UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Determining Multivariate Association between Multiple Data Sets with Applications to Neuroscience and Acoustic Networks

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

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)
vorgelegte Dissertation

von

M.Sc. Tanuj Hasija

Erster Gutachter: Prof. Dr. Peter J. Schreier
Zweite Gutachterin: Prof. Dr. Tülay Adalı

Tag der mündlichen Prüfung: 15.06.2021

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Abstract

Determining Multivariate Association between Multiple Data Sets with Applications to Neuroscience and Acoustic Networks

by **Tanuj Hasija**

Analyzing multivariate association among multiple data sets is essential in various fields like biomedicine, image processing, robotics and wearable technology. Although there are several ways to measure association between data sets, this thesis deals with correlation, a function of the second-order moments of the data, and thus analyzes linear association among different data sets. Statistical tools like (multiset) canonical correlation analysis can be employed to extract maximally correlated components from different data sets. However, whether the estimated correlation among these components is significant or is spurious due to limited data (or noise) if often ignored. To completely characterize the linear association, estimating the complete correlation structure, i.e., which of the extracted components are correlated and across which data sets, is necessary.

The two most critical challenges in this context are the large number of combinations in which components can be correlated and limited number of observations compared to the dimensionality of the data sets. When analyzing correlation between two data sets, it can be assumed without loss of generality that a component in the first data set can be correlated with only one other component in the second data set. This is not true for more than two data sets. Moreover, there can be several ways in which the components can be correlated in more than two data sets. Some components can be correlated across all data sets, some across different subsets of data sets and some completely uncorrelated. The second challenge is that when the number of observations is comparable to or smaller than the dimensions of the data, the correlation among the components is highly overestimated making the analysis and the subsequent inference extremely dubious.

This thesis addresses the aforementioned challenges by developing novel techniques for reliably determining the complete linear association between multiple data sets. First, a special correlation structure among the components is assumed, and the knowledge and tools existing for two data sets are applied. Then the more challenging problem of arbitrary correlation structure is addressed. The necessary and sufficient conditions under which the complete correlation structure can be identified are theoretically derived. The proposed techniques are based on statistical methods and thus allow interpretability, while at the same time, require minimal assumptions and thus are designed to be data driven. Their advantages over the state-of-the-art are demonstrated using extensive numerical examples. The developed techniques are also applied on real-world data from the fields of wireless acoustic networks, sports science and epilepsy, where estimating the complete correlation structure and quantifying the strength of association between multiple modalities lead to significant performance gains and identification of potential biomarkers.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

FAKULTÄT FÜR
ELEKTROTECHNIK,
INFORMATIK UND
MATHEMATIK

Zusammenfassung der Dissertation

Determining Multivariate Association between Multiple Data Sets with Applications to Neuroscience and Acoustic Networks

des **Herrn Tanuj Hasija**

Die Analyse des multivariaten Zusammenhangs zwischen unterschiedlichen Datensätzen ist in diversen Anwendungsgebieten wie Biomedizin, Bildverarbeitung, Robotik und Wearable Technology von grundlegender Bedeutung. Obwohl mehrere Möglichkeiten existieren, den Zusammenhang zwischen Datenätzen zu messen, befasst sich diese Arbeit mit der Korrelation. Dies ist eine Funktion der Momente zweiter Ordnung von den Daten und analysiert daher den linearen Zusammenhang zwischen unterschiedlichen Datensätzen. Statistische Methoden wie die kanonische Korrelationsanalyse (für mehrere Datensätze) können eingesetzt werden, um stark korrelierte Komponenten aus verschiedenen Datensätzen zu extrahieren. Ob die geschätzte Korrelation zwischen diesen Komponenten jedoch signifikant ist oder nur aufgrund von begrenzten Daten (oder Rauschen) auftritt, wird oft nicht berücksichtigt. Um den linearen Zusammenhang vollständig charakterisieren zu können, ist eine Schätzung der vollständigen Korrelationsstruktur - d.h. welche der extrahierten Komponenten in welchen Datensätzen korreliert sind - erforderlich.

Die beiden größten Herausforderungen in diesem Zusammenhang sind zum einen die große Anzahl von Kombinationen, in denen die Komponenten korreliert sein können, und zum anderen die begrenzte Anzahl von Beobachtungen im Vergleich zur Dimensionalität der Datensätze. Bei der Analyse der Korrelation zwischen zwei Datensätzen kann ohne Verlust der Allgemeingültigkeit angenommen werden, dass eine Komponente des ersten Datensatzes nur mit einer anderen Komponente des zweiten Datensatzes korreliert sein kann. Bei mehr als zwei Datensätzen ist dies nicht der Fall. Außerdem kann es mehrere Möglichkeiten geben, wie die Komponenten zwischen mehr als zwei Datensätzen korreliert sein können. Einige Komponenten können zwischen allen Datensätzen korreliert sein, einige zwischen verschiedenen Teilgruppen von Datensätzen und einige können vollkommen unkorreliert sein. Die zweite Herausforderung besteht darin, dass, wenn die Anzahl der Beobachtungen vergleichbar zu oder sogar kleiner als die Dimension der Daten ist, die Korrelation zwischen den Komponenten deutlich zu hoch geschätzt wird, was die Analyse und die anschließende Schlussfolgerung sehr fragwürdig erscheinen lässt.

Diese Arbeit befasst sich mit den oben genannten Herausforderungen, indem neue Methoden zur zuverlässigen Bestimmung des vollständigen linearen Zusammenhangs zwischen mehreren Datensätzen entwickelt werden. Zunächst wird eine spezielle Korrelationsstruktur zwischen den Komponenten angenommen und die Erkenntnisse und Methoden für zwei Datensätze werden angewendet. Darauf aufbauend wird das anspruchsvollere Problem einer beliebigen Korrelationsstruktur angegangen. Die notwendigen und hinreichenden Bedingungen zur Ermittlung der vollständigen Korrelationsstruktur werden theoretisch hergeleitet. Die entwickelten Methoden basieren auf statistischer Theorie und schaffen somit eine Möglichkeit zur Interpretation, während sie gleichzeitig nur wenige Annahmen erfordern und daher datengetrieben konzipiert sind. Anhand umfangreicher numerischer Beispiele werden die Vorteile dieser neuen Methoden gegenüber dem Stand der Technik demonstriert. Die entwickelten Techniken werden darüber hinaus auch auf reale Daten aus den Bereichen drahtlose akustische Netzwerke, Sportwissenschaft und Epilepsie angewendet, wo die Schätzung der vollständigen Korrelationsstruktur und der Stärke der Zusammenhänge zwischen mehreren Modalitäten zu signifikanten Verbesserungen der Ergebnisse und zur Identifizierung potentieller Biomarker führen.

Acknowledgements

I cannot fully express my gratitude towards everyone who has inspired and supported me during my doctoral days. First and foremost, I would like to thank my supervisor, Prof. Peter J. Schreier for his endless guidance, complete support and for providing a great research culture during my entire doctoral work. His vast knowledge and in-depth understanding in the field of statistical signal processing has always inspired me to learn more and not be content. His emphasis on the importance of bottom-up thinking, simple and concise writing, and effective presentation has greatly improved me as a researcher. I also deeply value his effort to organize informal team retreats and outings which has created a wonderful team environment not strictly limited to a professional relationship.

I am thankful to my second examiner, Prof. Tülay Adalı who has been a wonderful collaborator and whose invaluable guidance has helped me in different stages of my doctorate. Her endless passion for research has been a source of great motivation till this day. I would also like to thank her for providing me with an opportunity to visit her lab at UMBC for two weeks in the summer of 2017. I am also grateful to my committee members for their valuable time and effort in evaluating my thesis.

I would like to pay special regards to my collaborators: Dr. Solveig Vieluf, Dr. Michael Muma and Dr. Yuri Levin-Schwartz for the insightful discussions about the open problems in various real-world applications and for our ongoing research to address some of them. I also wish to thank my current and former group members: Stefanie Horstmann, Isabell Lehmann, Anna Merle, Mohammad Soleymani, Dr. Yu-Hang Xiao, Artur Lamm, Stefan Pede, Aaron Pries, Stefan Reimer, Dr. Tim Marrinan, Dr. Christian Lameiro, Dr. David Ramírez and Dr. Yang Song for joint research and for the countless discussions specially during lunch and coffee breaks. I would also like to thank my cricket team and friends in Germany for providing me with a family away from family.

Finally I wish to express my deepest gratitude to my entire family for always believing in

me. My parents for their selfless love, support and for always standing besides me. My sister who has been a source of limitless energy. My wife who has been a pillar of support for me day-in and day-out and has always challenged me to become a better person. My newborn daughter who has been wonderful in sleeping calmly through the nights. Without her support, I would not have been fresh enough to complete my dissertation in time. This thesis would not have been possible without all of them.

Contents

Acronyms

AIC	Akaike information criterion
ANS	autonomic nervous system
AR	autoregressive
BIC	Bayesian information criterion
BPSK	binary phase-shift keying
CAN	central autonomic network
CCA	canonical correlation analysis
CFAR	constant false alarm rate
DOA	direction-of-arrival
EDA	electrodermal activity
EEG	electroencephalography
EVD	eigenvalue decomposition
fMRI	functional magnetic resonance imaging
GENVAR	generalized variance
GLR	generalized likelihood ratio
GLRT	generalized likelihood ratio test
GSMM-NICA	group-sparse median-based multiplicative nonnegative independent component analysis
HR	heart rate

HRV	heart rate variability
HT	hypothesis testing
ICA	independent component analysis
ITC	information theoretic criterion
IVA	independent vector analysis
JBSS	joint blind source separation
KL	Kullback-Leibler
KPD	knee-point detector
LASSO	least absolute shrinkage and selection operator
LRT	likelihood ratio test
mCCA	multiset canonical correlation analysis
MAXVAR	maximum variance
MDL	minimum description length
M-NICA	multiplicative nonnegative independent component analysis
MRI	magnetic resonance imaging
ncPCA	noncircular principal component analysis
PCA	principal component analysis
sMRI	structural magnetic resonance imaging
SCV	source component vector
SMM-NICA	sparse median-based multiplicative nonnegative independent component analysis
SNR	signal-to-noise-ratio
STFT	short-term Fourier transform
SVD	singular value decomposition
Temp	skin temperature at wrist
VAD	voice activity detection

VAP	voice activity pattern
WASN	wireless acoustic sensor network

List of notations

General notations

x	scalar
$\hat{x}$	estimate of x
$\mathbf{x}$	column vector
$\|\mathbf{x}\|$	Euclidean norm of $\mathbf{x}$
$\mathbf{X}$	matrix
$\mathbf{X}^*$	complex conjugate of $\mathbf{X}$
$\mathbf{X}^T$	transpose of $\mathbf{X}$
$\mathbf{X}^H = (\mathbf{X}^T)^*$	Hermitian transpose of $\mathbf{X}$
$\mathcal{X}$	set
$\|\mathcal{X}\|$	cardinality of $\mathcal{X}$
$\mathcal{X}\backslash\mathcal{Y}$	set of elements of $\mathcal{X}$ not contained in $\mathcal{Y}$

Commonly used symbols and operators

$\mathbf{A}_p$	mixing matrix of pth data set
blkdiag$(\mathbf{X}_1, \ldots, \mathbf{X}_n)$	block diagonal matrix with diagonal blocks $\mathbf{X}_1, \ldots, \mathbf{X}_n$
$\mathbb{C}$	set of complex numbers
$\mathbf{C}_{pq}$	coherence matrix of pth and qth data sets
$\mathbf{C}$	composite coherence matrix
χ^2_ν	chi-squared distribution with ν degrees of freedom
diag$(x_1, \ldots, x_n)$	diagonal matrix with diagonal elements $x_1, \ldots, x_n$
det$(\mathbf{X})$	determinant of $\mathbf{X}$
d_{i}	number of improper components in a complex-valued data set
d_{pq}	number of correlated components between pth and qth data sets

d_{all}	number of components correlated across all pairs of data sets
$\epsilon_p^{(i)}$	ith canonical variable of pth data set
$E[x]$	expectation of x
η	generalized likelihood ratio
H_0	null hypothesis
H_1	alternative hypothesis
$\mathbf{I}$	identity matrix
m_p	dimension of pth signal vector, $\mathbf{s}_p$
M	number of samples
n_p	dimension of pth data vector, $\mathbf{x}_p$
$\mathbf{n}_p$	noise vector in pth data set
P	number of data sets
P_{fa}	probability of false alarm
$\rho_{pq}^{(i)}$	correlation coefficient between ith signal components of pth and qth data sets
$\mathbb{R}$	set of real numbers
$\mathbf{R}_{pp}$	covariance matrix of pth data set
$\mathbf{R}_{pq}$	cross-covariance matrix of pth and qth data sets
$\hat{\mathbf{R}}_{pp}$	sample covariance matrix of pth data set
$\hat{\mathbf{R}}_{pq}$	sample cross-covariance matrix of pth and qth data sets
$\mathbf{R}$	composite covariance matrix
$\mathbf{R}_{s_p s_q}$	cross-covariance matrix of $\mathbf{s}_p$ and $\mathbf{s}_q$
$s_p^{(i)}$	ith signal component of pth data set
$\mathbf{s}_p$	signal component vector of pth data set with elements $s_p^{(i)}$
$\mathbf{x}_p$	pth data vector
${}_b x$	x associated with bth bootstrap resample
$(\mathbf{X})^{-\frac{1}{2}}$	square root inverse of $\mathbf{X}$
$\mathbf{1}$	column vector with all entries one
$\mathbf{0}$	column vector with all entries zero

List of figures

List of tables

Part I.

Introduction and background

1. Introduction

1.1. Motivation

There are numerous applications where analyzing and characterizing multivariate association between different sets of data is vital. One such application which is easy to understand is fusion of data from different sensors. The idea of acquiring and combining the data from different sensors is natural to living organisms. To fully interact with the environment, humans rely on complementary information from visual, audio, tactile and other stimuli, and a loss to process even one of them can significantly change their interaction. Similarly, different artificial sensors (also commonly called *modalities*) provide complementary data. For example, lidar, radar and visual cameras provide unique information about an object in a 3D-space [1]. In biomedicine, each brain imaging modality like functional magnetic resonance imaging (fMRI), electroencephalography (EEG), provides a unique way of measuring the brain activity. The fMRI measures the brain activity with a high spatial resolution while the EEG does it with a very high temporal resolution [2]. Another example is in the field of epilepsy. Various studies have shown that an epileptic seizure induces changes in the autonomic nervous system (ANS) of the brain [3], [4]. The ANS is a complex system controlling different organs of the body and is monitored using data from different modalities such as sweat activity, heart rate, respiratory rate and body temperature [5].

Although data from different modalities can be analyzed separately, as in the natural world, their joint analysis brings significant advantages. For example, joint analysis of fMRI and EEG data leads to understanding of brain activity at a highly resolved spatial and temporal scale [6]. Similarly, jointly analyzing data from different ANS modalities has shown to improve detection and prediction of an epileptic seizure [7], [8]. Other examples include improved object detection for autonomous driving [9], finding coupled patterns between sea surface temperature and rainfall in oceanography [10] and extracting gene clusters in

genomics [11]. However, data might not always be generated from different modalities. Another application where evaluating multivariate association produces benefits is multisensor data analysis from same modality. For instance, enhanced voice activity detection (VAD) from multiple microphones in wireless acoustic sensor network (WASN) [12], efficient object detection using multiple spatially separated cameras [13] and an improved analysis of brain activity in multi-subject fMRI data [14]. *In all these applications, an essential step for performing the joint analysis is to study how the data sets (obtained either from multiple modalities or from same modality) are statistically associated with each other, or in other words, do they share some common information which can be quantified and examined?*

There are various techniques which examine the association between multiple data sets. When the availability of data points, also called *observations* or *samples*, is not an issue, a possible solution is to employ completely data-driven approaches like [15], [16]. However, these approaches are difficult to interpret especially in applications where the ground truth is unknown. Semi data-driven approaches like canonical correlation analysis (CCA) [17], independent vector analysis (IVA) [18], factor analysis [19], which assume each data set to be generated by a simple linear mixing of underlying components, are well-suited in these scenarios. This is because these approaches limit the model search space to avoid overfitting and at the same time rely on minimal assumptions to let the heterogenous multiple data sets fully interact [20]. Nonetheless in many fields, the number of samples is limited due to various constraints. For instance, the number of participants in a typical biomedical study is in the order of ten to hundred. Similarly, in oceanography studies measuring yearly variability, the number of samples (time points) depends on when the data collection started since we cannot collect future data in advance. With limited data and large number of parameters to estimate, measuring higher-order dependencies among the underlying components in multiple data sets can lead to unreliable results. In this case, CCA, multiset CCA (mCCA) [21], and their variants [13], [22], [23] which work with second-order statistics are more suitable. These tools extract components from each data set, referred to as canonical variables, that are highly correlated across different data sets. The canonical variables can then be used in a plethora of applications including the ones mentioned above. Despite the abundance of interest in extracting correlated components, one key question often remains overlooked. *Is the estimated correlation of these components actually present in the underlying system, or is it an artifact due to noise or an insufficient number of samples?*

Sometimes the answer to this question is assumed to be known apriori from domain-specific knowledge. However, outside of that limited realm, most applications employ methods for

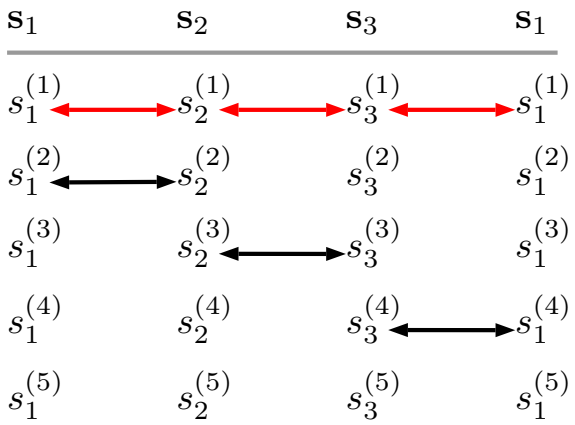


Figure 1.1.: Example of a correlation structure between the latent signal components of three data sets, $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$. The first component is correlated across all the data sets, the next three are correlated across two data sets only and the fifth component in each data set is uncorrelated with the other components.

thresholding the correlation coefficients as a way to determine the significant correlated components. These solutions are generally heuristic and often fail for one of two main reasons: 1) when the number of samples is limited, the correlation coefficient among the estimated components is overestimated, even to the point of identifying nonexistent, spurious correlations, 2) the number of possible correlation structures among the extracted components combinatorially increases as the number of data sets and their dimensions increase. It becomes even more challenging to set heuristic thresholds when the data sets are high-dimensional as the problems in 1) and 2) exacerbate.

Some techniques in the past have aimed to solve this as a *model-order selection* problem. In signal processing, model order is the term used for the dimension of a parameter vector, i.e., the number of parameters of the data model [24]. Thus, estimating the number of correlated components can be posed as a model-order selection problem. For two data sets, the signal components are either correlated or uncorrelated across both sets, and the model-order selection problem is well-defined. In this case, counting the number of correlated components is sufficient to completely characterize the linear association among the two sets. Some of the techniques for estimating the model order for two data sets are [25]–[29].

For more than two data sets, the model-order selection problem is not well defined. It is possible for the components to be correlated across no data sets, all data sets, or some subset of the collection. Figure 1.1 illustrates an example of correlation structure between the latent signal components of three data sets, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$. In most cases, however, we observe linear mixtures of these latent components instead of observing them directly. Nonetheless, Figure 1.1 can be used as a reference example to differentiate between the possible defini-

tions of model-order selection for multiple data sets. Each column of the figure indicates the components of one data set and thus components of each column have the same subscript. Each row represents the individual signal components that can be correlated between different pairs of data sets. In this example, each data set contains five signal components that are mutually uncorrelated within their set. The set $\mathbf{s}_1$ is repeated again in the last column to illustrate the correlation between the first and the third data set. Here, the first signal component of each data set is correlated with the first component of all other data sets. This type of correlation is indicated with red arrows. The next three components are correlated only between a pair of data sets as indicated with black arrows. The fifth component of each data set is uncorrelated with all other components.

In a special case, when the components are either correlated across all data sets or completely uncorrelated, the model order completely characterizes the correlation information in multiple data sets [30], [31]. In this case, it is assumed that the components indicated with black arrows in Figure 1.1 are not present. However, apart from this special case, many generalizations of model-order selection are valid (for example [13], [32]–[34]), and the problem must be more precisely defined. For example, one formulation as in [34] is to determine the number of components indicated with red arrows in Figure 1.1, while another as in [13] is to determine the number of components indicated with both the red and black arrows in Figure 1.1. Moreover, determining only the model order is insufficient for characterizing the complete second-order association in multiple data sets. This summary statistic only provides the knowledge that the components exhibit correlation. This knowledge, although sufficient for two data sets, is incomplete for multiple data sets.

Limited number of available samples presents another challenge here as the estimated correlation between the components is highly overestimated when the number of samples is not large compared to the dimensions of the data sets [13], [35]. This is commonly called as *sample-poor regime* or *small-sample support*. In this regime, traditional CCA and mCCA are unreliable for inferring the true linear relationship between the data sets, and a pre-processing step or some form of regularization must be applied [36], [37]. However, this pre-processing (or regularization) must be carefully designed to ensure that all or most of the correlation information is included in this step [28]. For two data sets, the issue of small-sample support is lately receiving attention [28], [29], [37]–[39]. However, in multiple data sets, this issue is scarcely addressed and that too either for model-order selection [13] or for estimating the latent signals themselves [40], but not yet for characterizing the complete correlation structure.

1.2. Contributions

The goal of this thesis is to characterize the complete linear association among multiple data sets with limited number of samples, and reap benefits in various real-world applications by employing this vital information. The association among different data sets is represented through correlation between the latent components of the data sets. Since there are different combinations in which the components can be correlated with each other, there are different ways to characterize the joint-correlation information among all the sets.

I first propose two techniques to estimate the model order identifying the joint-correlation information among all the data sets. This can be achieved by determining the number of components correlated across all data sets. In case of Figure 1.1, this corresponds to determining the number of components indicated with red arrows. The first technique assumes that the components are either correlated across all data sets or uncorrelated. In this case, the model order completely characterizes the linear association in multiple data sets. To this end, a generalized likelihood ratio test (GLRT)-based technique is designed specifically for the sample-poor regime. It employs principal component analysis (PCA)-based dimensionality reduction. The reduced PCA dimensions and the model-order are jointly estimated so that the estimated PCA rank retains all the correlated components in each data set and at the same time discards uncorrelated and noise components having smaller variance than correlated components.

However, assuming an apriori correlation structure restricts the applicability of the previous technique in many applications such as biomedicine, where the data sets are heterogenous and can contain components correlated with arbitrary correlation structure. In this case, the model order provides an incomplete summary as it only identifies that components are correlated. In order to achieve the goal of completely characterizing the joint-linear association, the complete correlation structure has to be determined, i.e., which components are correlated and across which data sets. For the example in Figure 1.1, the complete solution is not just determining that the first four components are correlated but also that the first component in each data set is correlated, and the successive components are correlated between data sets 1 and 2, 2 and 3, and 1 and 3, respectively. I then formulate and solve a more general *model selection problem* by utilizing the joint information in the composite coherence (whitened covariance) matrix of all data sets [41], denoted by $\mathbf{C}$. A one-to-one relationship between the number of eigenvalues of $\mathbf{C}$ greater than one and the number of correlated components, and the between the eigenvectors of $\mathbf{C}$ and correlation structure of the correlated components is theoretically established using tools from graph theory. To deal with the small-sample sup-

port, a simple extension of the technique is proposed using PCA, where the PCA dimensions are chosen to keep all correlated components.

I would like to point out that in this thesis I estimate the correlation structure among the latent signals in the data sets, and not the latent signals themselves. There exist various joint blind source separation (JBSS) techniques such as mCCA, IVA, joint independent component analysis (ICA) [18], [20], [23], [42], [43], which estimate the underlying correlated (or common) signals. However, it is not always necessary to estimate the correlated signals first (or to estimate them at all) and later infer their correlation structure. I will show in the upcoming chapters that under certain assumptions, there is a one-to-one relationship between certain properties of the data matrices (e.g., the multiset canonical correlations in the GLRT-based technique in Chapter 5 or the eigenvalues and eigenvectors of the composite coherence matrix in Chapter 6) and the correlation structure among the underlying components. Therefore, it is desirable and efficient to avoid the challenges and ambiguities involved in estimating the latent signals if they are not required to characterize the correlation structure.

Throughout this thesis, extensive simulation examples are provided to demonstrate the effectiveness of the proposed techniques over the competing techniques present in the literature. However, the results are not limited to just simulations. Since the association between multiple data sets is inherent in several fields, these techniques are applied in diverse applications of sensory array processing, WASN, sports science and epilepsy, and benefits of quantifying and utilizing this association are presented. Applying the proposed techniques to real-world applications was mostly done in joint collaboration with other researchers and I have tried my best to acknowledge and specify their valuable contributions in the respective chapters. The application-specific contributions of this thesis are the following.

First, two techniques based on information theoretic criterion (ITC) and GLRT in two data sets are adapted for determining the dimension of the improper subspace in complex-valued data. Improper signals are a class of complex-valued signals which are correlated with their complex conjugate and are useful in numerous applications such as communications, oceanography and biomedicine [44], [45]. The proposed techniques effectively exploit the correlation information between the data and its complex conjugate and thus, are able to work even in presence of additive colored noise, which the competing technique [46] is not able to handle. I also present numerical results for the application of sensor array processing, where the number of improper sources, e.g., binary phase-shift keying(BPSK)-modulated sources, in a sensor array with large number of array elements and small number of samples

is determined.

I have also applied the developed technique for estimating the complete correlation structure in WASN to determine the unknown number of active speakers and the cluster of nodes which dominantly hear each speaker. This is done blindly by only observing the mixture of signals from multiple speakers. Both these estimates are used to detect the voice activity pattern of each speaker efficiently. Only the nodes assigned to the cluster of the corresponding speaker are used to detect the voice activity, thus providing a better SNR and an improved performance compared to the state-of-the-art [47], which uses all microphones at a time.

In the field of sports science, I apply the existing techniques for two data sets and the developed techniques for multiple data sets to quantify the changes in association among different ANS modalities in subjects undergoing two different physical tasks. In the first study, the subjects underwent through an intense exercise, and in the other, they ran an ultramarathon. Supported by the fact the the correlation structure of the extracted components and their correlation strength changes after each physical load, the analysis helps in a better understanding of the ANS central control and its subsystems.

Finally, the techniques are applied to the peripheral ANS data collected from a wearable sensor in epileptic patients and the association between the modalities is analyzed before and after a seizure. The number of correlated components, their strength and the modalities across which they are correlated increase right before the seizure and decrease right after the seizure, offering a possibility of a potential biomarker for seizure detection and more importantly seizure prediction.

1.3. Overview of thesis

This thesis is divided into four parts. We will continue the first part of the thesis by discussing the tools for analyzing linear dependencies between multiple random vectors and the model-selection techniques in Chapter 2. Specifically, we will discuss CCA, mCCA, and three different model-selection techniques of ITC, GLRT and bootstrap, all of which are most relevant for the techniques proposed in this thesis.

In the next part, we will focus on model selection in two data sets. With the goal of making this thesis as complete as possible for the reader, two existing techniques based on ITC and GLRT for estimating the model-order jointly in two data sets along with their modifications

in the sample-poor regime are reviewed in Chapter 3. We will frequently refer to these techniques in the later chapters. In Chapter 4, two novel methods to estimate the number of improper signals in complex-valued data are proposed. These methods are adaptations of the techniques described in Chapter 3 and employ substantial modifications due to the fact that the second-order dependence between a random vector and its complex conjugate is analyzed. The proposed methods are designed specifically for data with small-sample support and corrupted by white or colored noise. An example of determining the number of sources impinging on a sensor array with large number of array elements is also presented.

Part 3 of this thesis focuses on characterizing the linear association among multiple (more than two) data sets. Chapter 5 aims to characterize this association through the model order identifying the number of components correlated across all data sets. In the first half of this chapter, it is assumed that the components are either correlated across all pairs of data sets or are completely uncorrelated. A GLRT-based detector for the model order is designed in Section 5.3 and is extended for data sets with relatively small number of samples. The second half of Chapter 5 proposes a new technique based on bootstrap for determining the model order with arbitrary correlation structure. In this case, the model order only characterizes the linear dependencies present across all the data sets and ignores the dependencies across subsets of them. This is tackled in the next chapter of this part, which introduces a more general model selection problem and includes model-order selection as its subproblem. Two techniques for estimating the complete correlation structure are proposed in Chapter 6. These techniques complement each other. The first technique is based on model order estimates from pairs of data sets while the second technique uses only the joint information from all the data sets.

The next part focusses on the applications of the proposed techniques on real-world data from three different fields. In Chapter 7, the joint approach for estimating the complete correlation structure of Chapter 6 is applied in WASN to estimate the number of speakers and their cluster of nodes. This vital information is then used to detect the voice activity of each speaker using a group-sparse constraint, which offers a robust solution against impulsive noise sources in the network. Chapter 8 applies the techniques reviewed in Chapter 3 and developed in Chapter 6 in the field of sports science to measure the changes in association between multiple ANS modalities in response to physical stressors. The adapted methods, results and their implication for a likely reorganization of the ANS after the physical load are further discussed. Later, the focus of Chapter 9 is in the field of epilepsy. The time evolution of the overall linear dependency estimated among four different modalities of ANS is analyzed around an epileptic seizure. The plausibility and challenges of employing the results

as a promising biomarker for seizure detection and prediction are also briefly discussed. We will thereafter conclude the thesis and discuss some potential ideas for future research in Chapter 10.

2. Background

In this chapter, we will introduce the fundamental methods which are used throughout in this thesis. We will start by describing CCA, which measures the linear dependence among two different sets of data in Section 2.1. In Section 2.2, we will introduce the extension of CCA for more than two data sets, commonly referred to as mCCA. Later in Section 2.3, we will discuss about the two commonly used model selection techniques: ITC and hypothesis testing. Two variants of hypothesis testing, the GLRT and the bootstrap, are also discussed.

2.1. Canonical correlation analysis

Let us consider two zero-mean random variables x_1 and x_2. The most commonly used statistical quantity to measure the linear relationship between these two variables is the Pearson correlation coefficient [48] defined as

$$k = \frac{E[x_1 x_2]}{\sqrt{E[x_1^2]E[x_2^2]}}, \tag{2.1}$$

where $E[\cdot]$ denotes the expectation operator. Now let us consider two zero-mean random vectors $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_2}$. How do we describe the linear dependence among $\mathbf{x}_1$ and $\mathbf{x}_2$? One way is to compute and infer the $n_1 \times n_2$ correlation coefficients between the variables of $\mathbf{x}_1$ and $\mathbf{x}_2$. However, this might not lead to the true linear relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$ as some of the variables within $\mathbf{x}_1$ and $\mathbf{x}_2$ might be correlated with each other. CCA is a powerful tool for finding the *canonical* linear relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$ [17]. CCA projects $\mathbf{x}_1$ and $\mathbf{x}_2$ using a linear transformation and aims to maximize the correlation between the projections [48]. It thus seeks to find the coordinate system where most of the correlation between the two random vectors is concentrated in a few dimensions.

Let us define the covariance matrices of $\mathbf{x}_1$ and $\mathbf{x}_2$ as $\mathbf{R}_{11} = E[\mathbf{x}_1\mathbf{x}_1^T]$, and $\mathbf{R}_{22} = E[\mathbf{x}_2\mathbf{x}_2^T]$, where the superscript T denotes the transpose. Similarly, the cross-covariance matrix is defined as $\mathbf{R}_{12} = E[\mathbf{x}_1\mathbf{x}_2^T]$. CCA seeks vectors $\mathbf{w}_1^{(1)}$ and $\mathbf{w}_2^{(1)}$, such that the correlation coefficient between the projections $\epsilon_1^{(1)} = \mathbf{w}_1^{(1)T}\mathbf{x}_1$ and $\epsilon_2^{(1)} = \mathbf{w}_2^{(2)T}\mathbf{x}_2$ is maximized [49]. That is,

$$\underset{\mathbf{w}_1^{(1)},\mathbf{w}_2^{(1)}}{\arg\max} \frac{\mathbf{w}_1^{(1)T}\mathbf{R}_{12}\mathbf{w}_2^{(1)}}{\sqrt{\mathbf{w}_1^{(1)T}\mathbf{R}_{11}\mathbf{w}_1^{(1)}}\sqrt{\mathbf{w}_2^{(1)T}\mathbf{R}_{22}\mathbf{w}_2^{(1)}}}. \tag{2.2}$$

Since the correlation coefficient is scale-invariant, any scaled versions of $\epsilon_1^{(1)}$ and $\epsilon_2^{(1)}$, will lead to the same correlation coefficient between them. Therefore, we constrain $\epsilon_1^{(1)}$ and $\epsilon_2^{(1)}$ to be of unit-variance, i.e., $\mathbf{w}_1^{(1)T}\mathbf{R}_{11}\mathbf{w}_1^{(1)} = \mathbf{w}_2^{(1)T}\mathbf{R}_{22}\mathbf{w}_2^{(1)} = 1$. Assuming that that $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are non-singular and defining new projection vectors, $\widetilde{\mathbf{w}}_1^{(1)} = \mathbf{R}_{11}^{\frac{1}{2}}\mathbf{w}_1^{(1)}$ and $\widetilde{\mathbf{w}}_2^{(1)} = \mathbf{R}_{22}^{\frac{1}{2}}\mathbf{w}_2^{(1)}$, the CCA cost function in (2.2) can be redefined as [38]

$$\begin{aligned} &\underset{\widetilde{\mathbf{w}}_1^{(1)},\widetilde{\mathbf{w}}_2^{(1)}}{\arg\max}\, \widetilde{\mathbf{w}}_1^{(1)T}\mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{R}_{12}\mathbf{R}_{22}^{-\frac{1}{2}}\widetilde{\mathbf{w}}_2^{(1)}, \\ &\text{such that } \widetilde{\mathbf{w}}_1^{(1)T}\widetilde{\mathbf{w}}_1^{(1)} = \widetilde{\mathbf{w}}_2^{(1)T}\widetilde{\mathbf{w}}_2^{(1)} = 1. \end{aligned} \tag{2.3}$$

The correlation coefficient, $k^{(1)} = E[\epsilon_1^{(1)}\epsilon_2^{(1)}]$ is called the first canonical correlation, and $\epsilon_1^{(1)}$ and $\epsilon_2^{(1)}$ are called the first pair of canonical variables. To obtain the next pair of canonical variables, the same process of maximizing the correlation coefficient with respect to new projection vectors $\mathbf{w}_1^{(2)}$ and $\mathbf{w}_2^{(2)}$ is followed, subject to additional constraints that the canonical variables belonging to the same random vector are uncorrelated [49]. This procedure can be repeated to compute $n_{12} = \min(n_1, n_2)$ pairs of canonical variables.

Closed Form Solution: The CCA procedure can be visualized as transforming $\mathbf{x}_1$ and $\mathbf{x}_2$ to n_{12}-dimensional random vectors $\boldsymbol{\epsilon}_1 = \mathbf{W}_1\mathbf{x}_1$ and $\boldsymbol{\epsilon}_2 = \mathbf{W}_2\mathbf{x}_2$, as shown in Figure 2.1, where the projection matrices $\mathbf{W}_1 = [\mathbf{w}_1^{(1)}, \ldots, \mathbf{w}_1^{(n_{12})}]^T$ and $\mathbf{W}_2 = [\mathbf{w}_2^{(1)}, \ldots, \mathbf{w}_2^{(n_{12})}]^T$, contain the n_{12} projection vectors [50]. The random vectors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are called canonical vectors and are constrained to be white, i.e.,

$$E[\boldsymbol{\epsilon}_1\boldsymbol{\epsilon}_1^T] = E[\boldsymbol{\epsilon}_2\boldsymbol{\epsilon}_2^T] = \mathbf{I}. \tag{2.4}$$

Let us define the coherence matrix (or the whitened covariance matrix) $\mathbf{C}_{12}$ [41] and its singular value decomposition (SVD) as

$$\mathbf{C}_{12} = \mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{R}_{12}\mathbf{R}_{22}^{-\frac{1}{2}} = \mathbf{F}_1\mathbf{K}_{12}\mathbf{F}_2^T. \tag{2.5}$$

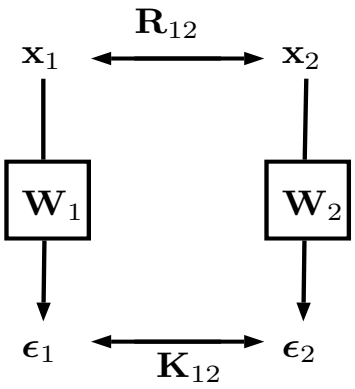


Figure 2.1.: Principle of CCA.

Here $\mathbf{F}_1$ and $\mathbf{F}_2$ are the left and right singular vector matrices of $\mathbf{C}_{12}$. The complete solution to the CCA optimization problem is given by the SVD of $\mathbf{C}_{12}$ with the projection matrices as

$$\mathbf{W}_1 = \mathbf{F}_1^T \mathbf{R}_{11}^{-\frac{1}{2}} \tag{2.6}$$

$$\mathbf{W}_2 = \mathbf{F}_2^T \mathbf{R}_{22}^{-\frac{1}{2}}, \tag{2.7}$$

and the canonical correlations as the singular values of $\mathbf{C}_{12}$ [41], i.e., the canonical correlation matrix

$$\mathbf{K}_{12} = E[\boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_2^T] = \text{diag}(k_{12}^{(1)}, \cdots, k_{12}^{(n_{12})}). \tag{2.8}$$

It is interesting to see that the expression of $\mathbf{C}_{12}$ for random vectors in (2.5) is similar to the expression of the Pearson correlation coefficient for random variables in (2.1).

An important property of CCA is that the canonical correlations are maximally invariant under non-singular linear transformations of $\mathbf{x}_1$ and $\mathbf{x}_2$ [48]. Therefore, $\mathbf{T}_1\mathbf{x}_1$ and $\mathbf{T}_2\mathbf{x}_2$ will also yield the same canonical correlation matrix $\mathbf{K}_{12}$, where $\mathbf{T}_1 \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{T}_2 \in \mathbb{R}^{n_2 \times n_2}$ are assumed to be of full rank. This property makes CCA applicable in numerous real-world applications where the relationship among two sets of underlying latent variables, observed in the measurement space through linear mixing is typically of interest.

2.2. Multiset canonical correlation analysis

CCA is limited to two random vectors only. There are several ways of extending it to more than two random vectors which are summarized in [21]. All of these extensions fall under the common term of *multiset CCA*. To understand why generalizing CCA for

multiple random vectors is not unique, let us now consider P zero-mean random vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_P$. As in CCA, we project these P vectors on to the projections, $\epsilon_p^{(1)} = \mathbf{w}_p^{(1)T}\mathbf{x}_p$, for $p = 1, \ldots, P$ and aim to jointly maximize the correlation among these P projections [21]. However, unlike CCA, we now have $\frac{P(P-1)}{2}$ correlation coefficients among the pair $\{\epsilon_p^{(1)}, \epsilon_q^{(1)}\}$, $p, q = 1, \ldots, P$, $p \neq q$, to be maximized. Since there is no unique way of jointly maximizing all of the $\frac{P(P-1)}{2}$ correlation coefficients, different versions of mCCA exist each based on a different cost function of these correlation coefficients [51]. Let us constrain each $\epsilon_p^{(1)}$ to be of unit variance and define their covariance matrix as

$$\widetilde{\mathbf{R}}^{(1)} = \begin{bmatrix} 1 & k_{12}^{(1)} & \cdots & k_{1P}^{(1)} \\ k_{12}^{(1)} & 1 & \cdots & k_{2P}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1P}^{(1)} & k_{2P}^{(1)} & \cdots & 1 \end{bmatrix}, \tag{2.9}$$

where $k_{pq}^{(1)}$ is the correlation coefficient of $\epsilon_p^{(1)}$ and $\epsilon_q^{(1)}$. [21] presented five different versions to perform mCCA. These versions are as follows:

1. Maximum variance (MAXVAR),
2. Minimum variance (MINVAR),
3. Generalized variance (GENVAR),
4. Sum of correlations (SUMCORR)
5. Sum of squared correlations (SSQCORR).

To obtain the first set of canonical variables, MAXVAR maximizes the largest eigenvalue of $\widetilde{\mathbf{R}}^{(1)}$, MINVAR minimizes the smallest eigenvalue of $\widetilde{\mathbf{R}}^{(1)}$ and GENVAR minimizes the determinant of $\widetilde{\mathbf{R}}^{(1)}$. On the other hand, SUMCORR and SSQCORR maximize the sum of absolute values and the sum of squared values of the $\frac{P(P-1)}{2}$ correlation coefficients in $\widetilde{\mathbf{R}}^{(1)}$, respectively. A useful property of these different mCCA versions is that all five of them reduce to CCA for $P = 2$. In the remaining part of this section, we will briefly explain the MAXVAR version and show how it extends the traditional CCA presented in Section 2.1 in a natural way. For a complete overview of all the cost functions, the reader is referred to [21] and [51].

Consider the composite vector $\mathbf{x}$ obtained by vertically concatenating the individual random vectors,

$$\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T, \tag{2.10}$$

and the composite covariance matrix $\mathbf{R} = E[\mathbf{x}\mathbf{x}^T]$. Let $\mathbf{R}_D = \text{blkdiag}(\mathbf{R}_{11}, \ldots, \mathbf{R}_{PP})$ be a block-diagonal matrix with $\mathbf{R}_{pp} = E[\mathbf{x}_p\mathbf{x}_p^T]$. As in CCA, let us define the new projection vectors $\widetilde{\mathbf{w}}_p^{(1)} = \mathbf{R}_{pp}^{\frac{1}{2}}\mathbf{w}_p^{(1)}$ and the composite projection vector, $\widetilde{\mathbf{w}}^{(1)} = [\widetilde{\mathbf{w}}_1^{(1)T}, \ldots, \widetilde{\mathbf{w}}_P^{(1)T}]$. The optimization problem of mCCA MAXVAR to obtain the first set of canonical variables can now be formulated as

$$\begin{aligned} &\arg\max_{\widetilde{\mathbf{w}}^{(1)}} \widetilde{\mathbf{w}}^{(1)T}\mathbf{R}_D^{-\frac{1}{2}}\mathbf{R}\mathbf{R}_D^{-\frac{1}{2}}\widetilde{\mathbf{w}}^{(1)}, \\ &\text{such that } \widetilde{\mathbf{w}}^{(1)T}\widetilde{\mathbf{w}}^{(1)} = 1. \end{aligned} \tag{2.11}$$

The definition of the coherence matrix for two random vectors can be generalized in a natural way for multiple random vectors as

$$\mathbf{C} = \mathbf{R}_D^{-\frac{1}{2}}\mathbf{R}\mathbf{R}_D^{-\frac{1}{2}}, \tag{2.12}$$

where $\mathbf{C}$ denotes the composite coherence matrix. Rewriting this problem as a Lagrangian makes it clear that (2.11) is maximized when $\widetilde{\mathbf{w}}^{(1)}$ is the eigenvector associated with the largest eigenvalue of $\mathbf{C}$. The canonical variables are obtained as $\epsilon_p^{(1)} = \widetilde{\mathbf{w}}_p^{(1)T}\mathbf{R}_{pp}^{-\frac{1}{2}}\mathbf{x}_p$. We refer to them as the first stage canonical variables. Solving (2.11) is equivalent to maximizing the largest eigenvalue of $\widetilde{\mathbf{R}}^{(1)}$ in (2.9) under the constrains that $\epsilon_p^{(1)}$ are of unit variance.

