly correlated with the stimuli on the left.

8.3.3 IVA results

We run IVA-G 20 times with random initializations and present the results of the most consistent run [Lon18]. In Figure 8.3(a), we see the estimated active brain networks of IVA for each slice, and in Figure 8.3(b), we see the correlation of the IVA time courses with the expected stimulus responses. We can directly see that C1 and C2 in ICA at -3.36 mm to -4.56 mm are merged in C1 in IVA, which makes more sense as they belong to the same anatomical region (left V1) and again reflect the contra-lateral activation of the visual pathway. However, C3 and C5 are completely missing in IVA. C7 and C8 are very similar for ICA and IVA, and also the brain networks of C1 and C6 for



(a) Active brain networks estimated by IVA for the six probe positions. Each brain network is visualized in a different color within a slice, and the corresponding brain networks are visualized in the same color across slices.



(b) Correlation of IVA time courses and expected stimulus responses.

Figure 8.3: IVA results. The active brain networks of C1 and C7 belong to the left and right V1, respectively. The time course of C1 is more highly correlated with the stimuli on the right side, and the time course of C7 is more highly correlated with the stimuli on the left side. The time courses of C4, corresponding to the RSA/RSG regions, are anti-correlated with all stimuli.

probe positions -4.76 mm and -4.96 mm look almost identical for both methods. The time courses of C1 (left side of brain) are again more highly correlated with the right stimuli, and the time courses of C7 (right side of brain) with the left stimuli. We see a strong negative correlation between the time course of C4 and all stimuli, which is not present in ICA. The activations of C4 in the RSA/RSG regions for probe positions -3.16 mm and -3.36 mm, which are associated with the mouse default mode network, do not correspond to C4 but to C2 in ICA. Thus, IVA seems to extract the default mode network, which ICA seems to merge with another component.

8.4 Summary

In this chapter, we have proposed the use of ICA and IVA to track the 3D visual information processing pathway in multi-slice 2D fUS imaging data. We demonstrated the suitability of these two methods by showing that the identified functio-anatomical brain networks match the activations in the correlation images for our specific task dataset. Furthermore, we have shown that the ICA/IVA time courses of the estimated active brain networks are correlated with the expected stimulus responses of those stimuli for which the corresponding brain network is present in the correlation images.

In terms of comparison of the methods ICA and IVA, IVA has the clear advantage of automatically aligning the identified brain networks across slices by making use of their dependence when estimating them. However, because IVA aims to make the components dependent across all slices, it tends to miss components that are present only in a few slices. On the other hand, ICA might split or merge brain networks. Choosing the proper analysis method involves a trade-off between obtaining a detailed understanding and minimizing the amount of time spent. We suggest the use of IVA if one is interested in tracking the active brain networks over slices and finds it acceptable if brain networks may be missed, and the use of ICA if one is interested to identify as many brain networks as possible while taking the risk that one brain network might be split in two components.

As we have demonstrated the successful application of ICA and IVA for multi-slice fUS task data, the next step can be to analyze resting state data using one of these two methods, where the gold standard of correlation images cannot be used as there is no expected stimulus response. Furthermore, instead of IVA-G, IVA-L-SOS [Bhi19] can be used to estimate the sources, as it also takes higher-order statistics into account and thus does not only look at correlations like IVA-G does. Lastly, the quality of the estimated components can be automatically determined by, e.g., analyzing which

percentage of pixels in a brain region according to the atlas is significant.

9 Conclusion

9.1 Summary

In this thesis, we have addressed two main challenges in the field of JBSS, which are 1) choosing a JBSS method for a specific application, and 2) identifying the relationship structure among multiple datasets using JBSS. In order to address the first challenge, we have established the connections between matrix and tensor decomposition methods for JBSS in Chapter 4 and discussed their source identification conditions in Chapter 5. In Chapter 4, we have proven that the canonical variables estimated by the all-at-once analytical mCCA-sumcor method [Nie95] and those estimated by the all-at-once analytical mCCA-maxvar method [Car68] are scaled versions of each other. Furthermore, we have shown that the canonical variables estimated by the all-at-once analytical sumcor method [Nie95] and by the deflationary analytical maxvar method [Ket71] are closely related, with their main difference being that the canonical variables within a dataset are constrained to be uncorrelated in maxvar but not in sumcor. Thereafter, we have shown that the PARAFAC2 and IVA models are intimately related: both can be formulated as the multiplication of a mixing matrix with a source matrix for each dataset, with the difference that PARAFAC2 imposes stronger constraints on the mixing matrices and IVA-G imposes stronger constraints on the source matrices. In simulations, we have demonstrated that also PARAFAC2 is able to achieve JBSS. In Chapter 5, we have summarized the source identification conditions of ICA, IVA, CCA, and mCCA-maxvar. To complete the comparison of the mCCA methods, we have derived and proven the source identification conditions of mCCA-sumcor, conjectured those of mCCA-minvar, and discussed theoretical considerations for those of mCCA-genvar and mCCA-ssqcor. Based on our experiments, which have substantiated the proposed conditions, genvar seems to have the least restrictive source identification conditions among all mCCA methods and thus can identify sources that the other methods may not recover. This is especially notable for a finite number of samples, where we observed in our experiments that genvar is the most statistically efficient method. According to our findings from Chapters 4 and 5, we generally recommend mCCA-genvar [Ket71]

as the preferred mCCA method for JBSS based on our experiments. However, if it is reasonable to assume that the true SCVs have an effective rank of one, meaning that the source components of all datasets within each SCV have the same underlying factor, we recommend the use of maxvar [Ket71], which is computationally much faster, especially when the dimension or number of datasets is high. Based on our comparison of IVA-G and PARAFAC2 in simulations, we recommend the use of PARAFAC2 when the columns of the mixing matrices across different datasets are expected to differ only up to a scaling, and of IVA-G when different mixing matrices are expected in different datasets.