For successive stages, as in CCA, the canonical variables from within a particular data set are constrained to be uncorrelated, i.e.,

$$E[\epsilon_p^{(j)}\epsilon_p^{(k)}] = 0 \tag{2.13}$$

for $p = 1, \ldots, P$ and for all $k < j$. This is enforced with a deflationary procedure where $\mathbf{C}$ is recomputed at stage j after projecting each data set onto the orthogonal complement of its existing canonical variables, $[\epsilon_p^{(1)}, \ldots, \epsilon_p^{(j-1)}]$. The optimization problem (2.11) with constraints (2.13) is optimized when the jth stage projection vector, $\widetilde{\mathbf{w}}^{(j)}$ is the dominant eigenvector of this updated $\mathbf{C}$.

We can observe that the cost function for CCA in (2.3) and for MAXVAR mCCA in (2.11) are very similar. The solution of CCA is obtained via the SVD of the coherence matrix $\mathbf{C}_{12}$, while for MAXVAR mCCA, the solution is obtained via the eigenvalue decomposition (EVD) of the composite coherence matrix $\mathbf{C}$. This is why MAXVAR mCCA is commonly referred as the natural extension of CCA for multiple random vectors.

2.3. Model selection techniques

Model selection is one of the most fundamental problem in statistical signal processing. For a given data, how to select a model that best fits the data? One of the most commonly used example in this context is that of curve fitting. Given a set of data points, how to select a curve that best describes the underlying function that generated those points? For M number of points, one can always chose an $(M-1)$th order polynomial function which fits all the points perfectly. However, such a complex model is usually not accurate as it overfits to the noise present in the system and does not generalize well for an unseen set of points generated from the same underlying function. On the other hand, a very simple model (very low-order polynomial function) might underfit and would also not generalize well for the unseen data. A preferred model is the one which balances the goodness-of-fit with the model complexity [52]. There are different techniques in the literature for model selection. We will focus on the two statistical techniques: the information theoretic criterion (ITC) and the hypothesis testing.

2.3.1. Information theoretic criterion

The information theoretic criterion introduced by Akaike [53], Schwartz [54] and Rissanen [55] states that, given a set of observations and a family of models, select the model that best fits the observation data, while also making sure that the model is not too complex. The principle of all the criteria is to compute an ITC score, which is the sum of model fit score and penalty function [24],

$$\text{ITC} = \text{model fit score} + \text{penalty function}. \tag{2.14}$$

The first term in (2.14), the model fit score, measures the goodness-of-fit of the observation data to the model. The second term, the penalty function, is dependent on the number of free parameters in the parameter space of the model. In general, with an increasing number of free parameters or the model order, the observation data better fits the model and the model fit score decreases. This is shown by the blue curve in Figure 2.2. However, if the model order is increased without any check, the model tends to overfit the data. Overfitting occurs when a model has too many parameters than required. In this case, the model tends to follow the random noise in the data rather than the underlying relationship. Hence, to penalize overfitting, the penalty function always increases with an increase in the model order as

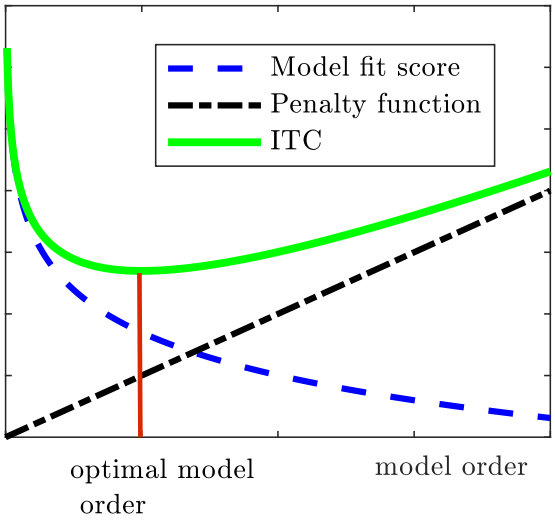


Figure 2.2.: Principle of information theoretic criterion.

shown by the black line in Figure 2.2. The ITC score is shown using the green line. The optimal model parameter is chosen as the one which minimizes the ITC score.

Let us now give the terms defined in (2.14) a probabilistic interpretation. Consider M independent and identically distributed (i.i.d.) samples of a random vector $\mathbf{x}$, which are stacked as columns of the data matrix $\mathbf{X}$. The M samples of $\mathbf{x}$ are assumed to be generated from one of the family of probability densities $f(\mathbf{X}|\mathbf{\Theta}_d)$ parameterized by the parameter space $\mathbf{\Theta}_d$ for some real-valued parameter d which can take values from 1 to n. In signal processing, d is commonly referred to as the model order [24]. The general expression of ITC in this case is

$$\text{ITC}(d) = -\ln f(\mathbf{X}|\hat{\mathbf{\Theta}}_d) + \alpha(M)C(d), \tag{2.15}$$

where $\hat{\mathbf{\Theta}}_d$ is the maximum likelihood estimate of $\mathbf{\Theta}_d$, $C(d)$ is the number of free parameters in $\mathbf{\Theta}_d$ and the term $\alpha(M)$ depends on the chosen ITC. For Akaike information criterion (AIC), $\alpha(M) = 1$, for minimum description length (MDL) and Bayesian information criterion (BIC), $\alpha(M) = \frac{\ln(M)}{2}$. The estimate of the parameter d or the model $f(\mathbf{\Theta}_d|\mathbf{X})$ which best fits the data $\mathbf{X}$, is the one that minimizes (2.15),

$$\hat{d} = \underset{d=1,\ldots,n}{\arg\min}\, \text{ITC}(d). \tag{2.16}$$

2.3.2. Hypothesis testing

Hypothesis testing forms the backbone of statistical detection theory and can also be employed for model selection as an alternative to the ITC. Given a set of models to choose from, one way to select the model that best fits the data is using multiple hypothesis testing where each hypothesis supports one model. Thus, for n possible models, n hypotheses $H_1, \ldots, H_n$ are defined as

$$\begin{aligned} H_1 &: \text{Model-1} \\ H_2 &: \text{Model-2} \\ &\vdots \\ H_n &: \text{Model-}n \end{aligned} \tag{2.17}$$

However, when the hypotheses are nested, i.e., when a model corresponding to one hypothesis is a special case of the model in the other hypothesis, and the model parameter is a scalar, a sequence of binary hypothesis tests can be applied instead. Although this is suboptimal compared to the multiple hypothesis tests in (2.17), it prevents the need of strict corrections for controlling the family-wise error rate and false discovery rate required in multiple hypothesis testing [56], [57].

Let us revisit the problem defined in Section 2.3.1, where from a number of models $f(\boldsymbol{\Theta}_d|\mathbf{X})$ parameterized by $\boldsymbol{\Theta}_d$, we need to select one of the model, or in other words, choose the appropriate model order d from the range, $1, \ldots, n$. In this case, a sequence of binary hypothesis tests can be performed one at a time until a stopping condition is met [27], [28], [58]. This means starting with a counter $i = 1$ and performing the following binary test of null hypothesis H_0 and alternative H_1

$$\begin{aligned} H_0 &: d = i, \\ H_1 &: d > i. \end{aligned} \tag{2.18}$$

If H_0 is rejected, i is incremented and another test of H_0 vs. H_1 is run. This is repeated until H_0 is not rejected or i reaches $n - 1$. The binary test in (2.18) requires a statistic whose (asymptotic) distribution under H_0 is known. We will now explain two methods: the generalized likelihood ratio test (GLRT), where the distribution of the statistic under H_0 is theoretically derived, and the bootstrap, where the distribution is empirically estimated from the data.

2.3.2.1. Generalized likelihood ratio test

When H_0 and H_1 are simple, and the likelihood function under both the hypotheses is known, the likelihood ratio test (LRT, also known as Neyman-Pearson detector) is the optimal detector. This means that the LRT maximizes the probability of detection for a given value of probability of false alarm (P_{fa}). However, in (2.18), H_0 is simple and H_1 is composite. Moreover, in most applications, the likelihood functions under H_0 and H_1 are not completely known and depend on the unknown parameters. In this case, the GLRT, where the unknown parameters are replaced by their maximum likelihood estimates, is the most commonly used detector. The generalized likelihood ratio (GLR) for the test in (2.18) is

$$\eta = \frac{f(\mathbf{X}|\hat{\boldsymbol{\Theta}}_i, d=i)}{f(\mathbf{X}|\hat{\boldsymbol{\Theta}}_{i+}, d>i)}, \tag{2.19}$$

where $f(\mathbf{X}|\hat{\boldsymbol{\Theta}}_i, d=i)$ and $f(\mathbf{X}|\hat{\boldsymbol{\Theta}}_{i+}, d>i)$ are the maximum likelihood functions under H_0 and H_1, respectively, $\hat{\boldsymbol{\Theta}}_i$ is the ML estimate of $\boldsymbol{\Theta}_i$ assuming H_0 is true and $\hat{\boldsymbol{\Theta}}_{i+}$ is the ML estimate of $\boldsymbol{\Theta}_{i+}$ assuming H_1 is true.

Wilks' theorem: An important result which makes GLRT applicable in various scenarios is the Wilks' theorem [59]. It states that under some mild conditions, the statistic $T(i) = -2\ln\eta$ under H_0 is asymptotically (as $M \to \infty$) χ^2_ν-distributed when $d = i$. Here, ν is the number of degrees of freedom (d.f.) and is equal to the difference between the number of free parameters under H_1 and H_0

$$\nu = C_{H_1}(i) - C_{H_0}(i). \tag{2.20}$$

If ν is independent of the unknown parameters, the distribution of $T(i)$ can be computed and a threshold $\tau(i)$ can be set to maintain a certain P_{fa}. This type of detector is also referred to as the constant false-alarm rate (CFAR) detector since P_{fa} is independent of the unknown parameters. However, it is important to note that P_{fa} is guaranteed (as $M \to \infty$) only for the given test of H_0 and H_1, and not for the entire sequence of binary tests. The model order d (and the corresponding model $f(\boldsymbol{\Theta}_d|\mathbf{X})$) can be selected as the smallest value of i for which the null hypothesis is not rejected, i.e.,

$$\hat{d} = \min_{i=1,\ldots,n-1} \{i : T(i) < \tau(i)\}. \tag{2.21}$$

Relationship between ITC and GLRT for model selection: From (2.15), (2.19) and (2.20), it can be seen that the GLRT and ITC are related to each other. Both of them are based on the

maximum likelihood functions and require the knowledge of the number of free parameters. This has also been shown in [60]. However, there is a distinct difference between the two approaches. The performance of the GLRT-based method depends on the probability of false alarm, P_{fa}. In general, if the number of samples is large enough, the detector with smaller P_{fa} performs better than the detector with larger P_{fa}. On the other hand, if the number of samples is small, the detector with larger P_{fa} performs better. This is a general rule because a detector with larger P_{fa} tends to overfit, whereas a detector with smaller P_{fa} tends to underfit. However, the best choice of P_{fa} for a given number of samples cannot be determined in general as it depends on the particular scenario. The advantage of the ITC-based method is that it does not require choosing a value for P_{fa} because it does the trade-off between underfitting and overfitting automatically. Nevertheless, this does not mean that the ITC-based method will always outperform the GLRT-based detector in every scenario. We will further discuss about the choice of P_{fa} and compare the performance of ITC and GLRT-based detectors when we present the numerical results in Sections 3.6 and 4.5.

2.3.2.2. Bootstrap

Both ITC and GLRT-based methods require the likelihood function to be maximized with respect to the unknown parameters in the parameter space $\boldsymbol{\Theta}_d$ as well as the number of free parameters in $\boldsymbol{\Theta}_d$ (or the degrees of freedom for the χ^2_ν-distribution). This is not always possible. In many cases, the dependence of the likelihood function on the unknown parameters might not be explicit, making it difficult to find the maximum likelihood function [33]. In other cases, determining the number of free parameters which depend on d might be challenging for complex models or the distribution of the data could be unknown [61]. In all of these scenarios, ITC and GLRT-based methods cannot be employed.

Bootstrap is a resampling technique which can be used for model selection under the framework of hypothesis testing. In bootstrap, it is assumed that the sample distribution of the data can be used as a reliable estimate of the true distribution [62]. Under this assumption, the samples are uniformly drawn from the available sample set with replacement to generate a bootstrap resample set of the same size as that of the original sample set. This process is repeated to generate many bootstrap resamples. These resamples are then employed for estimating the distribution of a parameter, obtaining confidence intervals, and performing hypothesis testing [63]. We use bootstrap to estimate the unknown distribution of the statistic under the null hypothesis. The advantage of bootstrap is that the true distribution of the data does not have to be known or estimated. This makes bootstrap a useful tool when only

a limited number of samples are available or for non-Gaussian distributed data. For these two scenarios, the traditional ITC and GLRT-based hypothesis test are ill-suited since they are predominantly based on asymptotic properties of Gaussian distributed data [61].

Let us revisit the binary hypothesis test in (2.18) and define the statistic $T(i)$ with a known value under the null hypothesis as $T_0(i)$. The test in (2.18) can be reformulated as

$$\begin{aligned} H_0 &: T(i) = T_0(i) \\ H_1 &: T(i) \neq T_0(i). \end{aligned} \tag{2.22}$$

To test whether our sample was generated under H_0, we estimate the distribution of $|T(i) - T_0(i)|$. This distribution is estimated via the bootstrap as follows. Given the sample matrix, compute $T(i)$. Resample the data by randomly choosing M indices from $\{1, \ldots, M\}$ (with uniform distribution and with replacement) to create a bootstrap data of the same size as the original data set. Repeat the resampling procedure B times and compute the test statistic each time to produce ${}_bT(i)$ for $b = 1, \ldots, B$. The distribution of $|T(i) - T_0(i)|$ under the null is then approximated by the bootstrap distribution ${}_bT^*(i) = |{}_bT(i) - T(i)|$ [64]. The algorithm for testing (2.22) is described in Algorithm 1 which is inspired from [61]. The model order d is estimated as the smallest value of i for which the null hypothesis is not rejected, i.e.,

$$\hat{d} = \min_{i=1,\ldots,n-1} \{i : |T(i) - T_0(i)| < T_\tau(i)\}, \tag{2.23}$$

where $T_\tau(i)$ is the threshold to maintain the desired P_{fa}.

2.4. Summary

We provided a brief overview about the principle, optimization problem and solution for both CCA and mCCA. Unlike CCA, the solution for mCCA is not unique and can be found using different cost functions of the correlation coefficients between the multiset canonical variables. Three different model selection techniques based on ITC, GLRT and bootstrap were also presented. Both ITC and GLRT techniques require the expressions for the maximum likelihood function and the degrees of freedom in the model, and the asymptotic properties of detectors based on them are well defined. However, when the distributional assumptions do not hold well, bootstrap provides an alternative by empirically estimating the distribution of the test statistic to perform the hypothesis test. Specific details about how these model selection techniques are applied for joint correlation analysis in two and more data sets are

Algorithm 1 Binary hypothesis testing of (2.22) using bootstrap

Input $\mathbf{X}$: observations
B: number of bootstrap resamples
P_{fa}: probability of false alarm
$T_0(i)$: value of the statistic under H_0
function BOOTSTRAPTEST($\mathbf{X}, B, P_{\text{fa}}$)
  $T(i) \leftarrow g(\mathbf{X})$
  ▷ Compute test statistic from the data using a known function $g(\cdot)$
  for $b = 1, \ldots, B$ **do**
  ▷ bootstrap resamples indexed by left subscript
    for $l = 1, \ldots, M$ **do**
      ${}_bj_l \leftarrow$ random integer $[1, M]$
  ▷ resample indices chosen with replacement
    ${}_b\mathbf{X} \leftarrow \left[\mathbf{x}({}_bj_1), \ldots, \mathbf{x}({}_bj_M)\right]$
  ▷ Compute the bth resample
    ${}_bT(i) \leftarrow g({}_b\mathbf{X})$
  ▷ Compute bootstrap test statistic from the resampled data
    ${}_bT^*(i) \leftarrow |{}_bT(i) - T(i)|$
  ▷ estimate distribution of $|T(i) - T_0(i)|$ under H_0
  $\{{}_{(l)}T^*(i)\}_{l=1}^B \leftarrow \text{sort}\{{}_bT^*(i)\}_{b=1}^B$
  ▷ s.t. ${}_{(1)}T^*(i) \leq \ldots \leq {}_{(B)}T^*(i)$
  $q \leftarrow \lceil (1 - P_{\text{fa}})(B + 1) \rceil$
  ▷ index to select the threshold
  $T_\tau(i) \leftarrow {}_{(q)}T^*(i)$
  if $|T(i) - T_0(i)| < T_\tau(i)$ **then** H_0 is not rejected
  else H_0 is rejected

provided in the upcoming chapters.

Part II.

Model selection in two data sets

3. A review of model selection in two data sets

In this chapter, we will review the traditional ITC and hypothesis testing techniques for determining the number of correlated components in two data sets. This number is sufficient to completely characterize the linear association between any two data sets. We will also discuss their extensions in the regime where the number of observations is comparable or even smaller than the dimensions of the two sets and numerically show their usefulness over the traditional techniques.

3.1. Introduction

The most common tool for studying and interpreting the association between two data sets is CCA [17]. As explained in Section 2.1, for any two n_1 and n_2-dimensional sets of data denoted by $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively, CCA provides linear projections from the measurement space to the so-called canonical space, where the correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$ is maximized. The projections $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are the vectors containing the canonical variables, and the correlation coefficients among them are referred to as canonical correlations. CCA leads to a diagonal correlation matrix between $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, i.e., the ith canonical variable of the first set $\epsilon_1^{(i)}$ is only correlated to the corresponding ith canonical variable of the second set $\epsilon_2^{(i)}$. Moreover, the canonical variables are sorted in decreasing order of their canonical correlations. Therefore, knowing the number of non-zero canonical correlations, denoted as d_{12}, is sufficient to completely characterize the second-order dependence between $\mathbf{x}_1$ and $\mathbf{x}_2$. This is because there is no ambiguity about which of the canonical variables are correlated. One can extract the first d_{12} canonical variables and stop as the remaining canonical variables are uncorrelated.

As seen in Section 2.1, the canonical variables and the canonical correlations can be obtained from the SVD of the coherence matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ [41] defined as

$$\mathbf{C}_{12} = \mathbf{R}_{11}^{-\frac{1}{2}} \mathbf{R}_{12} \mathbf{R}_{22}^{-\frac{1}{2}}. \tag{3.1}$$

In practice, the population covariance matrices, $\mathbf{R}_{11}, \mathbf{R}_{12}, \mathbf{R}_{22}$, are unknown and have to be estimated from the samples of $\mathbf{x}_1$ and $\mathbf{x}_2$. Given M i.i.d. joint samples of $\mathbf{x}_1$ and $\mathbf{x}_2$, the data matrices can be constructed as

$$\begin{aligned} \mathbf{X}_1 &= [\mathbf{x}_1(1), \mathbf{x}_1(2), \ldots, \mathbf{x}_1(M)] \\ \mathbf{X}_2 &= [\mathbf{x}_2(1), \mathbf{x}_2(2), \ldots, \mathbf{x}_2(M)]. \end{aligned} \tag{3.2}$$

The term *joint samples* of $\mathbf{x}_1$ and $\mathbf{x}_2$ means that the lth sample of $\mathbf{x}_1$ is associated with the lth sample of $\mathbf{x}_2$ for $l = 1, \ldots, M$. This means for example, the lth samples of $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn at the same time point or belong to the same subject. Using (3.2), the sample covariance matrices can be estimated as $\hat{\mathbf{R}}_{11} = \frac{1}{M}\mathbf{X}_1\mathbf{X}_1^T$, $\hat{\mathbf{R}}_{22} = \frac{1}{M}\mathbf{X}_2\mathbf{X}_2^T$, and $\hat{\mathbf{R}}_{12} = \frac{1}{M}\mathbf{X}_1\mathbf{X}_2^T$. The sample coherence matrix $\hat{\mathbf{C}}_{12}$ can then be computed using (3.1) and the sample canonical correlations and sample canonical variables can be estimated from its SVD as (2.5)-(2.8). In this case, however, not only d_{12}, but all the sample canonical correlations will be non-zero and d_{12} has to be estimated. One method common in the literature is to set a threshold in a heuristic way and assume that all the sample canonical correlations above the threshold are significant, while all those below the threshold are insignificant [65], [66]. However, this method is application- or task-specific and can often fail as the sample canonical correlations are highly overestimated when the number of samples is limited [35]. To reliably estimate d_{12}, model-order selection techniques such as ITC and hypothesis testing should be applied. These techniques provide a tradeoff between overfitting and underfitting of the model with respect to the model order according to a certain statistical criterion (see Section 2.3 for basics of ITC and hypothesis testing), and thus avoid the use of heuristic user-defined thresholds, which could easily introduce bias in the analysis. There is substantial work on model-order selection for CCA based on ITC and hypothesis testing [25]–[28], [38], [39], [67]. However, most of these techniques work, when the number of samples M is large compared to n_1, n_2, or in the so-called sample-rich regime.

When M is comparable to n_1 and n_2, the sample canonical correlations are significantly overestimated. Moreover, when $M < n_1 + n_2$, then at least $n_1 + n_2 - M$ sample canonical correlations are equal to one irrespective of the population canonical correlations [35]. We

call the regime as sample poor when M is not significantly larger than n_1, n_2. In this regime, traditional CCA cannot be used to infer the linear relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$, and a pre-processing step or some other form of regularization must be incorporated either before applying CCA or jointly with CCA. Even though various extensions of CCA like regularized CCA [68], [69], sparse CCA [36], [70] have been proposed in the literature to deal with the sample-poor regime, most of them do not address the model-order selection problem. However, a few recent works which have proposed solutions in this context have been using principal component analysis (PCA) before applying CCA [37], [38], joint PCA-CCA [28], sparse CCA for model-order selection [39], random projections [29] and cross-validation [71].

In this chapter, we will briefly review the data model in Section 3.2 followed by the ITC and hypothesis testing approaches for determining d_{12} in Sections 3.3 and 3.4, respectively. We will then discuss their extensions to the sample-poor regime using a joint PCA-CCA technique proposed in [28]. The technique in [28] compared to the competing techniques, is shown to be effective in extensive simulation scenarios and has been successfully applied in numerous real-world applications [72]–[75].

3.2. Data model for two sets

We consider two data sets represented by M i.i.d. samples of two zero-mean real-valued random vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ with dimensions n_1 and n_2, respectively. We assume without loss of generality that $n_1 \leq n_2$. The random vectors are assumed to be generated by the following linear mixing of signals with additive noise

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{A}_1\mathbf{s}_1 + \mathbf{n}_1, \\ \mathbf{x}_2 &= \mathbf{A}_2\mathbf{s}_2 + \mathbf{n}_2. \end{aligned} \tag{3.3}$$

The signal vectors $\mathbf{s}_1 \in \mathbb{R}^{m_1}$, $\mathbf{s}_2 \in \mathbb{R}^{m_2}$ are zero-mean and contain $m_1(\leq n_1)$, $m_2(\leq n_2)$ signal components, respectively. Each signal component is denoted by $s_p^{(i)}$, where the superscript i denotes the component number and the subscript p denotes the data set. The mixing matrices $\mathbf{A}_1 \in \mathbb{R}^{n_1 \times m_1}$, $\mathbf{A}_2 \in \mathbb{R}^{n_2 \times m_2}$ are unknown (but deterministic) with full column rank. The noise vectors $\mathbf{n}_1 \in \mathbb{R}^{n_1}$, $\mathbf{n}_2 \in \mathbb{R}^{n_2}$ are zero-mean and uncorrelated with the signal vectors and with each other.

We assume without loss of generality two kinds of association among the signal compo-

nents:

1. Intraset independence: signal components within each data set are uncorrelated, i.e.,

$$\mathbf{R}_{s_p s_p} = E[\mathbf{s}_p \mathbf{s}_p^T] = \text{diag}\left(\left(\sigma_p^{(1)}\right)^2, \ldots, \left(\sigma_p^{(m_p)}\right)^2\right), \tag{3.4}$$

   for $p = 1, 2$.

2. Interset dependence: between the two data sets, components may be correlated only pairwise, i.e., component $s_1^{(i)}$ may only correlate with component $s_2^{(i)}$ for $1 \leq i \leq m_{12}$, where $m_{12} = \min(m_1, m_2)$. This means, the signal cross-covariance matrix between data sets is

$$\mathbf{R}_{s_1 s_2} = \text{diag}\left(\rho_{12}^{(1)}\sigma_1^{(1)}\sigma_2^{(1)}, \ldots, \rho_{12}^{(m_{12})}\sigma_1^{(m_{12})}\sigma_2^{(m_{12})}\right), \tag{3.5}$$

   where $\rho_{12}^{(i)}$ represents the unknown (possibly zero) correlation coefficient between their ith components.

The noise covariance matrices

$$\mathbf{R}_{n_p n_p} = E[\mathbf{n}_p \mathbf{n}_p^T], \tag{3.6}$$

for $p = 1, 2$, are unknown and arbitrary. Thus, we do not make any kind of prior assumption about the noise. There are d_{12} signal components correlated between $\mathbf{s}_1$ and $\mathbf{s}_2$. Thus, d_{12} correlation coefficients in (3.5) are non-zero, i.e., $d_{12} = |\{i : \text{ for which } \rho_{12}^{(i)} \neq 0\}|$.

Our aim is to completely characterize the second-order dependence between $\mathbf{x}_1$ and $\mathbf{x}_2$. As discussed in Section 3.1, determining d_{12} is sufficient to characterize this information for two data sets. We summarize our goal as follows.

Goal: *Given M i.i.d. samples of $\mathbf{x}_1$ and $\mathbf{x}_2$ from the model in* (3.3)*, determine the number d_{12} of correlated components.*

Since d_{12} is a model order, determining d_{12} is a *model-order selection* problem. We will first discuss the traditional ITC and hypothesis testing techniques for solving the given model-order selection problem and later discuss their extensions for the sample-poor regime.

3.3. ITC-based model-order selection

Using (3.3), the cross-covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\mathbf{R}_{12} = E[\mathbf{x}_1 \mathbf{x}_2^T] = \mathbf{A}_1 \mathbf{R}_{s_1 s_2} \mathbf{A}_2^T. \tag{3.7}$$

Using (3.7), and under the assumption that $\mathbf{A}_1$ and $\mathbf{A}_2$ have full rank, the rank of $\mathbf{R}_{12}$ is equal to the rank of $\mathbf{R}_{s_1 s_2}$. Thus, for d_{12} correlated components,

$$\text{rank}(\mathbf{R}_{12}) = d_{12}. \tag{3.8}$$

Let $\mathbf{y}_1 = \mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{R}_{22}^{-\frac{1}{2}}\mathbf{x}_2$ be the whitened versions of $\mathbf{x}_1$ and $\mathbf{x}_2$. Assuming that $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are non-singular, the canonical correlations of $\mathbf{x}_1$ and $\mathbf{x}_2$ are same as the canonical correlations of $\mathbf{y}_1$ and $\mathbf{y}_2$. We will derive the ITC expression to estimate d_{12} using the transformed vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ as it is simpler to derive the ML function. However, the ITC has been earlier derived in [26] using $\mathbf{x}_1$, $\mathbf{x}_2$ and we show an alternative derivation for the sake of completeness. Under the assumption that $\mathbf{x}_1$ and $\mathbf{x}_2$ are jointly-Gaussian distributed, the composite vector, $\mathbf{y} = [\mathbf{y}_1^T, \mathbf{y}_2^T]^T$ is also Gaussian distributed with zero-mean and covariance matrix,

$$\mathbf{C}_{yy} = \begin{bmatrix} \mathbf{I} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{I} \end{bmatrix}. \tag{3.9}$$

Using (3.8), the rank of $\mathbf{C}_{12}$ is also equal to d_{12}.

The ITC score as explained in Section 2.3.1 depends on the log-likelihood function maximized with respect to the model order and on the number of free parameters in the model. The log-likelihood function for M i.i.d samples of $\mathbf{y}$ parameterized by $\mathbf{C}_{yy}$ is [48]

$$\ln f\left(\mathbf{y}(1), \ldots, \mathbf{y}(M) | \mathbf{C}_{yy}\right) = C - \frac{M}{2} \ln \det\left(\mathbf{C}_{yy}\right) - \frac{1}{2} \sum_{l=1}^{M} \left(\mathbf{y}^T(l) \mathbf{C}_{yy}^{-1} \mathbf{y}(l) \right), \tag{3.10}$$

where the constant C is independent of the parameter space $\mathbf{C}_{yy}$. Let $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(M)]$. The maximum-likelihood (ML) estimate of $\mathbf{C}_{yy}$ under the constraint that $\text{rank}(\mathbf{C}_{12}) = d_{12}$ is [26]

$$\hat{\mathbf{C}}_{yy} = \frac{1}{M} \mathbf{Y}\mathbf{Y}^T = \begin{bmatrix} \mathbf{I} & \hat{\mathbf{C}}_{12} \\ \hat{\mathbf{C}}_{12}^T & \mathbf{I} \end{bmatrix}, \tag{3.11}$$

where the sample coherence matrix $\hat{\mathbf{C}}_{yy}$ is the ML estimate of $\mathbf{C}_{yy}$ and $\hat{\mathbf{C}}_{12}$ has the SVD

$$\hat{\mathbf{C}}_{12} = \hat{\mathbf{F}}_1 \hat{\mathbf{K}}_{12} \hat{\mathbf{F}}_2^T, \tag{3.12}$$

with $\hat{\mathbf{K}}_{12} = \text{diag}(\hat{k}^{(1)}, \ldots, \hat{k}^{(d_{12})}, 0, \ldots, 0)$. Substituting (3.11) in (3.10), the maximum log-likelihood expression, to within a constant, is

$$\ln f(\mathbf{Y} | \hat{\mathbf{C}}_{yy}) \propto -\frac{M}{2} \ln \det(\hat{\mathbf{C}}_{yy}). \tag{3.13}$$

Using (3.11) and (3.12), $\hat{\mathbf{C}}_{yy}$ can be partitioned as [76]

$$\hat{\mathbf{C}}_{yy} = \begin{bmatrix} \mathbf{I} & \hat{\mathbf{F}}_1\hat{\mathbf{K}}_{12}\hat{\mathbf{F}}_2^T \\ \hat{\mathbf{F}}_2\hat{\mathbf{K}}_{12}^T\hat{\mathbf{F}}_1^T & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{F}}_1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{F}}_2 \end{bmatrix} \begin{bmatrix} \mathbf{I} & \hat{\mathbf{K}}_{12} \\ \hat{\mathbf{K}}_{12}^T & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{F}}_1^T & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{F}}_2^T \end{bmatrix}. \tag{3.14}$$

The determinant of $\hat{\mathbf{C}}_{yy}$ can thus, be simplified as

$$\begin{aligned} \det(\hat{\mathbf{C}}_{yy}) &= \det(\hat{\mathbf{F}}_1)\det(\hat{\mathbf{F}}_2)\det(\mathbf{I} - \hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{12}^T)\det(\hat{\mathbf{F}}_1^T)\det(\hat{\mathbf{F}}_2^T), \\ &= \det(\mathbf{I} - \hat{\mathbf{K}}_{12}\hat{\mathbf{K}}_{12}^T), \\ &= \prod_{i=1}^{d_{12}} \left(1 - \left(\hat{k}^{(i)}\right)^2\right). \end{aligned} \tag{3.15}$$

Here we have used the fact that $\det(\hat{\mathbf{F}}_p^T) = \frac{1}{\det(\hat{\mathbf{F}}_p)}$ since $\hat{\mathbf{F}}_p$ is orthogonal for $p = 1, 2$. The maximum log-likelihood in (3.13) as a function of d_{12} is thus [25], [26], [77],

$$f(\mathbf{Y}|\hat{\mathbf{C}}_{yy}, d_{12}) \propto -\frac{M}{2}\ln\left(\prod_{i=1}^{d_{12}} 1 - \left(\hat{k}^{(i)}\right)^2\right). \tag{3.16}$$

The number of free parameters in $\mathbf{C}_{yy}$ is equal to the number of free parameters in $\mathbf{C}_{12}$. That can be determined by counting the number of free parameters in the SVD of $\mathbf{C}_{12}$ as follows. The number of free parameters in the singular vectors, $\mathbf{F}_1$ and $\mathbf{F}_2$ is $n_1 d_{12}$ and $n_2 d_{12}$, respectively. However, not all of them are freely adjustable. There are d_{12} and $\frac{d_{12}(d_{12}-1)}{2}$ constraints on the elements of the singular vectors in both $\mathbf{F}_1$ and $\mathbf{F}_2$ due to normality and orthogonality, respectively. The number of free parameters in $\mathbf{K}_{12}$ is d_{12}. The total number of free parameters C_{12} is

$$\begin{aligned} C_{12} &= n_1 d_{12} + n_2 d_{12} - 2d_{12} - d_{12}(d_{12} - 1) + d_{12}, \\ &= n_1 d_{12} + n_2 d_{12} - d_{12}^2. \end{aligned} \tag{3.17}$$

The simplified ITC score using (2.15) is

$$\text{ITC}(d_{12}) = \frac{M}{2}\ln\left(\prod_{i=1}^{d_{12}} 1 - \left(\hat{k}^{(i)}\right)^2\right) + \frac{\ln(M)}{2}(n_1 d_{12} + n_2 d_{12} - d_{12}^2), \quad \text{for MDL}, \tag{3.18}$$

$$= \frac{M}{2}\ln\left(\prod_{i=1}^{d_{12}} 1 - \left(\hat{k}^{(i)}\right)^2\right) + (n_1 d_{12} + n_2 d_{12} - d_{12}^2), \quad \text{for AIC}. \tag{3.19}$$

The estimate of d_{12} is the one which minimizes (3.18) for MDL, and (3.19) for AIC.

3.4. GLRT-based model order selection

As discussed in Section 2.3.2, the model order can also can be estimated via a sequence of binary hypothesis tests. To estimate d_{12} using hypothesis testing, set $s = 0$ and perform the following test

$$\begin{aligned} H_0 &: d_{12} = s, \\ H_1 &: d_{12} > s, \end{aligned} \tag{3.20}$$

and increment s until H_0 is not rejected or $d_{12} = n_1 - 1$. When $\mathbf{x}_1$ and $\mathbf{x}_2$ are jointly Gaussian distributed, the GLR for (3.20) is given as

$$\eta = \frac{f(\mathbf{Y}|\hat{\mathbf{C}}_{yy}, d_{12} = s)}{f(\mathbf{Y}|\hat{\mathbf{C}}_{yy}, d_{12} > s)}, \tag{3.21}$$

where $f(\mathbf{Y}|\hat{\mathbf{C}}_{yy}, d_{12} = s)$ and $f(\mathbf{Y}|\hat{\mathbf{C}}_{yy}, d_{12} > s)$ is the ML function under H_0 and H_1, respectively. Using (3.16), η can be simplified as

$$\eta = \frac{\left(\prod_{i=1}^{s} 1 - \left(\hat{k}^{(i)}\right)^2 \right)^{-\frac{M}{2}}}{\left(\prod_{i=1}^{n_1} 1 - \left(\hat{k}^{(i)}\right)^2 \right)^{-\frac{M}{2}}}, \tag{3.22}$$

where we have used the fact that the parameter space $s = n_1$ is sufficient to parameterize all the possibilities when $d_{12} > s$. Thus,

$$\eta = \left(\prod_{i=s+1}^{n_1} 1 - \left(\hat{k}^{(i)}\right)^2 \right)^{-\frac{M}{2}}. \tag{3.23}$$

Wilks' statistic: According to the Wilks' theorem, the statistic $W(s) = -2 \ln \eta$ is asymptotically χ^2_ν distributed when $s = d_{12}$. The d.f. of this distribution can be computed using

(2.20) and the number of free parameters computed in (3.17) as

$$\begin{aligned}\nu &= C_{H_1} - C_{H_0} \\ &= n_1^2 + n_1 n_2 - n_1^2 - n_1 s - n_2 s + s^2 \\ &= (n_1 - s)(n_2 - s). \end{aligned} \tag{3.24}$$

$W(s)$ approaches the χ^2_ν distribution as $M \to \infty$. However, for finite sample regime, Bartlett [58] and Lawley [78] modified $W(s)$ to match the moments of the χ^2_ν distribution. The two statistics,

Bartlett statistic:

$$B(s) = -\left(M - \frac{n_1 + n_2 + 1}{2}\right) \ln\left(\prod_{i=s+1}^{n_1} 1 - \left(\hat{k}^{(i)}\right)^2\right), \tag{3.25}$$

and

Bartlett-Lawley statistic:

$$C(s) = -\left(M - s - \frac{n_1 + n_2 + 1}{2} + \sum_{i=1}^{s}\left(\hat{k}^{(i)}\right)^2\right) \ln\left(\prod_{i=s+1}^{n_1} 1 - \left(\hat{k}^{(i)}\right)^2\right), \tag{3.26}$$

provide a better approximation of χ^2_ν distribution than $W(s)$ for small sample size. For a given probability of false alarm P_{fa}, the threshold $\tau(s)$ can be set using the χ^2_ν-distribution and d_{12} can be estimated as

$$\hat{d}_{12} = \min_{s=0,\ldots,n_1-1}\{s : W(s) < \tau(s)\}, \tag{3.27}$$

where $W(s)$ can be replaced by $B(s)$ or $C(s)$ when using Bartlett or Bartlett-Lawley statistic, respectively.

3.5. Joint PCA-CCA detectors

When M is not large compared to the dimensions n_1, n_2, the sample canonical correlations are overestimated. Moreover, when $M < n_1 + n_2$, some of the sample canonical correlations are one irrespective of the true canonical correlations [35]. In this case, CCA gives a false impression that there exists perfect correlation among the components of the data sets even though there might be none. To analyze this closely, let us revisit the sample composite

coherence matrix defined in (3.14). The largest and smallest eigenvalues of $\hat{\mathbf{C}}_{yy}$ exist in pairs, and it can be shown using the factorization (3.14), that they are related to the sample canonical correlations as [79]

$$\mathbf{\Lambda}(\hat{\mathbf{C}}_{yy}) = \left\{1 + \hat{k}^{(1)}, \ldots, 1 + \hat{k}^{(d_{12})}, 1, \ldots, 1, 1 - \hat{k}^{(d_{12})}, \ldots, 1 - \hat{k}^{(1)}\right\}, \tag{3.28}$$

where $\mathbf{\Lambda}(\hat{\mathbf{C}}_{yy})$ denotes the set of eigenvalues of $\hat{\mathbf{C}}_{yy}$ arranged in the descending order. Thus, if the smallest eigenvalue of $\hat{\mathbf{C}}_{yy}$ is zero (or correspondingly the largest eigenvalue is two), at least one sample canonical correlation is one. Since the rank of $\hat{\mathbf{C}}_{yy}$ is minimum of M and $n_1 + n_2$, when $M < n_1 + n_2$, at least $n_1 + n_2 - M$ smallest eigenvalues of $\hat{\mathbf{C}}_{yy}$ are equal to zero (or correspondingly at least $n_1 + n_2 - M$ largest eigenvalues are equal to two). Thus, at least $n_1 + n_2 - M$ sample canonical correlations are equal to one. The sample canonical correlations in this regime are commonly regarded as *defective* [35]. In this case, the ITC and hypothesis testing techniques based on traditional CCA derived in Sections 3.3 and 3.4 will overestimate d_{12}. A possible solution is to reduce the dimension of the data sets so that M is large compared to the reduced dimensions. The most common and widely applicable dimension reduction tool is PCA. However, a complication with PCA is that it selects the components which have the most variance within a data set and these might not correspond to the components which are highly correlated among the two sets. Thus, the PCA ranks r_1 and r_2 corresponding to the dimension-reduced sets should be carefully chosen.

Let the EVD of the covariance matrices be

$$\mathbf{R}_{pp} = \mathbf{U}_p \mathbf{\Lambda}_p \mathbf{U}_p^T, \quad p = 1, 2. \tag{3.29}$$

The reduced rank PCA descriptions for the two sets with reduced dimensions r_1 and r_2, respectively is

$$\widetilde{\mathbf{x}}_p = \left[\mathbf{U}_p(:, 1 : r_p)\right]^T \mathbf{x}_p, \quad p = 1, 2, \tag{3.30}$$

where $\mathbf{U}_p(:, 1 : r_p)$ denotes the first r_p columns of $\mathbf{U}_p$. Using (3.29) and (3.30), the reduced rank data model is

$$\begin{aligned} \widetilde{\mathbf{x}}_1 &= \left[\mathbf{U}_1(:, 1 : r_1)\right]^T \mathbf{A}_1 \mathbf{s}_1 + \left[\mathbf{U}_1(:, 1 : r_1)\right]^T \mathbf{n}_1, \\ &= \widetilde{\mathbf{A}}_1 \mathbf{s}_1 + \widetilde{\mathbf{n}}_1, \end{aligned} \tag{3.31}$$

and similarly,

$$\widetilde{\mathbf{x}}_2 = \widetilde{\mathbf{A}}_2 \mathbf{s}_2 + \widetilde{\mathbf{n}}_2, \tag{3.32}$$

where $\widetilde{\mathbf{A}}_1 \in \mathbb{R}^{r_1 \times m_1}$, $\widetilde{\mathbf{A}}_2 \in \mathbb{R}^{r_2 \times m_2}$ are the mixing matrices and $\widetilde{\mathbf{n}}_1 \in \mathbb{R}^{r_1}$, $\widetilde{\mathbf{n}}_2 \in \mathbb{R}^{r_2}$ are the noise vectors of the reduced-rank model.

Maxmin MDL ITC - The MDL-ITC expression in (3.18) can be modified for the reduced-rank model in (3.31) and (3.32) as

$$\text{ITC}(d_{12}, r_1, r_2) = \frac{M}{2} \ln \left(\prod_{i=1}^{d_{12}} 1 - \left(\hat{k}^{(i)}(r_1, r_2) \right)^2 \right) + \frac{\ln(M)}{2} (r_1 d_{12} + r_2 d_{12} - d_{12}^2). \quad (3.33)$$

We choose the MDL-ITC since it results in a consistent estimator of d_{12} [80]. Note that the sample canonical correlations now depend on r_1 and r_2 and would significantly change on the basis of the choice of r_1 and r_2. If r_1, r_2 are small, not all the correlated components will likely be included in $\widetilde{\mathbf{x}}_1$, $\widetilde{\mathbf{x}}_2$ and d_{12} will most likely be underestimated. However, if r_1, r_2 are large enough to include all the correlated and stronger independent components, d_{12} will typically be estimated correctly. If r_1, r_2 are further increased such that even weaker independent components and noise are added in $\widetilde{\mathbf{x}}_1$, $\widetilde{\mathbf{x}}_2$, then d_{12} will not be overestimated since the MDL is consistent. In this case, however, it is possible that d_{12} is underestimated if M is not large enough. An effective solution as shown in [28] is to chose the maximum value of d_{12} estimated over all possible ranks, i.e.,

$$\hat{d}_{12} = \max_{r_1, r_2 = 1, \ldots, r_{\max}} \; \arg\min_{d_{12} = 0, \ldots, \min(r_1, r_2)} \text{ITC}(d_{12}, r_1, r_2). \quad (3.34)$$

For the technique in (3.34) to work, $r_{\max}$ should be chosen to be small compared to M. Typically, $r_{\max} = \min(n_1, n_2, \frac{M}{3})$ has been shown to work best using extensive simulations [28] and also for real-world data [72]–[74]. For more details about (3.34), please refer to [28].

Maxmin GLRT - The reduced-rank Bartlett-Lawley statistic for the model in (3.31) and (3.32) is

$$C(s, r_1, r_2) = -\left(M - s - \frac{r_1 + r_2 + 1}{2} + \sum_{i=1}^{s} \left(\hat{k}^{(i)}(r_1, r_2) \right)^2 \right) \times \ln \left(\prod_{i=s+1}^{\min r_1, r_2} 1 - \left(\hat{k}^{(i)}(r_1, r_2) \right)^2 \right), \quad (3.35)$$

Again, the test statistic $C(s, r_1, r_2)$ depends on the PCA ranks r_1, r_2. Assuming that r_1, r_2 are small compared to M and not large enough to include all the correlated components, $C(s, r_1, r_2)$ will approximately be $\chi^2_{\nu'}$-distributed with $\nu' = (r_1 - s)(r_2 - s)$ for $s = \widetilde{d}_{12}$,

where $\widetilde{d}_{12} < d_{12}$. Thus, d_{12} will typically be underestimated. However, when r_1, r_2 are large enough to include all the correlated components, d_{12} will typically not be overestimated and $C(s, r_1, r_2)$ will be approximately $\chi^2_{\nu'}$-distributed for $s = d_{12}$. [28] proposes the following decision rule to estimate d_{12},

$$\hat{d}_{12} = \max_{r_1,r_2=1,\ldots,r_{\max}} \min_{s=0,\ldots,\min(r_1,r_2)-1} \{s : C(s, r_1, r_2) < T(s, r_1, r_2)\}, \tag{3.36}$$

where $T(s, r_1, r_2)$ is the threshold chosen from the $\chi^2_{\nu'}$ distribution with the given P_{fa}. The rule (3.36) is motivated by the fact that if r is not chosen optimally, the min-step might return a number smaller than d_{12}. Because the min-step will not overfit, we can take the maximum result for all r from 1 up to $r_{\max}$. As in the maxmin ITC detector, $r_{\max} = \min(n_1, n_2, \frac{M}{3})$ works well. For more details about (3.36), please refer to [28].

3.6. Numerical results

In this section, we evaluate the performance of the traditional ITC and hypothesis testing techniques along with the joint PCA-CCA detectors using Monte-Carlo trials. We also compare the performance with the informative CCA (ICCA) technique of [38], where the most varying components in two data sets are first retained using separate PCA steps followed by estimating d_{12} using the Tracy-Widom approximation [81] for the largest canonical correlation due to noise. ICCA proposes [82] to estimate the number of components retained by the PCA step, which assumes that some noise-only samples of the data are available. Since we do not assume to have this knowledge, we instead use [83] for the PCA step. We employ two performance measures defined as follows:

a) *Mean accuracy of detecting* $\hat{d}_{12}$ - number of correct estimates of d_{12} divided by the total number of trials, and

b) *Mean value of* $\hat{d}_{12}$ - average value of $\hat{d}_{12}$ over all trials.

The results for two different scenarios are presented. The first scenario is the so called sample-rich regime, where M is large compared to the dimensions n_1, n_2. To show the effectiveness of the joint PCA-CCA approach, in the second scenario (sample-poor regime), M is comparable to n_1, n_2. The simulation setup common to both scenarios is explained below.

There are $m_1 = m_2 = 5$ signals out of which $d_{12} = 3$ signals are correlated with correlation

coefficients of 0.9, 0.8 and 0.7. All signals are Gaussian distributed with the variance of independent signals $\sigma_i^2 = 2$ and the variance of correlated signals $\sigma_c^2 = 1$. The mixing matrices are randomly generated orthogonal matrices. Each data set is corrupted by additive colored Gaussian noise. The noise is colored by applying the second order autoregressive (AR) filter with filter coefficients $[1, 0.33]$ to white noise. The variance of the white noise σ_n^2 is chosen according to the signal-to-noise-ratio (SNR) which is defined per component as

$$\text{SNR (dB)} = 10 \log_{10} \left(\frac{\sigma_c^2}{\sigma_n^2} \right). \tag{3.37}$$

The results are averaged over 500 independent trials.

i. **Sample-rich regime**, with $n_1 = 10, n_2 = 15$: Figure 3.1 shows the mean accuracy of $\hat{d}_{12}$ as a function of M for all the techniques. The SNR is 10dB. The traditional GLRT detector is applied using two different P_{fa} values. It can be seen that for a small values of M, the detector with higher P_{fa} outperforms the detector with smaller P_{fa}. The joint PCA-CCA-based detectors and the ICCA technique of [38] need less samples to correctly estimate d_{12} compared to the traditional detectors. Both the AIC and MDL-based ITC detectors are shown to illustrate their differences. The AIC-ITC detector works well when M is small compared to the MDL-ITC detector which outperforms the former for large M. This can be explained with the help of Figure 3.2 which shows the mean value of $\hat{d}_{12}$ for all detectors. The AIC overestimates on average and thus works well for small M, whereas the MDL underestimates for small M and approaches the true value of $d_{12} = 3$ for large M.

   Another thing to note in Figures 3.1 and 3.2 is that there is no clear winner between the ITC and the corresponding GLRT detector. For traditional techniques without any rank reduction, the GLRT-based detectors outperform the ITC detectors while for the reduced-rank case, the ITC-based PCA-CCA detector outperforms its GLRT counterpart for very small M.

   Additionally in Figure 3.3, we also see the mean accuracy as a function of the SNR with $M = 250$. It can be seen that when M is large enough, all techniques work very similarly with the PCA-CCA detectors slightly outperforming the other techniques.

ii. **Sample-poor regime**, with $n_1 = n_2 = 40$: The mean accuracy and the mean value of $\hat{d}_{12}$ for all detectors can be seen in Figures 3.4 and 3.5, respectively. In this case, the PCA-CCA detectors significantly outperform the traditional competitors. Even in the defective regime, i.e., when $M < n_1 + n_2$, the PCA-CCA detectors perform reasonably

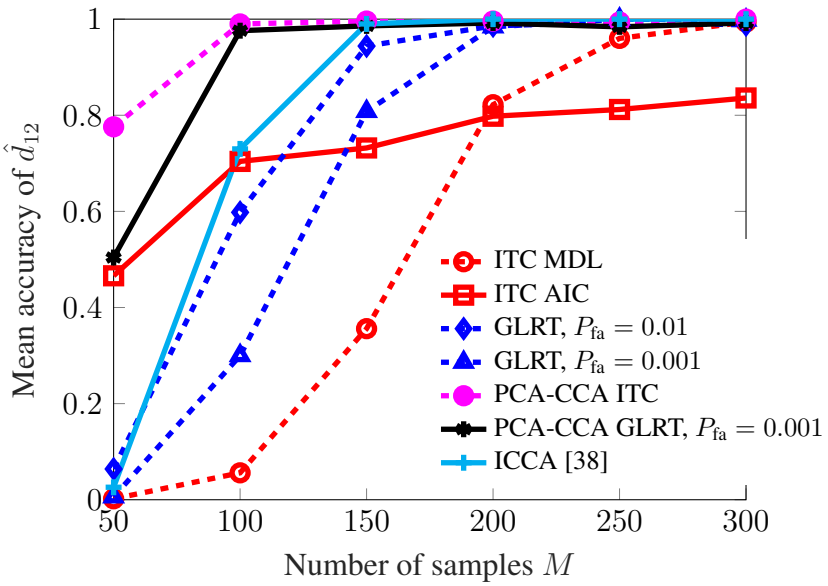


Figure 3.1.: Mean accuracy of $\hat{d}_{12}$ in scenario i) for the traditional and PCA-CCA detectors.

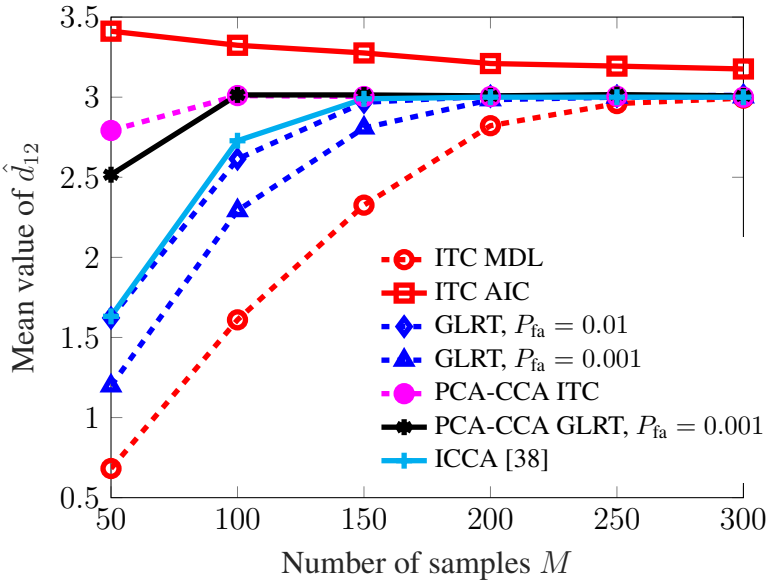


Figure 3.2.: Mean value of $\hat{d}_{12}$ in scenario i).

well. Their sample-rich counterparts require $M \gg n_1, n_2$ and therefore, do not work even for $M = 160$. Moreover, the PCA-CCA detectors significantly outperform the ICCA technique. This is due to the fact that the ICCA employs a separate PCA step before estimating d_{12}. When the data sets contain uncorrelated components with higher variance compared to the correlated components and/or the additive noise is colored, the PCA step retains too many components than desired. On the other hand, the PCA-CCA detectors jointly estimate the PCA ranks and d_{12} and are not likely to retain the undesired

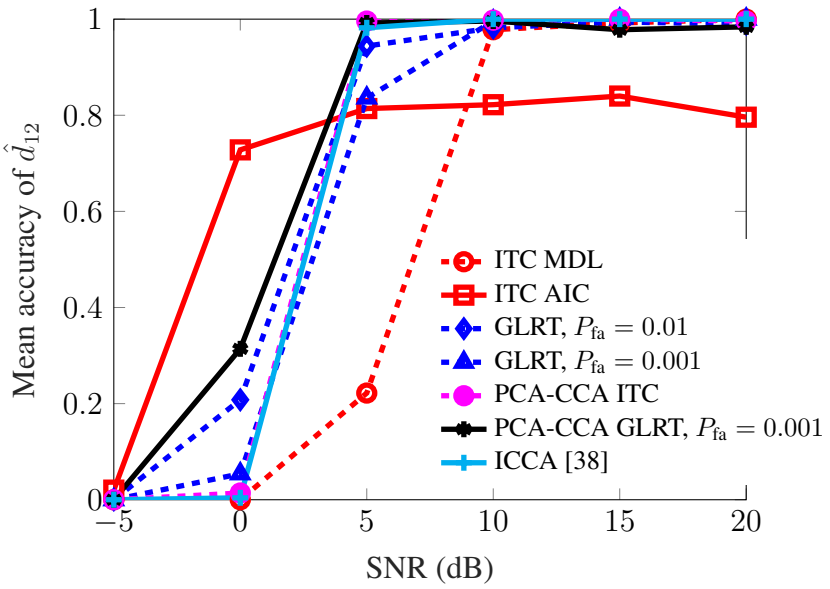


Figure 3.3.: Mean accuracy of $\hat{d}_{12}$ in scenario i) as a function of SNR.

uncorrelated and noise components. Hence, the ICCA technique requires more samples than the PCA-CCA detectors to correctly estimate d_{12}, as seen in Figure 3.4.

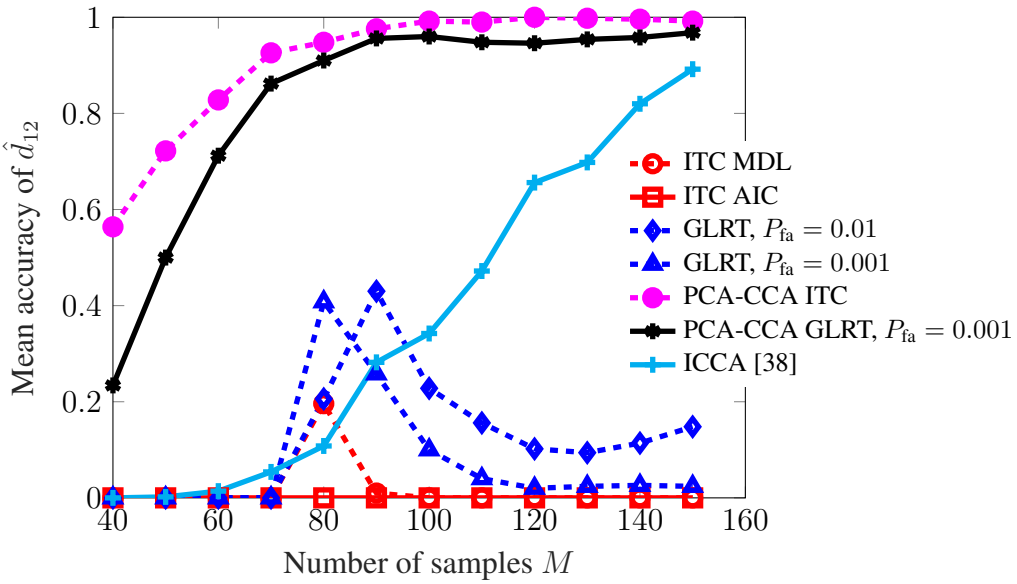


Figure 3.4.: Mean accuracy of $\hat{d}_{12}$ in scenario ii).

Finally, the small peaks for the traditional detectors in Figure 3.4 can be explained with the help of Figure 3.5. When $M < n_1 + n_2$, at least $n_1 + n_2 - M$ sample canonical correlations are one. In this regime, the traditional detectors highly overestimate d_{12}. As M increases, the value of their estimates decreases on average and coincidentally they pick the correct d_{12} for a certain M.

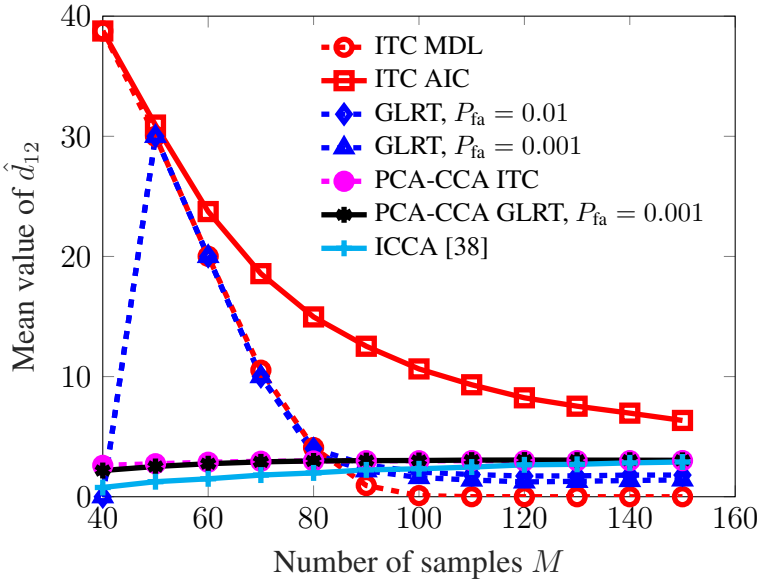


Figure 3.5.: Mean value of $\hat{d}_{12}$ in scenario ii).

3.7. Summary

We have reviewed the traditional techniques based on ITC and GLRT to estimate the model order jointly in two data sets. However, these techniques assume the number of samples to be very large compared to the dimensions of the data sets. We also reviewed their modifications based on the joint PCA-CCA analysis, which is specifically designed to tackle the small-sample regime. However, there is no free lunch. The joint PCA-CCA detectors assume that the PCA rank in both data sets required to include all the correlated components is small compared to the number of samples.

4. Determining the dimension of improper subspace in complex-valued data

A complex-valued signal is improper if it is correlated with its complex conjugate. The dimension of the improper signal subspace, i.e., the number of improper components in a complex-valued measurement, is an important parameter and is unknown in most applications. In this chapter, we introduce two approaches to estimate this dimension, one based on an information-theoretic criterion and one based on hypothesis testing. We also present reduced-rank versions of these approaches that work for scenarios where the number of observations is comparable to or even smaller than the dimension of the data. Unlike other techniques for determining model orders, our techniques also work in the presence of additive colored noise [1].

4.1. Introduction

Complex-valued signals (or complex random variables) are used in various fields like communications, oceanography, geophysics, speech processing [84]. Modelling two real-valued signals as one complex-valued signal leads not only to new insights and interpretations but also to economical and efficient algorithms [44]. For a zero-mean complex random variable x, the variance is defined as $r_{xx} = E[xx^*]$, where x^* denotes the complex conjugate of x. However, r_{xx} alone is an incomplete characterization of the second-order statistics of

[1] This chapter is based on the paper: "Determining the dimension of the improper signal subspace in complex-valued data, T. Hasija, C. Lameiro, and, P. J. Schreier, *IEEE Signal Processing Letters*, 2017."

x. Another second-order moment of x is $\tilde{r}_{xx} = E[x^2]$, commonly referred to as the complementary variance of x [44]. Both r_{xx} and $\tilde{r}_{xx}$ together provide a complete second-order characterization of x. If $\tilde{r}_{xx} = 0$, i.e., if x and x^* are uncorrelated, x is called proper, and otherwise improper. While propriety is a common assumption, improper signals arise in numerous areas in engineering such as communications and also in applied sciences such as oceanography and biomedicine [44], [45], [85].

Let us now extend our discussion to a zero-mean complex random vector $\mathbf{x}$ containing n complex random variables. The number of improper signals in $\mathbf{x}$ is an important parameter in various applications. For instance, detecting the number of improper signal components is often a prerequisite before performing further steps like estimating the direction-of-arrival (DOA) in array processing or blind source separation in biomedicine [86]–[89]. This detection problem can be solved as part of the more general problem of partitioning the observation space into signal and noise subspaces. The standard approach to achieve this partition is based on PCA and ITC [80]. However, this approach is suboptimal when some or all of the signals in the observed data are improper. This is because this technique only takes into account the statistics of the covariance matrix $\mathbf{R}_{xx} = E[\mathbf{x}\mathbf{x}^H]$ and ignores the complementary covariance matrix $\widetilde{\mathbf{R}}_{xx} = E[\mathbf{x}\mathbf{x}^T]$.

Noncircular PCA (ncPCA) introduced in [46] improves on PCA by also taking into account the information about impropriety contained in the complementary covariance matrix. Based on ITC, [46] determines the dimensions of both the proper and improper signal subspaces from noisy observations. However, in some applications, we might only be interested in the dimension of the improper subspace, for instance, when we know that all signal components are improper [88]. This is the problem we solve in this chapter. Even though the technique in [46] can be used for this scenario as well, it is to be expected that a specialized technique works better than a more general one. Indeed, by determining the number of improper signal components only, we are able to reduce the number of required samples and relax the assumption on the noise structure. We only need to assume that the noise is proper, but unlike typical PCA-based methods, it does not have to be white.

We introduce two alternative approaches: one that is based on the MDL-ITC (see Section 4.3), and one that is based on a sequence of GLRTs (see Section 4.4). The proposed approaches are designed specifically for applications with high-dimensional data but small number of samples. They build on a more general technique of joint PCA-CCA, which we have reviewed in Chapter 3 that determines the dimension of the signal subspace correlated between two different data sets [28]. This chapter specializes these techniques to the case

where the two data sets are $\mathbf{x}$ and its complex conjugate $\mathbf{x}^*$. This, however, is not straightforward and requires special care when counting the number of free parameters in the ITC and deriving the approximating distributions in the hypothesis tests.

4.2. Data model for complex-valued data

Consider a linear signal-plus-noise model for the generation of the observed data vector $\mathbf{x} \in \mathbb{C}^n$

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}, \tag{4.1}$$

where $\mathbf{s} \in \mathbb{C}^{d_\mathrm{i}+f}$ is a zero-mean complex Gaussian source vector, $\mathbf{A} \in \mathbb{C}^{n\times(d_\mathrm{i}+f)}$ is an unknown but fixed mixing matrix with full column rank, and $\mathbf{n} \in \mathbb{C}^n$ is a zero-mean complex Gaussian noise vector independent from the source vector. The following additional assumptions are made:

1. The source vector contains d_i improper and f proper signal components. This means that

$$\begin{aligned} &\mathrm{rank}(E[\mathbf{s}\mathbf{s}^H]) = d_\mathrm{i} + f, \\ &\mathrm{rank}(E[\mathbf{s}(\mathbf{s}^*)^H]) = \mathrm{rank}(E[\mathbf{s}\mathbf{s}^T]) = d_\mathrm{i}. \end{aligned} \tag{4.2}$$

   We also allow $f = 0$, i.e., all the signal components may be improper. All signal components are independent, and the dimensions d_i and f are unknown with $d_\mathrm{i} + f \leq n$.

2. The noise vector $\mathbf{n}$ is proper and possibly colored with an arbitrary covariance matrix $\mathbf{R}_{nn}$. This is a more general noise model than the one used in [46] where the noise vector is assumed to be white.

Under the above assumptions, the covariance and the complementary covariance matrices of $\mathbf{x}$ are

$$\begin{aligned} \mathbf{R}_{xx} &= E[\mathbf{x}\mathbf{x}^H] = \mathbf{A}E[\mathbf{s}\mathbf{s}^H]\mathbf{A}^H + \mathbf{R}_{nn}, \\ \widetilde{\mathbf{R}}_{xx} &= E[\mathbf{x}\mathbf{x}^T] = \mathbf{A}E[\mathbf{s}\mathbf{s}^T]\mathbf{A}^T. \end{aligned} \tag{4.3}$$

Let us define the complex augmented vector $\underline{\mathbf{x}} = [\mathbf{x}^T, \mathbf{x}^H]^T$ obtained by stacking $\mathbf{x}$ on top of its complex conjugate $\mathbf{x}^*$. The covariance matrix of $\underline{\mathbf{x}}$ is the augmented covariance matrix

[44]

$$\underline{\mathbf{R}}_{xx} = E[\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H] = \begin{bmatrix} \mathbf{R}_{xx} & \widetilde{\mathbf{R}}_{xx} \\ \widetilde{\mathbf{R}}_{xx}^* & \mathbf{R}_{xx}^* \end{bmatrix}, \tag{4.4}$$

which is a convenient way of keeping track of both $\mathbf{R}_{xx}$ and $\widetilde{\mathbf{R}}_{xx}$. In this chapter, we are interested in estimating the dimension of the improper signal subspace d_{i}, which is equal to the rank of $\widetilde{\mathbf{R}}_{xx}$.

We consider M independent and identically distributed (i.i.d.) samples of $\underline{\mathbf{x}}$, arranged as the M columns of the data matrix $\underline{\mathbf{X}} = \Big[\underline{\mathbf{x}}(1), \underline{\mathbf{x}}(2), \ldots, \underline{\mathbf{x}}(M)\Big]$, where $\underline{\mathbf{x}}(l)$ denotes the lth sample of $\underline{\mathbf{x}}$. When $\widetilde{\mathbf{R}}_{xx}$ is estimated from $\underline{\mathbf{X}}$, its rank, in general, will not be equal to d_i. In Sections 4.3 and 4.4 we introduce two ways of estimating d_i, which are both based on the *circularity coefficients* of $\mathbf{x}$ [44]. These are the canonical correlations between $\mathbf{x}$ and $\mathbf{x}^*$, which can be computed as the singular values of the coherence matrix $\mathbf{C}_{xx} = \mathbf{R}_{xx}^{-\frac{1}{2}} \widetilde{\mathbf{R}}_{xx} \mathbf{R}_{xx}^{-\frac{T}{2}}$. The circularity coefficients are normalized to take values between 0 and 1, and they measure the degree of impropriety of each signal component. A maximally improper component leads to a circularity coefficient of 1, and a proper component to a zero circularity coefficient. When working with samples, the following complication arises. Unless the number of samples is significantly greater than the dimension of the data, the sample circularity coefficients are significantly greater than the (true) population circularity coefficients. As we would like to be able to handle the sample-poor scenario, this requires the use of a dimension-reducing preprocessing step.

Note: Another term common in complex-valued signal processing is *circularity*, which is a stronger version of propriety. For the Gaussian distribution, propriety implies circularity and noncircularity implies impropriety. As we have assumed $\mathbf{x}$ to be zero-mean Gaussian, the improper signal subspace is the noncircular signal subspace. However, in general noncircularity does not imply impropriety [44].

4.3. Detector based on ITC

The goodness-of-fit is measured by the likelihood function for M samples of $\underline{\mathbf{x}}$, which is parameterized by $\underline{\mathbf{R}}_{xx}$

$$f(\underline{\mathbf{X}}|\underline{\mathbf{R}}_{xx}) = \prod_{l=1}^{M} \frac{1}{\pi^n \sqrt{\det \underline{\mathbf{R}}_{xx}}} \exp\left[-\frac{\underline{\mathbf{x}}^H(l) \underline{\mathbf{R}}_{xx}^{-1} \underline{\mathbf{x}}(l)}{2} \right]. \tag{4.5}$$

Using (2.15), the ITC score is

$$\text{ITC}(d_\text{i}) = -\ln f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}) + \alpha(M)C(d_\text{i}), \tag{4.6}$$

where $\hat{\underline{\mathbf{R}}}_{xx}$ (which is simply the sample augmented covariance matrix) is the maximum likelihood estimate of $\underline{\mathbf{R}}_{xx}$, and the second term in (4.6) is the penalty function. In our case, the model order is the number of improper signals d_i. In the penalty term, $C(d_\text{i})$ is the number of free parameters in the parameter space of the model, i.e., in $\underline{\mathbf{R}}_{xx}$. The term $\alpha(M)$ depends on the chosen ITC. We use the MDL criterion as it leads to a consistent estimator of d_i [80], for which $\alpha(M) = \frac{\ln(M)}{2}$. The MDL-ITC chooses the d_i that minimizes (4.6), that is

$$\hat{d}_\text{i} = \underset{d_\text{i}=0,\ldots,n}{\arg\min}\, \text{ITC}(d_\text{i}). \tag{4.7}$$

The ITC expression in (4.6) can be simplified as follows.

Model fit score: The maximization of the log-likelihood is performed under the constraint that $\text{rank}(\widetilde{\mathbf{R}}_{xx}) = d_\text{i}$. The maximum log-likelihood is [44]

$$-\ln f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}) \propto \frac{M}{2}\ln\left(\prod_{i=1}^{d_\text{i}} 1 - \left(\hat{k}^{(i)}\right)^2\right), \tag{4.8}$$

where $\hat{k}^{(i)}$ are the sample circularity coefficients of $\mathbf{x}$.

Number of free parameters: Since only the complementary covariance matrix of $\mathbf{x}$, $\widetilde{\mathbf{R}}_{xx}$, depends on d_i, only $\widetilde{\mathbf{R}}_{xx}$ instead of the entire $\underline{\mathbf{R}}_{xx}$ is considered when calculating the number of free parameters. To do this, we perform the Takagi factorization for complex symmetric matrices [76] given as

$$\widetilde{\mathbf{R}}_{xx} = \mathbf{F}\mathbf{K}\mathbf{F}^T. \tag{4.9}$$

Here, $\mathbf{F}$ is a complex unitary matrix that contains the singular vectors, and $\mathbf{K} = \text{diag}(k^{(1)}, k^{(2)}, \ldots, k^{(d_\text{i})}, 0, \ldots, 0)$ contains the d_i non-zero circularity coefficients. Since $\text{rank}(\widetilde{\mathbf{R}}_{xx}) = d_\text{i}$, there are $2nd_\text{i}$ and d_i free parameters in $\mathbf{F}$ and $\mathbf{K}$, respectively. However, not all of these parameters are freely adjustable. There are d_i and $d_\text{i}(d_\text{i}-1)$ constraints on the elements of the singular vectors in $\mathbf{F}$ due to normality and orthogonality, respectively. Therefore,

$$\begin{aligned} C(d_\text{i}) &= 2nd_\text{i} + d_\text{i} - (d_\text{i} + d_\text{i}(d_\text{i}-1)), \\ &= 2nd_\text{i} - d_\text{i}^2 + d_\text{i}. \end{aligned} \tag{4.10}$$

The simplified MDL-ITC expression is thus given as

$$\text{ITC}(d_{\text{i}}) = \frac{M}{2} \ln \left(\prod_{i=1}^{d_{\text{i}}} 1 - \left(\hat{k}^{(i)} \right)^2 \right) + \frac{\ln M}{2} (2nd_{\text{i}} - d_{\text{i}}^2 + d_{\text{i}}). \tag{4.11}$$

4.3.1. Sample poor scenario

Unless the number of samples M is significantly larger than the dimension n, the number of improper components d_{i} cannot be correctly estimated using (4.7) because the sample circularity coefficients $\hat{k}^{(i)}$ are significantly larger than the population circularity coefficients. Since in this chapter we are dealing with the correlation between n-dimensional $\mathbf{x}$ and $\mathbf{x}^*$, when $M < 2n$, at least $2n - M$ sample circularity coefficients are equal to one, independently of the underlying model generating them [35]. This calls for a pre-processing step before or alongside the estimation of d_{i}. We follow the approach of Section 3.5 and use PCA as this pre-processing step.

The rank-r PCA description of $\mathbf{x}$ is

$$\mathbf{x} = \mathbf{U}_r^H \mathbf{x}, \tag{4.12}$$

where $\mathbf{U}_r$ denotes the matrix containing as its columns the first r principal eigenvectors of $\mathbf{R}_{xx}$. Of course, PCA retains the signal components that have maximum variance within the data. These do not necessarily correspond to the most improper signals, which have maximum covariance between $\mathbf{x}$ and $\mathbf{x}^*$. Nevertheless, following the joint PCA-CCA approach explained in Section 3.5, we can choose r large enough to include all the improper signals, while eliminating much of the noise and those proper components whose variance is smaller than that of the weakest improper component. This can be done based on the reduced-rank version of the ITC shown in the following result.

Result 4.1. *The reduced-rank ITC for estimating the number of improper signals in complex-valued data* $\mathbf{X}$ *with PCA rank* r *is*

$$\text{ITC}(d_{\text{i}}, r) = \frac{M}{2} \ln \left(\prod_{i=1}^{d_{\text{i}}} 1 - \left(\hat{k}^{(i)}(r) \right)^2 \right) + \frac{\ln M}{2} (2rd_{\text{i}} - d_{\text{i}}^2 + d_{\text{i}}). \tag{4.13}$$

The circularity coefficients $\hat{k}^{(i)}(r)$ are computed from the rank-r PCA description (4.12) of the data and thus depend on the rank r. They can change significantly depending on how r is chosen. The optimal rank is the one that includes all the improper signal components, but not more than that. The maxmin ITC detector of Section 3.5 allows us to *jointly* choose the optimum rank r and estimate d_{i} number of improper components [28]. The decision rule for

d_{i} is

$$\hat{d}_{\text{i}} = \max_{r=1,\ldots,r_{\max}} \arg\min_{d_{\text{i}}=0,\ldots,r} \text{ITC}(d_{\text{i}}, r), \tag{4.14}$$

and the r that leads to $\hat{d}_{\text{i}}$ is the chosen PCA rank. Here, $r_{\max}$ is the maximum allowable rank and as in Section 3.5, is chosen to be sufficiently smaller than M (typically $\frac{M}{3}$) [28]. This is a much more relaxed condition than requiring n to be sufficiently smaller than M.

It is to be noted that although the decision rule (4.14) corresponds to maxmin ITC detector in (3.34), the expression for $\text{ITC}(d_{\text{i}}, r)$ in this chapter differs because the number of free parameters are different when analyzing correlation between $\mathbf{x}$ and $\mathbf{x}^*$ rather than two different data sets.

4.4. Detector based on GLRT

The number d_{i} can also be estimated by performing a series of binary hypothesis tests [58], [78]. Starting with a counter $s = 0$, each binary test is:

$$\begin{aligned} H_0 &: d_{\text{i}} = s \\ H_1 &: d_{\text{i}} > s \end{aligned} \tag{4.15}$$

If H_0 is rejected, s is incremented and another test of H_0 vs. H_1 is run until H_0 is not rejected or $s = n - 1$. Each binary test is a likelihood ratio test. Since the unknown parameters are replaced by their ML estimates, this leads to a GLRT. The GLR is

$$\eta = \frac{f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}, d_{\text{i}} = s)}{f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}, d_{\text{i}} > s)}, \tag{4.16}$$

where $f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}, d_{\text{i}} = s)$ and $f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}, d_{\text{i}} > s)$ are the ML functions under the null and the alternative hypothesis, respectively.

Since the parameter space for $d_{\text{i}} = n$ is sufficient to parametrize all the possibilities when $d_{\text{i}} > s$, we have

$$f(\underline{\mathbf{X}}|\hat{\underline{\mathbf{R}}}_{xx}, d_{\text{i}} > s) \propto \left(\prod_{i=1}^{n} 1 - \left(\hat{k}^{(i)}\right)^2 \right)^{-\frac{M}{2}}, \tag{4.17}$$

and thus

$$\eta = \left(\prod_{i=s+1}^{n} 1 - \left(\hat{k}^{(i)}\right)^2 \right)^{\frac{M}{2}}. \tag{4.18}$$

According to Wilks' theorem, under H_0 the statistic $W(s) = -2\ln\eta$ is asymptotically χ^2_ν-distributed with d.f. ν equal to the difference between the numbers of free parameters under H_1 and H_0 [59]. Under H_0, the d.f. are given by (4.10). Under H_1, the d.f. are obtained from (4.10) by setting $s = n$. Hence, for $M \to \infty$, $W(s)$ is χ^2_ν-distributed with $\nu = (n-s)(n-s+1)$ d.f.

4.4.1. Sample poor scenario

As discussed in Section 4.3.1, sample poor scenarios require rank reduction to correctly estimate the number of improper signals.

Result 4.2. *A reduced-rank version of the test statistic $W(s)$ is the Box statistic [90] given by*

$$D(s,r) = -(M-r)\ln\left(\prod_{i=s+1}^{r} 1 - \left(\hat{k}^{(i)}\right)^2\right), \tag{4.19}$$

and is asymptotically $\chi^2_{\nu'}$-distributed with $\nu' = (r-1)(r-s+1)$ d.f. under the true H_0, i.e., when $s = d_i$.

The correction term $(M-r)$ in (4.19) provides a better approximation of the $\chi^2_{\nu'}$ distribution than the Wilks statistic for much smaller number of samples [91]. It can be shown numerically that $D(s,r)$ approximately follows a $\chi^2_{\nu'}$ distribution as long as r is large enough to capture all the improper components and is also sufficiently small compared to M (as in Section 3.5, $r < M/3$ seems to work well). A decision rule can thus be formulated as

$$\hat{d}_{\text{i}} = \max_{r=1,\ldots,r_{\max}} \min_{s=0,\ldots,r-1} \{s : D(s,r) < T(s,r)\}, \tag{4.20}$$

where $T(s,r)$ is the threshold chosen to maintain a specified probability of false alarm P_{fa}, which can be obtained from the $\chi^2_{\nu'}$-approximation. This is the PCA-CCA GLRT detector of Section 3.5 specialized to the case of detecting the number of correlated components between $\mathbf{x}$ and $\mathbf{x}^*$. The motivation behind it is similar to the that of (3.36). While the PCA-CCA detector uses a Bartlett-Lawley approximation of the test statistic $W(s)$, the fact that here we are analyzing correlations between $\mathbf{x}$ and $\mathbf{x}^*$ means that the Box statistic with different d.f. needs to be used instead [28] [90].

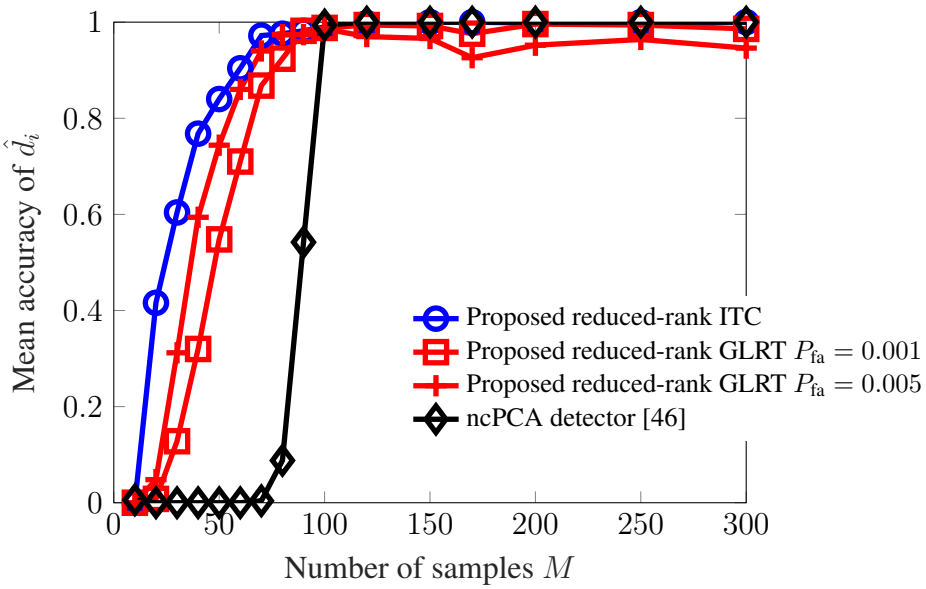
4.5. Numerical results

In this section, we evaluate the performance of the proposed detectors based on ITC and GLRT for the application of sensor array processing. We consider the case when $f = 0$, i.e. the entire signal subspace is improper. This is the scenario used in [87], [88], [92], [93], which show that utilizing the complementary covariance matrix for DOA estimation can lead to significant performance improvement when improper signals such as BPSK-modulated sources impinge on the sensor array. DOA estimation techniques typically assume that the dimension of the signal subspace is known. In practice, this is not the case. If it is known apriori that all sources are improper, then our technique can be employed to find the number of sources.

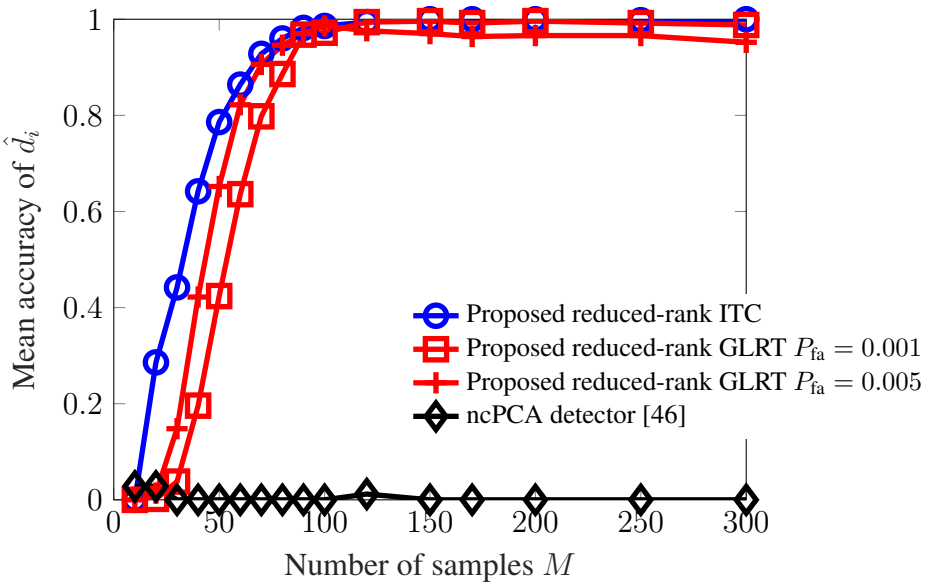
The simulation setup is as follows. We use a uniform linear array with $n = 60$ sensors with half-wavelength inter-sensor spacing. There are 4 far-field, narrowband Gaussian sources that impinge on the array at angles $\Theta = [10^\circ, 15^\circ, 20^\circ, 25^\circ]$. The mth column of the mixing matrix $\mathbf{A}$ is $[1, \exp(j\frac{\pi}{2}\cos(\theta_m)), \ldots, \exp(j\frac{\pi}{2}(n-1)\cos(\theta_m))]^T$ for $m = 1, \ldots, 4$. Each source has variance 5 and the circularity coefficients for the sources are 1, 0.9, 0.8, and 0.6. Two scenarios are presented: i) the additive noise is white and Gaussian with unit variance; ii) the noise is filtered through an autoregressive (AR) filter of order 4 and filter coefficients $[1/2, \sqrt{7}/4, 1/2, 1/4]$. The variance of the noise components before filtering is $1/4$.