We have also proposed a method to identify the relationship structure among multiple datasets using JBSS methods in Chapter 6, being the first to address this challenge to the best of our knowledge. The proposed method consists of three steps: 1) estimating latent sources from observed datasets using IVA-L-SOS, 2) identifying structured SCVs based on bootstrap-based hypothesis testing, and 3) extracting features from the structured SCVs, which are then used by hierarchical clustering to identify the relationship structure among the datasets. We have first verified the superior performance of the second step of the proposed method against competing techniques and then demonstrated in simulations that the complete method successfully identifies the relationship structure among multiple datasets.

After having finished the theoretical discussions of this thesis, we have applied the JBSS methods to real-world functional imaging data. In Chapter 7, we have shown that IVA-G and PARAFAC2 can perform data fusion—the joint analysis of multiple related datasets such that they can interact and inform each other—of multi-task fMRI data collected from patients with schizophrenia and healthy controls. Both methods identify potential biomarkers for schizophrenia, i.e., source components corresponding to brain networks that discriminate between patients with schizophrenia and healthy controls, and provide information about the relationship among the task datasets in the SCV covariance matrices. Building on these findings, we have successfully applied the proposed method from Chapter 6, which identifies a meaningful relationship structure among the multi-task fMRI datasets, as the method groups together the task datasets from similar tasks. Furthermore, it again reveals brain networks with significantly different activations in patients with schizophrenia and healthy controls that might serve as potential biomarkers for the disorder. After having demonstrated the suitability of JBSS methods for fMRI data analysis, in Chapter 8, we have studied their application for deriving 3D functional brain networks from multi-slice fUS data, an alternative measure to fMRI for inferring brain activity with higher spatio-temporal resolution. We have shown that ICA and IVA successfully estimate active brain networks that

are similar to those found by the gold standard correlation images. While ICA might split or merge brain networks and requires (time-consuming) manual alignment, IVA might miss brain networks but automatically aligns them by leveraging the information across slices. Finally, we would like to emphasize that the JBSS methods studied in this thesis, as well as the proposed method for identifying the relationship structure among multiple datasets, are not limited to the analysis of functional imaging data but can also be applied in different fields.

9.2 Future Work

Deriving further connections and theoretical understanding of the JBSS methods would help to ultimately give a recommendation on which JBSS method should be used for which type of application. A good start in this direction is, for example, to formally derive the connection between mCCA-genvar and a deflationary version of IVA-G with orthogonality constraint, which have similar cost functions [And10], and thereby be able to also connect the mCCA methods with IVA and PARAFAC2. Building on these connections, the next steps can be to formally prove the conjecture for mCCA-minvar and to derive and prove the source identification conditions of mCCA-genvar, mCCA-ssqcor, and PARAFAC2. Thereafter, the JBSS performance of IVA-G, IVA-L-SOS, all mCCA methods, and PARAFAC2 can be tested together in more extensive simulations, e.g., in simulations where the source identification conditions of the different methods are violated or in simulations inspired by multiple different applications.

Furthermore, the proposed 3-step method for identifying the relationship structure among multiple datasets can be applied to more interesting datasets to solve further medical challenges, of which we list a few possibilities in the following. For example, by identifying the relationship structure among subject datasets of, e.g., patients with schizophrenia, bipolar disorder, and their subtypes, the proposed method could be used for subgroup identification in precision medicine. Applying the proposed method to ANS measures from several time instances before and after epileptic seizures, and thereby identifying the changes in the relationship structure among these measures over time, might allow the development of novel techniques for seizure detection or prediction. Also, directly applying the suitable JBSS methods to analyze resting-state fMRI or fUS data is of great importance to obtain information about the functional organization of the brain [Urb15; PR21], which cannot be achieved with correlation images, as there exists no external stimulus.

Lastly, but just as important, is a comparison between the matrix and tensor

decomposition-based JBSS methods and deep learning methods. Both of them learn a representation from observed data in a data-driven manner, depending on their assumptions. A deep neural network trained for a classification task, e.g., the detection of schizophrenia, learns the representation that leads to the best classification accuracy and therefore outperforms matrix and tensor decompositions, which learn representations that are easier to interpret (because of their connection with the physical model [Ada22]) but typically do not achieve as high classification accuracies as the black-box models. With the advancement of eXplainable Artificial Intelligence (XAI) methods, strategies have been developed to explain the predictions or classifications of neural networks, e.g., SHapley Additive exPlanations (SHAP) [Lun17]. However, XAI methods in the end just perform a post-processing to try to understand what the network learns, but do not make the networks learn something interpretable in the first place. In contrast to this, matrix and tensor decomposition-based JBSS methods find interpretable components by their design (if the JBSS model matches the underlying physics of the data [Ada22] and the source identification conditions of a specific JBSS method are satisfied), which is of advantage in fields like medicine where interpretability is of importance. A compromise for finding interpretable representations but still achieving high classification accuracies can be found in deep-learning extensions of matrix and tensor decomposition methods, e.g., Deep CCA (DCCA) [And13b] or Deep Canonically Correlated AutoEncoders (DCCAE) [Wan15]. A strategy to explain DCCA, eXplanations for DCCA (XDCCA), is proposed in [Kus25]. A promising future direction in JBSS is to extend this strategy to deep-learning based extensions of mCCA like DGCCA [Ben17], which is an extension of maxvar [Car68]. Furthermore, we see a high potential in developing a deep-learning extension of genvar, as it may outperform DGCCA in JBSS of real-world data where source components in an SCV are assumed to be nonlinearly dependent and generated by more than one underlying factor. However, deep-learning based CCA methods typically come at the cost that it is not clear under which conditions the true underlying sources can be identified [Lyu20]. Identification conditions for a nonlinear multiview analysis method similar to DCCAE are derived and proven in [Lyu20], but they assume that the non-linearity is applied to each dimension separately, which usually is not the case in real-world applications. Therefore, to combine the advantages of both approaches, matrix and tensor decompositions can be explored to find identifiable and interpretable sources, e.g., biomarkers, and then, the deep learning methods can be used to perform downstream tasks such as classification or regression based on these biomarkers, as done in, e.g., [Vie21] and [Vie23; Has25].

List of publications

Journal articles

I. Lehmann, T. Hasija, B. Gabrielson, M. A. B. S. Akhonda, V. D. Calhoun, and T. Adali, "Identifying the Relationship Structure among Multiple Datasets Using Independent Vector Analysis: Application to Multi-task fMRI Data," *IEEE Access*, vol. 12, pp. 109 443–109 456, 2024.