We compared the performance of our proposed detectors (4.14) and (4.20) with the ncPCA detector in [46]. Figure 4.1 shows the mean accuracy of $\hat{d}_i$ as a function of the number of samples. For each data point, we ran 500 independent Monte-Carlo trials. For the white noise case shown in subplot (i), all the detectors perform well for a sufficiently large number of samples, but our detectors reach their best performance for a smaller number of samples than the ncPCA detector. In the case of colored noise shown in subplot (ii), the ncPCA detector fails while our detectors continue to work well. This is because the ncPCA detector detects both the proper and improper signal subspaces, and hence must assume white noise to distinguish between signal and noise. Since we only identify improper signal components, we only need to assume proper noise, but it does not have to be white. As far as we know, there is no competing detector that works in colored noise.

The performance of the detector (4.20), which is based on hypothesis tests, depends on P_{fa}. We observe from the plots that, if the number of samples is *large* enough, the detector with *smaller* probability of false alarm $P_{\text{fa}} = 0.001$ performs better than the detector with $P_{\text{fa}} = 0.005$. On the other hand, if the number of samples is *small*, the detector with *larger*



(i)



(ii)

Figure 4.1.: Mean accuracy of correctly detecting $d_{\text{i}} = 4$ improper signal components for the proposed detectors and the ncPCA detector in [46] when i) the additive noise is white Gaussian ii) the additive noise is colored AR(4) Gaussian.

P_{fa} performs better. The advantage of the ITC-based detector (4.14) is that it does not require choosing a value for P_{fa}. It automatically does the trade-off between underfitting and overfitting of the detector. Nevertheless, this does not guarantee that the ITC-based detector will always outperform the hypothesis-test-based detector in every scenario.

4.6. Summary

We have presented two techniques, based on ITC and hypothesis testing, for detecting the dimension of the improper signal subspace in high-dimensional complex data with additive noise. There is no assumption made on the structure of the covariance matrix of the noise, and we have shown that the proposed detectors work well even in the presence of colored noise. We have also introduced reduced-rank detectors, which work reliably even for small number of samples.

Part III.

Model selection in multiple data sets

5. Model order selection in multiple data sets

In this chapter, we address the problem of estimating the model order that identifies the number of signals correlated across all data sets. We present two different techniques for estimating the model order. The first technique assumes a special correlation structure among the underlying components, which enables to derive the GLR and its distribution in closed form. This technique is further extended for high-dimensional data sets with a small number of samples, where the PCA rank and the model order are jointly determined. The second technique works for arbitrary correlation structure and employs bootstrap to estimate the unknown distribution of the test statistic under the null hypothesis [1].

5.1. Introduction

When dealing with two or more data sets, one particularly important model-order selection problem is to detect the dimension of the subspace common across multiple data sets. The main challenge when dealing with multiple data sets is the number of possible correlation structures among the latent signals. For two data sets, this problem does not arise as the two data set scenario can be reduced without loss of generality to the pairwise correlation structure between the latent signals. Thus, the individual signals are either correlated or uncorrelated between the data sets. That is not possible in the multiple-data set case. Even if we assume that signals are correlated in a pairwise fashion, there are numerous combina-

[1] Section 5.3 of this chapter is based on the paper: "Detecting the dimension of the subspace correlated across multiple data sets in the sample poor regime, T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, *IEEE Signal Processing Workshop*, 2016". Section 5.4 of this chapter is based on the paper: "Bootstrap-based Detection of the Number of Signals Correlated Across Multiple Data Sets, T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, *Asilomar conference on Signals, Systems and Computers*, 2016".

tions how the signals can be correlated. Some signals might be independent among the data sets, while some are shared only among a subset of all data sets. There can also be signals correlated across all the data sets. This can be seen in Figure 5.1, which is very similar to Figure 1.1 introduced in Chapter 1. In this example, the first component is correlated across all pairs of data sets. The second components of the first and second data sets, and of the second and third data sets are correlated. The third and fourth components are each pairwise correlated. All other components are uncorrelated. For more than three data sets, the number of pairwise combinations combinatorially increases and illustrating the correlation structure using a figure similar to Figure 5.1 becomes complicated. In this chapter, we are interested in those signals that are common or correlated across all data sets, i.e., the signals indicated with red arrows in Figure 5.1.

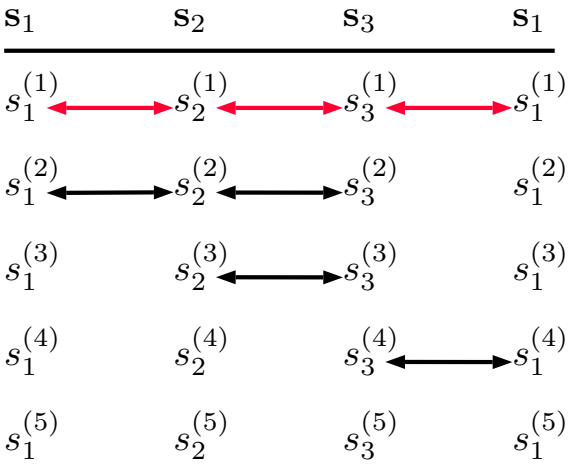


Figure 5.1.: Example of a correlation structure for three data sets. Arrows indicate correlated components, and red arrows indicate components correlated across all data sets.

In the literature, model-order selection for multiple data sets has not yet received the attention that it deserves. While the problem with two data sets has been dealt with in numerous works, e.g., [25]–[28], [37], [38], only a few studies have addressed this for multiple data sets [13], [30], [34], [94]. The paper [94] used an ad hoc approach for detecting the number of sources in multiple arrays. A detection technique based on ITC was derived in [30] using a similar data model as in [94].

In this chapter, we first propose and investigate a GLRT-based technique for detecting the number of correlated source signals in multiple data sets. Like [30], we assume that the number of correlated signals is the same for any pair of data sets. This enables us to derive the GLR and its distribution, and extend them to the case when the number of samples is comparable to or even smaller than the dimension of the data sets. In the second part,

we address the model-order selection problem without assuming any particular correlation structure between the components. We show that the rank of the product of coherence matrices (normalized cross-covariance matrices) of all possible pairs of data sets is equal to the number of signals correlated between all the data sets, provided the SNR is sufficiently large. We then employ bootstrap-based hypothesis testing to estimate the rank of this product of coherence matrices.

5.2. Data model for multiple data sets

We consider P data sets consisting of zero-mean, real-valued random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_P$ with dimensions $n_1, \ldots, n_P$, respectively. Without loss of generality, it is assumed that $n_1 \leq n_2 \leq \ldots \leq n_P$. The data sets are generated by an unknown linear mixing of underlying real-valued signal component vectors $\mathbf{s}_1, \ldots, \mathbf{s}_P$ with uncorrelated additive noise. The generating data model is

$$\mathbf{x}_p = \mathbf{A}_p \mathbf{s}_p + \mathbf{n}_p, \quad p = 1, 2, \ldots, P, \tag{5.1}$$

where $\mathbf{A}_p \in \mathbb{R}^{n_p \times m_p}$ is an unknown but fixed mixing matrix with full column rank. The noise vector $\mathbf{n}_p \in \mathbb{R}^{n_p}$ is zero-mean and uncorrelated with the signal vectors and also with the noise vectors of other data sets. The signal vector $\mathbf{s}_p$ contains $m_p (\leq n_p)$ signal components. The ith signal component of the pth data set is denoted by $s_p^{(i)}$. These components are assumed to be zero-mean and unit variance without loss of generality, i.e.,

$$E[s_p^{(i)}] = 0, \qquad \text{and} \tag{5.2}$$

$$E[(s_p^{(i)})^2] = 1, \qquad \text{for } i = 1, \ldots, m_p. \tag{5.3}$$

In the multiset model defined in (5.1), it is common to assume two kinds of association among the signal components:

1. Intraset independence: signal components within each data set are uncorrelated, i.e.,

$$\mathbf{R}_{s_p s_p} = E[\mathbf{s}_p \mathbf{s}_p^T] = \mathbf{I}, \tag{5.4}$$

   where $\mathbf{I}$ is an identity matrix, and

2. Interset dependence: between any two data sets p and q, components may be correlated only pairwise, i.e., component $s_p^{(i)}$ may only correlate with component $s_q^{(i)}$ for $1 \leq i \leq \min(m_p, m_q)$. This means, the signal cross-covariance matrix between data sets p and

q ($p \neq q$) is

$$\mathbf{R}_{s_p s_q} = \text{diag}(\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(\min(m_p, m_q))}), \tag{5.5}$$

where $\rho_{pq}^{(i)}$ represents the unknown (possibly zero) correlation coefficient between their ith components.

When analyzing the correlations between two data sets only, $\mathbf{R}_{s_p s_q}$ can be assumed to be diagonal without loss of generality. For more than two sets, diagonal cross-covariance matrices are a restriction on the problem, as they do not represent all possible correlation structures. However, this assumption is common in the literature since it makes the multiset correlation structure uniquely identifiable based on observations of linear mixtures [2], [18], [23], [42], [95]. The noise covariance matrix of the pth data set,

$$\mathbf{R}_{n_p n_p} = E[\mathbf{n}_p \mathbf{n}_p^T], \qquad \text{for } p = 1, \ldots, P, \tag{5.6}$$

is unknown and not necessarily white. However, noise vectors of any two data sets are assumed to be uncorrelated, i.e., $E[\mathbf{n}_p \mathbf{n}_q^T] = \mathbf{0}$, for $p \neq q$.

There is an unknown number d_{pq} of components correlated between the pth and qth signal vectors corresponding to the d_{pq} nonzero entries of $\mathbf{R}_{s_p s_q}$. An unknown number of d_{all} components are correlated across all the data sets, i.e., $d_{\text{all}} = |\{i : \rho_{pq}^{(i)} \neq 0 \; \forall p, q\}|$. The goal of this chapter is the following:

Goal: *Given M i.i.d. samples of $\mathbf{x}_1, \ldots, \mathbf{x}_P$ from the model in* (5.1)*, determine the number d_{all} of correlated components.*

We determine d_{all} using two different correlation structures as follows:

- In Section 5.3, we assume a special correlation structure where $d_{pq} = d_{\text{all}}$ $\forall (p, q) \in \{1, \ldots, P\}$, $p \neq q$.
- In Section 5.4, we do not make any assumption on the correlation structure, i.e., we allow d_{pq} to be different for different pairs of data sets and also different than d_{all}.

5.3. Order selection with special correlation structure

In this section, GLRT-based detectors for d_{all} are derived for both sample-rich and sample-poor regimes by assuming a special correlation structure between the latent signal components. Fig. 5.2 illustrates one such example for three data sets. The first two components are correlated among all pairs of data sets. Therefore, $d_{12} = d_{23} = d_{31} = d_{\text{all}} = 2$.

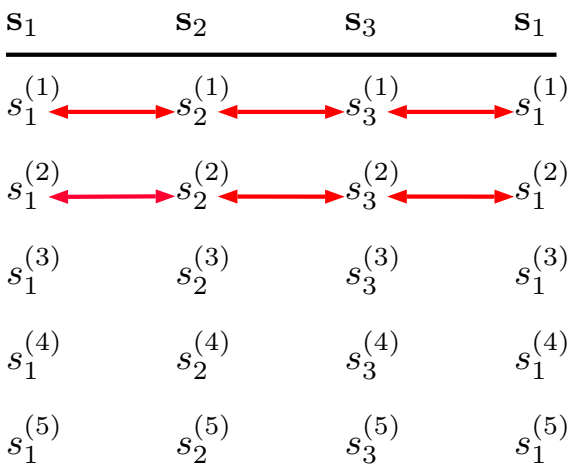


Figure 5.2.: Example for three data sets with the special correlation structure. Arrows indicate correlated components. Here, $d_{12} = d_{23} = d_{31} = d_{\text{all}} = 2$.

Let us revisit the composite data vector $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T$ with composite covariance matrix

$$\mathbf{R} = E[\mathbf{x}\mathbf{x}^T] = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1P} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{P1} & \mathbf{R}_{P2} & \cdots & \mathbf{R}_{PP} \end{bmatrix}, \tag{5.7}$$

where

$$\mathbf{R}_{pp} = \mathbf{A}_p \mathbf{A}_p^T + \mathbf{R}_{n_p n_p}, \tag{5.8}$$

$$\mathbf{R}_{pq} = \mathbf{A}_p \mathbf{R}_{s_p s_q} \mathbf{A}_q^T. \tag{5.9}$$

Under the assumption that $d_{pq} = d_{\text{all}}\ \forall(p, q)$,

$$\text{rank}(\mathbf{R}_{pq}) = d_{\text{all}}, \qquad \forall(p, q) \in \{1, \ldots, P\},\ p \neq q. \tag{5.10}$$

Similarly, from the definition of coherence matrix of two data sets in (3.1),

$$\text{rank}(\mathbf{C}_{pq}) = d_{\text{all}}, \qquad \forall(p, q) \in \{1, \ldots, P\},\ p \neq q. \tag{5.11}$$

Thus, d_{all} canonical correlations between all pairs of data sets are non-zero. In this case, d_{all} can also be estimated using any pair of data sets. However, in the upcoming sections, we will derive a method to estimate d_{all} jointly from all P data sets and show that there is a

significant performance improvement compared to the techniques in Chapter 3, which work with only a pair of data sets.

Consider M i.i.d. samples of the composite data vector $\mathbf{x}$, arranged as the M columns of the data matrix $\mathbf{X} = \Big[\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(M)\Big]$. We first introduce a GLRT for multiple data sets in the sample rich regime in Section 5.3.1, and then propose its reduced-rank version for small sample support in Section 5.3.2.

5.3.1. Sample rich regime

To estimate d_{all}, we use a sequence of binary hypothesis tests starting with $s = 0$ and performing the test

$$\begin{aligned} H_0 &: d_{\text{all}} = s, \\ H_1 &: d_{\text{all}} > s, \end{aligned} \tag{5.12}$$

and increment s until H_0 is not rejected. Under the assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_P$ are jointly Gaussian-distributed, $\mathbf{x}$ is also Gaussian distributed with zero-mean and covariance matrix $\mathbf{R}$. The GLR for the test in (5.12) is

$$\eta = \frac{f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} = s)}{f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} > s)}, \tag{5.13}$$

where $f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} = s)$ is the ML function of $\mathbf{X}$ under H_0 and $f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} > s)$ is the ML function of $\mathbf{X}$ under H_1. The ML function $f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} = s)$ under the constraints (5.10) is derived in [30]. It is the product of $P - 1$ functions and is given by

$$f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} = s) \propto \left(\prod_{p=1}^{P-1} \prod_{i=1}^{s} 1 - \left(\hat{k}^{(i)}(p)\right)^2 \right)^{-\frac{M}{2}}. \tag{5.14}$$

The term with $p = 1$ can be interpreted as the ML with respect to the first data set $\mathbf{X}_1$ and the remaining data sets $\mathbf{Z}_1 = [\mathbf{X}_2^T, \ldots, \mathbf{X}_P^T]^T$ and is a function of their sample canonical correlations denoted as $\hat{k}^{(i)}(1)$. The maximization of the likelihood function is performed under the constraint that $\text{rank}(E[\mathbf{x}_1 \mathbf{z}_1^T]) = \text{rank}([\mathbf{R}_{12}, \ldots, \mathbf{R}_{1P}]) = s$. The term with $p = 2$ is similarly a function of the sample canonical correlations $\hat{k}^{(i)}(2)$ between the second data set $\mathbf{X}_2$ and $\mathbf{Z}_2 = [\mathbf{X}_3^T, \ldots, \mathbf{X}_P^T]^T$, and so on.

Since the parameter space $s = n_p$ is sufficient to parameterize all the possibilities in $d_{\text{all}} > s$,

the ML function under H_1 is

$$f(\mathbf{X}|\hat{\mathbf{R}}, d_{\text{all}} > s) \propto \left(\prod_{p=1}^{P-1} \prod_{i=1}^{n_p} 1 - \left(\hat{k}^{(i)}(p) \right)^2 \right)^{-\frac{M}{2}}. \tag{5.15}$$

Using (5.14) and (5.15), η can be simplified to

$$\eta = \left(\prod_{p=1}^{P-1} \prod_{i=s+1}^{n_p} 1 - \left(\hat{k}^{(i)}(p) \right)^2 \right)^{\frac{M}{2}}. \tag{5.16}$$

Bartlett statistic - According to Wilks' theorem, the statistic $W(s) = -2 \ln \eta$ is asymptotically χ^2_ν distributed when H_0 is true, i.e. $s = d_{\text{all}}$ [59]. Since the GLR is the product of $P-1$ functions of pairwise canonical correlations, the d.f. for the χ^2_ν distribution can also be expressed as the sum of $P-1$ d.f. for two data sets. Using (3.24),

$$\nu = \sum_{p=1}^{P-1} (a_1(p) - s)(a_2(p) - s). \tag{5.17}$$

Here, $a_1(p)$ and $a_2(p)$ depend on the dimensions of the data sets used to compute the pth set of canonical correlations. For $p = 1$, $a_1(1) = n_1$ and $a_2(1) = n_2 + n_3 + \ldots + n_P$. Similarly, for $p = 2$, $a_1(2) = n_2$ and $a_2(2) = n_3 + n_4 + \ldots + n_P$ and so on. For small sample size, Bartlett's statistic [67] provides a better approximation of the χ^2_ν distribution. After applying Bartlett's correction to each of the $P-1$ terms in (5.16), we obtain the following result.

Result 5.1. *The Bartlett statistic for testing s components correlated across all P data sets is*

$$B(s) = -\sum_{p=1}^{P-1} \left(M - \frac{(a_1(p) + a_2(p) + 1)}{2} \right) \ln \left(\prod_{i=s+1}^{a_1(p)} 1 - \left(\hat{k}^{(i)}(p) \right)^2 \right), \tag{5.18}$$

and is asymptotically χ^2_ν-distributed under H_0 with ν computed in (5.17).

5.3.2. Sample poor regime

As we have seen in Chapter 3, for any pair of data sets $\mathbf{x}_p$ and $\mathbf{x}_q$, when the number of samples is smaller than the sum of their corresponding dimensions, i.e., $M < n_p + n_q$, at least $n_p + n_q - M$ sample canonical correlations will be equal to one, irrespective of the model from which $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated [35]. Moreover, even if $M > n_p + n_q$, but not

significantly greater, the sample canonical correlations significantly overestimate the true canonical correlations. This problem further exacerbates in multiple data sets since the GLR in (5.16) is the function of the canonical correlations computed by concatenating different data sets. As we have assumed $n_1 \leq \ldots \leq n_P$, the defective regime in this case is defined by the term $p = 1$ which computes the canonical correlations between $\mathbf{x}_1 \in \mathbb{R}^{a_1(1)}$ and $\mathbf{z}_1 \in \mathbb{R}^{a_2(1)}$ where $a_1(1) = n_1$ and $a_2(1) = n_2 + \ldots + n_P$. Let $n_{\text{total}} = \sum_{p=1}^{P} n_p$. In this case, when $M < n_{\text{total}}$, at least $n_{\text{total}} - M$ number of sample canonical correlations are equal to one. Therefore, it can be seen that for a fixed M, adding another data set to the analysis leads to a further overestimation of the sample canonical correlations. This calls for rank reduction either before or alongside the detection of the number of correlated signals.

When the signal vectors contain only the correlated components or in the case when all the independent components have smaller variance compared to the correlated components, then the same PCA rank can be applied to all the data sets. This rank is denoted by r. In this case, r will be sufficient to keep all the correlated components and the reduced-rank version of Bartlett's statistic for multiple data sets in (5.18) is

$$B(s,r) = -\sum_{p=1}^{P-1} \left(M - \frac{(2r+1)}{2} \right) \ln \left(\prod_{i=s+1}^{r} 1 - \left(\hat{k}^{(i)}(r,p) \right)^2 \right). \tag{5.19}$$

Here, the PCA rank r is applied in the p^{th} term to the data sets $\mathbf{X}_p$ and $\mathbf{Z}_p = [\mathbf{X}_{p+1}^T, \ldots, \mathbf{X}_P^T]^T$, and the source counter s can take values from 0 to $r-1$. Similar to $B(s)$, $B(s,r)$ is also a sum of $P-1$ statistics for two data sets. The first statistic for $p = 1$ involves the sample canonical correlations of the rank-reduced versions of $\mathbf{X}_1$ and $\mathbf{Z}_1$. Similarly, the second statistic for $p = 2$ in $B(s,r)$ involves the sample canonical correlations of the rank-reduced $\mathbf{X}_2$ and $\mathbf{Z}_2$, and so forth.

However, when there is different number of independent components with stronger variance compared to the correlated components in different data sets, reducing all the data sets to rank r is suboptimal. In this case, the statistic,

$$B(s, r_1, \widetilde{r}_1, r_2, \widetilde{r}_2, \ldots, r_{P-1}, \widetilde{r}_{P-1}) = -\sum_{p=1}^{P-1} \left(M - \frac{(r_p + \widetilde{r}_p + 1)}{2} \right) \times \ln \left(\prod_{i=s+1}^{\min(r_p, \widetilde{r}_p)} 1 - \left(\hat{k}^{(i)}(r_p, \widetilde{r}_p, p) \right)^2 \right), \tag{5.20}$$

is optimal where every data set has a different PCA rank. Here, r_p and $\widetilde{r}_p$ are the PCA

ranks applied to $\mathbf{X}_p$ and $\mathbf{Z}_p$, respectively. However, the statistic in (5.20) depends on $2P-1$ parameters and does not computationally scale well even for small P. For this reason, we will use the statistic $B(s,r)$ and numerically show in Section 5.5 that our proposed detector with $B(s,r)$ performs well even when different data sets contain different number of stronger independent components.

Using (5.17), $B(s,r)$ is $\chi^2_{\nu'}$-distributed with

$$\nu' = (P-1)(r-s)^2. \tag{5.21}$$

Thus, the reduced-rank version of Result 5.1 is as follows.

Result 5.2. *The reduced-rank Bartlett statistic for testing s components correlated across all P rank-reduced data sets each with PCA rank of r is*

$$B(s,r) = -\sum_{p=1}^{P-1}\left(M - \frac{(2r+1)}{2}\right)\ln\left(\prod_{i=s+1}^{r} 1 - \left(\hat{k}^{(i)}(r,p)\right)^2\right), \tag{5.22}$$

and is asymptotically $\chi^2_{\nu'}$ distributed under H_0 with ν' computed in (5.21)*.*

It was shown in Section 3.5 that under the null hypothesis $s = d_{\text{all}}$, each of these $P-1$ statistics is approximately $\chi^2_{\tilde{\nu}}$-distributed with $\tilde{\nu} = (r-s)^2$, as long as the PCA rank r is large enough to capture all correlated components, yet sufficiently smaller than the number of samples M (this is typically the case when $r < M/3$). As long as M is large with respect to r, these statistics are close to the $\chi^2_{\tilde{\nu}}$ distribution. Under the same conditions, $B(s,r)$ is also approximately $\chi^2_{\nu'}$-distributed with ν' d.f. computed in (5.21). The $\chi^2_{\nu'}$ approximation for $B(s,r)$ in sample-poor regime is demonstrated in Figure 5.3 for an example of four data sets with $n_1 = 40$, $n_2 = 50$, $n_3 = 55$ and $n_4 = 60$, and $M = 80$ samples. The number of correlated sources is $d_{\text{all}} = 4$, all of which have equal signal power. The correlation coefficients $\rho^{(1)}_{pq}, \rho^{(2)}_{pq}, \rho^{(3)}_{pq}, \rho^{(4)}_{pq}$ are chosen as 0.9, 0.9, 0.8 and 0.7, respectively, for all data sets. The noise is white with small power compared to the signal power. The empirical distribution of $B(s,r)$ is shown in Figure 5.3 along with the $\chi^2_{\nu'}$ distribution for $s = 4$ and $r = 5$. It can be seen that when H_0 is true, i.e. $s = d_{\text{all}} = 4$, $B(s,r)$ closely follows the $\chi^2_{\nu'}$ distribution.

This means that in a series of binary tests of H_0 vs H_1 based on $B(s,r)$, d_{all} is generally not overestimated. It is likely, however, to be underestimated if r is not chosen large enough. If r is too small, then the reduced-rank PCA descriptions do not capture all of the correlated components and thus the series of binary tests decides for too small a dimension d_{all}. This

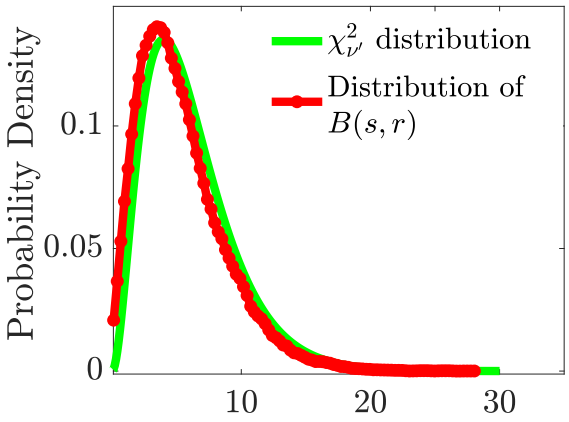


Figure 5.3.: Empirical distribution of $B(s,r)$ for $s = d_{\text{all}} = 4$ and $r = 5$.

reasoning leads to the decision rule

$$\hat{d}_{\text{all}} = \max_{r=1,\ldots,r_{\max}} \min_{s=0,\ldots,r-1} \{s : B(s,r) < T(s,r)\}, \tag{5.23}$$

and the r that leads to $\hat{d}_{\text{all}}$ is the PCA rank. In (5.23), $r_{\max}$ should be chosen sufficiently smaller than $M/2$ (typically $M/3$) and $T(s,r)$ is the threshold value chosen to maintain a specific probability of false alarm P_{fa}. The min-operator choses the smallest s such that $B(s,r) < T(s,r)$. If there is no such s, it chooses $s = r$. More details about the detector in (5.23) for the case of two data sets are provided in Section 3.5.

5.4. Order selection with arbitrary correlation structure

In this section, we do not make any assumption on the correlation structure among the components in the data sets defined in Section 5.2. More precisely, we allow d_{pq} to be different for different pairs of data sets and also different than d_{all}. Let us now consider the scenario in Figure 5.1, where the correlation structure does not satisfy the assumption made in Section 5.3. In this case, $d_{12} = 2$, $d_{23} = 3$, $d_{31} = 2$ and $d_{\text{all}} = 1$. Thus, the technique proposed in Section 5.3, which assumes the special correlation structure where $d_{pq} = d_{\text{all}}\ \forall p, q$, will not work. As the number of data sets increases, the number of possible correlation structures increases rapidly, which makes it clear that we require a detector that works for an arbitrary correlation structure.

In the upcoming subsections, we propose a novel technique to estimate d_{all}. In Section 5.4.1, we will show that the rank of the product of coherence matrices of all possible pairs of data

sets is equal to d_{all}, provided the SNR is sufficiently large. The problem thus comes down to estimating the rank of this product of matrices. For this, we employ a standard procedure based on a series of binary hypothesis tests [58], [78]. Since the distribution of the utilized test statistic under the null hypothesis is difficult to derive analytically, we estimate it using the bootstrap technique in Section 5.4.2.

5.4.1. Product of coherence matrices

The motivation behind the proposed method is that the rank of the product of all possible signal cross-covariance matrices is equal to the number d_{all} of correlated sources,

$$\text{rank}\left(\prod_{p,q}\mathbf{R}_{s_p s_q}\right) = d_{\text{all}}. \tag{5.24}$$

Here, the indices p and q are chosen in such a way that all $\frac{P(P-1)}{2}$ signal cross-covariance matrices are considered at least once, and the dimensions of the matrices match. The procedure of generating the indices p, q to obtain this product for an arbitrary number of data sets is explained in detail in the appendix in Section 5.7. For instance, for three data sets with correlation structure shown in Figure 5.1, where $d_{\text{all}} = 1$,

$$\text{rank}\left(\mathbf{R}_{s_1 s_2}\mathbf{R}_{s_2 s_3}\mathbf{R}_{s_3 s_1}\right) = 1.$$

However, the true signals are unobservable and d_{all} has to be estimated from the observed data. Let us introduce the following result.

Result 5.3. *The rank of the product of coherence matrices for all possible pairs of data sets is equal to d_{all}, provided that the SNR is large enough. That is,*

$$\textit{rank}\left(\prod_{p,q}\mathbf{C}_{pq}\right) \approx d_{all}. \tag{5.25}$$

Here the indices p and q are chosen as described by the procedure in Section 5.7, and the approximation in (5.25) holds for large SNR values. We will now prove Result 5.3 for three data sets. For more than three data sets, the proof can be trivially extended. Consider the

product of three coherence matrices $\mathbf{C}_{12}$, $\mathbf{C}_{23}$ and $\mathbf{C}_{31}$,

$$\begin{aligned}\mathbf{C}_{123} &= \mathbf{C}_{12}\mathbf{C}_{23}\mathbf{C}_{31} \\ &= \mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{R}_{12}\mathbf{R}_{22}^{-\frac{1}{2}}\mathbf{R}_{22}^{-\frac{1}{2}}\mathbf{R}_{23}\mathbf{R}_{33}^{-\frac{1}{2}}\mathbf{R}_{33}^{-\frac{1}{2}}\mathbf{R}_{31}\mathbf{R}_{11}^{-\frac{1}{2}} \\ &= \mathbf{R}_{11}^{-\frac{1}{2}}\mathbf{A}_1\mathbf{R}_{s_1s_2}\underbrace{\mathbf{A}_2^T\mathbf{R}_{22}^{-1}\mathbf{A}_2}_{\mathbf{P}}\mathbf{R}_{s_2s_3}\underbrace{\mathbf{A}_3^T\mathbf{R}_{33}^{-1}\mathbf{A}_3}_{\mathbf{Q}}\mathbf{R}_{s_3s_1}\mathbf{A}_1^T\mathbf{R}_{11}^{-\frac{1}{2}}. \end{aligned} \tag{5.26}$$

The cross-covariance matrices $\mathbf{R}_{12}$, $\mathbf{R}_{23}$, and $\mathbf{R}_{31}$ do not include any noise terms as the noise is uncorrelated with the signals and also between any two data sets according to our modelling assumptions in (5.2). Let us expand the expression for matrix $\mathbf{P}$ using (5.8) as

$$\begin{aligned}\mathbf{P} &= \mathbf{A}_2^T\mathbf{R}_{22}^{-1}\mathbf{A}_2 \\ &= \mathbf{A}_2^T(\mathbf{A}_2\mathbf{A}_2^T + \mathbf{R}_{n_2n_2})^{-1}\mathbf{A}_2. \end{aligned} \tag{5.27}$$

Applying the matrix inversion lemma [96],

$$\begin{aligned}\mathbf{P} &= \mathbf{A}_2^T(\mathbf{R}_{n_2n_2}^{-1} - \mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2(\mathbf{I} + \mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1})\mathbf{A}_2 \\ &= \mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2 - \mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2(\mathbf{I} + \mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2)^{-1}\mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2. \end{aligned} \tag{5.28}$$

Let $\mathbf{B} = \mathbf{A}_2^T\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2$. Therefore,

$$\begin{aligned}\mathbf{P} &= \mathbf{B} - \mathbf{B}(\mathbf{I} + \mathbf{B})^{-1}\mathbf{B} \\ &= \mathbf{B} - \mathbf{B}(\mathbf{I} + \mathbf{B})^{-1}(\mathbf{B} + \mathbf{I} - \mathbf{I}) \\ &= \mathbf{B} - \mathbf{B}(\mathbf{I} - (\mathbf{I} + \mathbf{B})^{-1}) \\ &= \mathbf{B} - \mathbf{B} + \mathbf{B}(\mathbf{I} + \mathbf{B})^{-1} \\ &= \mathbf{B}(\mathbf{I} + \mathbf{B})^{-1}. \end{aligned} \tag{5.29}$$

Typically the matrix $\mathbf{B} \gg \mathbf{I}$ when the signal to noise ratio is high. Then, $(\mathbf{I} + \mathbf{B})^{-1} \approx \mathbf{B}^{-1}$, hence,

$$\mathbf{P} \approx \mathbf{I}. \tag{5.30}$$

Using the same derivation, it can be shown that

$$\mathbf{Q} \approx \mathbf{I}. \tag{5.31}$$

Inserting the approximate values of $\mathbf{P}$ and $\mathbf{Q}$ in (5.26), we get

$$\mathbf{C}_{123} \approx \mathbf{R}_{11}^{-\frac{1}{2}} \mathbf{A}_1 \mathbf{R}_{s_1 s_2} \mathbf{R}_{s_2 s_3} \mathbf{R}_{s_3 s_1} \mathbf{A}_1^T \mathbf{R}_{11}^{-\frac{1}{2}}.$$

Since all other matrices are full rank, the rank of $\mathbf{C}_{123}$ will be equal to d_{all} when the approximations in (5.30) and (5.31) apply

$$\text{rank}(\mathbf{C}_{123}) = d_{\text{all}}, \quad \text{when } \mathbf{P} = \mathbf{I} \text{ and } \mathbf{Q} = \mathbf{I}. \tag{5.32}$$

Thus, the singular values of $\mathbf{C}_{123}$ are of the form

$$\gamma^{(1)} \geq \gamma^{(2)} \geq \ldots \geq \gamma^{(d_{\text{all}})} > \gamma^{(d_{\text{all}}+1)} = \ldots = \gamma^{(n_1)} = 0, \tag{5.33}$$

where the d_{all} largest singular values are non-zero and the rest are equal to zero. The matrices $\mathbf{P}$ and $\mathbf{Q}$ will approach identity matrices as the SNR approaches infinity. However, when these approximations are not valid, the rank of $\mathbf{C}_{123}$ will generally be greater than d_{all}.

5.4.2. Hypothesis testing using bootstrap

One approach to estimate d_{all} is to perform a series of binary hypothesis tests as in (5.12). This approach, however, requires a statistic whose (asymptotic) distribution under the null hypothesis is known. In Section 5.3, we derived a GLRT by assuming that the signals correlated between any two data sets are also correlated across the remaining data sets. This assumption makes the problem of maximizing the likelihood under the unknown parameters tractable [30]. However, deriving a GLRT with arbitrary correlation structure among the signals becomes challenging. The bootstrap is a resampling technique that can be used to estimate the distribution of a parameter of interest, particularly when the underlying distribution of the data is unknown or is too complicated to derive [61].

Based on the result in (5.33), the null hypothesis H_0 checks if the first s singular values of the product of coherence matrices are non-zero, i.e.,

$$H_0 : \gamma^{(1)} \geq \gamma^{(2)} \geq \ldots \geq \gamma^{(s)} > \gamma^{(s+1)} = \ldots = \gamma^{(n_1)} = 0 \tag{5.34}$$

A test statistic to check this is the difference between arithmetic and geometric mean [97],

[98]

$$T(s) \quad = \quad \left(\frac{1}{n_1 - s} \sum_{i=s+1}^{n_1} \hat{\gamma}^{(i)} \right) \; - \; \left(\prod_{i=s+1}^{n_1} \hat{\gamma}^{(i)\frac{1}{n_1-s}} \right) \quad \text{for } s = 0, \ldots, n_1 - 1. \quad (5.35)$$

The value of this statistic under the null hypothesis will be close to zero as the sample singular values $\hat{\gamma}^{(i)}$ are close to zero. The distribution of $T(s)$ under the null hypothesis is estimated using the bootstrap. The complete procedure to test H_0 vs H_1 using bootstrap is given in Algorithm 1 on page 24. The estimate of d_{all} is chosen as the minimum value of s for which H_0 is not rejected. If there is no such s, it chooses $\hat{d}_{\text{all}} = n_1$.

5.5. Numerical results

In this section, we evaluate the two techniques proposed in Sections 5.3 and 5.4, and compare their performance with the competing approaches in the literature. We will first show the performance of the techniques in the scenario where the signal components exhibit the special correlation structure where $d_{pq} = d_{\text{all}} \forall p, q$. This scenario is further divided into a sample-rich and a sample-poor experiment to show the effectiveness of the proposed GLRT-based detector in Section 5.3 in both cases. In the second scenario, we generate a correlation structure where the components are also correlated across a subset of data sets. We have compared the proposed techniques with the following competitors: a) joint PCA-CCA technique for two data sets [28] reviewed in Chapter 3, b) ITC for multiple data sets of [30] and c) MCCA knee-point detector (KPD) of [34].

The simulation setup common to all scenarios is explained below. $P = 4$ data sets are considered. The number of samples, $M = 350$ unless otherwise stated. There are $n_p = 7$ Gaussian distributed signals in all the sets out of which two are two uncorrelated signals. The variance of the correlated and uncorrelated signals is one and two, respectively. The correlation coefficients for the correlated signals are $\{0.9, 0.85, 0.8, 0.75, 0.7\}$. Each data set is corrupted by additive Gaussian colored noise. The noise is colored by applying the second order autoregressive (AR) filter with filter coefficients, $[1, 0.33]$, to the white noise. The variance of the white noise, σ_n^2 is chosen according to the SNR defined in (3.37). The results are averaged over 500 independent Monte-Carlo trials.

i. **Special correlation structure:**

   A. **Sample-rich regime-** In this case, the dimensions $n_1 = 10, n_2 = 15, n_3 = 15, n_4 =$

20 are chosen. There are $d_{\text{all}} = 5$ correlated signals. Figure 5.4 shows the mean accuracy as a function of SNR. Since $d_{pq} = d_{\text{all}}$, PCA-CCA detectors for two data sets can also be used to estimate d_{all}. To show a fair comparison, we select the first two data sets with the smallest dimensions for the PCA-CCA detectors. However, the proposed multiset GLRT technique performs better than the PCA-CCA detectors since the former incorporates the joint correlation information in all data sets and not just from any two sets. The ITC multiset technique of [30] also works under the assumption of special correlation structure but requires higher SNR than all the other techniques to estimate d_{all} with high accuracy. The proposed product of coherence matrices (PCM) detector also estimates d_{all} accurately but requires a higher SNR compared to the multiset GLRT and PCA-CCA detectors. The MCCA-KPD technique of [34] performs well at low SNR but is not consistent at medium and high SNR values. However, it should be noted that the proposed PCM technique and the MCCA- KPD technique of [34] do not assume an apriori correlation structure among the signal components.

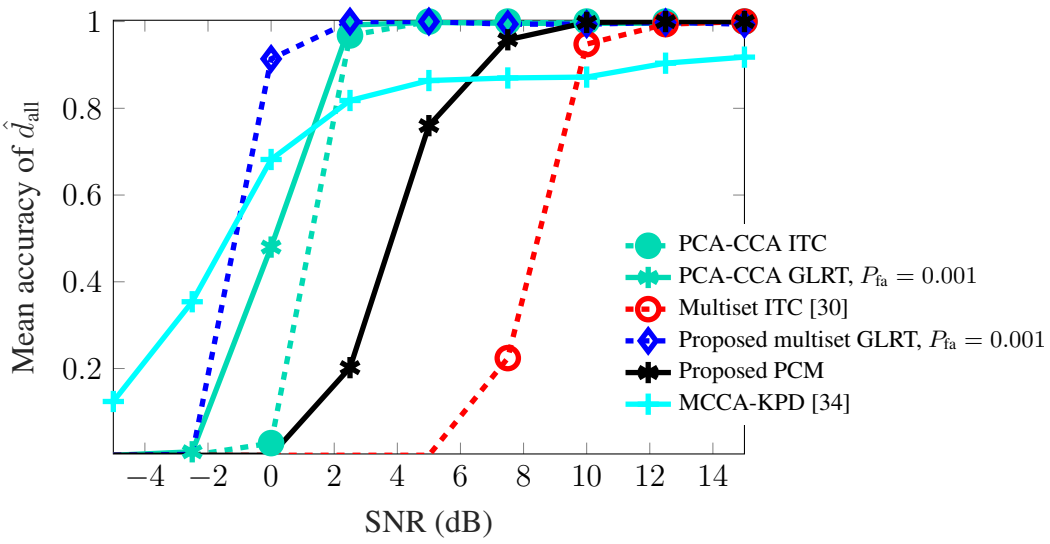


Figure 5.4.: Mean accuracy of $\hat{d}_{\text{all}}$ as a function of SNR for four data sets in scenario i)A. There are $d_{\text{all}} = 5$ components correlated across all five sets.

B. **Sample-poor regime-** In this setup, the data sets are high-dimensional with $n_1 = n_2 = n_3 = n_4 = 40$. The mean accuracy of d_{all} as a function of M is shown in Figure 5.5. The SNR is 5dB. In this case, the proposed multiset GLRT and the PCA-CCA detectors perform well as they are specialized for the sample-poor regime, while the other techniques do not work. The multiset GLRT technique significantly outperforms the PCA-CCA detectors in this case. It is able to estimate d_{all} accurately

even in the defective regime when $M < n_{\text{total}}$, where $n_{\text{total}} = 160$ in this case.

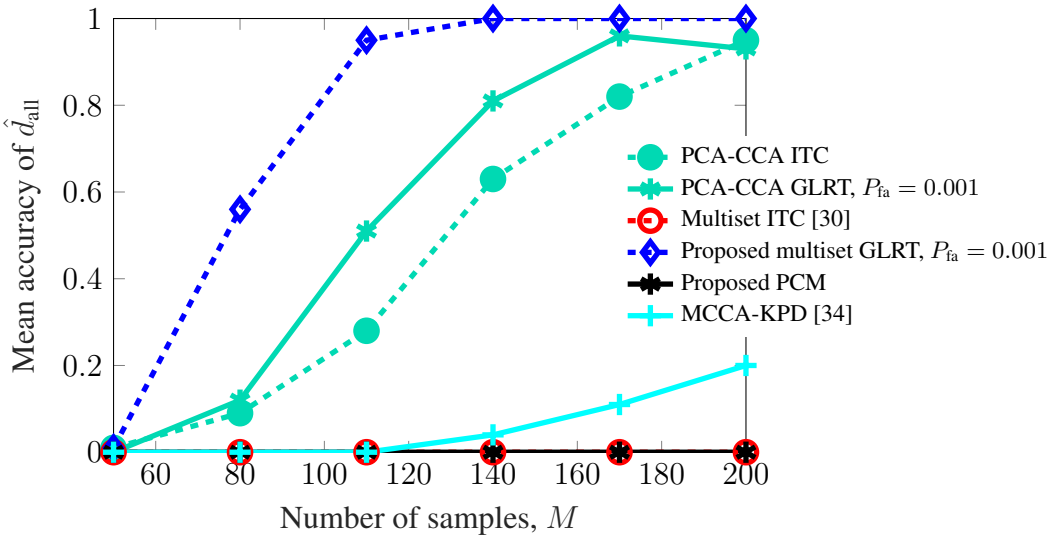


Figure 5.5.: Mean accuracy of $\hat{d}_{\text{all}}$ as a function of M for four data sets in scenario i)B. The data set dimensions are $n_1 = n_2 = n_3 = n_4 = 40$. There are $d_{\text{all}} = 5$ components correlated across all five sets.