B. Gabrielson, M. A. B. S. Akhonda, **I. Lehmann**, and T. Adali, "An Efficient Analytic Solution for Joint Blind Source Separation," *IEEE Transactions on Signal Processing*, vol. 72, pp. 2436–2449, 2024.

I. Lehmann*, B. Gabrielson*, T. Hasija, and T. Adali, "A Comprehensive Guide to Multiset Canonical Correlation Analysis and its Application to Joint Blind Source Separation," *accepted for IEEE Transactions on Signal Processing*, 2025.

Conference articles

I. Lehmann, E. Acar, T. Hasija, M. A. B. S. Akhonda, V. D. Calhoun, P. Schreier, and T. Adali, "Multi-task fMRI Data Fusion Using IVA and PARAFAC2," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1466–1470.

R. A. Borsoi, **I. Lehmann**, M. A. B. S. Akhonda, V. D. Calhoun, K. Usevich, D. Brie, and T. Adali, "Coupled CP tensor decomposition with shared and distinct components for multi-task fMRI data fusion," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

I. Lehmann, T. Adali, P. Kruizinga, and B. Hunyadi, "Deriving 3D Functional Brain Regions from Multi-Slice Functional Ultrasound Data Using ICA and IVA," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2023, pp. 1484–1490.

Acronyms

2D 2-dimensional.

3D 3-dimensional.

ALS Alternating Least Squares.

AMI Adjusted Mutual Information.

ANS Autonomic Nervous System.

AOD Auditory Oddball.

BOLD Blood Oxygenation Level Dependent.

BSS Blind Source Separation.

CCA Canonical Correlation Analysis.

CCV Canonical Component Vector.

CP CANDECOMP/PARAFAC.

CSF Cerebrospinal Fluid.

DCCA Deep CCA.

DCCAE Deep Canonically Correlated AutoEncoders.

DGCCA Deep GCCA.

DMN Default Mode Network.

ECG electrocardigram.

EHR Electronic Health Records.

EVD eigenvalue decomposition.

fMRI functional Magnetic Resonance Imaging.

fUS functional Ultrasound.

GCCA Generalized Canonical Correlation Analysis.

GEVD Generalized EVD.

GLM General Linear Model.

GTCS Generalized Tonic-Clonic Seizure.

HRF Hemodynamic Response Function.

ICA Independent Component Analysis.

ICA-EBM ICA by Entropy Bound Minimization.

IVA Independent Vector Analysis.

IVA-G IVA with a Gaussian model.

IVA-L-SOS IVA with a Laplacian model and second-order statistics.

JBSS Joint Blind Source Separation.

jISI joint Inter-Symbol-Interference.

mCCA multiset Canonical Correlation Analysis.

MCIC MIND Clinical Imaging Consortium.

MEG magnetoencephalography.

MGGD Multivariate Generalized Gaussian Distribution.

OLS Ordinary Least Squares.

PCA Principal Component Analysis.

pdf probability density function.

PDI Power Doppler Image.

SCV Source Component Vector.

SHAP SHapley Additive exPlanations.

SIRP Sternberg Item Recognition Paradigm.

SM Sensory Motor.

sMRI structural Magnetic Resonance Imaging.

SPM Statistical Parametric Mapping.

SVD Singular Value Decomposition.

SVM Support Vector Machine.

UVA Uncorrelated Vector Analysis.

w.l.o.g. without loss of generality.

w.r.t. with respect to.

XAI eXplainable Artificial Intelligence.

XDCCA eXplanations for DCCA.

List of symbols

The first column denotes the symbol name, the second column the meaning, and the third column the dimensions.

Symbol	Meaning	Dimension
N	Dimension of datasets	1
K	Number of datasets	1
V	Number of samples	1
$\mathbf{x}^{[k]}$	k^{th} dataset	N
$\mathbf{X}^{[k]}$	k^{th} observed dataset	$N \times V$
$\mathbf{C}_{\mathbf{x}}^{[k]}$	true covariance matrix of $\mathbf{x}^{[k]}$	$N \times N$
$\widehat{\mathbf{C}}_{\mathbf{x}}^{[k]}$	estimated covariance matrix of $\mathbf{X}^{[k]}$	$N \times N$
$\mathbf{C}_{\mathbf{x}}^{[k,l]}$	true cross-covariance matrix of $\mathbf{x}^{[k]}$ and $\mathbf{x}^{[l]}$	$N \times N$
$\mathbf{x}$	concatenated $\mathbf{x}^{[1]}, \ldots, \mathbf{x}^{[K]}$	NK
$\mathbf{C}_{\mathbf{x}}$	true covariance matrix of $\mathbf{x}$	$NK \times NK$
$\mathbf{D}_{\mathbf{x}}$	block-diagonal matrix of $\mathbf{C}_{\mathbf{x}}^{[1]}, \ldots, \mathbf{C}_{\mathbf{x}}^{[K]}$	$NK \times NK$
$\mathbf{y}^{[k]}$	k^{th} whitened dataset	N
$\mathbf{Y}^{[k]}$	k^{th} observed whitened dataset	$N \times V$
$\mathbf{y}$	concatenated $\mathbf{y}^{[1]}, \ldots, \mathbf{y}^{[K]}$	NK
$\mathbf{C}_{\mathbf{y}}$	covariance matrix of $\mathbf{y}$	$NK \times NK$
$s_n^{[k]}$	n^{th} true source component of the k^{th} dataset	1
$\mathbf{s}^{[k]}$ / $\mathbf{S}^{[k]}$	k^{th} true source vector	N / $N \times V$
$\widehat{\mathbf{s}}^{[k]}$ / $\widehat{\mathbf{S}}^{[k]}$	k^{th} estimated source vector	N / $N \times V$
$\mathbf{C}_{\mathbf{s}}^{[k]}$	true covariance matrix of $\mathbf{s}^{[k]}$	$N \times N$

$\widehat{\mathbf{C}}_{\mathbf{s}}^{[k]}$	estimated covariance matrix of $\widehat{\mathbf{S}}^{[k]}$	$N \times N$
$\mathbf{s}_n$	n^{th} source component vector (SCV)	K
$\mathbf{C}_{\mathbf{s}_n}$	true covariance matrix of the n^{th} SCV $\mathbf{s}_n$	$K \times K$
$\boldsymbol{\lambda}_n$	eigenvalues of $\mathbf{C}_{\mathbf{s}_n}$	K
$\mathbf{C}_{\mathbf{s}}$	true covariance matrix of concatenated $\mathbf{s}^{[1]}, \ldots, \mathbf{s}^{[K]}$	$NK \times NK$
$\mathbf{C}_{\mathbf{s}}^{(\text{SCV})}$	true covariance matrix of concatenated $\mathbf{s}_1, \ldots, \mathbf{s}_N$	$NK \times NK$
$\mathbf{A}^{[k]}$	mixing matrix of the k^{th} dataset	$N \times N$
$\mathbf{W}^{[k]}$	demixing matrix for the k^{th} dataset	$N \times N$
$\mathbf{P}$	permutation matrix	$N \times N$
$\boldsymbol{\Gamma}^{[k]}$	diagonal scale matrix for the k^{th} dataset	$N \times N$
$\widetilde{\boldsymbol{\Gamma}}^{[k]}$	diagonal matrix with diagonal elements equal to either -1 or 1, for the k^{th} dataset	$N \times N$
$u_n^{[k]}$	n^{th} canonical variable of the k^{th} dataset	1
$\mathbf{u}^{[k]}$ / $\mathbf{U}^{[k]}$	k^{th} canonical vector	N / $N \times V$
$\mathbf{u}_n$ / $\mathbf{U}_n$	n^{th} canonical component vector (CCV)	K / $K \times V$
$\mathbf{C}_{\mathbf{u}_n}$	covariance matrix of the n^{th} CCV $\mathbf{u}_n$	$K \times K$
$\boldsymbol{\ell}_n$	eigenvalues of $\mathbf{C}_{\mathbf{u}_n}$	K
$\left(\mathbf{T}^{[k]}\right)^{\mathsf{T}}$	transformation matrix for finding $\mathbf{u}^{[k]}$ / $\mathbf{U}^{[k]}$	$N \times N$

List of figures

List of tables

Bibliography

[Ada14] T. Adali, M. Anderson, and G. S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, 2014.

[Ada15] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal Data Fusion Using Source Separation: Two Effective Models Based on ICA and IVA and Their Properties," *Proceedings of the IEEE*, vol. 103, no. 9, 2015.

[Ada19] T. Adali, M. A. B. S. Akhonda, and V. D. Calhoun, "ICA and IVA for Data Fusion: An Overview and a New Approach Based on Disjoint Subspaces," *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, 2019.

[Ada22] T. Adali, F. Kantar, M. Akhonda, S. Strother, V. D. Calhoun, and E. Acar, "Reproducibility in matrix and tensor decompositions: focus on model match, interpretability, and uniqueness," *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 8–24, 2022.

[Akh21] M. Akhonda, B. Gabrielson, V. D. Calhoun, and T. Adali, "Complete model identification using independent vector analysis: application to the fusion of task FMRI data," in *2021 IEEE Data Science and Learning Workshop (DSLW)*, IEEE, 2021, pp. 1–6.

[And10] M. Anderson, X. L. Li, and T. Adali, "Nonorthogonal independent vector analysis using multivariate Gaussian model," in *International Conference on Latent Variable Analysis and Signal Separation*, Springer, 2010, pp. 354–361.

[And12] M. Anderson, T. Adali, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1672–1683, 2012.

[And13a] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adali, "Independent Vector Analysis, the Kotz Distribution, and Performance Bounds," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 3243–3247.

[And13b] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, PMLR, 2013, pp. 1247–1255.

[And14] M. Anderson, G. S. Fu, R. Phlypo, and T. Adali, "Independent Vector Analysis: Identification Conditions and Performance Bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, 2014.

[Ase15] N. A. Asendorf, "Informative Data Fusion: Beyond Canonical Correlation Analysis," Ph.D. dissertation, University of Michigan, 2015.

[Bas11] D. S. Bassett and M. S. Gazzaniga, "Understanding Complexity in the Human Brain," *Trends in Cognitive Sciences*, vol. 15, no. 5, pp. 200–209, May 2011.

[Bel95] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[Ben17] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[Ber23] A. Bertolo, J. Ferrier, S. Cazzanelli, S. Diebolt, M. Tanter, S. Pezet, M. Pernot, B.-F. Osmanski, and T. Deffieux, "High sensitivity mapping of brain-wide functional networks in awake mice using simultaneous multi-slice fUS imaging," *Imaging Neuroscience*, vol. 1, pp. 1–18, Nov. 2023.

[Bhi17] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Data-Driven Fusion of Multi-Camera Video Sequences: Application to Abandoned Object Detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 1697–1701.

[Bhi19] S. Bhinge, R. Mowakeaa, V. D. Calhoun, and T. Adali, "Extraction of Time-Varying Spatiotemporal Networks Using Parameter-Tuned Constrained IVA," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1715–1725, 2019.

[Boe24] V. L. Boerwinkle, M. A. Nowlen, J. E. Vazquez, M. A. Arhin, W. R. Reuther, E. G. Cediel, P. J. McCarty, I. Manjón, J. H. Jubran, A. C. Guest, K. D. Gillette, F. M. Nowlen, A. R. Pines, M. H. Kazemi, and B. F. Qaqish, "Resting-State fMRI Seizure Onset Localization Meta-Analysis: Comparing Rs-fMRI to Other Modalities Including Surgical Outcomes," *Frontiers in Neuroimaging*, vol. 3, p. 1 481 858, Dec. 2024.

[Bos14] D. J. Bos, T. R. van Raalten, B. Oranje, A. R. Smits, N. A. Kobussen, J. van Belle, S. A. R. B. Rombouts, and S. Durston, "Developmental Differences in Higher-Order Resting-State Networks in Autism Spectrum Disorder," *NeuroImage: Clinical*, vol. 4, pp. 820–827, Jan. 2014.