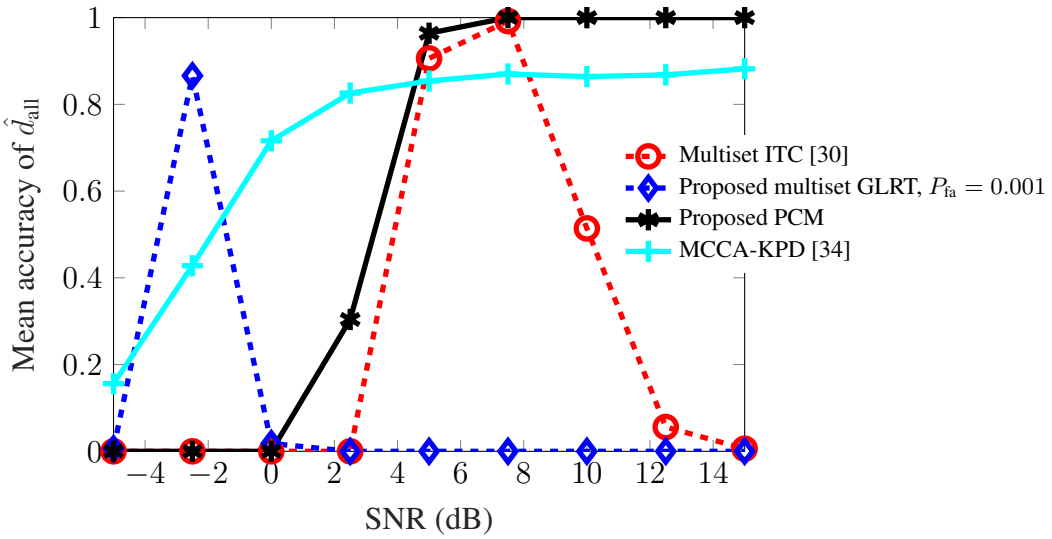


Figure 5.6.: Mean accuracy of $\hat{d}_{\text{all}}$ as a function of SNR for five data sets in scenario ii). There are $d_{\text{all}} = 2$ components correlated across all five sets. The other three correlated components are only correlated across a subset of data sets.

ii. **Arbitrary correlation structure:** In this scenario, there are some components that are correlated only across a subset of data sets. The first two components are correlated across all data sets, thus $d_{\text{all}} = 2$. The third component is correlated across three sets and the last two components are correlated across two sets only. Thus the assumption of special correlation structure is violated. In this case, the multiset GLRT and ITC

techniques cannot accurately estimate d_{all} as can be seen in Figure 5.6. The peaks in the figure can be explained using Figure 5.7, which shows the mean values of $\hat{d}_{\text{all}}$. For small SNR values, these two techniques underestimate d_{all}. As the SNR increases, the mean value of their estimates also increases and they coincidentally pick the correct d_{all} for a certain SNR. However, as the SNR further increases, they overestimate d_{all}. In contrast, the proposed PCM and MCCA-KPD [34] techniques estimate d_{all} correctly for medium and high SNR values. The MCCA-KPD however, overestimates d_{all} on average even for high SNR values, explaining its lower accuracy compared to the proposed PCM technique.

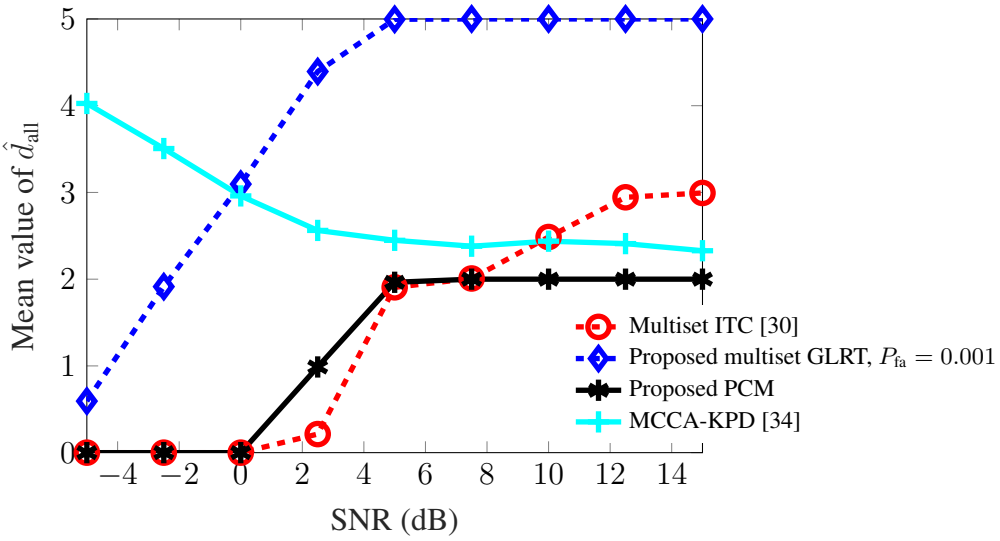


Figure 5.7.: Mean value of $\hat{d}_{\text{all}}$ as a function of SNR for five data sets in scenario ii).

5.6. Summary

In this chapter, two techniques for estimating the number of components correlated across multiple data sets have been presented. The first technique assumes a special correlation structure which leads to a GLRT-based solution. It is also modified for the sample-poor regime. The second technique is based on bootstrap and does not make any assumption on the correlation structure among the components of different data sets. Both techniques can be employed to infer the linear dependency described by the components correlated between all pairs of data sets. When components are either correlated across all pairs of data sets or completely uncorrelated, the model order estimate from the GLRT-based technique can be used in the subsequent JBSS framework to only estimate and infer the first $\hat{d}_{\text{all}}$ components.

In case of arbitrary correlation structure, the product of coherence matrices can also be used to estimate the $\hat{d}_{\text{all}}$ components similar to how the coherence matrix is used for estimating the correlated components in two data sets [99].

5.7. Appendix - Generating the product of coherence matrices

The indices p and q for generating the product of coherence matrices for P data sets used in the technique in Section 5.4 are chosen using the following procedure:

- If P is odd, $N = \frac{(P-1)}{2}$ groups of data pairs are formed.

  1. For $n = 1, 2, \ldots, N$, repeat the following steps to generate the nth group:

     For $r = 1, 2, \ldots,$ GCD(n, P), repeat:
     $\triangleright$ GCD(a, b) denotes the greatest common divisor of a and b.

     a) Set $p = r$.

     b) Set $q = 0$.

     c) While $q \neq r$, repeat:

     – Set

     $$q = \begin{cases} \text{mod}(p + n, P) & \text{if } p + n \neq P, \\ P & \text{if } p + n = P. \end{cases}$$

     $\triangleright$ mod(a, b) denotes the modulus operator on a and b.

     – Include $\mathbf{C}_{pq}$ in the nth group of product of coherence matrices.

     – Set $p = q$.

     d) If $r <$ GCD(n, P)

     – Include $\mathbf{C}_{r\,r+1}$ in the product of coherence matrices.

     Elseif $r \neq 1$

     – Include $\mathbf{C}_{r1}$ in the product of coherence matrices.

  2. Combine all n group of matrices to obtain the final product of coherence matrices.

- If P is even, perform this procedure for $P+1$ data sets with the following modification: For each group n, there will be two pairs $\{p, P+1\}$ and $\{P+1, q\}$. Remove these and instead include $\{p, q\}$ in the product.

For instance, for $P = 5$, the following $N = 2$ groups of data pairs can be formed.

- $n = 1 : \{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}, \{5, 1\}$.
- $n = 2 : \{1, 3\}, \{3, 5\}, \{5, 2\}, \{2, 4\}, \{4, 1\}$.

The two groups are combined to generate the indices for the product of coherence matrices. For $P = 4$, again the data pairs for $P = 5$ are generated. In the first group, $\{4, 5\}$ and $\{5, 1\}$ are replaced with $\{4, 1\}$, and in the second group, $\{3, 5\}$ and $\{5, 2\}$ are replaced with $\{3, 2\}$. Hence,

- $n = 1 : \{1, 2\}, \{2, 3\}, \{3, 4\}, \{\mathbf{4}, \mathbf{1}\}$.
- $n = 2 : \{1, 3\}, \{\mathbf{3}, \mathbf{2}\}, \{2, 4\}, \{4, 1\}$.

6. Complete model selection in multiple data sets

Detecting the components common or correlated across multiple data sets is challenging due to a large number of possible correlation structures among the components. Even more challenging is to determine the precise structure of these correlations. The techniques we have discussed until now have focused on determining only the model order, i.e., the dimension of the correlated subspace, a number that depends on how the model-order problem is defined. Moreover, identifying the model order is often not enough to completely characterize the linear relationship among the components in different data sets. In this chapter, we aim at solving the complete model-selection problem, i.e., determining which components are correlated across which data sets. We propose two different techniques to solve this problem [1].

6.1. Introduction

Let us revisit the example of correlation structure between the latent signal components of three data sets, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ illustrated by Figure 6.1. It is exactly the same as Figure 1.1 and repeated here for better readability. One variation of the model-order selection problem for multiple data sets is to determine the number of signal components correlated across every pair of data sets. In Figure 6.1, only the first signal components demonstrate this property, so

[1] Section 6.3 of this chapter is based on the paper: "Complete model selection in multiset canonical correlation analysis, T. Marrinan, T. Hasija, C. Lameiro, P.J. Schreier, *European Signal Processing Conference*, 2018". I specifically contributed in the development of the proposed method along with T. Marrinan. Parts introducing and analyzing the method have contributions from all authors. Section 6.4 of this chapter is based on the paper: "Determining the dimension and structure of the subspace correlated across multiple data sets, T. Hasija, T. Marrinan, C. Lameiro, P.J. Schreier, *Signal Processing*, 2020".

the model order by this definition would be one. Model orders of this type can be identified using the methods described in Chapter 5 and references therein.

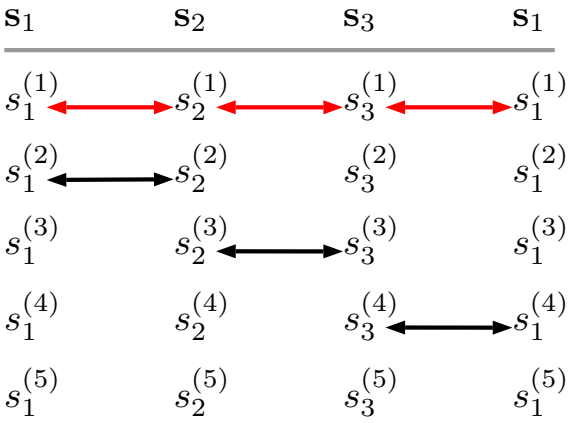


Figure 6.1.: Revisiting the example in 6.1 showing the correlation structure between latent signal components of three data sets, $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$.

Alternatively, signal components correlated across a subset of the collection of data sets might also be of interest. For instance, when tracking an object in videos recorded by spatially separated cameras, the object might not be visible in every frame of each camera [13]. Thus if multiset canonical correlation is used to measure the similarity of frames from the different views, it would be pertinent for the model order to represent the number of components correlated across all data sets or a subset of data sets. Similarly in brain imaging, estimating the number of signal components activated in the fMRI data of all the subjects is useful for multi-subject analysis [14], [100]. However, some brain regions may not appear active for some subjects, due to noise or other factors, even if biological intuition suggests that they should be. Knowing that correlations exist among the signal components corresponding to these regions is useful even if they are only present in a subset of subjects. These scenarios suggest that an appropriate definition of the model order should count the signals that demonstrate correlation across all or a subset of data sets. By this definition the model order of the example in Figure 6.1 is four. In [13], the authors propose a test statistic to estimate this model order and showed when this number can be correctly estimated for different settings of the signal-to-noise ratio (SNR) and the number of samples.

In the end, determining only the model order is insufficient to completely characterize the correlation structure in multiple data sets. This summary statistic only provides the knowledge that the components exhibit correlation. This knowledge, although sufficient for two data sets, is incomplete for multiple data sets as it is also required to determine which components are correlated across which data sets. For the example in Figure 6.1, the complete solution is not just determining that the first four components are correlated but also that the

first component in each data set is correlated, and the successive components are correlated between data sets 1 and 2, 2 and 3, and 1 and 3, respectively.

In this chapter, we formulate and solve a more general *model-selection* problem, which includes model-order selection as a subproblem. We propose two novel techniques in this chapter. The first technique applies a series of hypothesis tests to pairs of components extracted using CCA. It first determines the number of pairwise correlations, two data sets at a time, and amalgamates these detections using joint information from the complete collection of data sets obtained using mCCA. The second technique solves the model-selection problem using the eigenvalues and eigenvectors of the coherence matrix (normalized covariance matrix [41]) of the composite data set. To this end,

- We prove that, under fairly general conditions, the correlation structure in multiple data sets can be fully characterized from the eigenvector decomposition of the coherence matrix of the composite data.
- Using this theoretical result, we develop an algorithm that identifies the correlation structure effectively in practical scenarios.

Our program for this chapter is as follows. After defining the problem in Section 6.2, we formulate the first technique for complete model selection in Section 6.3. In Section 6.4, we formulate the second technique. The extensions of the two proposed techniques for high-dimensional data sets with relatively less number of samples is presented in Section 6.5. Finally, in Section 6.6, simulations show that our techniques reliably estimate the complete correlation structure of multiple data sets, and is competitive with the existing state-of-the-art approaches.

6.2. Noiseless data model for multiple data sets

We consider P data sets consisting of zero-mean, real-valued random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_P$, each of dimension n. The data sets are generated by an unknown linear mixing of underlying signal vectors $\mathbf{s}_1, \ldots, \mathbf{s}_P \in \mathbb{R}^n$. The generating data model is

$$\mathbf{x}_p = \mathbf{A}_p \mathbf{s}_p, \quad p = 1, 2, \ldots, P, \tag{6.1}$$

where $\mathbf{A}_p \in \mathbb{R}^{n\times n}$ is an unknown but fixed mixing matrix with full rank [2]. The signal vectors each contain n signal components. The assumptions about the dependence between the signal components in the multiset model (6.1) are same as defined in Section 5.2. However, due to notational changes because of square mixing matrices, we will redefine them in this section for the sake of completeness. The ith signal component of the pth data set is denoted by $s_p^{(i)}$. These components are assumed to be zero-mean and unit variance without loss of generality, i.e.,

$$E[s_p^{(i)}] = 0, \qquad \text{and} \tag{6.2}$$

$$E[(s_p^{(i)})^2] = 1, \qquad \text{for } i = 1, \ldots, n. \tag{6.3}$$

Two kinds of association among the signal components are assumed:

1. Intraset independence: signal components within each data set are uncorrelated, i.e.,

$$\mathbf{R}_{s_p s_p} = E[\mathbf{s}_p \mathbf{s}_p^T] = \mathbf{I}, \tag{6.4}$$

   where $\mathbf{I}$ is an identity matrix, and

2. Interset dependence: between any two data sets p and q, components may be correlated only pairwise, i.e., component $s_p^{(i)}$ may only correlate with component $s_q^{(i)}$ for $1 \leq i \leq n$. This means, the signal cross-covariance matrix between data sets p and q ($p \neq q$) is

$$\mathbf{R}_{s_p s_q} = \text{diag}(\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(n)}), \tag{6.5}$$

   where $\rho_{pq}^{(i)}$ represents the unknown (possibly zero) correlation coefficient between their ith components. We assume that all nonzero correlation coefficients are positive since their negative sign can be incorporated into the mixing matrix.

The goal of this chapter is:

Goal: *Given M i.i.d. joint samples of the data vectors $\mathbf{x}_p$, $p = 1, \ldots, P$, our aim is to completely determine the underlying correlation structure among the signal components $\mathbf{s}_p$. More precisely, we identify the all the components in $\mathbf{s}_p$ with index $i = 1, \ldots, n$ and all the data sets $p, q \in \{1, \ldots, P\}, p \neq q$, for which $\rho_{pq}^{(i)} \neq 0$.*

Let us define the model order d, which represents the total number of components that

[2] All the results in this work can be easily extended to the general case where the data sets have different dimensions and the mixing matrices are non-square with full column rank. We omit the general case as the extension is trivial when the inverses are replaced by pseudo-inverses, because the more cumbersome bookkeeping distracts from the actual result.

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{23}^{(i)}$
$i=1$	0.5	0.6	0.6
$i=2$	0.7	0	0
$i=3$	0	0	0.8
$i=4$	0	0.4	0
$i=5$	0	0	0

Table 6.1.: Example of the correlation structure in Figure 1.1 with three data sets each with five signal components. The entries are the correlation coefficients between signal components of different pairs of data sets.

demonstrate nonzero correlation i.e., $d = |\{i : \exists p, q \text{ for which } \rho_{pq}^{(i)} \neq 0\}|$. We denote the number of signal components correlated pairwise between data sets p and q as d_{pq}. The number of signals correlated across all data sets is denoted by d_{all}, i.e., $d_{\text{all}} = |\{i : \rho_{pq}^{(i)} \neq 0 \; \forall p, q\}|$. In this thesis, we have until now focused on either determining d_{pq} for two data sets in Chapter 3 or on estimating d_{all} for more than two data sets in Chapter 5. However, as discussed earlier, even knowing both d_{pq} and d_{all} is not enough to completely determine the underlying correlation structure (except with very special types of correlation structures, e.g., as in Section 5.3).

Examples: Let us revisit the example in Figure 6.1. Table 6.1 provides one example of correlation coefficients that match the structure presented in Figure 6.1. The entries are the correlation coefficients between signal components of different pairs of data sets. In this case, $d_{pq} = 2$ for all choices of p and q, $d_{\text{all}} = 1$, and $d = 4$. Now consider an example of 4 data sets each with 4 signal components as shown in Table 6.2. The first component of data sets $2, 3,$ and 4 and the second component of data sets $1, 3,$ and 4 are correlated. The third and fourth components of data sets 1 and 2 are correlated. Hence, $d_{\text{all}} = 0$, $d = 4$ and d_{pq} is the number of nonzero entries in the corresponding column. In both these examples, the techniques in [28], [30]–[34] provide solutions for either d_{pq} or d_{all}. The techniques proposed in this chapter, however, aim to identify which of the entries in Tables 6.1 and 6.2 are nonzero.

6.3. Technique based on pairwise model orders

In Section 2.2, we discussed about mCCA. For P data sets, it provides P sets of canonical variables which are chosen such that they are highly correlated with those from the other sets at each stage of the algorithm, but uncorrelated with the canonical variables of different

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{34}^{(i)}$
$i=1$	0	0	0	0.7	0.2	0.8
$i=2$	0	0.6	0.4	0	0	0.5
$i=3$	0.5	0	0	0	0	0
$i=4$	0.5	0	0	0	0	0

Table 6.2.: Example of correlation structure with four data sets each with four signal components.

stages from within a set [21]. These multiset canonical variables provides a solution to the joint blind source separation (JBSS) problem in certain scenarios, where the aim is to extract the underlying latent signals in (6.1) that exhibit joint information (or correlation) between them [23].

6.3.1. Sample canonical correlations in mCCA

For Gaussian-distributed sample matrices $\mathbf{X}_1, \ldots, \mathbf{X}_P$, the covariance matrices can be replaced with their ML estimates, $\hat{\mathbf{R}}_{pp} = \frac{1}{M}\mathbf{X}_p\mathbf{X}_p^T$ and $\hat{\mathbf{R}}_{pq} = \frac{1}{M}\mathbf{X}_p\mathbf{X}_q^T$. The sample coherence matrix can then be estimated using (2.12) and the set of sample canonical variables, $\{\hat{\epsilon}_1^{(j)}, \ldots, \hat{\epsilon}_P^{(j)}\}$, $j = 1, \ldots, n$ can be obtained by imposing the constraints defined in (2.13).

When these sets solve the JBSS problem, the jth set of sample canonical variates will approximate the ith set of signal components with some ambiguity about which i. This is commonly referred to as the permutation ambiguity and is inherent in BSS since for an unknown mixing matrix $\mathbf{A}_p$, different orderings of the signal components in $\mathbf{s}_p$ leads to the same data vector $\mathbf{x}_p$ [101]. Thus with the JBSS solution, the inner products of same-stage sample canonical variables provide the sample canonical correlations, $\hat{\rho}_{pq}^{(j)} = |\frac{1}{M}\hat{\epsilon}_p^{(j)}\hat{\epsilon}_q^{(j)}|$, and approximate the true correlations $\rho_{pq}^{(i)}$. The accuracy of this approximation is affected by numerous parameters of the observed data including the number of data sets, the number of sets across which a particular signal component is correlated, and the number of observed samples. To bring this technique into the model-selection paradigm, we determine which values of $\hat{\rho}_{pq}^{(j)}$ are significant for $j = 1, \ldots, n$.

In the case of two data sets, the sample canonical correlations are ML estimates of the true correlations. Thus the number of correlated components d_{pq} can be estimated through a series of binary hypothesis tests as explained in Section 3.4. The distribution of the Bartlett-Lawley

statistic $C(s)$ is χ^2_ν when H_0 is true, i.e., when $s = d_{pq}$, which can be used to set a threshold to test each binary test. However, in the multiset scenario, $\hat{\rho}_{pq}^{(j)}$ is not an ML estimate of the true pairwise correlations because the canonical variables have been jointly estimated for all data sets at once. This means that $C(s)$ for $s = d_{pq}$ is not guaranteed to follow the χ^2_ν distribution, and a test based on the Bartlett-Lawley statistic will not work. This hurdle is overcome by pairwise estimating the number of correlated components between each combination of data sets and then identifying which of the multiset sample canonical variables for those two data sets have the highest correlations.

6.3.2. Proposed model selection framework

The signal components correlated across any number of data sets can be identified using the following three-step approach.

1. Compute sample canonical variables $\{\hat{\epsilon}_1^{(j)}, \ldots, \hat{\epsilon}_P^{(j)}\}$ for stages $j = 1, \ldots, n$ using mCCA.

2. Estimate the pairwise model order for each combination of data sets. This is done by computing ML estimates of the canonical correlations between $\mathbf{x}_p$ and $\mathbf{x}_q$, and applying the sequence of hypothesis tests based on the Bartlett-Lawley statistic to find $\hat{d}_{pq}$.

3. Identify which sample canonical variables have the largest magnitude for each pair of data sets. If $|\frac{1}{M}\hat{\epsilon}_p^{(j)}\hat{\epsilon}_q^{(j)}|$ is one of the $\hat{d}_{pq}$ largest inner products for data set p and data set q, a nonzero correlation is identified between the ith signal components of $\mathbf{x}_p$ and $\mathbf{x}_q$. Note that these will not necessarily correspond to stages $1, \ldots, \hat{d}_{pq}$ of the mCCA method.

In practical situations, estimates of correlation coefficients between the canonical variables are affected by the number of observed samples. When there are few observations relative to the number of signal components, canonical correlations are overestimated. However, pairwise correlation estimates can be improved in small sample scenarios with PCA-CCA detectors reviewed in Section 3.5. The maxmin GLRT is one such technique, and the proposed algorithm employs this method in Step 2 to provide a better estimate of pairwise model order over the Bartlett-Lawley statistic without rank reduction.

6.3.2.1. Computational complexity

We compute the big-$\mathcal{O}$ complexity for the proposed technique. The dominating term in the complexity analysis is Step 2, which runs $\frac{P(P-1)}{2}$ tests in (3.36) for each pair of n-dimensional data sets. Each test computes the SVD of the pairwise coherence matrix, which can be computed (assuming $n \leq M$) in $\mathcal{O}(Mn^2)$ flops. The overall big-$\mathcal{O}$ complexity of the technique is thus, $\mathcal{O}(MP^2n^4)$ flops.

6.4. Technique based on joint information in all sets

The problem as stated in Section 6.2 requires joint knowledge of the relationships between all data sets. In this section, our main results demonstrate that the pertinent information can be found in the eigenvector decomposition of the coherence matrix of the concatenation of all data sets. We solve this problem by first estimating the model order d and then estimating the data sets across which these d components are correlated.

6.4.1. Correlated subspace in multiple data sets

Consider the composite data vector $\mathbf{x}$ obtained by vertically concatenating the individual data vectors,

$$\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T, \tag{6.6}$$

and the composite covariance matrix $\mathbf{R} = E[\mathbf{x}\mathbf{x}^T]$. Similarly, the composite signal vector is defined as $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_P^T]^T$, and the composite signal covariance matrix as $\mathbf{R}_{ss} = E[\mathbf{s}\mathbf{s}^T]$. The definition of the coherence matrix for two data sets [41] can be generalized in a natural way for this composite data as

$$\mathbf{C} = \mathbf{R}_D^{-\frac{1}{2}}\mathbf{R}\mathbf{R}_D^{-\frac{1}{2}}. \tag{6.7}$$

Here, $\mathbf{R}_D = \text{blkdiag}(\mathbf{R}_{11}, \ldots, \mathbf{R}_{PP})$ is a block-diagonal matrix with $\mathbf{R}_{pp} = E[\mathbf{x}_p\mathbf{x}_p^T]$, and exponent $-\frac{1}{2}$ on $\mathbf{R}_D$ denotes the inverse of the matrix square root. The composite coherence matrix $\mathbf{C}$ can be written in a block structure as

$$\mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1P} \\ \mathbf{C}_{21} & \mathbf{I} & \cdots & \mathbf{C}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{P1} & \mathbf{C}_{P2} & \cdots & \mathbf{I} \end{bmatrix}, \tag{6.8}$$

where the diagonal blocks are identity matrices and the off-diagonal blocks are the coherence matrices of two data sets. For each pair of data sets $\mathbf{x}_p$ and $\mathbf{x}_q$, the coherence matrix can be decomposed as

$$\begin{aligned}\mathbf{C}_{pq} &= \mathbf{R}_{pp}^{-\frac{1}{2}}\mathbf{R}_{pq}\mathbf{R}_{qq}^{-\frac{1}{2}}, \\ &= (\mathbf{A}_p\mathbf{R}_{s_ps_p}\mathbf{A}_p^T)^{-\frac{1}{2}}\mathbf{A}_p\mathbf{R}_{s_ps_q}\mathbf{A}_q^T(\mathbf{A}_q\mathbf{R}_{s_qs_q}\mathbf{A}_q^T)^{-\frac{1}{2}}, \\ &= (\mathbf{A}_p\mathbf{A}_p^T)^{-\frac{1}{2}}\mathbf{A}_p\mathbf{R}_{s_ps_q}\mathbf{A}_q^T(\mathbf{A}_q\mathbf{A}_q^T)^{-\frac{1}{2}}, \end{aligned} \tag{6.9}$$

since $\mathbf{R}_{s_ps_p} = \mathbf{R}_{s_qs_q} = \mathbf{I}$. Let $\bar{\mathbf{A}}_p = (\mathbf{A}_p\mathbf{A}_p^T)^{-1/2}\mathbf{A}_p$ and similarly $\bar{\mathbf{A}}_q = (\mathbf{A}_q\mathbf{A}_q^T)^{-1/2}\mathbf{A}_q$ so that $\bar{\mathbf{A}}_p\bar{\mathbf{A}}_p^T = \bar{\mathbf{A}}_q\bar{\mathbf{A}}_q^T = \mathbf{I}$, and we have

$$\mathbf{C}_{pq} = \bar{\mathbf{A}}_p\mathbf{R}_{s_ps_q}\bar{\mathbf{A}}_q^T. \tag{6.10}$$

Using (6.8) and (6.10), the composite coherence matrix $\mathbf{C}$ can be written as

$$\mathbf{C} = \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^T, \tag{6.11}$$

where $\mathbf{A} = \text{blkdiag}(\bar{\mathbf{A}}_1, \ldots, \bar{\mathbf{A}}_P)$. Based on the assumption that $\mathbf{R}_{s_ps_q}$ is diagonal, the elements of $\mathbf{R}_{ss}$ can be permuted to form a block-diagonal matrix whose ith block is the covariance matrix formed by the ith components of each data set. That is, there exists a permutation $\mathbf{P}$ where

$$\begin{aligned}\mathbf{C} &= \mathbf{A}\mathbf{P}^T\mathbf{P}\mathbf{R}_{ss}\mathbf{P}^T\mathbf{P}\mathbf{A}^T, \\ &= \mathbf{A}\mathbf{P}^T\widetilde{\mathbf{R}}_{ss}\mathbf{P}\mathbf{A}^T, \end{aligned} \tag{6.12}$$

such that $\widetilde{\mathbf{R}}_{ss} = \mathbf{P}\mathbf{R}_{ss}\mathbf{P}^T$ is a block-diagonal matrix defined as

$$\widetilde{\mathbf{R}}_{ss} = \text{blkdiag}(\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(n)}), \tag{6.13}$$

and $\mathbf{R}^{(i)} \in \mathbb{R}^{P\times P}$ is the covariance matrix of the ith components of each data set.

6.4.1.1. Eigenvalues of $\mathbf{C}$

We now illustrate an explicit relationship between the dimension of the correlated subspace d and the eigenvalues of $\mathbf{C}$. It can be observed from (6.8) that $\mathbf{C}$ is an identity matrix when all signal components are uncorrelated, and thus all eigenvalues of $\mathbf{C}$ are one. However,

when some components are correlated, $\mathbf{C}$ has eigenvalues that are different from one. More specifically, when data sets demonstrate correlations across d components, $\mathbf{C}$ has at least d eigenvalues greater than one. A key question then is: *when is the dimension of the correlated subspace, d, exactly equal to the number of eigenvalues of $\mathbf{C}$ greater than one?*

The answer is not as straightforward as one would hope. However, we can identify a set of sufficient conditions under which this property holds. The property also often holds when these conditions are not met and the algorithm proposed in this work is robust to the violation of these assumptions. The proof of the sufficiency of these conditions relies on three things:

i) The composite coherence matrix $\mathbf{C}$ is similar (through a similarity transformation) to a block diagonal matrix where each block contains a non-identity matrix corresponding to a collection of data sets whose ith components are all correlated with each other and an identity matrix corresponding to data sets whose ith components are uncorrelated.

ii) If the ith components of four or more data sets are correlated with each other, all nonzero correlations are greater than a prescribed threshold.

iii) All correlations are transitive. That is, if a signal component is correlated between data sets p and q, and between data sets q and r, then it is also correlated between data sets p and r.

For simplicity, we prove the following result (Result 6.1) when there is only one block of the form described in condition i) for the ith components of all data sets, but of course the result holds for any number of such blocks. We will discuss the requirement of conditions ii) and iii) during the proof of Result 6.1.

Result 6.1. *Let $\mathbf{C}$ be the composite coherence matrix of P data sets constructed according to the linear mixing model in* (6.1) *with pairwise diagonal signal cross-covariance matrices. Let $k^{(i)}$ be the number of data sets whose ith components are correlated. Assume that correlations are transitive, and for $k^{(i)} \geq 4$, each correlation coefficient is either $\rho_{pq}^{(i)} = 0$ or $\rho_{pq}^{(i)} > \xi^{(i)} = (\frac{k^{(i)}-1}{k^{(i)}})^2$ for all p, q. Let $\mathcal{I} = \{1, \ldots, n\}$ be an index set for the signals. $\mathbf{C}$ has exactly d eigenvalues greater than one if and only if there exists a subset of signals $\mathcal{D} \subseteq \mathcal{I}$ with $|\mathcal{D}| = d$, and for each $i \in \mathcal{D}$ there exists a $p \neq q$ such that $s_p^{(i)}$ and $s_q^{(i)}$ are correlated.*

We begin by showing that if there exists a subset of signals $\mathcal{D} \subseteq \mathcal{I}$ with $|\mathcal{D}| = d$, and for each $i \in \mathcal{D}$ there exists a p, q with $s_p^{(i)}$ and $s_q^{(i)}$ correlated, $\mathbf{C}$ has d eigenvalues greater than one. Let $\widetilde{\mathbf{R}}_{ss}$ be an $nP \times nP$ matrix with the structure defined in (6.13). The diagonal blocks of $\widetilde{\mathbf{R}}_{ss}$ can be indexed by the associated signal component. That is, $\mathbf{R}^{(i)} \in \mathbb{R}^{P \times P}$ is the

covariance matrix of the ith components of each data set with the form

$$\mathbf{R}^{(i)} = \begin{bmatrix} \mathbf{B}^{(i)} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{6.14}$$

$\mathbf{B}^{(i)} \in \mathbb{R}^{k^{(i)} \times k^{(i)}}$ is a symmetric matrix with diagonal elements equal to one and off-diagonal elements equal to the correlation coefficients between the correlated ith components. The indices associated with the subset of data sets whose ith components are correlated are $\mathcal{K}^{(i)} \subseteq \mathcal{P} = \{1, \dots P\}$, and the dimensions of $\mathbf{B}^{(i)}$ are the size of this subset $|\mathcal{K}^{(i)}| = k^{(i)}$. As $\widetilde{\mathbf{R}}_{ss}$ is block-diagonal, its eigenvalues are equal to the eigenvalues of the blocks $\mathbf{R}^{(i)}$. Since $\mathbf{AP}^T$ is an orthogonal matrix, the eigenvalues of $\mathbf{C}$ are equal to the eigenvalues of $\widetilde{\mathbf{R}}_{ss}$ and therefore, equal to the eigenvalues of $\mathbf{R}^{(i)}$.

Let $\mathbf{B}^{(i)}$ be decomposed as $\mathbf{B}^{(i)} = \mathbf{I} + \mathbf{H}^{(i)}$ for each i. $\mathbf{H}^{(i)}$ is a hollow (with zeros on the diagonal) symmetric matrix whose off-diagonal elements are the nonzero correlation coefficients corresponding to the ith component. We show that the matrix $\mathbf{H}^{(i)}$ has exactly one positive eigenvalue for each i as follows:

Case 1 ($k^{(i)} = 2$): $\mathbf{H}^{(i)}$ has exactly one positive eigenvalue for all values of $\rho_{pq}^{(i)} > 0$, because

$$\mathbf{H}^{(i)} = \begin{bmatrix} 0 & \rho_{pq}^{(i)} \\ \rho_{pq}^{(i)} & 0 \end{bmatrix}, \tag{6.15}$$

which has eigenvalues $\{\rho_{pq}^{(i)}, -\rho_{pq}^{(i)}\}$.

Case 2 ($k^{(i)} = 3$): When the ith components of exactly three data sets, p, q and r, are correlated, $\mathbf{H}^{(i)}$ is

$$\mathbf{H}^{(i)} = \begin{bmatrix} 0 & \rho_{pq}^{(i)} & \rho_{pr}^{(i)} \\ \rho_{pq}^{(i)} & 0 & \rho_{qr}^{(i)} \\ \rho_{pr}^{(i)} & \rho_{qr}^{(i)} & 0 \end{bmatrix}. \tag{6.16}$$

The characteristic polynomial of $\mathbf{H}^{(i)}$ is

$$y_{\mathbf{H}^{(i)}}(\lambda) = -\lambda^3 + \lambda\left(\left(\rho_{pq}^{(i)}\right)^2 + \left(\rho_{pr}^{(i)}\right)^2 + \left(\rho_{qr}^{(i)}\right)^2\right) + 2\rho_{pq}^{(i)}\rho_{pr}^{(i)}\rho_{qr}^{(i)}. \tag{6.17}$$

Using Descartes' rule of sign change, $y_{\mathbf{H}^{(i)}}(\lambda)$ has only one positive root for any $\rho_{pq}^{(i)}, \rho_{pr}^{(i)}, \rho_{qr}^{(i)} > 0$ [102]. Therefore, $\mathbf{H}^{(i)}$ has only one positive eigenvalue.

Case 3 ($k^{(i)} \geq 4$): [103, Theorem 3.5] can be used to show that $\mathbf{H}^{(i)}$ has exactly one positive eigenvalue if all the off-diagonal elements of $\mathbf{H}^{(i)}$ are greater than $\xi^{(i)} = (\frac{k^{(i)}-1}{k^{(i)}})^2$. This result is demonstrated in Appendix 6.8. Without the assumption of transitive correlations,

$\mathbf{H}^{(i)}$ cannot be guaranteed to have all the positive off-diagonal elements as required. Thus, $\mathbf{H}^{(i)}$ has exactly one positive eigenvalue in each case as desired.

Let $\mathcal{D} \subseteq \mathcal{I}$ be the d values of i for which correlation exists. For each $i \in \mathcal{D}$, $\mathbf{H}^{(i)}$ has one positive eigenvalue and $k^{(i)} - 1$ non-positive eigenvalues. Let $\mathbf{H}^{(i)} = \mathbf{U}\mathbf{\Lambda}^{(i)}\mathbf{U}^T$ be the EVD of $\mathbf{H}^{(i)}$ with $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. The EVD of $\mathbf{B}^{(i)}$ can be written as $\mathbf{B}^{(i)} = \mathbf{U}(\mathbf{I}+\mathbf{\Lambda}^{(i)})\mathbf{U}^T$ so that, for each $i \in \mathcal{D}$, $\mathbf{B}^{(i)}$ has one eigenvalue greater than one and the remaining $k^{(i)} - 1$ eigenvalues less than or equal to one. Using (6.14), the maximum eigenvalue of $\mathbf{B}^{(i)}$ is also the maximum eigenvalue of $\mathbf{R}^{(i)}$, implying that $\mathbf{R}^{(i)}$ has exactly one eigenvalue greater than one. Hence $\mathbf{C}$ has exactly d eigenvalues greater than one as desired.

We now show that the converse is also true. Let $\mathbf{C}$ has d eigenvalues greater than one. There exists a permutation as described by (6.12) where

$$\mathbf{R}^{(i)} = \begin{bmatrix} \mathbf{B}^{(i)} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{6.18}$$

and $\mathbf{B}^{(i)}$ is the sum of an identity matrix and a hollow symmetric matrix of correlation coefficients, i.e., $\mathbf{B}^{(i)} = \mathbf{I} + \mathbf{H}^{(i)}$, as described above. When no data sets are correlated for a given i, $\mathbf{R}^{(i)} = \mathbf{I}$. Let $\mathcal{D}' \subseteq \mathcal{I}$ be the indices for which nonzero correlation exists (and by assumption is greater than $\xi^{(i)}$ for $k^{(i)} \geq 4$). Thus,

$$\mathbf{R}^{(i)} = \begin{bmatrix} \mathbf{I} + \mathbf{H}^{(i)} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{6.19}$$

for $i \in \mathcal{D}'$ and $\mathbf{R}^{(i)} = \mathbf{I}$ for $i \in \mathcal{I}\backslash\mathcal{D}'$. Clearly $\mathbf{R}^{(i)}$ has no eigenvalues greater than one for $i \in \mathcal{I}\backslash\mathcal{D}'$.

Suppose that $|\mathcal{D}'| < d$. Then there exists an i for which $\mathbf{H}^{(i)}$ has more than one eigenvalue greater than zero. However, this contradicts our proof that $\mathbf{H}^{(i)}$ has exactly one positive eigenvalue for each i. Similarly, suppose that $|\mathcal{D}'| > d$. Then there exists an i for which $\mathbf{H}^{(i)}$ has no eigenvalues greater than zero. This also contradicts our proof, and thus $|\mathcal{D}'| = d$ and there are exactly d values of i for which $\rho_{pq}^{(i)} \neq 0$ for some $p \neq q$. This concludes our proof for Result 6.1.

Result 6.1 guarantees that, when d eigenvalues of $\mathbf{C}$ are greater than one, d signal components are correlated across at least a pair of data sets. Therefore, the number of correlated signal components can be determined by testing for the number of eigenvalues of $\mathbf{C}$ that are greater than one. Such a test is formulated in Section 6.4.2.1 using bootstrap-based hypoth-

esis testing.

One of the assumptions in Result 6.1 that needs further discussion is that if the ith components of four or more data sets are correlated (i.e., $k^{(i)} \geq 4$), then the correlation coefficient between any pair of ith signal components must be either zero or greater than $\xi^{(i)} = (\frac{k^{(i)}-1}{k^{(i)}})^2$. This assumption guarantees that only one eigenvalue of $\mathbf{C}$ corresponding to the ith component is greater than one. The threshold, $\xi^{(i)}$, is derived in the appendix in Section 6.8 and is a restrictive threshold since $\lim_{k^{(i)}\to\infty} \xi^{(i)} = 1$. However, the proof does not claim to represent all matrices with the desired eigenvalue structure. That is, there is a nonempty set of real positive hollow symmetric matrices that have exactly one positive eigenvalue but do not meet this element-wise threshold. One example is the following: Suppose the nonzero correlation coefficients associated with the ith component are equal for all $k^{(i)}(\geq 4)$ data sets. That is, $\rho_{pq}^{(i)} = \rho^{(i)}\ \forall p, q \in \mathcal{K}^{(i)}$, where $\mathcal{K}^{(i)}$ is the subset of indices associated with the data sets whose ith components are correlated. In this case,

$$\mathbf{H}^{(i)} = \begin{bmatrix} 0 & \rho^{(i)} & \cdots & \rho^{(i)} \\ \rho^{(i)} & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho^{(i)} & \cdots & \cdots & 0 \end{bmatrix}, \tag{6.20}$$

This can be simplified as $\mathbf{H}^{(i)} = \rho^{(i)}\mathbf{1}\mathbf{1}^T - \rho^{(i)}\mathbf{I}$, where $\mathbf{1} \in \mathbb{R}^{k^{(i)}}$ is a vector with all elements equal to one. The maximum eigenvalue of $\mathbf{H}^{(i)}$ is $(k^{(i)} - 1)\rho^{(i)}$ and the remaining $k^{(i)} - 1$ eigenvalues are $-\rho^{(i)}$. Therefore, $\mathbf{H}^{(i)}$ has one positive eigenvalue for any $\rho^{(i)} > 0$. In this example, the relationship between the eigenvalues of $\mathbf{C}$ and the number of signals with nonzero correlations described by Result 6.1 holds true for $0 < \rho_{pq}^{(i)} \leq 1$.