[Bou21] Z. Boukouvalas, M. Puerto, D. C. Elton, P. W. Chung, and M. D. Fuge, "Independent Vector Analysis for Molecular Data Fusion: Application to Property Prediction and Knowledge Discovery of Energetic Materials," in *2020 28th European Signal Processing Conference (EUSIPCO)*, Jan. 2021, pp. 1030–1034.

[Bru21] C. Brunner, M. Grillet, A. Urban, B. Roska, G. Montaldo, and E. Macé, "Whole-brain functional ultrasound imaging in awake head-fixed mice," *Nature Protocols*, vol. 16, no. July, pp. 3547–3571, 2021.

[Cal01] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. J. Pekar, "A Method for Making Group Inferences from Functional MRI Data Using Independent Component Analysis," *Human Brain Mapping*, vol. 14, pp. 140–151, 2001.

[Cal04] V. D. Calhoun, K. A. Kiehl, P. F. Liddle, and G. D. Pearlson, "Aberrant localization of synchronous hemodynamic activity in auditory cortex reliably characterizes schizophrenia," *Biological psychiatry*, vol. 55, no. 8, pp. 842–849, 2004.

[Cal06] V. D. Calhoun, T. Adali, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson, "A method for multitask fMRI data fusion applied to schizophrenia," *Human Brain Mapping*, vol. 27, no. 7, pp. 598–610, 2006.

[Cal12] V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 60–73, 2012.

[Car68] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proceedings of the 76th annual convention of the American Psychological Association*, 1968, pp. 227–228.

[Car70] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[Cer01] M. Cercignani, M. Inglese, M. Siger-Zajdel, and M. Filippi, "Segmenting Brain White Matter, Gray Matter and Cerebro-Spinal Fluid Using Diffusion Tensor-MRI Derived Indices," *Magnetic Resonance Imaging*, vol. 19, no. 9, pp. 1167–1172, Nov. 2001.

[Cha14] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavsky, and A. Yeredor, "Joint Matrices Decompositions and Blind Source Separation: A Survey of Methods, Identification, and Applications," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 34–43, May 2014.

[Che07] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 143–152.

[Che18] M. M. Cheriyan, P. A. Michael, and A. Kumar, "Blind Source Separation with Mixture Models – A Hybrid Approach to MR Brain Classification," *Magnetic Resonance Imaging*, vol. 54, pp. 137–147, Dec. 2018.

[Che19] J. Chen, G. Wang, and G. B. Giannakis, "Graph multiview canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 2826–2838, 2019.

[Com94] P. Comon, "Independent Component Analysis, A New Concept?" *Signal Processing*, Higher Order Statistics, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[Con22] Connectome Coordination Facility, "Components of the Human Connectome Project: Task fMRI," *https://www.humanconnectome.org/study/hcp-young-adult/project-protocol/task-fmri*, 2022.

[Cor10] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 39–50, 2010.

[Dam24] L. P. Damasceno, E. Rexhepi, A. Shafer, I. Whitehouse, N. Japkowicz, C. C. Cavalcante, R. Corizzo, and Z. Boukouvalas, "Exploiting Sparsity and Statistical Dependence in Multivariate Data Fusion: An Application to Misinformation Detection for High-Impact Events," *Machine Learning*, vol. 113, no. 4, pp. 2183–2205, Apr. 2024.

[Dav06] R. Davidson and J. G. MacKinnon, "The power of bootstrap and asymptotic tests," *Journal of Econometrics*, vol. 133, no. 2, pp. 421–441, 2006.

[de 18] F. de Vos, M. Koini, T. M. Schouten, S. Seiler, J. van der Grond, A. Lechner, R. Schmidt, M. de Rooij, and S. A. R. B. Rombouts, "A Comprehensive Analysis of Resting State fMRI Measures to Classify Individual Patients with Alzheimer's Disease," *NeuroImage*, vol. 167, pp. 62–72, Feb. 2018.

[DL06] L. De Lathauwer, "A Link between the Canonical Decomposition in Multilinear Algebra and Simultaneous Matrix Diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 3, pp. 642–666, Jan. 2006.

[Du12] W. Du, V. D. Calhoun, H. Li, S. Ma, T. Eichele, K. A. Kiehl, G. D. Pearlson, and T. Adali, "High classification accuracy for schizophrenia with rest and task fmri data," *Frontiers in Human Neuroscience*, vol. 6, p. 145, 2012.

[Ero22] A. Erol, C. Soloukey, B. Generowicz, N. Van Dorp, S. Koekkoek, P. Kruizinga, and B. Hunyadi, "Deconvolution of the Functional Ultrasound Response in the Mouse Visual Pathway Using Block-Term Decomposition," *Neuroinformatics*, pp. 1–19, 2022.

[Fal14] N. Falco, J. A. Benediktsson, and L. Bruzzone, "A Study on the Effectiveness of Different Independent Component Analysis Algorithms for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2183–2199, Jun. 2014.

[Fu17] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M.-C. So, "Scalable and flexible multiview MAX-VAR canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4150–4165, 2017.

[Gab24] B. Gabrielson, M. A. Akhonda, I. Lehmann, and T. Adali, "An Efficient Analytic Solution for Joint Blind Source Separation," *IEEE Transactions on Signal Processing*, vol. 72, pp. 2436–2449, 2024.

[Gao17] L. Gao, L. Qi, E. Chen, and L. Guan, "Discriminative multiple canonical correlation analysis for information fusion," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1951–1965, 2017.

[Gie04] J. N. Giedd, "Structural Magnetic Resonance Imaging of the Adolescent Brain," *Annals of the New York Academy of Sciences*, vol. 1021, no. 1, pp. 77–85, 2004.

[Gol13] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, D. O. Leary, C Ho, S. Brauns, and D. S. Manoach, "The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, 2013.

[Góm98] E. Gómez, M. Gomez-Viilegas, and J. M. Marín, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[Gre04] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-Mode Network Activity Distinguishes Alzheimer's Disease from Healthy Aging: Evidence from Functional MRI," *Proceedings of the National Academy of Sciences*, vol. 101, no. 13, pp. 4637–4642, Mar. 2004.

[Har70] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[Har72] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," *UCLA working papers in phonetics*, vol. 22, no. 10, pp. 30–44, 1972.

[Has20] T. Hasija, T. Marrinan, C. Lameiro, and P. J. Schreier, "Determining the dimension and structure of the subspace correlated across multiple data sets," *Signal Processing*, vol. 176, p. 107 613, 2020.

[Has25] T. Hasija, M. Kuschel, M. Jackson, S. Dailey, H. Menne, C. Reinsberger, S. Vieluf, and T. Loddenkemper, *Improving Wearable-Based Seizure Prediction by Feature Fusion Using Growing Network*, Jan. 2025.

[Hen19] J. Henríquez and W. Kristjanpoller, "A Combined Independent Component Analysis–Neural Network Model for Forecasting Exchange Rate Variation," *Applied Soft Computing*, vol. 83, p. 105 654, Oct. 2019.

[Her86] J. Herault and C. Jutten, "Space or Time Adaptive Signal Processing by Neural Network Models," *AIP Conference Proceedings*, vol. 151, no. 1, pp. 206–211, Aug. 1986.

[Hik23] K. Hikishima, T. Tsurugizawa, K. Kasahara, R. Takagi, K. Yoshinaka, and N. Nitta, "Brain-wide mapping of resting-state networks in mice using high-frame rate functional ultrasound," *NeuroImage*, vol. 279, p. 120 297, 2023.

[Hor61] P. Horst, *Generalized canonical correlations and their application to experimental data*. Journal of clinical psychology, 1961.

[Hot36] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[Hu17] M.-L. Hu, X.-F. Zong, J. J. Mann, J.-J. Zheng, Y.-H. Liao, Z.-C. Li, Y. He, X.-G. Chen, and J.-S. Tang, "A review of the functional and anatomical default mode network in schizophrenia," *Neuroscience bulletin*, vol. 33, no. 1, pp. 73–84, 2017.

[Hub85] P. J. Huber, "Projection pursuit," *The annals of Statistics*, pp. 435–475, 1985.

[Hun17] B. Hunyadi, P. Dupont, W. Van Paesschen, and S. Van Huffel, "Tensor decompositions and data fusion in epileptic electroencephalography and functional magnetic resonance imaging data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 1, pp. 1–15, 2017.

[Hyv00] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411–430, 2000.

[Jut91] C. Jutten and J. Herault, "Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, Jul. 1991.

[Kal21] Y. Kaloga, P. Borgnat, S. P. Chepuri, P. Abry, and A. Habrard, "Variational graph autoencoders for multiview canonical correlation analysis," *Signal Processing*, vol. 188, p. 108 182, 2021.

[Kan18] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and M. Hong, "Structured SUMCOR multiview canonical correlation analysis for large-scale data," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 306–319, 2018.

[Kar11] P. Karunanayaka, K. K. Kim, S. K. Holland, and J. P. Szaflarski, "The Effects of Left or Right Hemispheric Epilepsy on Language Networks Investigated with Semantic Decision fMRI Task and Independent Component Analysis," *Epilepsy & Behavior*, vol. 20, no. 4, pp. 623–632, Apr. 2011.

[Kay98] S. M. Kay, *Fundamentals of statistical signal processing: detection theory*. Prentice-Hall, Inc., 1998.

[Kes18] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, 2018.

[Ket71] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[Kie99] H. A. Kiers, J. M. Ten Berge, and R. Bro, "PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model," *Journal of Chemometrics*, vol. 13, no. 3-4, pp. 275–294, 1999.

[Kim06] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. 40th Asilomar Conference on Signals, Systems and Computers*, 2006, pp. 1393–1396.

[Kim07] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind Source Separation Exploiting Higher-Order Frequency Dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, Jan. 2007.

[Kim09] D. I. Kim, D. Mathalon, J. Ford, M Mannell, J. Turner, G. Brown, A Belger, R Gollub, J Lauriello, C Wible, *et al.*, "Auditory oddball deficits in schizophrenia: an independent component analysis of the fMRI multisite function BIRN study," *Schizophrenia bulletin*, vol. 35, no. 1, pp. 67–81, 2009.

[Kim23] S.-G. Kim and P. A. Bandettini, "Principles of BOLD Functional MRI," in *Functional Neuroradiology: Principles and Clinical Applications*, S. H. Faro and F. B. Mohamed, Eds., Cham: Springer International Publishing, 2023, pp. 461–472.

[Koc12] W. Koch, S. Teipel, S. Mueller, J. Benninghoff, M. Wagner, A. L. W. Bokde, H. Hampel, U. Coates, M. Reiser, and T. Meindl, "Diagnostic Power of Default Mode Network Resting State fMRI in the Detection of Alzheimer's Disease," *Neurobiology of Aging*, vol. 33, no. 3, pp. 466–478, Mar. 2012.

[Kol09] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[Kos19a] M. R. Kosorok and E. B. Laber, "Precision Medicine," *Annual Review of Statistics and Its Application*, vol. 6, no. Volume 6, 2019, pp. 263–286, Mar. 2019.

[Kos19b] J. Kossaifi, Y. Panagakis, A. Anandkumar, and M. Pantic, "Tensorly: Tensor learning in python," *Journal of Machine Learning Research*, vol. 20, no. 26, pp. 1–6, 2019.

[Kot01] S. Kotz, T. Kozubowski, and K. Podgórski, *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2001.

[Kus25] M. Kuschel, A. Alkhatib, T. Hasija, and H. Boström, "Explaining Representations in Correlation-based Deep Multiview Representation Learning," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[Lac08] C Lacadie, R. Fulbright, J Arora, R Constable, and X Papademetris, "Brodmann areas defined in mni space using a new tracing tool in bioimage suite," in *Proceedings of the 14th annual meeting of the organization for human brain mapping*, vol. 771, 2008.

[Lah15] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[Lan20] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–6.

[Le24] T. T. Le, K. Abed-Meraim, P. Ravier, O. Buttelli, and A. Holobar, "Tensor Decomposition Meets Blind Source Separation," *Signal Processing*, vol. 221, p. 109 483, Aug. 2024.

[Leh22] I. Lehmann, E. Acar, T. Hasija, M. Akhonda, V. D. Calhoun, P. Schreier, and T. Adali, "Multi-task fMRI Data Fusion Using IVA and PARAFAC2," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1466–1470.