In the general case, even though $\xi^{(i)}$ is restrictive, an element-wise threshold like this is perhaps the best that can be hoped for without imposing further constraints on the structure of the correlation among the components. As noted in [103], for any $k \in \mathbb{N}$ with $k \geq 3$, there exists a positive hollow symmetric $\mathbf{H} \in \mathbb{R}^{k\times k}$ such that $\mathbf{H}$ has only two nonpositive eigenvalues. That is to say, without the element-wise constraint there will always be feasible correlation structures for which $\mathbf{C}$ has more eigenvalues greater than one than signals with nonzero correlations. Moreover, as we will see in our numerical examples later, our hypothesis-test based techniques presented in Section 6.4.2 may still perform satisfactorily even in cases where the assumptions of Result 6.1 are violated.

As an immediate consequence of Result 6.1, any eigenvalue of $\mathbf{C}$ that is equal to the max-

imum possible value P identifies a signal component where all P data sets are perfectly correlated. This is shown in Result 6.2.

Result 6.2. *If any eigenvalue of* $\mathbf{C}$*,* $\lambda^{(i)}$*, is equal to* P*, there exists an* $i \in \mathcal{D}$ *such that the correlation between* $s_p^{(i)}$ *and* $s_q^{(i)}$ *is one for all* $p, q = 1, \ldots, P$*.*

Result 6.2 is proved as follows. By Result 6.1, if $\lambda^{(i)} > 1$ is an eigenvalue of $\mathbf{C}$, there exists an $\mathbf{R}^{(i)} \neq \mathbf{I}$ whose largest eigenvalue is equal to $\lambda^{(i)}$. The diagonal elements of $\mathbf{R}^{(i)}$ are equal to one by definition, so trace$(\mathbf{R}^{(i)}) = P = \sum_{j=1}^{P} \lambda_j^{(i)}$, where $\lambda_j^{(i)}$ is the jth largest eigenvalue of $\mathbf{R}^{(i)}$. Since the largest eigenvalue of $\mathbf{R}^{(i)}$, $\lambda_1^{(i)} = \lambda^{(i)} = P$, we have $\lambda_2^{(i)} = \ldots = \lambda_P^{(i)} = 0$, and thus, the rank of $\mathbf{R}^{(i)}$ is one.

Let $\mathbf{R}^{(i)} = \mathbf{w}\mathbf{w}^T$ be a rank-one matrix, where $\mathbf{w} = [w_1, \ldots, w_P]^T \in \mathbb{R}^P$. The diagonal elements of $\mathbf{R}^{(i)}$ are equal to one, implying that $w_p^2 = 1$ for all p. Since the off-diagonal elements are bounded by zero and one, w_p is positive for $p = 1, \ldots, P$. Thus, $w_1 = w_2 = \ldots = w_P = 1$ is the only solution for $\mathbf{w}$, and $\mathbf{R}^{(i)} = \mathbf{1}\mathbf{1}^T$. Therefore, $\rho_{pq}^{(i)} = 1 \forall p, q$ and the ith component of each data set is perfectly correlated with the ith component of all other data sets.

6.4.1.2. Eigenvectors of C

The eigenvalues of $\mathbf{C}$ provide information about the dimension of the correlated subspace, but identifying exactly which data sets demonstrate correlation in a particular component requires more information than this summary contains. The eigenvectors of $\mathbf{C}$, on the other hand, contain as their elements the coefficients for constructing the correlated signals from each data set that correspond to the associated eigenvalue (that is greater than one). Data sets connected to the nonzero elements of an eigenvector are then the ones whose components are correlated among the associated group of components. Let us introduce the following result.

Result 6.3. *Let* $\mathbf{C}$ *be the composite coherence matrix of* P *data sets constructed according to the linear mixing model in* (6.1) *with pairwise diagonal signal cross-covariance matrices. Let* $k^{(i)}$ *be the number of data sets whose* i*th components are correlated. Assume that correlations are transitive, and for* $k^{(i)} \geq 4$*, each correlation coefficient is either* $\rho_{pq}^{(i)} = 0$ *or* $\rho_{pq}^{(i)} > \xi^{(i)} = (\frac{k^{(i)}-1}{k^{(i)}})^2$ *for all* p, q*. Let* $\mathbf{C}\mathbf{u}^{(i)} = \lambda^{(i)}\mathbf{u}^{(i)}$ *such that* $\lambda^{(i)} > 1$ *is an eigenvalue with algebraic multiplicity of one, and let the eigenvector* $\mathbf{u}^{(i)}$ *be partitioned into* P *subvectors,* $\mathbf{u}^{(i)} = [\mathbf{u}_1^{(i)T}, \mathbf{u}_2^{(i)T}, ..., \mathbf{u}_P^{(i)T}]^T$*, where* $\mathbf{u}_p^{(i)} \in \mathbb{R}^n$ *contains the elements of* $\mathbf{u}^{(i)}$ *associated with the* p*th data set. Then the* i*th signal component in the* p*th data set is among the group*

of correlated ith components if and only if $\mathbf{u}_p^{(i)} \neq \mathbf{0}$.

To prove Result 6.3, let the ith components of $k^{(i)}$ data sets be correlated, and let $\mathcal{K}^{(i)} \subseteq \mathcal{P} = \{1, \ldots, P\}$ be the subset of indices associated with these correlated data sets so that $|\mathcal{K}^{(i)}| = k^{(i)}$. Suppose the ith signal component in the pth data set is among the correlated components, i.e., p is an element of $\mathcal{K}^{(i)}$. Then there exists a permutation as described by (6.12) where $\mathbf{R}^{(i)} \in \mathbb{R}^{P \times P} \neq \mathbf{I}$ such that

$$\mathbf{R}^{(i)} = \begin{bmatrix} \mathbf{B}^{(i)} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{6.21}$$

where $\mathbf{B}^{(i)} \in \mathbb{R}^{k^{(i)} \times k^{(i)}}$ is a symmetric and element-wise positive matrix that contains the correlation coefficients between the correlated ith components. According to Result 6.1, the largest eigenvalue of $\mathbf{B}^{(i)}$ is greater than one and satisfies $\mathbf{B}^{(i)}\mathbf{v}^{(i)} = \lambda_{\text{max}}^{(i)}\mathbf{v}^{(i)}$, where $\mathbf{v}^{(i)}$ is the eigenvector of $\mathbf{B}^{(i)}$ associated with $\lambda_{\text{max}}^{(i)}$. According to the Perron-Frobenius theorem, all the entries of $\mathbf{v}^{(i)}$ are positive [104]. Since $\mathbf{R}^{(i)}$ is block-diagonal with blocks $\mathbf{B}^{(i)}$ and $\mathbf{I}$, $\lambda_{\text{max}}^{(i)}$ is also the largest eigenvalue of $\mathbf{R}^{(i)}$ with the eigenvector

$$\tilde{\mathbf{v}}^{(i)} = \begin{bmatrix} \mathbf{v}^{(i)} \\ \mathbf{0} \end{bmatrix}, \tag{6.22}$$

where $\mathbf{0}$ is a zero vector of dimensions $P - k^{(i)}$.

$\mathbf{R}^{(i)}$ is a block of $\widetilde{\mathbf{R}}_{ss}$ as defined in (6.13), and thus $\lambda_{\text{max}}^{(i)}$ is also an eigenvalue of $\widetilde{\mathbf{R}}_{ss}$. Each eigenvector $\tilde{\mathbf{u}}^{(i)}$ of $\widetilde{\mathbf{R}}_{ss}$ can be partitioned into n subvectors where the elements of the ith subvector correspond to the ith block of $\widetilde{\mathbf{R}}_{ss}$. Since $\lambda_{\text{max}}^{(i)}$ has an algebraic multiplicity of one, the eigenvector of $\widetilde{\mathbf{R}}_{ss}$ associated with $\lambda_{\text{max}}^{(i)}$ has $\widetilde{\mathbf{v}}^{(i)}$ at the ith position and zeros everywhere else and can be written as $\tilde{\mathbf{u}}^{(i)} = [\mathbf{0}^T, \ldots, \tilde{\mathbf{v}}^{(i)T}, \ldots, \mathbf{0}^T]^T$.

Using (6.12), the eigenvector of $\mathbf{C}$ associated with $\lambda_{\text{max}}^{(i)}$ is related to $\tilde{\mathbf{u}}^{(i)}$ by

$$\mathbf{u}^{(i)} = \mathbf{A}\mathbf{P}^T\tilde{\mathbf{u}}^{(i)}. \tag{6.23}$$

For each $p \in \mathcal{K}^{(i)}$, the element of $\tilde{\mathbf{v}}^{(i)}$ corresponding to the pth data set is strictly positive. Therefore, from the definition of $\tilde{\mathbf{u}}^{(i)}$ and (6.23), $\mathbf{u}_p^{(i)} \neq \mathbf{0}$ as desired.

To show the other implication, suppose that the ith signal component in the pth data set is uncorrelated among the ith group of components. Thus, p is not an element of $\mathcal{K}^{(i)}$. Therefore, the pth data set is not represented in $\mathbf{v}^{(i)}$ but rather corresponds to one of the

elements in the zero vector of $\tilde{\mathbf{v}}^{(i)}$. Using the definition of $\tilde{\mathbf{u}}^{(i)}$ and (6.23), it is easy to see that the pth part of $\mathbf{u}^{(i)}, \mathbf{u}_p^{(i)} = \mathbf{0}$. This implies that if the pth part of $\mathbf{u}^{(i)}$ associated with $\lambda^{(i)} > 1$, $\mathbf{u}_p^{(i)} \neq \mathbf{0}$, then the ith signal component in the pth data set is among the group of correlated ith components as desired. This concludes our proof for Result 6.3.

Result 6.3 assumes that any eigenvalue of $\mathbf{C}$ that is greater than one has an algebraic multiplicity of one. This is not a necessary but a sufficient condition. Section 6.4.1.3 discusses its sufficiency and also scenarios in which the correlation structure can be completely determined using Result 6.3 even when this assumption is not true.

Due to Results 6.1 and 6.3, if the ith eigenvalue of $\mathbf{C}$ is greater than one and is unique as an eigenvalue, the existence of correlation associated with the ith component of the pth data set can be determined by testing the hypothesis $\mathbf{u}_p^{(i)} = \mathbf{0}$. A bootstrap-based hypothesis test for this purpose is proposed in Section 6.4.2.2.

6.4.1.3. Identifiability of the underlying correlation structure

Results 6.1 and 6.3 state the conditions that allow us to determine the correlated components along with their correlation structure using the eigenvalue decomposition of $\mathbf{C}$. One additional assumption in Result 6.3 is that the eigenvalues of $\mathbf{C}$ greater than one are distinct, i.e., have algebraic multiplicity of one. In this section, we will briefly discuss why this assumption is needed. We will also mention the scenarios in which the correlation structure can still be completely determined using Result 6.3 even if the assumption is not true.

Let $\lambda^{(i)}$ and $\lambda^{(j)}$ be the two eigenvalues of $\mathbf{C}$ with $\lambda^{(i)} > 1$ and $\lambda^{(j)} > 1$. Let $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(j)}$ be the eigenvectors associated with $\lambda^{(i)}$ and $\lambda^{(j)}$, respectively. Let $\mathbf{u} = a\mathbf{u}^{(i)} + b\mathbf{u}^{(j)}$ be a vector formed by a linear combination of $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(j)}$, and a and b are scalars. If $\lambda^{(i)} = \lambda^{(j)}$, any linear combination of $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(j)}$ is an eigenvector of $\lambda^{(i)}$ or $\lambda^{(j)}$. In this case, if the ith and jth group of components are correlated across different data sets, their correlation structure, i.e., across which data sets the components are correlated, cannot always be determined using Result 6.3. For instance, if the ith components are correlated across all data sets except the pth data set, then according to Result 6.3, the pth part of $\mathbf{u}^{(i)}$, $\mathbf{u}_p^{(i)} = \mathbf{0}$. Similarly, $\mathbf{u}_q^{(j)} = \mathbf{0}$ if the jth components are correlated across all data sets except the qth data set. When $\lambda^{(i)} = \lambda^{(j)}$, then $a\mathbf{u}^{(i)} + b\mathbf{u}^{(j)}$ can also be an eigenvector of $\lambda^{(i)}$ or $\lambda^{(j)}$ for any a, b. Therefore, $\mathbf{u}_p^{(i)}$ or $\mathbf{u}_q^{(j)}$ are not necessarily equal to zero.

However, if the ith and jth components are correlated across the same subset of data sets, even when $\lambda^{(i)} = \lambda^{(j)}$, their correlation structure can be determined using Result 6.3. This

is due to the fact that the zeros in $\mathbf{u} = a\mathbf{u}^{(i)} + b\mathbf{u}^{(j)}$ are at the same positions as those of $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(j)}$ for any a, b.

To conclude, Result 6.3 can completely identify the correlation structure of the components when the eigenvalues associated with the components that are correlated across different subset of data sets are distinct.

6.4.2. Bootstrap-based tests for eigenvalues and eigenvectors of $\mathbf{C}$

6.4.2.1. Test for detecting eigenvalues of $\mathbf{C}$ associated with the correlated subspace

Result 6.1 gives conditions when the number of eigenvalues of $\mathbf{C}$ greater than one is equal to the dimension of the correlated subspace d. In practice, however, the composite coherence matrix $\mathbf{C}$ is unknown and has to be estimated from the samples. As a result, the number of eigenvalues of the sample composite coherence matrix that are greater than one will often not equal the dimension of the correlated subspace d. This inconsistency is addressed in the related model-order selection literature by setting a threshold for the eigenvalues that is determined with an ITC or via hypothesis testing.

Using the eigenvalues of a covariance matrix to estimate the dimension of a signal subspace is a well-studied paradigm [24]. However, most of these techniques are based on one or two data sets [26], [27], [80], [83]. In these cases, the population eigenvalues that do not belong to the signal subspace are assumed to satisfy a certain property that is independent of the unknown parameters. For example, with one data set, it is assumed that the eigenvalues of the covariance matrix of observed data with a signal subspace dimension of d satisfy

$$\lambda^{(1)} \geq \ldots \geq \lambda^{(d)} > \lambda^{(d+1)} = \ldots = \lambda^{(n)},$$

i.e., all eigenvalues following the largest d eigenvalues are equal (in most cases, assumed to be the noise variance). In the proposed multiset model, the eigenvalues of $\mathbf{C}$ following the d largest eigenvalues are not equal since some of the eigenvalues are less than one and the number of such eigenvalues depends on the unknown correlation structure of the d correlated components. Thus, there is no immediate generalization of the assumptions used in standard techniques based on ITC or hypothesis testing and we must formulate a novel approach.

To fill this void in the literature, we present a novel algorithm for determining d, which

also uses the hypothesis testing framework. As is common in model-selection literature, a sequence of binary hypothesis tests is performed one at a time until a stopping condition is met [27], [28], [58]. In this context, this means starting with a counter $s = 0$ and performing the following binary test of null hypothesis H_0 and alternative H_1

$$\begin{aligned} H_0 &: d = s, \\ H_1 &: d > s. \end{aligned} \tag{6.24}$$

If H_0 is rejected, s is incremented and another test of H_0 vs. H_1 is run. This is repeated until H_0 is not rejected or s reaches its maximum possible value. The binary test in (6.24) requires a statistic whose (asymptotic) distribution under H_0 is theoretically known or estimated from samples. In Section 5.3, the distribution of the statistic is derived by assuming a special correlation structure where the components are either correlated across all pairs of data sets or completely uncorrelated. This assumption might be applicable in some tasks, for example, in sensor array processing as in [30] where the same source vector is received by multiple arrays. However, in many other applications, this assumption is too restrictive [72], [73]. Since the distribution of the statistic for arbitrary correlation structures is unknown, we use the bootstrap technique to estimate this distribution. Another advantage of bootstrap is that it works well when only a limited number of samples are available. It also provides good results for non-Gaussian distributed data; a scenario for which the traditional ITC and hypothesis testing methods are ill-suited because they are predominantly based on asymptotic properties of Gaussian distributed data [61].

Let the eigenvalues of $\mathbf{C}$ be arranged in nonincreasing order so that $\lambda^{(1)} \geq \lambda^{(2)} \geq \ldots \geq \lambda^{(nP)}$. For each signal component that is independent of all other signal components, there is at least one eigenvalue of $\mathbf{C}$ equal to one. To estimate d, we propose a statistic based on the assumption that there is at least one independent component among all of the data sets. The statistic measures how the $(s+1)$st largest eigenvalue of $\mathbf{C}$ differs from one, and we estimate the significance of this deviation to test the null hypothesis, $H_0 : \lambda^{(s+1)} = 1$. In order to increase the power and stability of the test, we make the stronger assumption that each data set has at least one signal component that is completely uncorrelated in this manner, i.e., $d_{pq} < n$ for all p, q. Therefore, $\mathbf{C}$ has at least P eigenvalues equal to one and the null hypothesis for each test is

$$H_0 : \lambda^{(s+1)} = \lambda^{(s+2)} = \ldots = \lambda^{(s+P)} = 1. \tag{6.25}$$

Note that we cannot include all $nP - s$ eigenvalues following the s largest eigenvalues in the test since an unknown number of them are less than one. The proposed test statistic is

$$T(s) = \sum_{i=s+1}^{s+P} (\lambda^{(i)} - 1)^2, \tag{6.26}$$

and the null hypothesis is rejected when $T(s)$ is sufficiently greater from zero.

Under the null hypothesis, H_0, the statistic $T_0(s) = \sum_{i=s+1}^{s+P} (\lambda^{(i)} - 1)^2 = 0$. However, given our sample coherence matrix, it is unlikely that the sample statistic $T(s)$ is exactly equal to zero. To test whether our sample was generated under H_0 we need to estimate the distribution of $T(s) - T_0(s)$ [3].

This distribution is estimated via the bootstrap as follows. Given the sample matrices, compute $T(s)$. Resample the data by randomly choosing M indices from $\{1, \ldots, M\}$ (with uniform distribution and with replacement) to create a bootstrap dataset of the same size as the original data set. Repeat the resampling procedure B times and compute the test statistic in (6.26) each time to produce ${}_bT(s)$ for $b = 1, \ldots, B$. The distribution of $T(s) - T_0(s)$ under the null is then approximated by the bootstrap distribution ${}_bT^*(s) = {}_bT(s) - T(s)$ [61].

Algorithm 2 describes the complete technique for estimating d. The algorithm takes as input the sample matrices, the number of bootstrap resamples B and the probability of false alarm P_{fa}.

6.4.2.2. Test for eigenvectors of C corresponding to correlated components

In addition to identifying the dimension of the correlated subspace, d, our stated goal is to estimate the structure of the correlations between the collection of data sets. As a consequence of Result 6.3, we need to identify the values of i and p for which the subvector $\mathbf{u}_p^{(i)} = \mathbf{0}$ in order to determine which data sets have an uncorrelated ith signal component. However, we still do not have direct access to the composite coherence matrix. When $\mathbf{C}$ is estimated from samples, these subvectors will not be exactly zero. Thus we propose a novel method for identifying multiset correlation structure that uses a bootstrap-based test to detect zero subvectors in the eigenvectors of $\mathbf{C}$.

Assuming d correlated components (which can be estimated via Algorithm 2), the technique

[3] Compared to (2.22), we have omitted the absolute sign here as we are testing H_0 against a one-sided H_1.

Algorithm 2 Estimator for the dimension of the correlated subspace of P data sets

Input $\{\mathbf{X}_p\}_{p=1}^{P}$: sample matrices
B: number of bootstrap resamples
P_{fa}: probability of false alarm
Output $\hat{d}$: dimension of correlated subspace
function CORRDIM($\{\mathbf{X}_p\}_{p=1}^{P}, B, P_{\text{fa}}$)
  $\widehat{\mathbf{R}}_D \leftarrow \frac{1}{M}\text{blkdiag}\big(\mathbf{X}_1\mathbf{X}_1^T, \ldots, \mathbf{X}_P\mathbf{X}_P^T\big)$
  $\mathbf{X} \leftarrow \big[\mathbf{X}_1^T, \ldots, \mathbf{X}_P^T\big]^T$
  $\widehat{\mathbf{R}} \leftarrow \frac{1}{M}\mathbf{X}\mathbf{X}^T$
  $\widehat{\mathbf{C}} \leftarrow \widehat{\mathbf{R}}_D^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{R}}_D^{-1/2}$
  $\hat{\boldsymbol{\lambda}} \leftarrow \text{eigenvalues}\big(\widehat{\mathbf{C}}\big)$
    ▷ s.t. $\hat{\lambda}^{(1)} \geq \cdots \geq \hat{\lambda}^{(nP)}$
  for $b = 1, \ldots, B$ **do**
    ▷ bootstrap resamples indexed by left subscript
    for $m = 1, \ldots, M$ **do**
      ${}_bj_m \leftarrow$ random integer $[1, M]$
    ▷ resample indices chosen with replacement
    for $p = 1, \ldots, P$ **do**
      ${}_b\mathbf{X}_p \leftarrow \big[\mathbf{x}_p({}_bj_1), \ldots, \mathbf{x}_p({}_bj_M)\big]$
    ${}_b\widehat{\mathbf{R}}_D \leftarrow \frac{1}{M}\text{blkdiag}\big({}_b\mathbf{X}_1\,{}_b\mathbf{X}_1^T, \ldots, {}_b\mathbf{X}_P\,{}_b\mathbf{X}_P^T\big)$
    ${}_b\mathbf{X} \leftarrow \big[{}_b\mathbf{X}_1^T, \ldots, {}_b\mathbf{X}_P^T\big]^T$
    ${}_b\widehat{\mathbf{R}} \leftarrow \frac{1}{M}\,{}_b\mathbf{X}\,{}_b\mathbf{X}^T$
    ${}_b\widehat{\mathbf{C}} \leftarrow {}_b\widehat{\mathbf{R}}_D^{-1/2}\,{}_b\widehat{\mathbf{R}}\,{}_b\widehat{\mathbf{R}}_D^{-1/2}$
    ${}_b\hat{\boldsymbol{\lambda}} \leftarrow \text{eigenvalues}\big({}_b\widehat{\mathbf{C}}\big)$
    ▷ s.t. ${}_b\hat{\lambda}^{(1)} \geq \cdots \geq {}_b\hat{\lambda}^{(nP)}$
  $s_{\max} \leftarrow n - 1$
  for $s = 0, \ldots, s_{\max}$ **do**
    $T(s) \leftarrow \sum_{i=s+1}^{s+P}(\hat{\lambda}^{(i)} - 1)^2$
    for $b = 1, \ldots, B$ **do**
      ${}_bT(s) \leftarrow \sum_{i=s+1}^{s+P}\big({}_b\hat{\lambda}^{(i)} - 1\big)^2$
      ${}_bT^*(s) \leftarrow {}_bT(s) - T(s)$
    $\{{}_{(i)}T^*(s)\}_{i=1}^{B} \leftarrow \text{sort}\{{}_bT^*(s)\}_{b=1}^{B}$
    ▷ s.t. ${}_{(1)}T^*(s) \leq \ldots \leq {}_{(B)}T^*(s)$
    $\eta \leftarrow \lceil (1 - P_{fa})(B + 1) \rceil$
    ▷ index to select the threshold
    $T_\tau(s) \leftarrow {}_{(\eta)}T^*(s)$
return $\hat{d} \leftarrow \min\big\{\underset{s=0,\ldots,s_{\max}}{\arg\min}\ T(s) < T_\tau(s), n-1\big\}$

tests the d eigenvectors associated with the d largest eigenvalues of the sample composite coherence matrix. For $i = 1 \dots d$ and $p = 1 \dots P$ we test the hypotheses

$$\begin{aligned} H_0 &: \mathbf{u}_p^{(i)} = \mathbf{0}, \\ H_1 &: \mathbf{u}_p^{(i)} \neq \mathbf{0}. \end{aligned} \tag{6.27}$$

The null hypothesis is rejected when the squared Euclidean norm of $\mathbf{u}_p^{(i)}$,

$$T = \|\mathbf{u}_p^{(i)}\|^2, \tag{6.28}$$

is sufficiently far from zero. Under the null hypothesis, the statistic $T_0 = \|\mathbf{u}_p^{(i)}\|^2$ is zero. We estimate the distribution of $T - T_0$ using bootstrap. The resampling procedure is the same as in Section 6.4.2.1, but the bootstrap statistic is the resampled version of (6.28) so that the distribution under the null is approximated as ${}_bT^* = {}_bT - T$. The full procedure is given in Algorithm 3.

Combining Algorithms 2 and 3 leads to an effective method for determining the correlation structure among multiple data sets. Algorithm 2 determines how many signal components d have nonzero correlations, and Algorithm 3 reveals the data sets across which these d components are correlated. The final output is a binary matrix $\mathbf{Z}$, which is similar to Tables 6.1 and 6.2 except that nonzero correlation coefficients are represented by ones. We refer to this as a correlation map, an example of which can be seen in Figure 6.2 for the correlation structure in Table 6.2.

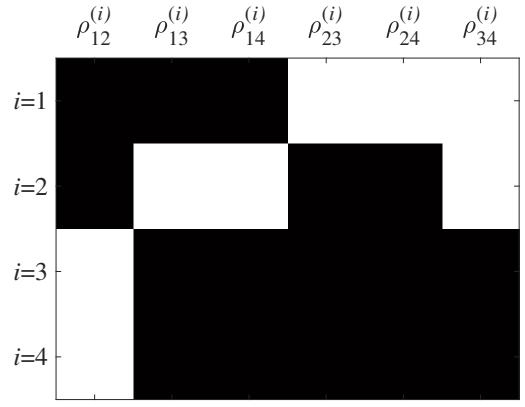


Figure 6.2.: The correlation map of four components correlated in four data sets with correlation coefficients given in Table 6.2. The white blocks represent nonzero correlation coefficients and the black blocks represent zero correlation coefficients.

Algorithm 3 Estimator for the correlation structure of P data sets

Input $\{\mathbf{X}_p\}_{p=1}^{P}$: sample matrices
$\hat{d}$: dimension of correlated subspace
B: number of bootstrap resamples
P_{fa}: probability of false alarm
Output $\widehat{\mathbf{Z}}$: correlation map
function CORRSTRUC($\{\mathbf{X}_p\}_{p=1}^{P}, \hat{d}, B, P_{\text{fa}}$)
  $\widehat{\mathbf{Z}} \leftarrow [\mathbf{1}] \in \mathbb{R}^{\hat{d}\times\binom{P}{2}}$
  ▷ rows indexed by signal components, columns by pairs of data sets in lexicographical order
  $\widehat{\mathbf{R}}_D \leftarrow \frac{1}{M}\text{blkdiag}\big(\mathbf{X}_1\mathbf{X}_1^T, \ldots, \mathbf{X}_P\mathbf{X}_P^T\big)$
  $\mathbf{X} \leftarrow \big[\mathbf{X}_1^T, \ldots, \mathbf{X}_P^T\big]^T$
  $\widehat{\mathbf{R}} \leftarrow \frac{1}{M}\mathbf{X}\mathbf{X}^T$
  $\widehat{\mathbf{C}} \leftarrow \widehat{\mathbf{R}}_D^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{R}}_D^{-1/2}$
  for $i = 1, \ldots, \hat{d}$ **do**
    $\hat{\mathbf{u}}^{(i)} \leftarrow \text{eigenvector}(\widehat{\mathbf{C}})$
    ▷ ordered by associated eigenvalue s.t. $\hat{\lambda}^{(1)} \geq \cdots \geq \hat{\lambda}^{(\hat{d})}$
    $\hat{\mathbf{u}}^{(i)} = \Big[\hat{\mathbf{u}}_1^{(i)T}, \ldots, \hat{\mathbf{u}}_P^{(i)T}\Big]^T$ with $\hat{\mathbf{u}}_p^{(i)} \in \mathbb{R}^n \ \forall p$
  for $b = 1, \ldots, B$ **do**
    ▷ bootstrap resamples indexed by left subscript
    for $m = 1, \ldots, M$ **do**
      ${}_bj_m \leftarrow$ random integer $[1, M]$
    ▷ resample indices chosen with replacement
    for $p = 1, \ldots, P$ **do**
      ${}_b\mathbf{X}_p \leftarrow \big[\mathbf{x}_p({}_bj_1), \ldots, \mathbf{x}_p({}_bj_M)\big]$
    ${}_b\widehat{\mathbf{R}}_D \leftarrow \frac{1}{M}\text{blkdiag}\big({}_b\mathbf{X}_1\,{}_b\mathbf{X}_1^T, \ldots, {}_b\mathbf{X}_P\,{}_b\mathbf{X}_P^T\big)$
    ${}_b\mathbf{X} \leftarrow \big[{}_b\mathbf{X}_1^T, \ldots, {}_b\mathbf{X}_P^T\big]^T$
    ${}_b\widehat{\mathbf{R}} \leftarrow \frac{1}{M}\,{}_b\mathbf{X}\,{}_b\mathbf{X}^T$
    ${}_b\widehat{\mathbf{C}} \leftarrow {}_b\widehat{\mathbf{R}}_D^{-1/2}\,{}_b\widehat{\mathbf{R}}\,{}_b\widehat{\mathbf{R}}_D^{-1/2}$
    for $i = 1, \ldots, \hat{d}$ **do**
      ${}_b\hat{\mathbf{u}}^{(i)} \leftarrow \text{eigenvector}({}_b\widehat{\mathbf{C}})$
    ▷ ordered by eigenvalue s.t. ${}_b\hat{\lambda}^{(1)} \geq \cdots \geq {}_b\hat{\lambda}^{(\hat{d})}$
      ${}_b\hat{\mathbf{u}}^{(i)} = \Big[{}_b\hat{\mathbf{u}}_1^{(i)T}, \ldots, {}_b\hat{\mathbf{u}}_P^{(i)T}\Big]^T$ with ${}_b\hat{\mathbf{u}}_p^{(i)} \in \mathbb{R}^n$
  for $i = 1, \ldots, \hat{d}$ **do**
    for $p = 1, \ldots, P$ **do**
      $T \leftarrow \|\hat{\mathbf{u}}_p^{(i)}\|^2$
      for $b = 1, \ldots, B$ **do**
        ${}_bT \leftarrow \|{}_b\hat{\mathbf{u}}_p^{(i)}\|^2$
        ${}_bT^* \leftarrow {}_bT - T$
      $\{{}_{(i)}T^*\}_{i=1}^{B} \leftarrow \text{sort}\{{}_bT^*\}_{b=1}^{B}$
    ▷ s.t. ${}_{(1)}T^* \leq \ldots \leq {}_{(B)}T^*$
      $\eta \leftarrow \lceil(1 - P_{\text{fa}})(B + 1)\rceil$
    ▷ index to select the threshold
      $T_\tau \leftarrow {}_{(\eta)}T^*$
      if $T < T_\tau$ **then**
        $\widehat{\mathbf{Z}}(i, j\{p, q\}) \leftarrow 0 \ \forall q$
    ▷ $j\{p, q\}$ gets linear index of p, q in lexicographical order
return $\widehat{\mathbf{Z}}$

6.4.2.3. Computational complexity

We compute the big-$\mathcal{O}$ complexity for the proposed technique. The EVD of $\hat{\mathbf{C}}$ can be computed using the right singular vectors of the data sets as follows. Let $\mathbf{X}_p = \mathbf{F}_p\mathbf{K}_p\mathbf{G}_p^T$ be the economy SVD of $\mathbf{X}_p$, which can be computed for all P data sets (assuming $n \leq M$) in $\mathcal{O}(MPn^2)$ flops. Let $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_P]$. The eigenvalues of $\hat{\mathbf{C}}$ are the squared singular values of $\mathbf{G}$, and the eigenvectors of $\hat{\mathbf{C}}$ are the right singular vectors of $\mathbf{G}$, which can be computed in $\mathcal{O}(MP^2n^2)$ flops. Thus, assuming $Pn \leq M$, the eigenvalues and eigenvectors of $\hat{\mathbf{C}}$ can be computed in $\mathcal{O}(MP^2n^2 + Pn)$ flops, where the additional Pn flops are required for squaring the singular values of $\mathbf{G}$. Since we compute the EVD of $\hat{\mathbf{C}}$ once for the original data and once for each bootstrap resample, the final complexity is $B + 1$ times larger. Reporting only the dominant term, the complexity of the proposed technique is $\mathcal{O}(BMP^2n^2)$. Finally, it is worth noting that some of the operations, e.g., computing the SVD for P data sets, bootstrap resamples etc., can be parallelized to reduce the overall complexity of the technique.

6.5. Towards complete model selection in multiple high-dimensional data sets

In this section, we will briefly address the problem of estimating the complete correlation structure when the number of samples M is not large relative to the dimension n and the number of data sets P. To be concise and clear, from now on we will call the technique proposed in Section 6.3 as the mCCA-HT technique since it combines multiset CCA and hypothesis testing for each pair of data sets. We will call the technique proposed in Section 6.4 as the joint-EVD technique since it is based on the eigenvalues and eigenvectors of the composite coherence matrix which measure the joint-correlation information among all sets. Both these techniques are based on the composite coherence matrix $\mathbf{C}$.

Let us consider the regime where $M < nP$. In this case, the sample estimate of $\mathbf{C}$, denoted by $\hat{\mathbf{C}}$ and computed in line 7 of Algorithm 2, is not full rank. Since the maximum rank of $\hat{\mathbf{C}}$ is $\min(M, nP)$, where $\min(\cdot)$ denotes the minimum operator, $\hat{\mathbf{C}}$ has the maximum rank of M. Therefore, $nP - M$ smallest eigenvalues of $\hat{\mathbf{C}}$ are equal to zero. To briefly motivate how this regime affects the proposed techniques, consider the mCCA-HT technique. If the mCCA-HT technique uses either the MINVAR or GENVAR cost function to perform mCCA, the estimates of the canonical variables will be highly unreliable in this regime. This is because

the MINVAR is based on the smallest eigenvalue of $\hat{\mathbf{C}}$ while the GENVAR is based on the determinant of $\hat{\mathbf{C}}$ [21]. Both these quantities are deterministically zero when $M < nP$. As the performance of all mCCA versions is similar as shown in [23], mCCA with MAXVAR, SUMCORR and SSQCORR will also be unreliable in this regime. It is shown in [13] that the largest eigenvalue of $\hat{\mathbf{C}}$ is highly overestimated when M is small compared to nP. Thus, the joint-EVD technique will also perform poorly when $M < nP$. It is interesting to see that for $P = 2$, this is the same regime where the sample canonical correlations are defective [35]. Similar to CCA, even if $M > nP$ but not significantly large, the sample eigenvalues of $\hat{\mathbf{C}}$ will be far from their population counterparts, thus making the correlation structure estimate unreliable. This calls for rank reduction either before or jointly with the correlation structure estimation.

In [13], a PCA pre-processing step is proposed before performing mCCA. However, as discussed before in Chapters 3 and 5, PCA retains the components with most variance within a data set and these components are not necessarily the ones that are correlated across multiple data sets. If the components are retained only on the basis of their variance, then the PCA step before the multiset correlation analysis will most likely retain unnecessary uncorrelated components which have smaller variance compared to the correlated components.

We propose an improved solution where the PCA rank for each data set is determined using the joint PCA-CCA technique of Section 3.5. The PCA dimensions for a particular data set are chosen as the maximum of the estimated PCA ranks using the joint PCA-CCA technique from all pair combinations for that data set. This approach is based on the fact that the components correlated across all data sets are also correlated across a given pair of data sets. Hence, these components are retained in the dimension-reduced data sets using this approach. The proposed dimension reduction technique is summarized in Algorithm 4. The mCCA-HT or the joint-EVD technique is applied to the dimension-reduced data sets to estimate the correlation structure.

The proposed dimension reduction approach, however, is suboptimal as the estimated PCA rank for each data set is a function of all the PCA ranks estimated in a pairwise fashion with that data set. Therefore, an inaccurate rank estimation in one pair of data sets can lead to an inaccurate overall rank and thus, an incorrect correlation structure estimation. In comparison, a joint approach for estimating the PCA ranks and the correlation structure could lead to significant performance improvement. However, the two main challenges with such an approach are a) the number of PCA ranks to estimate increases with the number of sets, thereby increasing the number of rank combinations to search for the correct correlation

Algorithm 4 Estimator for the PCA ranks to determine the correlation structure in P high-dimensional data sets

Input $\{\mathbf{X}_p\}_{p=1}^{P}$: sample matrices
Output $\widehat{\mathbf{r}}$: P-dimensional vector containing estimated PCA rank for each data set
function MULTISETRANK($\{\mathbf{X}_p\}_{p=1}^{P}$)
    $\mathbf{r}_{\mathrm{m}} \leftarrow [\mathbf{0}] \in \mathbb{R}^{P \times P}$
    ▷ rows and columns indexed by data sets
    $\mathbf{comb} \leftarrow \binom{P}{2}$-dimensional cell
    ▷ each cell element is a two-dimensional vector containing indices of a pair of data sets in lexicographical order
    for $c = 1, \ldots, \binom{P}{2}$ **do**
        $p \leftarrow \mathbf{comb}\{c\}(1)$
        ▷ first data set in pair c
        $q \leftarrow \mathbf{comb}\{c\}(2)$
        ▷ second data set in pair c
        $\hat{r}_p, \hat{r}_q \leftarrow$ PCACCADetector($\mathbf{X}_p, \mathbf{X}_q$)
        ▷ joint PCA-CCA technique of Section 3.5, $\hat{r}_p, \hat{r}_q$ are PCA ranks estimated for $\mathbf{X}_p, \mathbf{X}_q$ that keep all the components correlated between $\mathbf{X}_p, \mathbf{X}_q$
        $\mathbf{r}_{\mathrm{m}}(p, q) \leftarrow \hat{r}_p$
        $\mathbf{r}_{\mathrm{m}}(q, p) \leftarrow \hat{r}_q$
    $\widehat{\mathbf{r}} = \mathrm{Rowmax}(\mathbf{r}_{\mathrm{m}})$
    ▷ returns a vector containing the maximum value from each row
return $\widehat{\mathbf{r}}$

structure, and b) two different rank combinations can lead to the same model order d but a different correlation structure. Therefore, such a joint approach is left for future work.

6.6. Numerical results

In this section, we use Monte-Carlo simulations to demonstrate the performance of the proposed techniques. Initially, we compare these techniques with those proposed in Chapter 5 and their competitors [30], [34], which aim to estimate the number of components correlated across all data sets, d_{all}. To estimate d_{all} for mCCA-HT and joint-EVD techniques, we simply count the number of components that are correlated across all the data sets. Next, we investigate the behavior of the joint-EVD technique when the pairwise correlation coefficients are not above the threshold necessary for the proof of Result 6.1.We show that the method is robust to the violation of this assumption and that the accuracy remains high for many correlation structures. Finally, we compare the performance of the mCCA-HT and joint-EVD techniques for determining the complete correlation structure in multiple data sets. These comparisons highlight the quantitative and qualitative differences between the two techniques. We use the MAXVAR cost function for the mCCA-HT technique. This is

because the solution of MAXVAR mCCA is closed-form and is a function of the eigenvectors of the composite coherence matrix, which are also employed in the joint-EVD technique. This helps in an accurate analysis of the differences between the two proposed techniques in principle and eliminates the differences that can arise due to convergence issues in other mCCA cost functions like GENVAR, SSQCOR and SUMCOR. Moreover, all five MCCA cost functions are closely linked and perform similarly as reported in [23].

We present results with different correlation structures.The following simulation settings are common to all of them. The signal components in each data set have unit variance. For all scenarios except iii) and iv), the signal components are Gaussian distributed. For scenarios iii) and iv), the signals are generated from a Laplacian distribution to demonstrate that the techniques also perform well with non-Gaussian signals. The mixing matrices are randomly generated orthogonal matrices. Each data set is corrupted by additive white Gaussian noise. The variance of noise components is chosen according to the SNR that is defined per component in equation (3.37). The SNR is the same for all data sets. The number of bootstrap resamples is $B = 1000$ and the probability of false alarm is $P_{\text{fa}} = 0.05$. The performance plots are mostly shown either as a function of SNR, which is varied from -10 dB to 15 dB or as a function of the number of samples. The results are averaged over 500 independent trials. The performance of each method for determining model order is measured by the mean accuracy (number of correct estimates divided by number of trials) or the mean value (average value over all trials). The performance in estimating the complete correlation structure is measured using precision, i.e., the percentage of correctly detected correlations among all the detected correlations, and recall, i.e., the percentage of correctly detected correlations among all actual correlations.

The four different scenarios are the following:

i) **Evaluation of model-order selection with special correlation structure**, for $P = 4$ data sets with $d = d_{\text{all}} = 3$: Each data set is of dimension $n = 7$ and the number of samples is $M = 350$. The components are either correlated across all data sets or are uncorrelated. Thus, the number of components that are correlated between at least a pair of data sets, d, is equal to the number of components correlated across all data sets, d_{all}. This type of correlation structure satisfies the special correlation assumption of Section 5.3. The pairwise correlation coefficients for the three correlated components are shown in Table 6.3, all of which exceed the $\xi^{(i)} = (3/4)^2 = 0.5625$ threshold as assumed by Result 6.1.

   Figure 6.3 shows the mean accuracy of $\hat{d}_{\text{all}}$ as a function of SNR for the proposed and

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{34}^{(i)}$
$i=1$	0.63	0.78	0.69	0.81	0.64	0.91
$i=2$	0.62	0.67	0.74	0.71	0.82	0.91
$i=3$	0.84	0.81	0.72	0.57	0.71	0.62

Table 6.3.: Correlation structure of the three correlated components in four data sets used in scenario i).