[Li08] Y.-O. Li, W. Wang, T. Adali, and V. D. Calhoun, "CCA for joint blind source separation of multiple datasets with application to group fMRI analysis," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1837–1840.

[Li09] Y. O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[Li10] X.-L. Li and T. Adali, "Blind Spatiotemporal Separation of Second and/or Higher-Order Correlated Sources by Entropy Rate Minimization," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 1934–1937.

[Lip17] I. Lipkovich, A. Dmitrienko, and R. B. D'Agostino, "Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials," *Statistics in Medicine*, vol. 36, no. 1, pp. 136–196, 2017.

[Loh19] W.-Y. Loh, L. Cao, and P. Zhou, "Subgroup identification for precision medicine: A comparative review of 13 methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 5, e1326, 2019.

[Lon18] Q. Long, C. Jia, Z. Boukouvalas, B. Gabrielson, D. Emge, and T. Adali, "Consistent run selection for independent component analysis: Application to fmri analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2581–2585.

[Lon20] Q. Long, S. Bhinge, V. D. Calhoun, and T. Adali, "Independent vector analysis for common subspace analysis: Application to multi-subject fMRI data yields meaningful subgroups of schizophrenia," *NeuroImage*, vol. 216, p. 116 872, 2020.

[LS17] Y. Levin-Schwartz, V. D. Calhoun, and T. Adali, "Quantifying the Interaction and Contribution of Multiple Datasets in Fusion: Application to the Detection of Schizophrenia," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1385–1395, 2017.

[Lun17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[Luo18] Z. Luo, C. Li, and L. Zhu, "A comprehensive survey on blind source separation for wireless adaptive processing: Principles, perspectives, challenges and new research directions," *IEEE Access*, vol. 6, pp. 66 685–66 708, 2018.

[Lyu20] Q. Lyu and X. Fu, "Nonlinear Multiview Analysis: Identifiability and Neural Network-Assisted Implementation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2697–2712, 2020.

[Mac11] E. Macé, G. Montaldo, I. Cohen, M. Baulac, M. Fink, and M. Tanter, "Functional ultrasound imaging of the brain," *Nature Methods*, vol. 8, no. 8, pp. 662–664, 2011.

[Mad17] K. H. Madsen, N. W. Churchill, and M. Mørup, "Quantifying functional connectivity in multi-subject fmri data using component models," *Human Brain Mapping*, vol. 38, no. 2, pp. 882–899, 2017.

[Mar21] M. Markicevic, I. Savvateev, C. Grimm, and V. Zerbi, "Emerging imaging methods to study whole-brain function in rodent models," *Translational Psychiatry*, vol. 11, no. 1, 2021.

[Mem20] Members and collaborators of the Wellcome Centre for Human Neuroimaging, *Statistical Parametric Mapping Toolbox: SPM12*, 2020. [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/.

[Mic09] A. M. Michael, S. A. Baum, J. F. Fries, B. C. Ho, R. K. Pierson, N. C. Andreasen, and V. D. Calhoun, "A method to fuse fMRI tasks through spatial correlations: Applied to schizophrenia," *Human Brain Mapping*, vol. 30, no. 8, pp. 2512–2529, 2009.

[Mij12] B. Mijović, K. Vanderperren, N. Novitskiy, B. Vanrumste, P. Stiers, B. Van den Bergh, L. Lagae, S. Sunaert, J. Wagemans, S. Van Huffel, and M. De Vos, "The "why" and "how" of JointICA: Results from a visual detection task," *NeuroImage*, vol. 60, no. 2, pp. 1171–1185, 2012.

[Mum06] J. Mumford and T. Nichols, "Modeling and Inference of Multisubject fMRI Data," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 2, pp. 42–51, Mar. 2006.

[Mwa17] T. E. Mwansisya, A. Hu, Y. Li, X. Chen, G. Wu, X. Huang, D. Lv, Z. Li, C. Liu, Z. Xue, J. Feng, and Z. Liu, "Task and resting-state fMRI studies in first-episode schizophrenia: A systematic review," *Schizophrenia Research*, vol. 189, pp. 9–18, 2017.

[Nat11] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease, *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, 2011.

[Nie02] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[Nie95] A. A. Nielsen, "Analysis of Regularly and Irregularly Sampled Spatial, Multivariate, and Multi-temporal Data," Ph.D. dissertation, Institute of Mathematical Modelling, Technical University of Denmark, 1995.

[Osm14] B. F. Osmanski, S. Pezet, A. Ricobaraza, Z. Lenkei, and M. Tanter, "Functional ultrasound imaging of intrinsic connectivity in the living rat brain with high spatiotemporal resolution," *Nature Communications*, vol. 5, 2014.

[Ped11] 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*, vol. 12, pp. 2825–2830, 2011.

[Pen24] M. Peng, M. Motagh, Z. Lu, Z. Xia, Z. Guo, C. Zhao, and Q. Liu, "Characterization and Prediction of InSAR-derived Ground Motion with ICA-assisted LSTM Model," *Remote Sensing of Environment*, vol. 301, p. 113 923, Feb. 2024.

[Pes14] L. Pessoa, "Understanding Brain Networks and Brain Organization," *Physics of Life Reviews*, vol. 11, no. 3, pp. 400–435, Sep. 2014.

[PR21] P. Pais-Roldán, C. Mateo, W. J. Pan, B. Acland, D. Kleinfeld, L. H. Snyder, X. Yu, and S. Keilholz, "Contribution of animal models toward understanding resting state functional connectivity," *NeuroImage*, vol. 245, Sep. 2021.

[Qi18] S. Qi, X. Yang, L. Zhao, V. D. Calhoun, N. Perrone-Bizzozero, S. Liu, R. Jiang, T. Jiang, J. Sui, and X. Ma, "MicroRNA132 associated multimodal neuroimaging patterns in unmedicated major depressive disorder," *Brain*, vol. 141, no. 3, pp. 916–926, 2018.

[Ram15] M. Ramezani, K. Marble, H. Trang, I. S. Johnsrude, and P. Abolmaesumi, "Joint sparse representation of brain activity patterns in multi-task fMRI data," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 2–12, 2015.

[Roa20] M. Roald, S. Bhinge, C. Jia, V. Calhoun, T. Adalı, and E. Acar, "Tracing network evolution using the PARAFAC2 model," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1100–1104.

[Ror04] C. Rorden and H.-O. Karnath, "Using Human Brain Lesions to Infer Function: A Relic from a Past Era in the fMRI Age?" *Nature Reviews Neuroscience*, vol. 5, no. 10, pp. 812–819, Oct. 2004.

[San20] N. Sanford, J. C. Whitman, and T. S. Woodward, "Task-merging for finer separation of functional brain networks in working memory," *Cortex*, vol. 125, pp. 246–271, 2020.

[Sar14] M. Sarfraz, A. A. Khan, and F. F. Li, "Using Independent Component Analysis to Obtain Feature Space for Reliable ECG Arrhythmia Classification," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2014, pp. 62–67.

[Sas13] K Sasirekha and P Baby, "Agglomerative hierarchical clustering algorithm – a review," *International Journal of Scientific and Research Publications*, vol. 83, no. 3, p. 83, 2013.

[Sch08] P. J. Schreier, "A unifying discussion of correlation analysis for complex random vectors," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1327–1336, 2008.

[Smi04] S. M. Smith, "Overview of fMRI Analysis," *British Journal of Radiology*, vol. 77, no. 2, pp. 167–175, Dec. 2004.

[Sor21] M. Sorensen, C. I. Kanatsoulis, and N. D. Sidiropoulos, "Generalized Canonical Correlation Analysis: A Subspace Intersection Approach," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2452–2467, 2021.

[Sta09] J. H. Stapleton, *Linear statistical models*. John Wiley & Sons, 2009, p. 31.

[Ste51] R. G. Steel, "Minimum generalized variance for a set of linear functions," *The Annals of Mathematical Statistics*, pp. 456–460, 1951.

[Sui13] J. Sui, H. He, G. D. Pearlson, T. Adali, K. A. Kiehl, Q. Yu, V. P. Clark, E. Castro, T. White, B. A. Mueller, *et al.*, "Three-way (N-way) fusion of brain imaging data based on mCCA+ jICA and its application to discriminating schizophrenia," *NeuroImage*, vol. 66, pp. 119–132, 2013.

[Tre21] A. H. Treacher, P. Garg, E. Davenport, R. Godwin, A. Proskovec, L. G. Bezerra, G. Murugesan, B. Wagner, C. T. Whitlow, J. D. Stitzel, J. A. Maldjian, and A. A. Montillo, "MEGnet: Automatic ICA-based Artifact Removal for MEG Using Spatiotemporal Convolutional Neural Networks," *NeuroImage*, vol. 241, p. 118 402, Nov. 2021.

[Urb15] A. Urban, C. Dussaux, G. Martel, C. Brunner, E. Mace, and G. Montaldo, "Real-time imaging of brain activity in freely moving rats using functional ultrasound," *Nature Methods*, vol. 12, no. 9, pp. 873–878, 2015.

[Vía11] J. Vía, M. Anderson, X. L. Li, and T. Adali, "Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2520–2523.

[Vie21] S. Vieluf, T. Hasija, P. J. Schreier, R. El Atrache, S. Hammond, F. Mohammadpour Touserkani, R. A. Sarkis, T. Loddenkemper, and C. Reinsberger, "Generalized Tonic-Clonic Seizures Are Accompanied by Changes of Interrelations within the Autonomic Nervous System," *Epilepsy & Behavior*, vol. 124, p. 108 321, Nov. 2021.

[Vie23] S. Vieluf, T. Hasija, M. Kuschel, C. Reinsberger, and T. Loddenkemper, "Developing a Deep Canonical Correlation-Based Technique for Seizure Prediction," *Expert Systems with Applications*, vol. 234, p. 120 986, Dec. 2023.

[Vin09] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" In *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1073–1080.

[Vir20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[Wan15] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On Deep Multi-View Representation Learning," in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 1083–1092.

[Wan20a] Q. Wang, S.-L. Ding, Y. Li, J. Royall, D. Feng, P. Lesnar, N. Graddis, M. Naeemi, B. Facer, A. Ho, *et al.*, "The Allen mouse brain common coordinate framework: a 3D reference atlas," *Cell*, vol. 181, no. 4, pp. 936–953, 2020.

[Wan20b] Y. Wang, Y. Zhao, T. M. Therneau, E. J. Atkinson, A. P. Tafti, N. Zhang, S. Amin, A. H. Limper, S. Khosla, and H. Liu, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *Journal of Biomedical Informatics*, vol. 102, pp. 1–10, Dec. 2020.

[Wax85] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[WG09] S. Whitfield-Gabrieli, H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette, *et al.*, "Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with

schizophrenia," *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1279–1284, 2009.

[Wij23] R. Wijnands, J. Dauwels, I. Serra, P. Kruizinga, A. Badura, and B. Hunyadi, "Modeling and inference of sparse neural dynamic functional connectivity networks underlying functional ultrasound data," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, IEEE, 2023, pp. 1–5.

[Wil32] S. S. Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, pp. 471–494, 1932.

[Yan22] H. Yang, M. A. B. S. Akhonda, F. Ghayem, Q. Long, V. D. Calhoun, and T. Adali, "Independent Vector Analysis Based Subgroup Identification from Multisubject fMRI Data," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 1471–1475.

[Yu08] S.-N. Yu and K.-T. Chou, "Integration of Independent Component Analysis and Neural Networks for ECG Beat Classification," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2841–2846, May 2008.

[Zia23] S. Ziani, Y. Farhaoui, and M. Moutaib, "Extraction of Fetal Electrocardiogram by Combining Deep Learning and SVD-ICA-NMF Methods," *Big Data Mining and Analytics*, vol. 6, no. 3, pp. 301–310, Sep. 2023.

[Zou04] A. M. Zoubir and D. R. Iskander, *Bootstrap techniques for signal processing*. Cambridge University Press, 2004, ch. Hypothesis Testing with the Bootstrap.