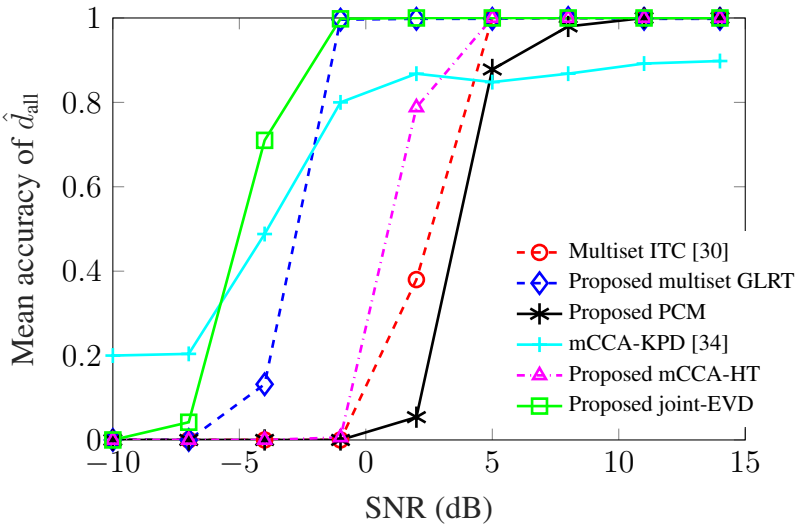


Figure 6.3.: Mean accuracy of $\hat{d}_{\text{all}}$ in scenario i) for the proposed and the competing techniques in detecting three components correlated across all four data sets.

the competing techniques. All the techniques correctly estimate d_{all} when the SNR is high. When the SNR is low, the joint-EVD approach outperforms all other techniques.

ii) **Evaluation of model-order selection with arbitrary correlation structure**, for $P = 4$ data sets with $d = 3, d_{\text{all}} = 1$: In this setting all the parameters are the same as in the previous scenario except that the first component of each data set is correlated with all other data sets but the other two components are only correlated across a subset of data sets. The second component is correlated across all except the first data set and the third component is correlated between data sets two and four. The pairwise correlation coefficients are shown in Table 6.4. The proposed multiset GLRT and the multiset ITC method of [30] are not evaluated as they are inapplicable in this setting.

Figure 6.4 shows that the mCCA-HT and joint-EVD techniques work better than the PCM and mCCA-KPD technique of [34] in estimating the model order d_{all} for low values of SNR. It is also worth noting that the PCM and mCCA-KPD techniques estimate only d_{all} while the methods proposed in this chapter also detect the components corre-

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{34}^{(i)}$
$i=1$	0.63	0.78	0.69	0.81	0.64	0.91
$i=2$	0	0	0	0.71	0.82	0.91
$i=3$	0	0	0	0	0.71	0

Table 6.4.: Correlation structure of the three correlated components in four data sets used in scenario ii).

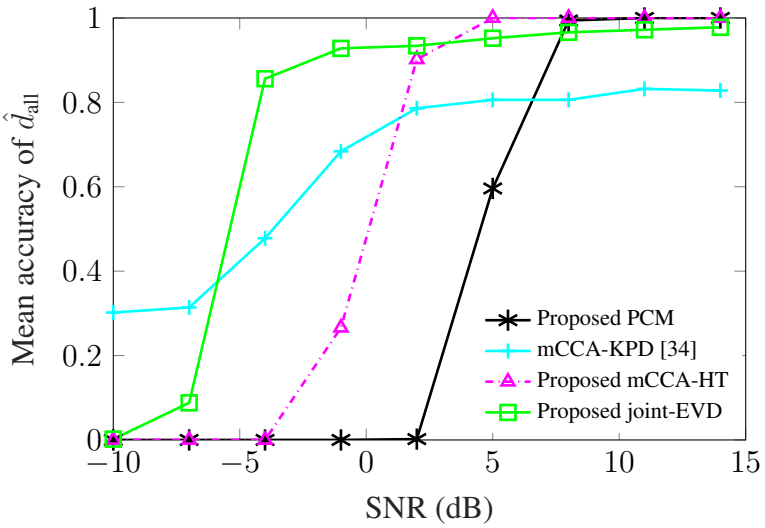


Figure 6.4.: Mean accuracy of $\hat{d}_{\text{all}}$ in scenario ii) for the proposed and the competing techniques in detecting $d_{\text{all}} = 1$ component correlated across all four data sets, in presence of two signal components correlated across a subset of the data sets, i.e., $d = 3$.

lated across subsets of the data sets along with their correlation structure.

iii) **Performance of the joint-EVD method when the element-wise threshold is not met**, for $P = 5$ data sets with $d = d_{\text{all}} = 2$: We also investigate the performance of the joint-EVD technique for determining the number of correlated components when some of the pairwise correlation coefficients do not meet the threshold required for Result 6.1. For this, we assume that the first two components in each data set are correlated across all data sets. Thus, the threshold for the pairwise correlation coefficients given by Result 6.1 is $\xi^{(1)} = \xi^{(2)} = \xi = (\frac{4}{5})^2 = 0.64$. We keep some of the pairwise correlation coefficients above the threshold ξ and vary the remaining ones. More precisely, we set $\rho_{pq}^{(i)} = 0.75$ for $p > q$, $p = 2, 3, 4, 5$ and $q = 3, 4, 5$, which exceeds the threshold ξ, and jointly vary the remaining correlation coefficients $\rho_{12}^{(i)} = \rho_{13}^{(i)} = \rho_{14}^{(i)} = \rho_{15}^{(i)} = \rho$, for $i = 1, 2$. All the pairwise correlation coefficients are listed in Table 6.5.

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{15}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{25}^{(i)}$	$\rho_{34}^{(i)}$	$\rho_{35}^{(i)}$	$\rho_{45}^{(i)}$
$i=1$	ρ	ρ	ρ	ρ	0.75	0.75	0.75	0.75	0.75	0.75
$i=2$	ρ	ρ	ρ	ρ	0.75	0.75	0.75	0.75	0.75	0.75

Table 6.5.: Correlation structure of the two correlated components in five data sets used in scenario iii).

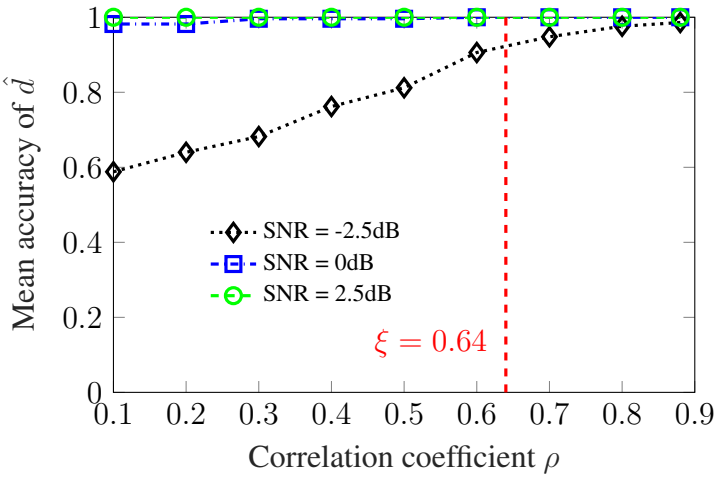


Figure 6.5.: Mean accuracy of the joint-EVD technique for estimating $d = 2$ components correlated in five data sets as a function of the correlation coefficient ρ in scenario iii).

To demonstrate the relative robustness of the method against violating the assumption in Result 6.1, we show the accuracy of $\hat{d}$ as a function of ρ in Figure 6.5 for different values of SNR. For $\rho < \xi$, the performance depends on the SNR. For low SNR, it becomes increasingly difficult to correctly estimate d for only weakly correlated components. On the other hand, as long as the SNR is high enough, violating the threshold in Result 6.1 does not present a problem, and d can still be correctly determined.

iv) **Evaluation of complete correlation structure**,

A. for $P = 5$ data sets with $d = 3$, $d_{\text{all}} = 1$: Finally, we compare the performance of the mCCA-HT and joint-EVD techniques for determining the complete correlation structure. Each data set has $n = 4$ components and the number of samples is $M = 250$. Of the $d = 3$ correlated components, $d_{\text{all}} = 1$ and the first component of each data set is correlated with the first component of each other data set. The second component of each data set is correlated across all the data sets except the fourth data set. Finally, the third components of data sets one, four and five are correlated. Each pairwise correlation coefficient is 0.7, thus exceeding the threshold as required

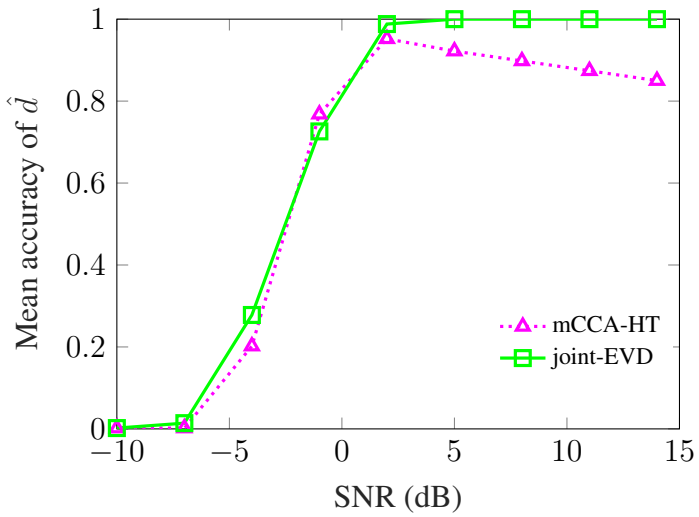
in Result 6.1.

Figure 6.6a shows the mean accuracy of estimating d for both techniques as a function of SNR. Both methods perform similarly for low values of SNR when estimating d. However, for medium and high SNR values the joint-EVD method outperforms the mCCA-HT method. The loss in performance can be described by the fact that the mCCA-HT is based on hypothesis test derived for Gaussian distributed data. This assumption is violated in this scenario as the signals are Laplacian distributed.

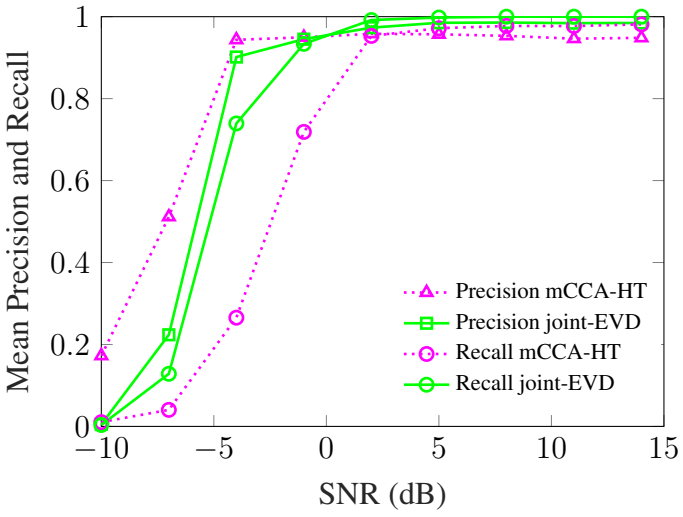
Figure 6.6b shows the mean precision and recall values for determining the complete correlation structure. The precision of the mCCA-HT method is better at low SNR while the recall for the joint-EVD method is better at both low and medium SNRs. The main reason for this difference is tied to the information that each method uses for hypothesis testing. The mCCA-HT method performs mCCA to extract the sets of signal components that are highly correlated via a deflationary approach. The components are extracted jointly from all data sets. Then, hypothesis tests are conducted on the extracted signal components from pairs of data sets to detect the underlying correlation structure. At low SNR values some of the correlations are missed while testing an individual pair of data sets, so its average recall is small in this regime.

On the other hand, the joint-EVD method detects the components and their correlation structure by applying hypothesis tests to the eigenvalues and eigenvectors of the composite coherence matrix directly. The eigenvalues and the eigenvectors are functions of *all* the pairwise correlation coefficients associated with the component and thus, this method tests on this joint information.

This is further illustrated in Figure 6.8, which shows heat maps of the average accuracy for the two methods in estimating the complete correlation structure for this scenario. The average accuracy for correlation structure is computed as the mean value of each element in the estimated correlation map across all trials. The true correlation structure is visualized in Figure 6.7 using a binary map. This map mirrors the structure of Tables 6.1, 6.2 and 6.3 but represents the nonzero correlation coefficients with white blocks and the zero correlation coefficients with black blocks. This binary map can be compared to heat maps of the simulation results to assess qualitative differences between the two methods. Ideally, these heat maps should look exactly like the binary map in Figure 6.7. The green star symbols in the heat maps indicate correlated components.



(a)



(b)

Figure 6.6.: Performance of the proposed techniques for determining the complete correlation structure in five data sets in scenario iv)A. a) Mean accuracy of estimating d, the total number of correlated signal components b) Precision and recall for determining the complete correlation structure of the detected components.

In the low SNR regime (shown in Figure 6.8a,b at SNR = -7dB), both techniques detect very few correlations as illustrated by the dark color of the boxes. Some of the boxes in the second row corresponding to $i = 2$ in Figure 6.8a that should be black (boxes without green star symbol) are dark brown, indicating that the joint-EVD method detects a few false positives, and therefore has low precision value compared to the mCCA-HT method.

The main differences between performance of the two methods start to appear from

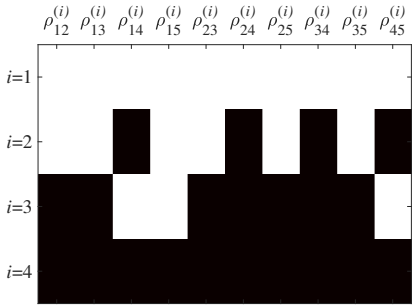


Figure 6.7.: The correlation map of four components correlated in five data sets in scenario iv)A.

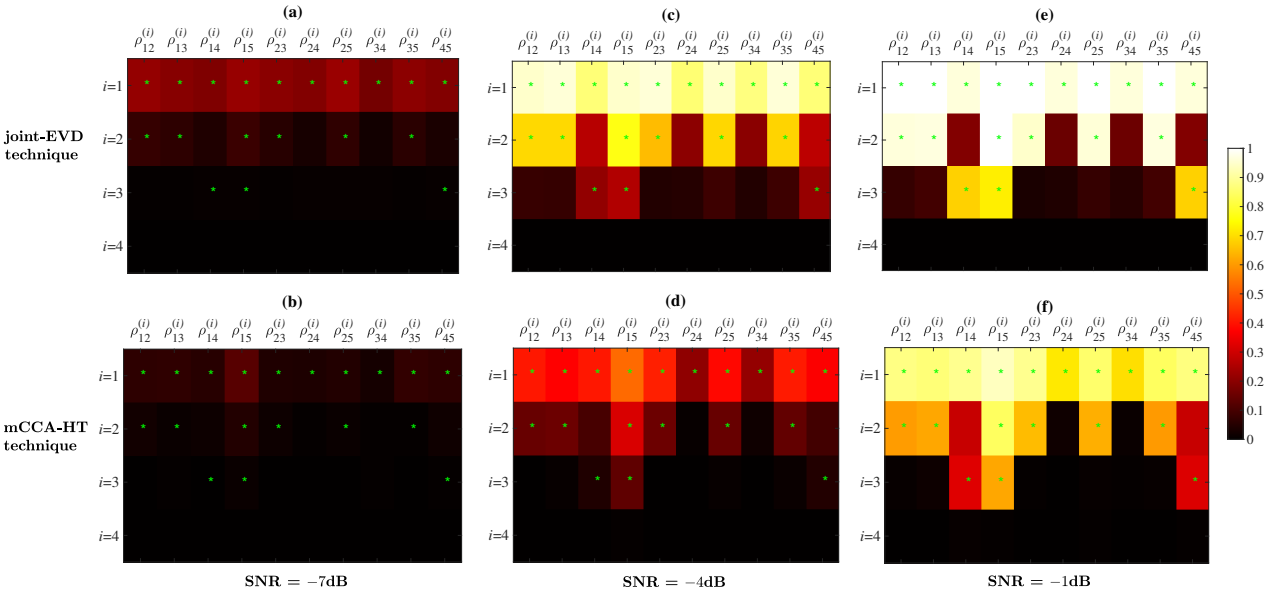


Figure 6.8.: Heat maps showing the mean accuracy of detecting individual correlations for the joint-EVD and mCCA-HT methods at three different SNR values of -7dB, -4dB and -1dB in scenario iv)A. The true correlation structure is shown in Figure 6.7. The green star symbols indicate correlated components.

SNR of -4dB. Figure 6.8c,d and 6.8e,f show the heat maps for SNR$=-4$dB and SNR$=-1$dB, respectively. In Figure 6.8d,f, boxes corresponding to the nonzero correlations of same component number have different colors. This is expected because mCCA-HT method conducts tests on pairs of data sets so it is possible to detect the correlation between one pair and miss it between another. On the other hand, the joint-EVD technique tests the eigenvalues and eigenvectors of the composite coherence matrix, which provides a summary of all the pairwise correlations. Therefore, the boxes of the nonzero correlations corresponding to components with same component number are more uniform in color in Figure 6.8c,e.

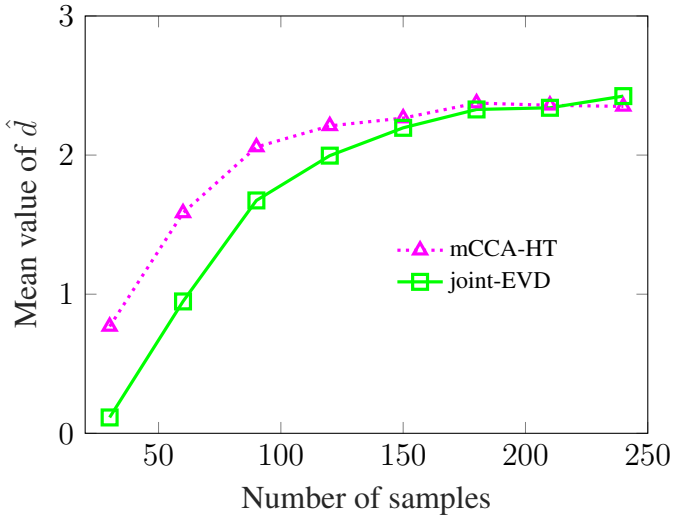
At low SNR values, the number of data sets correlated across a particular component affects the accuracy of the joint-EVD method. The more data sets that are correlated along a given component, the better the joint-EVD method performs. This can be observed in Figure 6.8c. The boxes of the first row ($i = 1$) corresponding to nonzero correlations are significantly brighter than those of second and third rows ($i = 2, 3$), indicating a higher accuracy when detecting the first component. Similarly in Figure 6.8e, there is a contrast between the high accuracy when detecting the correlations of first and second components and the relatively lower accuracy of detecting the third component. This is because the eigenvalue associated with the component correlated across more data sets is significantly greater than one and thus makes its detection possible even when the noise power is high. This stands in contrast with the mCCA-HT method where no advantage is gained by the number of data sets across which a component is correlated. In Figure 6.8d,f, we can see that boxes corresponding to the nonzero correlations of $i = 1, 2, 3$ have less variation in color, indicating similar accuracy across the board.

B. for $P = 5$ data sets with $d = 3$, $d_{\text{all}} = 0$: As seen from the previous experiment, compared to the mCCA-HT technique, the joint-EVD technique benefits from components being correlated across more sets. However, when the correlation structure is sparse, i.e., the components are correlated across a few data sets only, the mCCA-HT technique which detects pairwise model orders is likely to perform better. This is due to the fact that for limited number of samples, the benefits obtained from jointly detecting components correlated across a few data sets in the joint-EVD technique are outweighed by the disadvantage introduced by estimating large number of parameters in $\mathbf{C}$ and using these unreliable estimates for hypothesis testing. We validate this in the current experiment. The correlation structure for five data sets is shown in Table 6.6. The mean value of $\hat{d}$ and the average precision and recall as the function of number of samples are shown in Figure 6.9a and 6.9b, respectively. The SNR is 3dB. For low number of samples, the mCCA-HT technique outperforms the joint-EVD technique in estimating d and the correlation structure. For large number of samples, however, both techniques perform similarly.

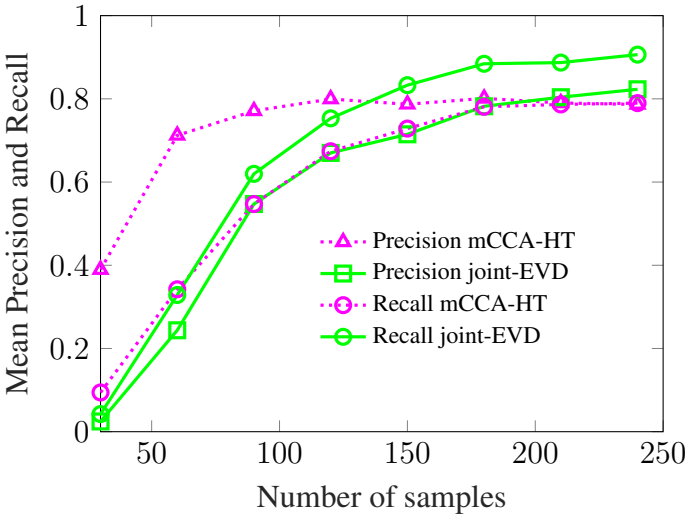
v) **Computational complexity** for $P = 6$ data sets with $d = d_{\text{all}} = 4$: The complexity analysis for the two proposed techniques is explained in Section 6.3.2.1 and Section 6.4.2.3. The complexity of the mCCA-HT technique grows with the dimension of the data sets at a higher rate compared the complexity of the joint-EVD technique. On the other hand, the complexity of the joint-EVD technique grows linearly with the number

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{15}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{25}^{(i)}$	$\rho_{34}^{(i)}$	$\rho_{35}^{(i)}$	$\rho_{45}^{(i)}$
$i=1$	0.9	0.9	0	0	0.9	0	0	0	0	0
$i=2$	0	0	0	0	0	0	0	0.7	0.7	0.7
$i=3$	0	0	0	0	0	0	0	0	0	0.8

Table 6.6.: Correlation structure of the three correlated components in five data sets used in scenario iv)B.



(a)



(b)

Figure 6.9.: Performance of the proposed techniques for determining the complete correlation structure in five data sets in scenario iv)B. a) Mean value of $\hat{d}$, the total number of correlated signal components b) Precision and recall for determining the complete correlation structure of the detected components.

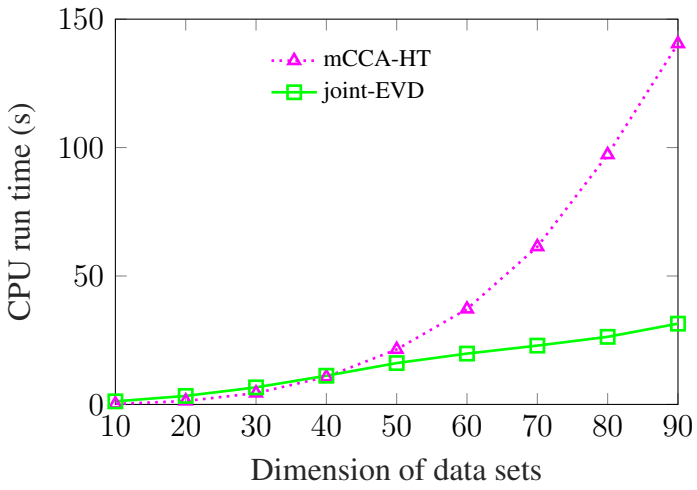


Figure 6.10.: Average CPU run time for estimating $d = 4$ components correlated in six data sets as a function of the dimension of the data sets.

of bootstrap resamples. In this experiment, we compare the complexities of the two techniques using the average CPU run time. It should be noted that the CPU run time is an approximate proxy for the computational complexity and our aim here is to compare the complexities of the two techniques and draw inferences, and not to draw conclusions on the absolute values of the CPU run time for each technique.

Figure 6.10 shows the average CPU run time in seconds for both techniques as a function of the dimension of the data sets. In this setting, $M = B = 500$. For smaller dimensions, the joint-EVD technique needs more time compared to mCCA-HT technique due to bootstrapping. For dimensions higher than 50, the mCCA-HT method needs significantly more computation time compared to the joint-EVD technique. We also present the precision and recall accuracies in estimating the correlation structure in Figure 6.11. For this scenario, both techniques perform well as the dimension of the data sets increase.

vi) **Evaluation of complete correlation structure in high-dimensional data sets** for $P = 5$ data sets with $n = 50, d = 4, d_{\text{all}} = 1$: In this setting, the data sets are high-dimensional and we show the performance of the techniques as a function of the number of samples M. The correlation structure of $d = 4$ correlated components is given in Table 6.7. There are two uncorrelated components with variance twice than the variance of the correlated components. The SNR is 10dB.

We applied Algorithm 4 on page 101 to estimate the PCA ranks and then used the mCCA-HT and joint-EVD techniques to estimate the correlation structure in the rank-

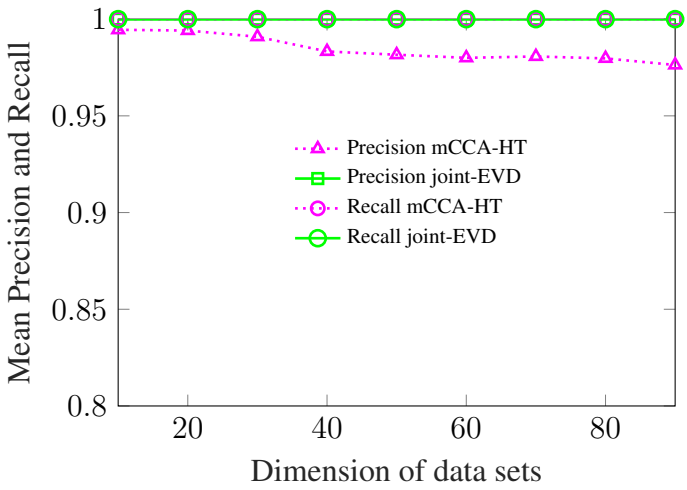


Figure 6.11.: Mean precision and recall for joint-EVD and mCCA-HT techniques for determining the correlation structure in scenario v) as a function of the dimension of the data sets.

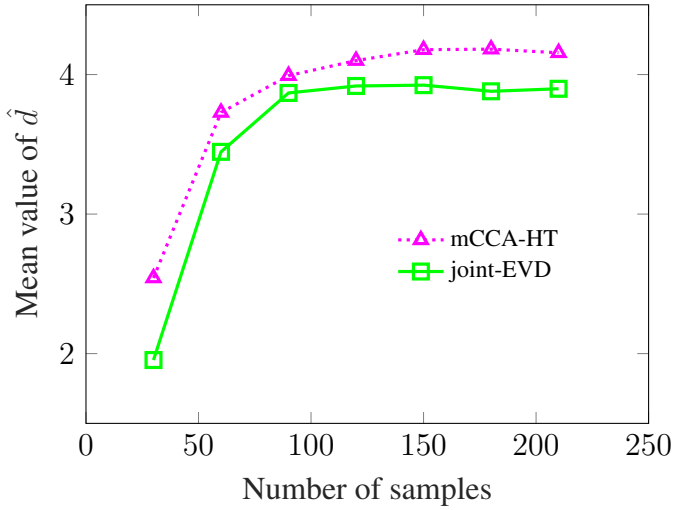
	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{15}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{25}^{(i)}$	$\rho_{34}^{(i)}$	$\rho_{35}^{(i)}$	$\rho_{45}^{(i)}$
$i=1$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
$i=2$	0	0	0	0	0.85	0.85	0.85	0.85	0.85	0.85
$i=3$	0	0.78	0	0.78	0	0	0	0	0.78	0
$i=4$	0	0	0.7	0.7	0	0	0	0	0	0.7

Table 6.7.: Correlation structure of four correlated components in five data sets used in scenario vi).

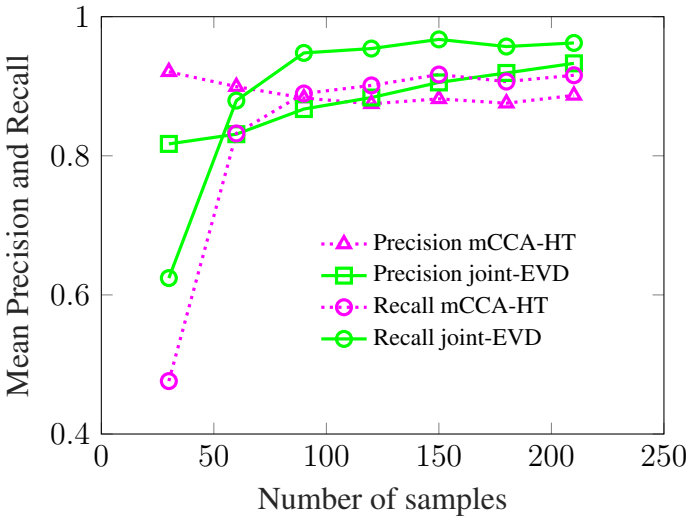
reduced data sets. The mean value of $\hat{d}$ and the mean precision and recall are shown in Figure 6.12. We specifically show the regime $M < nP$ ($nP = 250$ in this scenario), where the estimate of $\mathbf{C}$ without any rank reduction is highly unreliable. It can be observed that both techniques work well in this regime after the PCA rank reduction. However, as pointed out in Section 6.5, the dimension reduction step can be improved by jointly estimating the PCA ranks across all data sets compared to using a pair of data sets as in the current approach.

6.7. Summary

We have presented two techniques: the mCCA-HT and the joint-EVD techniques, which solve the model-selection problem to determine the complete correlation structure in multiple sets of data, i.e., identifying which components are correlated across which data sets.



(a)



(b)

Figure 6.12.: Performance of the proposed techniques for determining the complete correlation structure of $d = 4$ components in five high-dimensional data sets in scenario vi). a) Mean value of $\hat{d}$ b) Precision and recall for determining the complete correlation structure of the detected components.

For the joint-EVD technique, we also provide the necessary and sufficient conditions under which the correlation structure can be identified, and justify these conditions theoretically. Both techniques have shown competitive performance and broad applicability in various simulation scenarios. The two techniques are complementary in the sense that the mCCA-HT technique performs well compared to the joint-EVD technique with sparse correlation structure, i.e., correlation structure with only a few nonzero correlation coefficients. On the other hand, the joint-EVD technique performs better with dense correlation structures. An approach to combine the advantages of both these techniques is an interesting avenue for fu-

ture research. A straightforward solution for applying both techniques in presence of small number of samples is also presented and empirically evaluated. However, with multiple data sets, the required number of samples for the techniques to work highly depends on the number of data sets and correlation structure among the latent components.

6.8. Appendix - Number of positive eigenvalues of a hollow symmetric matrix

Crucial to the goal of identifying correlated signals via the eigenvalues of $\mathbf{C}$ is a requirement that the correlated subspace of signal components associated with the same component number correspond to a single eigenvalue. As we see in Result 6.1 on page 86, this requirement necessitates identifying classes of matrices that have exactly one positive eigenvalue. The most general result in this domain comes from [103] and characterizes the eigenvalues of hollow (zero-diagonal) symmetric nonnegative matrices.

To leverage this result, we must first define a generalized Ramsey number. Let $\{G_{a_1}, G_{a_2}, \ldots, G_{a_c}\}$ be a collection of simple graphs where a_i is the number of vertices of the ith graph. Suppose we wish to color the edges of a complete graph G with c colors. The *generalized Ramsey number*, $R(G_{a_1}, G_{a_2}, \ldots, G_{a_c})$, is the minimum number of vertices of the complete graph G such that for any c-coloring of G, there exists an $i \in \{1, 2, \ldots, c\}$ such that G_{a_i} is an induced subgraph of G with all edges of color i. By Ramsey's Theorem [105] such a number always exists.

Result 6.4. (Charles, Farber, Johnsons, and Kennedy-Schaffer [[103], Theorem 3.5]). *Let k and $2 \leq j \leq k-1$ be two positive integers, let G_{j+1} be a complete graph on $j+1$ vertices, and let $c > 1$ be the smallest integer for which*

$$k \leq R(\underbrace{G_{j+1}, G_{j+1}, \ldots, G_{j+1}}_{c\ times}).$$

Let $\epsilon = (\frac{j}{j+1})^c$. Then all hollow symmetric nonnegative matrices of order at least k and with off-diagonal entries from $(\epsilon, 1]$ have at least j nonpositive eigenvalues.

As an immediate consequence of this result we have the following relevant result.

Result 6.5. *Let $\epsilon = (\frac{k-1}{k})^2$. If $\mathbf{H} \in \mathbb{R}^{k \times k}$ is a hollow symmetric matrix with off-diagonal elements from $(\epsilon, 1]$, then the largest eigenvalue of $\mathbf{H}$ is positive and the other $k-1$ eigenvalues are nonpositive.*

The proof of Result 6.5 is as follows. In the special case of $j = k - 1$, we are trying find the smallest value of $c > 1$ for which the Ramsey number of c copies of G_k is greater than k, that is, the smallest value of c for which

$$k \leq R(\underbrace{G_k, G_k, \ldots, G_k}_{c\ times}).$$

For $c = 2$, the Ramsey number of $\{G_k, G_k\}$ is the minimum number of vertices needed for a complete graph such that any coloring results in an isomophic copy of G_k whose edges are all monochromatic (one color). When $k = 2$, any 2-coloring of G_2 contains a monochromatic copy of G_2, therefore $k = R(G_2, G_2)$. For $k > 2$, there exists a 2-coloring of G_k that does not contain a monochromatic isomorphic copy of G_k. For example, any 2-coloring that is not monochromatic will not contain a copy of G_k. Thus the number of vertices needed must be greater than k and $k \leq R(G_k, G_k) \forall k$. This implies that $\mathbf{H}$ will have at least $k - 1$ nonpositive eigenvalues when the off-diagonal elements are chosen from the interval $(\epsilon, 1]$ with $\epsilon = (\frac{k}{k+1})^2$. Since the trace of a symmetric matrix is the sum of its eigenvalues, the eigenvalues of $\mathbf{H}$ must sum to zero. This means that $k - 1$ eigenvalues are nonpositive, but not all identically zero, implying that the largest eigenvalue must be positive. Thus $\mathbf{H}$ has exactly one positive eigenvalue as desired.

Part IV.

Real-world applications

7. Source enumeration and voice activity detection in wireless acoustic sensor networks

In this chapter, we propose a robust technique for multi-speaker voice activity detection and source enumeration in wireless acoustic sensor networks (WASNs). We adapt the joint-EVD technique proposed in Section 6.4 to first cluster the nodes that observe the same speaker as the dominant source. We then estimate the voice activity of each speaker by introducing a block-sparsity penalizing term in the unmixing problem. The method is scalable in terms of the number of simultaneously active speakers, does not require setting empirical thresholds and is robust to impulsive noise sources. The results are validated using a WASN with four human speakers and two impulsive noise sources observed by 15 nodes [1].

7.1. Introduction

WASNs provide a next generation system with great potential for new services, e.g., in ambient assisted living, habitat monitoring, smart cities [106], [107]. They combine many comparatively low-resource, distributed nodes with sensing, computing and communication capabilities. Compared to traditional microphone arrays, the spacial field is sampled in a

[1] This chapter is based on the paper: "Source enumeration and robust voice activity detection in wireless acoustic sensor networks, T. Hasija, M. Gölz, M. Muma, P.J. Schreier and A.M. Zoubir, *Asilomar Conference on Signals, Systems, and Computers*, 2019" and is a result of a joint collaboration between Signal and System Theory group, Paderborn University and Signal Processing Group, Darmstadt Technical University. The block-sparse technique of Section 7.4, and its results in Table 7.2, Figures 7.3 and 7.4 are contributed by M. Gölz. Sections introducing the study, and interpreting and discussing the results have contributions from all authors.

larger area, which leads to a significant performance increase. WASNs, however, also provide new signal processing challenges. This chapter is concerned with multi-speaker voice activity detection (VAD) for WASNs, which is a prerequisite for many speech enhancement algorithms [108]. Multi-speaker VAD is challenging because hidden voice activity patterns must be recovered for all sources, given observed mixtures only.

Existing VAD approaches, e.g., [47], rely on energy separation by means of multiplicative nonnegative independent component analysis (M-NICA). However, they suffer from a significant performance loss as the number of sources increases, making the M-NICA problem more and more difficult. Further, non-active speech may yield a small but non-zero energy value, making the detection task non-trivial. Finally, existing methods are non-robust against impulsive noise.

We propose a multi-speaker VAD approach that addresses the above three challenges. To provide scalability in terms of the active number of sources, we propose a new method to identify the so-called *dominant source model* that has been introduced in [109]. Measurements from nodes that observe the same source are highly correlated. The number of such sources and their associated node clusters are, therefore, estimated based on correlation information. We formulate the task as model-selection problem in multiple data sets which we have addressed in Chapter 6. The joint-EVD technique is applied because it uses joint information from all nodes and is better suited for the WASN problem in this chapter compared to the mCCA-HT technique. This is because the joint-EVD technique can explicitly determine if a source is observed by a node by testing the eigenvector component corresponding to that node. Thus, it can be applied to determine $i = 1, \ldots, d$ node clusters which observe the same source. This divides the energy separation problem into d simpler energy separation tasks. The second challenge of thresholding voice activity has been recently addressed by incorporating sparseness constraints on the energy signatures [110], [111]. Shrinking small energy values to zero relieves the practitioner from the necessity of defining a heuristic voice activity threshold. However, such approaches are not robust against impulsive noise. Therefore, we introduce a group sparsity constraint that matches the characteristics of human speech and suppresses impulsive noise in the energy unmixing step. The proposed method outperforms existing approaches for a WASN in a 20×10 m room with four simultaneously active human speakers and two impulsive noise sources observed by 15 nodes.

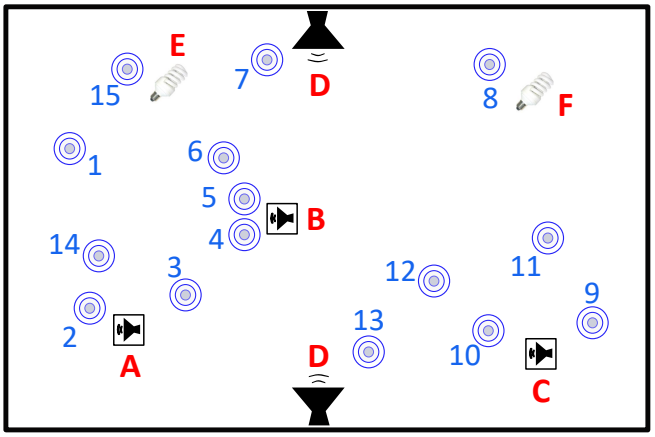


Figure 7.1.: An example of a WASN with 6 sources (A-F) and 15 sensor nodes in a 20m$\times$10m room.

7.2. Problem formulation

A WASN of P sensor nodes, each equipped with n microphones, is considered. Let q acoustic sources be active in the network. We assume there are $d(\leq q)$ clusters of nodes, where for each cluster, all nodes observe the same source as the dominant source. Figure 7.1 shows an example of a WASN with 6 sources (A-F, shown in red) and 15 sensor nodes (1-15, shown in blue). In this case, the cluster of nodes 2,3 and 14 observe speaker A as the dominant source. Similarly, speaker B is the dominant source for nodes 4,5 and 6, and so on. Thus, there are unknown number of d dominant sources and their corresponding node clusters in the WASN. The remaining $q-d$ sources are each observed by only a single sensor node. For example, in Figure 7.1, sources E and F (generating bulb-flickering noise) are observed by sensors 15 and 8, respectively. Our aim is to robustly estimate the voice activity of the unknown number of d dominant sources in WASN, given a received mixture at each sensor.

The proposed method consists of two processing steps: first, d along with the associated node clustering information $\mathbf{B}$ is computed. The binary matrix $\mathbf{B}$ is of size $d \times P$ whose $\{ij\}$th entry is 1 if the ith dominant source is observed by the jth node, and 0 otherwise. Based on the estimates $\hat{d}$ and $\hat{\mathbf{B}}$, in the second step, the voice activity pattern (VAP) is determined separately for the ith dominant source by only using the nodes marked as 1 in the ith row of $\hat{\mathbf{B}}$. The two steps are explained in detail in Sections 7.3 and 7.4, respectively.

7.3. Source enumeration and node clustering

7.3.1. Frequency-domain signal model

Let the observed vector $\mathbf{x}_p(f) \in \mathbb{C}^n$ at node p and frequency index f be modeled as

$$\mathbf{x}_p(f) = \mathbf{A}_p(f)\mathbf{s}_p(f), \qquad p = 1, \ldots, P, \tag{7.1}$$

where $\mathbf{A}_p(f) \in \mathbb{C}^{n \times m_p}$ is the full column rank mixing matrix (acoustic transfer function), and $\mathbf{s}_p(f) \in \mathbb{C}^{m_p}$ refers to the source vector. We assume an unknown $m_p (\leq n)$ number of uncorrelated and (locally) stationary sources in $\mathbf{s}_p$ that, without loss of generality, are assumed to be zero-mean and unit variance. From now on, we will drop the frequency index (f) from the sources, observed vectors and their covariance matrices for readability. The kth signal component of the pth data set is denoted by $s_p^{(k)} = u_p^{(k)} + jv_p^{(k)}$, where $u_p^{(k)}$ and $v_p^{(k)}$ are real and imaginary parts of $s_p^{(k)}$.

Between any two nodes p and q, sources may be correlated only pairwise, i.e., the source $s_p^{(k)}$ may only correlate with source $s_q^{(k)}$ for $1 \leq k \leq m_{pq} (= \min(m_p, m_q))$. There are two types of complex-valued correlation coefficients when analyzing the correlation between $s_p^{(k)}$ and $s_q^{(k)}$. The standard correlation coefficient is $\rho_{pq}^{(k)} = E[s_p^{(k)}(s_q^{(k)})^*]$, and the complementary correlation coefficient is $\tilde{\rho}_{pq}^{(k)} = E[s_p^{(k)} s_q^{(k)}]$. We make the following assumptions about these correlation coefficients.

1. Using $s_p^{(k)} = u_p^{(k)} + jv_p^{(k)}$ and $s_q^{(k)} = u_q^{(k)} + jv_q^{(k)}$, the expression for $\rho_{pq}^{(k)}$ is

$$\rho_{pq}^{(k)} = E[u_p^{(k)} u_q^{(k)}] + E[v_p^{(k)} v_q^{(k)}] + j(E[v_p^{(k)} u_q^{(k)}] - E[u_p^{(k)} v_q^{(k)}]). \tag{7.2}$$

   We assume that covariances of real and imaginary parts of correlated sources are equal, i.e., $E[v_p^{(k)} u_q^{(k)}] = E[u_p^{(k)} v_q^{(k)}]$. Thus, using (7.2), $\rho_{pq}^{(k)}$ is real-valued. This assumption is reasonable since the correlated sources in different nodes are generated from a common underlying speaker.

2. We assume that the complementary correlation coefficients between all sources are zero, i.e.,

$$\tilde{\rho}_{pq}^{(k)} = E[s_p^{(k)} s_q^{(k)}] = 0, \qquad \forall k = 1, \ldots, \leq m_{pq}, p, q = 1, \ldots, P, p \neq q. \tag{7.3}$$

   When $s_p^{(k)} = s_q^{(k)}$, this assumption is equivalent to assuming that all the information

is contained in the magnitude of $s_p^{(k)}$ and its phase carries no information, which is a common assumption in the speech processing literature [112], [113] [2]. Based on this assumption, the source complementary cross-covariance matrix $\widetilde{\mathbf{R}}_{s_p s_q} = E[\mathbf{s}_p \mathbf{s}_q^T] = 0$ for all p, q, $p \neq q$. Thus, all the correlation information is contained in the standard source cross-covariance matrices.

Using assumption 1, the standard source cross-covariance matrix between nodes p and q ($p \neq q$),

$$\mathbf{R}_{s_p s_q} = E[\mathbf{s}_p \mathbf{s}_q^H] = \text{diag}(\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(m_{pq})}), \tag{7.4}$$

is a diagonal and real-valued matrix. Let $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_P^T]^T$ be the composite data vector of P nodes and $\mathbf{R} = E[\mathbf{x}\mathbf{x}^H]$. Let $\mathbf{R}_D = \text{blkdiag}(\mathbf{R}_{11}, \ldots, \mathbf{R}_{PP})$ be a block diagonal matrix with $\mathbf{R}_{pp} = E[\mathbf{x}_p \mathbf{x}_p^H]$. The standard composite coherence matrix $\mathbf{C} \in \mathbb{C}^{nP \times nP}$ is defined as

$$\mathbf{C} = \mathbf{R}_D^{-\frac{1}{2}} \mathbf{R} \mathbf{R}_D^{-\frac{1}{2}}, \tag{7.5}$$

where the exponent $-\frac{1}{2}$ denotes the square-root matrix inverse (or square-root pseudo-inverse of a rank-deficient matrix). Let there be d number of dominant sources such that for $i = 1, \ldots, d$, there exist $P^{(i)}$ nodes whose kth sources are correlated with each other. We have shown in Section 6.4 that the EVD of $\mathbf{C}$ can completely characterize the correlation structure among multiple sets of data. In this context, d is equal to the number of correlated sources among the nodes, and determining the correlation structure is equivalent to finding the cluster of nodes for the d sources. However, the results in Section 6.4 are derived for a real-valued $\mathbf{C}$. Under the assumption that $\mathbf{R}_{s_p s_q}$ is real-valued, Results 6.1 and 6.3 derived in Section 6.4 for a real-valued C can be straightforwardly extended for a complex-valued C. More specifically,

i) $\mathbf{C}$ has exactly d eigenvalues greater than one if and only if there exist d dominant sources in the WASN, and

ii) Let $\mathbf{u}^{(i)}$ be the eigenvector associated with the ith largest eigenvalue of $\mathbf{C}$. $\mathbf{u}^{(i)}$ can be partitioned into P subvectors, $\mathbf{u}^{(i)} = [\mathbf{u}_1^{(i)T}, \mathbf{u}_2^{(i)T}, ..., \mathbf{u}_P^{(i)T}]^T$, where $\mathbf{u}_p^{(i)} \in \mathbb{C}^n$ contains the elements of $\mathbf{u}^{(i)}$ associated with the pth node. The ith source is observed by the pth node if and only if $\mathbf{u}_p^{(i)} \neq \mathbf{0}$.

[2] However, there is also substantial recent work in the literature that shows that phase is also useful for speech processing and ignoring it can lead to suboptimal performance [114], [115].

7.3.2. Bootstrap-based hypothesis testing

Let M available samples of each node form the columns of the sample matrices $\mathbf{X}_1, \ldots, \mathbf{X}_P$. The sample coherence matrix $\hat{\mathbf{C}}$ can be computed from the sample estimates of covariance matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}_D$ using (7.5). As stated in Section 6.4.2, the number of eigenvalues of $\hat{\mathbf{C}}$ that are greater than one will often not be equal to d. We use the same approach of a sequence of binary hypothesis tests as used in Section 6.4.2 to estimate d. We propose a statistic based on the assumption that there is at least one independent source or noise component among all nodes. This means that at least one eigenvalue of $\mathbf{C}$ is equal to one. The null hypothesis for each binary hypothesis test in (6.24) is

$$H_0 : \lambda^{(s+1)} = 1 \tag{7.6}$$

and the proposed statistic is

$$T(s) = (\lambda^{(s+1)} - 1)^2. \tag{7.7}$$

We use bootstrap to estimate the unknown distribution of $T(s)$ under H_0 since it has been shown to work well with non-Gaussian data and limited number of samples, both being relevant to this application [116]. The procedure is the same as in Algorithm 2 on page 96 except that the test statistic here is based on only one eigenvalue following the s eigenvalues of $\mathbf{C}$. However, as in Algorithm 2, a more robust test statistic based on P eigenvalues of $\mathbf{C}$ can be employed if it is assumed that the WASN contains small number of dominant sources compared to the total number of nodes and their dimensionality, i.e., if $d \ll nP$.

As pointed out in Section 6.4.2, the subvectors $\mathbf{u}_p^{(i)}$ will not be zero when computed using $\hat{\mathbf{C}}$. In this case, the norm of these subvectors is overestimated. In this section, we propose a more robust test for the eigenvectors compared to Section 6.4.2 by introducing a nonzero threshold u_0. For $i = 1 \ldots d$ and $p = 1 \ldots P$ we test the hypotheses

$$\begin{aligned} H_0 &: \|\mathbf{u}_p^{(i)}\| \leq u_0, \\ H_1 &: \|\mathbf{u}_p^{(i)}\| > u_0. \end{aligned} \tag{7.8}$$

The threshold u_0 controls the selection of nodes observing the dominant source. If the ith source is correlated across all nodes with equal pairwise correlation coefficients, $\|\mathbf{u}_p^{(i)}\| = \frac{1}{\sqrt{P}}, \forall p = 1, \ldots, P$. However, often in a WASN, a source is not observed by all nodes. Thus, some of the subvectors in $\mathbf{u}^{(i)}$ will be close to zero. Due to the constraint that $\|\mathbf{u}^{(i)}\| = 1$, this will push the subvectors corresponding to the nodes observing the ith dominant source

to be significantly higher than $\frac{1}{\sqrt{P}}$. For this reason, we chose $u_0 = \frac{1}{\sqrt{P}}$. When $u_0 = 0$, the test in (7.8) is the same as (6.27) tested using the statistic (6.28).

The distribution of the proposed statistic $T_p^{(i)} = \|\mathbf{u}_p^{(i)}\|$ under H_0 is estimated using the bootstrap as in Algorithm 3 on page 98 and in this case, $T_p^{(i)} - u_0$ is compared to the threshold ${}_{(\eta)}T^*$ in line 33 of Algorithm 3 to reject H_0. If H_0 is not rejected, $\mathbf{B}\{ip\} = 0$, otherwise $\mathbf{B}\{ip\} = 1$.

7.4. Group sparse voice activity detection

Because of the on/off behavior of human speech, the individual signal energies are block sparse. Thus, for each of the $\hat{d}$ clusters of nodes obtained, we employ an SVD on the received mixed energies and impose a block-sparsity constraint on the right rotation matrix. We improve upon recent work [110] that assumed only sparse energy sources. In contrast to non-sparse methods as M-NICA [47], sparse median-based M-NICA (SMM-NICA) [110] and also our proposed group-sparse median-based M-NICA (GSMM-NICA) perform VAD intrinsically as all non-zero entries in the reconstructed energy signature are automatically labeled as active speech. The entries of reconstructed energy signatures for non-sparse methods are, in general, all nonzero and an activity threshold τ has to be defined, which heavily depends on the deployed application scenario.

7.4.1. Time-domain energy model

Let the received energy vector $\mathbf{y}_p(t) \in \mathbb{R}^n$ for all n microphones at node p and frame index t be

$$\mathbf{y}_p(t) = \mathbf{H}_p\mathbf{e}(t) + \mathbf{w}_p(t), \tag{7.9}$$

where $\mathbf{H}_p \in \mathbb{R}^{n\times q}$ is the mixing matrix for node p, $\mathbf{e}(t) \in \mathbb{R}^q$ contains the signal energy from all q speakers in the WASN, i.e., the sum of squared received signal values during time frame t per speaker, and $\mathbf{w}_p(t) \in \mathbb{R}^n$ is additive noise. The individual node observations are summarized in a network received energy vector $\mathbf{y}(t) = \left[\mathbf{y}_1^T(t), \dots \mathbf{y}_P^T(t)\right]^T$. Finally, the observations at all frame indices $t = 1, \dots, N$ and nodes are expressed as

$$\mathbf{Y} = \mathbf{H}\mathbf{E} + \mathbf{W}, \tag{7.10}$$

where $\mathbf{Y} \in \mathbb{R}^{nP \times N}$, the WASN mixing matrix $\mathbf{H} \in \mathbb{R}^{nP \times q}$, energy matrix $\mathbf{E} \in \mathbb{R}^{q \times N}$ and noise matrix $\mathbf{W} \in \mathbb{R}^{nP \times N}$.

Since the initial source enumeration and node clustering algorithm divides the P nodes into $\hat{d}$ clusters, we define cluster-wise received energy matrices

$$\mathbf{Y}^{(i)} = \mathbf{h}^{(i)}\mathbf{e}^{(i)} + \mathbf{W}^{(i)}, \tag{7.11}$$

where $\mathbf{Y}^{(i)} \in \mathbb{R}^{n^{(i)} \times N}$, $\mathbf{h}^{(i)} \in \mathbb{R}^{n^{(i)} \times 1}$, $\mathbf{e}^{(i)} \in \mathbb{R}^{1 \times N}$ and $\mathbf{W}^{(i)} \in \mathbb{R}^{n^{(i)} \times N}$ with the number of nodes $P^{(i)}$ and the number of microphones $n^{(i)} = nP^{(i)}$ per cluster $i = 1, \ldots, \hat{d}$.

7.4.2. Group-sparse singular value decomposition

M-NICA [47] separates sources by applying an SVD to the energy matrix $\mathbf{Y}$, i.e., $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. $\mathbf{\Sigma}$ contains the singular values of $\mathbf{Y}$ on its diagonal, and $\mathbf{U}$ and $\mathbf{V}$ are composed of the left and right singular vectors of $\mathbf{Y}$. It is shown in [47] that the matrix $\mathbf{V}$ contains the required information about the time-domain shape of the underlying energies. Due to the previous decomposition of the WASN into clusters of microphones using the dominant source approach, we aim to extract exactly one source, the dominant source, from each of the $\hat{d}$ clusters. Hence, the SVD provides us with one singular value $\mathbf{\Sigma}^{(i)} = \sigma^{(i)}$ and one left/right singular vector $\mathbf{U}^{(i)} = \mathbf{u}^{(i)}$ and $\mathbf{V}^{(i)} = \mathbf{v}^{(i)}$, respectively. Since we only work with cluster-wise quantities in the following, we drop superscript (i) for readability.

To enforce group sparsity in the right singular vector $\mathbf{v}$, we first decompose $\mathbf{Y}$ by a standard SVD to obtain estimates for $\mathbf{u}, \sigma$ and $\mathbf{v}$. Then, we divide the cluster-wise received energy matrix $\mathbf{Y} \in \mathbb{R}^{n^{(i)} \times N}$ into L groups of frames, each of length N_g and extract the cluster-wise group received energy matrices $\mathbf{Y}_{\mathrm{g},l} \in \mathbb{R}^{n^{(i)} \times N_\mathrm{g}}$ such that $\mathbf{Y} = [\mathbf{Y}_{\mathrm{g},1}, \ldots, \mathbf{Y}_{\mathrm{g},L}]$. Equivalently, we define $\mathbf{v}_{\mathrm{g},l} = [v_{\mathrm{g},l}[1], \ldots, v_{\mathrm{g},l}[N_\mathrm{g}]]^T \in \mathbb{R}^{N_\mathrm{g}}$. We reformulate the SVD optimization problem w.r.t. the right singular vector $\mathbf{v}$ as

$$\arg\min_{\mathbf{v}} \left|\left| \sum_{l=1}^{L} \mathbf{Y}_{\mathrm{g},l} - \sigma \mathbf{u}\mathbf{v}_{\mathrm{g},l}^T \right|\right| + \lambda_\mathbf{v}\Omega(\mathbf{v}), \tag{7.12}$$

where $\Omega(\mathbf{v})$ is a sparsity-inducing penalty term and $\lambda_\mathbf{v}$ is a tuning parameter that determines the degree of sparsity of $\mathbf{v}$. In contrast to [110], we work with grouped quantities and define $\Omega(\mathbf{v}) = \sum_{l=1}^{L} \sqrt{\sum_{t=1}^{N_\mathrm{g}} |v_{\mathrm{g},l}[t]|^2}$ as a mixed $\ell 1/\ell 2$ norm. The optimization problem (7.12) is a model-adjusted variant of the problem in [117], whose solution is commonly referred

to as the group least absolute shrinkage and selection operator (group LASSO). Since $\sigma\mathbf{u}$ is orthonormal as $\mathbf{u}$ is a singular vector, the group LASSO [117] for our data model results in

$$\mathbf{v}_{\mathrm{g},l}^T = \begin{cases} \left(1 - \frac{\lambda_{\mathbf{v}}\sqrt{N_{\mathrm{g}}}}{||\mathbf{u}^T\mathbf{Y}_{\mathrm{g},l}||^2}\right)\mathbf{u}^T\mathbf{Y} & x \geq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{7.13}$$

where we iterate until convergence of $\mathbf{v}$ over all L groups.

The tuning parameter $\lambda_{\mathbf{v}}$ is selected as $\lambda_{\mathbf{v}} = \arg\min C_{\mathrm{N}}$, where Mallows' C_{N} [118] is equivalent to the AIC-ITC for our time-domain signal energy model. Thus,

$$\begin{aligned} C_{\mathrm{N}} = \frac{\left|\left|\mathbf{Y} - \mathbf{u}\mathbf{v}^T(\lambda_{\mathbf{v}})\right|\right|^2}{\sigma_{\mathbf{Y}}^2} - n^{(i)} \\ +2\sum_{l=1}^{L}(||\mathbf{v}_{\mathrm{g},l}|| > 0) + 2\sum_{l=1}^{L}\frac{||\mathbf{v}_{\mathrm{g},l}||}{\left|\left|\mathbf{v}_{\mathrm{g},l}^{\mathrm{LS}}\right|\right|}(N_{\mathrm{g}} - 1), \end{aligned} \tag{7.14}$$

where $\mathbf{v}^{\mathrm{LS}} = \left[\mathbf{v}_{\mathrm{g},1}^{\mathrm{LS}\,T}, \ldots, \mathbf{v}_{\mathrm{g},L}^{\mathrm{LS}\,T}\right]^T = ((\mathbf{u}^T\mathbf{u})\mathbf{u}^T\mathbf{Y})^T$ and $\sigma_{\mathbf{Y}}^2$ is the variance of $\mathbf{Y}$.

Proposed Algorithm- Let us briefly recapitulate our proposed GSMM-NICA. We decompose the cluster-wise received energies initially by a standard SVD. Then, we iteratively compute (7.13) for all L groups, which provides us with a group sparse energy signature shape estimate $\mathbf{v}$. $\lambda_{\mathbf{v}}^* = \arg\min_{\lambda_{\mathbf{v}}} C_{\mathrm{N}}$ is selected in each iteration individually. We stop iterating when the update on solution $\mathbf{v}$ falls below a convergence threshold. We then continue with de-correlation of $\mathbf{Y}$ using σ, $\mathbf{u}$ and (group-sparse) $\mathbf{v}$ as in SMM-NICA [110].

7.5. Results

We validate the proposed technique on a WASN generated from real speakers in a $20\mathrm{m}\times10\mathrm{m}$ room with a reverberation time of $0.3\mathrm{s}$ [3]. Each node is equipped with a uniform linear array of $n = 3$ microphones sampled at 16kHz. To have the knowledge of the ground truth, all nodes hear only one speaker at a time, and the observed signals from multiple speakers are added to generate a mixture of speech signals for each node. Each node is then corrupted by an additive white uncorrelated Gaussian noise with variance of 0.05. The variance of all speech signals before mixing is normalized to one. We set the duration of one energy time

[3]The WASN speech dataset has been generated within the EU FET-Open Project HANDiCAMS (GA no. 323944).

frame to 30ms, which fits well to speech characteristics. For source enumeration and node clustering, short-term Fourier transform (STFT) with a Hamming window and frame length of 1024 with 50% overlap is applied to the data obtained from each microphone. The number of frequency bins for the STFT is chosen as 32. The proposed technique provides an estimate of the number of dominant sources and the corresponding clusters for each frequency bin. The estimated number of sources are averaged over all frequency bins and rounded off to the nearest integer to obtain the final estimate of d. The top d majority voted clusters from all frequency bins are selected as the final clusters associated with the d dominant sources. We provide a brief summary of the selected competitors from the literature.

Competitors- We compare our proposed method with two algorithms from the literature, namely, the standard M-NICA [47] algorithm and SMM-NICA [110]. The latter one is deployed on a cluster-wise level, meaning that the presented results for SMM-NICA were obtained using the dominant source model presented in Section 7.3, to demonstrate the effect of the group sparsity constraint in GSMM-NICA. Also, for M-NICA, one has to select a voice activity threshold as all entries of the resulting energy signature are generally nonzero. To compare our method to the best possible results for M-NICA, we decided to perform VAD for a grid of possible energy threshold τ and select the one τ_{opt} that provides us with the largest number of correctly classified time samples. In a real-world scenario, the underlying ground truth VAP is unknown and determining τ_{opt} would be impossible. Therefore, we refer to this competitor as oracle M-NICA.

The results for two different scenarios are presented.

Scenario 1: A total of 10 nodes observe three spatially well-separated speakers for a duration of 15s as shown in Figure 7.2. The clustering result is shown in Table 7.1 for $\hat{d} = 3$ estimated speakers. The nodes 1, 7 and 8, which are comparatively far away from all speakers are not selected in any of the clusters.

Dominant Source	Cluster of Nodes (Scenario 1)	Cluster of Nodes (Scenario 2)
A	2 and 3	2, 3 and 14
B	4, 5 and 6	4, 5 and 6
C	9 and 10	9, 10 and 11
D	Not active	7, 12 and 13

Table 7.1.: The clustering result of the proposed technique for scenarios 1 and 2.

Based on the clustering results in Table 7.1, we run the group sparse VAD as presented in

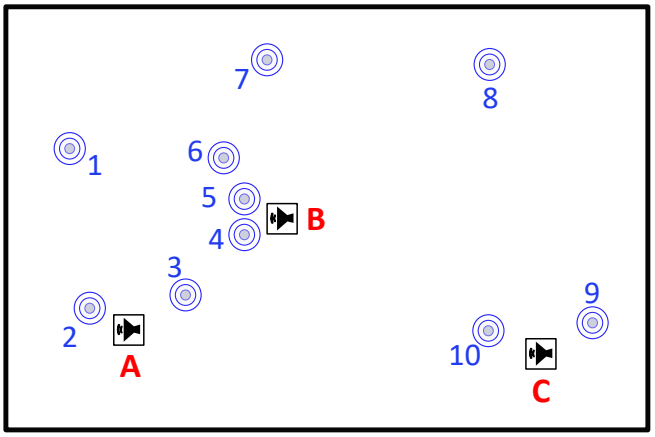


Figure 7.2.: An example of WASN with 3 sources (A-C) and 10 sensor nodes in a 20 $\times$ 10 m room.

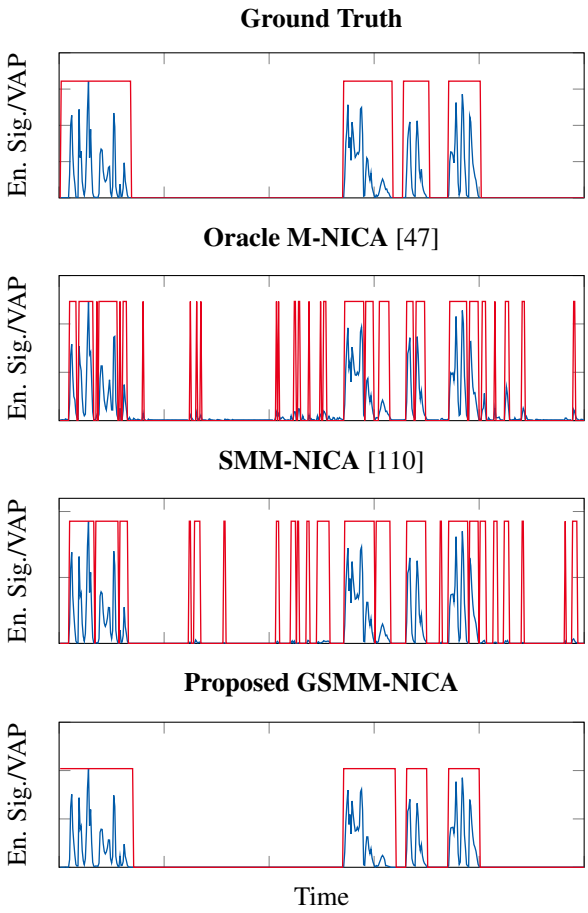


Figure 7.3.: Extracted energy signatures (blue) and voice activity patterns (red) for speaker B in Scenario 1. GSMM-NICA outperforms the competitors and the extracted VAP is closest to the ground truth VAP.

Section 7.4. As an exemplary result, we plot the obtained energy signatures and VAPs for speaker B in Figure 7.3. In general, we observed in simulations that a group length N_{g}

between 8 and 12 samples at 16 kHz seems to best fit the human speech characteristics. Thus, we choose $N_{\mathrm{g}} = 10$ samples.

The proposed clustered group sparse method clearly outperforms the competitors. The results for speakers A and C are similar to the ones displayed in the figure.

Scenario 2: In a more challenging scenario, a public announcement loudspeaker is introduced at two opposite sides of the room denoted as speaker D in Figure 7.1 along with two sources E and F which generate uncorrelated bulb-flickering impulsive noise. The six sources are observed by 15 nodes for 30s duration. The clustering result is listed in Table 7.1. The proposed technique estimates $\hat{d} = 4$ speakers correctly along with their node clusters. For loudspeaker D, nodes 7, 12 and 13 form one cluster even though they are are spatially far from each other. Nodes 1, 8 and 15, which either observe impulsive noise or are far from all speakers do not belong to any cluster.

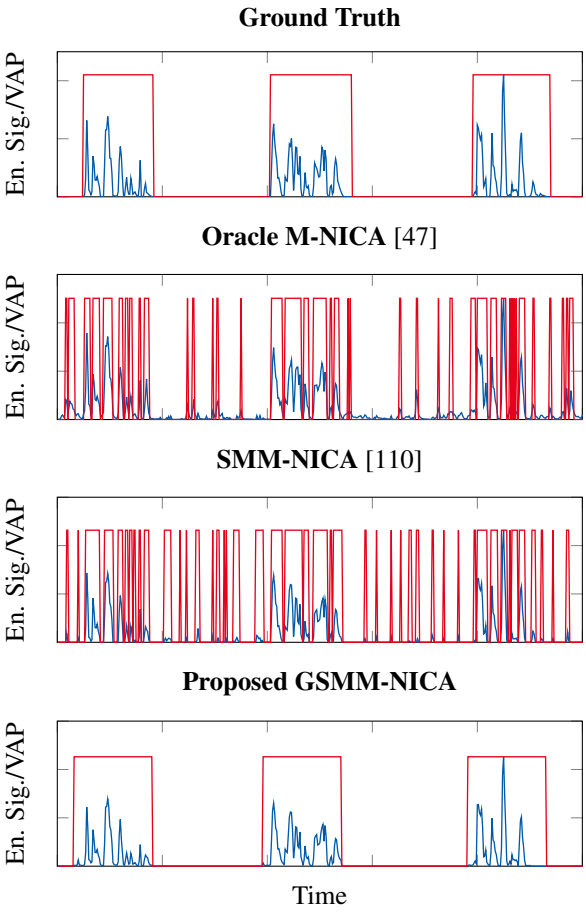


Figure 7.4.: Extracted energy signatures (blue) and voice activity patterns (red) for speaker C in Scenario 2. GSMM-NICA outperforms the competitors and the extracted VAP is closest to the ground truth VAP.

Again, we run the group sparse VAD on the cluster observation matrices. In this case, we suggest using a larger group length, which suppresses the flickering noise most effectively. The energy signatures and VAPs for speaker C in Figure 7.4 were obtained for $N_{\mathrm{g}} = 15$ samples. The proposed method again clearly outperforms the competitors, which are not able to treat the noise as efficiently. The results for speakers A, B and D are similar to the ones displayed in the figure.

	Source	Oracle M-NICA [47]	SMM-NICA [110]	Proposed GSMM-NICA
Scenario 1	A	81.4	86.8	84.2
	B	83.2	85.0	97.6
	C	84.2	86.0	89.8
Scenario 2	A	61.7	67.4	62.7
	B	83.3	81.6	81.5
	C	79.1	77.7	93.6
	D	75.1	71.8	78.7

Table 7.2.: The percentage of correctly labeled frames for all sources.

In Table 7.2, we provide the percentage of correctly labeled (speech/pause) frames for the two scenarios and all sources. In scenario 1, the proposed technique with group length $N_{\mathrm{g}} = 10$ samples outperforms the competitors for speakers B and C and shows similar performance for speaker A. For the second scenario, we decided for $N_{\mathrm{g}} = 15$ samples to completely suppress the flickering noise. However, the resulting VAPs are a bit longer for speakers that are not strongly disturbed by the flickering. Therefore, we observe a slight degradation in the performance for speakers A and B. A more sophisticated approach to determine the ideal group length for each speaker would allow for a general improvement over M-NICA and SMM-NICA.

7.6. Summary

We have devised a new method to perform multi-speaker voice activity detection and source enumeration in WASNs. Due to the clustering of nodes according to the dominant source model, the approach outperforms existing standard procedures that process all nodes in the network jointly. The sparseness constraint relieves the practitioner from the necessity of defining a heuristic voice activity threshold. The group sparseness constraint suits better to the characteristics of human speech than a simple sparseness constraint.

8. Analyzing sports-induced interactions in multiple modalities of the autonomic nervous system

In this chapter, we apply the techniques developed in this thesis to characterize the changes across the modalities of the human autonomic nervous system (ANS) in response to a physical load. Data from three peripheral modalties of the ANS were recorded in participants wearing the Empatica E4 wristband and undergoing two different physical stresses. The three selected modalities were electrodermal activity (EDA), heart rate (HR), and skin temperature at wrist (Temp). Bimodal and multimodal analysis revealed not only an increase in the number of components correlated across the modalities, but also an increase in their correlation strength after the physical load pointing towards a reorganization of central ANS control [1].

8.1. Introduction

ANS is a complex system which regulates the functioning of various internal organs. Its complex functionality is achieved by task-specific modulation of several organ-specific sub-

[1] This chapter is based on the papers: "Exercise-induced changes of multimodal interactions within the autonomic nervous network, S. Vieluf, T. Hasija, R. Jakobsmeyer, P. J. Schreier and C. Reinsberger, *Frontiers in physiology*, 2019" and "Multimodal approach towards understanding the changes in the autonomic nervous system induced by an ultramarathon, S. Vieluf, V. Scheer, T. Hasija, P. J. Schreier and C. Reinsberger, *Research in Sports Medicine*, 2019". It is a result of a joint collaboration between the Signal and System Theory group and the Institute of Sports Medicine at Paderborn University. The data was recorded and preprocessed at the Institute of Sports Medicine. I specifically implemented and presented bimodal and multimodal data analyses, and generated all the tables and figures used in this chapter. Sections introducing the study, and interpreting and discussing the results have contributions from all authors.

networks as well as their interrelation [5], [119]. The ANS consists of a network of cortical structures, such as left amygdala, right anterior and left posterior insula, and midcingulate cortices and subcortical structures such as thalamus and brainstem [120], [121]. Typically, the ANS activity is subject to high day-to-day variations with poor systematics [122]. However, more systematic central alterations of the ANS due to specific stressors like physical exercise may perturb organ functions, and thus alter ANS activity [123]. Indeed, the interactions across several ANS subsystems are indicative of various ANS states [124]. Therefore, the analysis of the interrelation of subsystems of the ANS may provide additional insights into the alteration of ANS control in response to physical exercise. In the context of sports and exercise, there are various studies which analyze one ANS subsystem at a time. For example, changes in the cardiac system are analyzed in [122], and the changes in the electrodermal and the thermoregulatory system have been described in [125]. However, a combined analysis of exercise-induced effects in different ANS subsystems or modalities is rare. [126] reported changes of heart rate variability (HRV) and EDA in a study on incremental exercise levels, and reported correlations between EDA and cardiac measures for low exercise intensities. Nevertheless, systematic approaches to analyze multimodal measures are still missing. In this study, we aim at detecting sports-induced changes in different ANS subsystems and the changes in their interactions after a physical exercise. We selected HR, EDA and skin temperature as relevant measures of ANS subnetworks because physiological mechanisms as well as practical applications are well described for each subsystem [125], [127], [128]. Moreover, these modalities are easy to measure.

Two different studies which induced different physical stress in participants were conducted. In the first study, the subjects exercised on a treadmill at different intensity levels. In the second study, the effect of a bigger physical load on the ANS was studied. In this case, the subjects completed a 65km ultramarathon. The data and results of the two studies are presented in Sections 8.2 and 8.3. For both these studies, the data was analyzed before and after the physical exercise. We hypothesize that the interactions among different modalities change before and after completing the physical task.

CCA and mCCA are the most common tools to analyze the bimodal and multimodal linear dependencies. However, the results obtained from these techniques are misleading when the number of observations (the number of participants in this study) is small compared to the dimensions of the data sets (the number of time points in the recorded ANS time series). To overcome this challenge, we applied the PCA-CCA technique of Section 3.5, which can reliably determine the number of correlated components and their correlation strength among two different ANS modalities. For the multimodal analysis, we applied the proposed mCCA-

HT technique of Section 6.3 to estimate the correlation structure among the components of all modalities after reducing the dimension of the data sets using PCA as proposed in Algorithm 4 on page 101.

8.2. Exercise-induced interactions in the ANS

8.2.1. Participants and dataset

Data of five minutes during rest, before and after exercise from 24 male subjects was recorded. The subjects ran on a treadmill with different intensity levels measured by VO_2max. Two different intensities of 60% and 95% VO_2max were analyzed. 60% VO_2max is defined as the transition zone between aerobic and anaerobic energy supply, which corresponds to the first ventilatory threshold and therefore represents moderate intensity [73]. 95% VO_2max is defined as a high-demanding intensity close to the maximum load, but at least performable over a longer period of time [73]. ANS signals were recorded by Empatica E4 multisensor device. The sensor was placed at the participants' left wrists. Successive data series of HR (sampling rate 1 Hz), EDA (sampling rate 4 Hz) and Temp (sampling rate 4 Hz) at the wrist were acquired during five minutes in supine position prior to exercise and 30 minutes post-exercise. More details about the experiment can be found in [73].

From each measurement, a time window of three continuous minutes was manually selected to control for data quality and avoid movement artifacts. Based on data quality, five participants were excluded from further analysis, either due to loss of sensor connection, an incomplete HR recording or lack of a movement-free segment. The data sets for each modality were generated such that the recorded time series from each subject forms a column of the data matrix. Thus, the size of the data matrix for each modality is the number of time points times the number of participants.

8.2.2. Bimodal interactions

Proposed analysis- As pointed out earlier, the number of samples (subjects) in this work is much smaller than the dimensions of the data sets (number of time points). In this case, as shown in Section 3.5, many estimated canonical correlations are defective, irrespective of their true values. Therefore, to reliably estimate the number of correlated components d_{pq} between the pth and qth modalities, we applied the joint PCA-CCA technique of Section 3.5.

In this work, the GLRT-based PCA-CCA detector was used since the P_{fa} can be adjusted which is crucial in such a low-sample regime. The P_{fa} for the PCA-CCA detector was set to 0.05. The maximum PCA rank for each modality was set to seven, which is approximately equal to one third of the number of subjects (as proposed in Section 3.5). The complexity of the interaction between the modalities is linked to the number of correlated components between them. If the interaction is limited to a linear relationship in a single dimension, this should indicate a rather simple interaction. On the other hand, multiple correlated components would indicate a more complex type of relationship between modalities. However, not only the number of correlated components but also their strength of correlation is of interest in this work. This can be measured with an overall correlation coefficient ρ_c [44], which can be computed as a function of the nonzero d_{pq} canonical correlations as

$$\rho_c = 1 - \left(\prod_{i=1}^{d_{pq}} 1 - \left(k^{(i)}(r_p, r_q) \right)^2 \right), \tag{8.1}$$

where $k^{(i)}(r_p, r_q)$ denotes the ith canonical correlation for the chosen PCA ranks r_p, r_q. The overall correlation coefficient in (8.1) relates to the mutual information between the two data sets if the two data sets are Gaussian distributed [44]. Thus, high values of ρ_c indicate that the two data sets share more information.

Modality pair	Exercise intensity	Pre-exercise		Post-exercise	
		Number of correlated components	Estimated canonical correlations	Number of correlated components	Estimated canonical correlations
EDA-HR	60%	2	0.8, 0.73	1	0.91
HR-Temp		3	0.97, 0.93, 0.84	1	0.88
Temp-EDA		1	0.96	2	0.97, 0.95
EDA-HR	95%	1	0.84	2	0.94, 0.89
HR-Temp		0	0	2	0.86, 0.78
Temp-EDA		1	0.76	2	0.88, 0.82

Table 8.1.: Number of correlated components and their canonical correlation values for each modality pair during pre- and post-exercise measures at moderate and high intensity.

Results- Model selection using the joint PCA-CCA approach revealed significant correlated components in all measures for most modality pairs. This is shown in Table 8.1. In the pre-exercise the number of correlated components differs between modality pairs and between intensities. For the post-60% intensity, HR shows one correlated component with EDA and Temp each, while EDA and Temp show two correlated components. For the 95% intensity,

the number of components increases to two for all modality pairs and the canonical correlations increase for all components from pre to post. This is summarized in Figure 8.1, which shows how the estimated overall correlation coefficient $\hat{\rho}_c$ varies from pre to post-exercise. The estimate $\hat{\rho}_c$ is computed from the canonical correlations in Table 8.1 for each modality pair, and at both moderate and high intensity. At moderate intensity, as seen in Figure 8.1a, exercise does not seem to have a clear and specific effect on $\hat{\rho}_c$. However, at high intensity, as seen in Figure 8.1b, there is a substantial and clear increase in $\hat{\rho}_c$ from pre- to post-exercise for all three modality pairs.

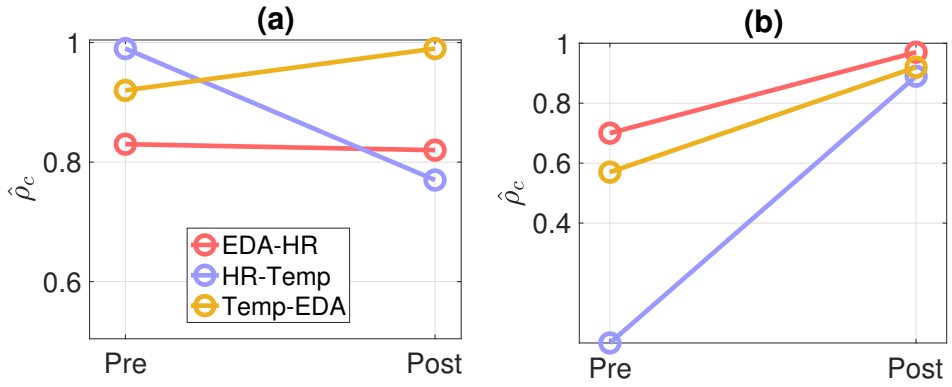


Figure 8.1.: The overall correlation coefficient $\hat{\rho}_c$ estimated for each modality pair at a) 60% exercise intensity and b) 95% exercise intensity.

8.2.3. Multimodal interactions

Proposed analysis- To jointly estimate the components correlated across multiple modalities, mCCA can be employed [21]. However, mCCA suffers from the same problems as CCA when the number of samples is small compared to the dimensions of the data sets. The correlation coefficients for the extracted components are significantly overestimated and do not reflect the true correlation structure among these components. There is no technique yet that jointly determines the required PCA dimensions and simultaneously performs mCCA. Our proposed solution is to estimate the PCA dimensions using Algorithm 4 on page 101, which although is suboptimal, it provides reasonable performance in the numerical examples shown in Section 6.6. In this work, the GENVAR cost function is used to perform mCCA. For GENVAR, the canonical variables are extracted such that they minimize the determinant of their correlation matrix [21]. We used the GENVAR mCCA since it has been widely applied in biomedicine, for instance, in analyzing fMRI data [20] and for fusing fMRI, EEG and structural MRI (sMRI) data [72].

However, analyzing the estimated correlation coefficients among the extracted components in mCCA is incomplete without having the knowledge of the correlation structure among the components. That is because the correlation structure indicates whether the correlation coefficients between the components are significant and also reveals the modalities across which the components exhibit those significant correlations. For this application, the mCCA-HT technique of Section 6.3 was applied to estimate the correlation structure. This is because the joint-EVD technique (explained in Section 6.4), which also estimates the correlation structure, employs hypothesis testing based on bootstrap. Since bootstrap resamples from the given sample set, it leads to inaccurate results as in this study the number of samples is extremely small.

Results- All the estimated canonical variables extracted in the ith stage of mCCA can be grouped together to form a source component vector (SCV) denoted by

$$\hat{\mathbf{E}}^{(i)} = \begin{bmatrix} \hat{\boldsymbol{\epsilon}}_{\text{EDA}}^{(i)} \\ \hat{\boldsymbol{\epsilon}}_{\text{HR}}^{(i)} \\ \hat{\boldsymbol{\epsilon}}_{\text{Temp}}^{(i)} \end{bmatrix}. \tag{8.2}$$

Here, $\hat{\boldsymbol{\epsilon}}_{\text{EDA}}^{(i)} \in \mathbb{R}^{1\times M}$ for example, is the ith stage estimated canonical variable corresponding to EDA, and M is the number of subjects. The SCV $\hat{\mathbf{E}}^{(i)}$ provides a convenient way to analyze correlations among all the canonical variables of the ith stage. The correlation matrix of the SCV shows the joint multimodal interactions among the extracted canonical variables. An example of the correlation matrix of the first SCV from pre-exercise data at 60% intensity is shown in Figure 8.2. The values are the absolute correlation coefficients between the canonical variables of different modalities. The corresponding heat map is also shown.

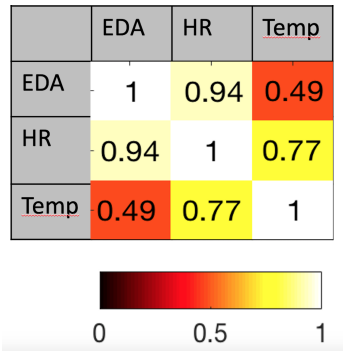


Figure 8.2.: The heat map of the correlation matrix of first SCV showing the absolute correlation coefficient values for pre-exercise data at 60% intensity. Color coding indicates the strength of correlation between pairs of modalities.

In Figure 8.3, the heat map of the absolute values of the correlation coefficients between the

first two SCVs from pre-exercise data is shown. The magenta-framed rectangle contains the correlations among the SCVs, $\hat{\mathbf{E}}^{(1)}$ and $\hat{\mathbf{E}}^{(2)}$ at 60% intensity. The 3 x 3 off-diagonal blocks within the magenta rectangle are almost zero as the second stage of canonical variables is constrained to be uncorrelated with the previous stage of canonical variables within each modality. Similarly, the green-framed rectangle displays the correlations among the two SCVs extracted at 95% intensity. In line with bimodal results, the correlations among the pre-exercise components at both intensities are not uniformly high and quite variable across the modalities. Finally, the blue-framed rectangle shows the correlation coefficients among the components at 60% and 95% intensities. These components are almost uncorrelated.

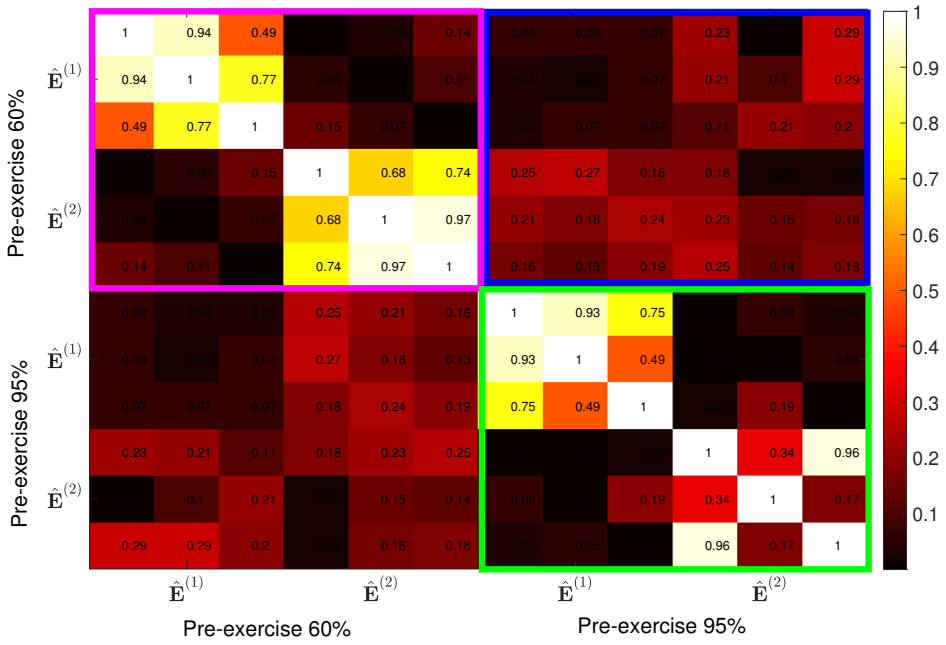


Figure 8.3.: Illustration of the absolute correlation coefficient values within pre-exercise measures. On the x and y axis, the first two SCVs $\hat{\mathbf{E}}^{(1)}$ and $\hat{\mathbf{E}}^{(2)}$ are depicted. Highlights indicate the correlation of the maximally correlated source components within (pink and green square) and between (blue square) intensities.

Similarly, Figure 8.4 shows the absolute values of the correlation coefficients between the extracted components from the post-exercise data. As in the bimodal results, the correlations between the SCVs at 95% intensity are high. However, the limitation of the bimodal analysis is that we cannot straightforwardly analyze if the components are correlated across all pairs of modalities or not. This can be done by the proposed mCCA-HT technique which estimates the correlation structure among the estimated SCVs. The estimated correlation structure at 95% intensity is shown in Figure 8.5. There are two detected correlated components at post-95% intensity (shown in Figure 8.5b), which are correlated across all modalities. This is in

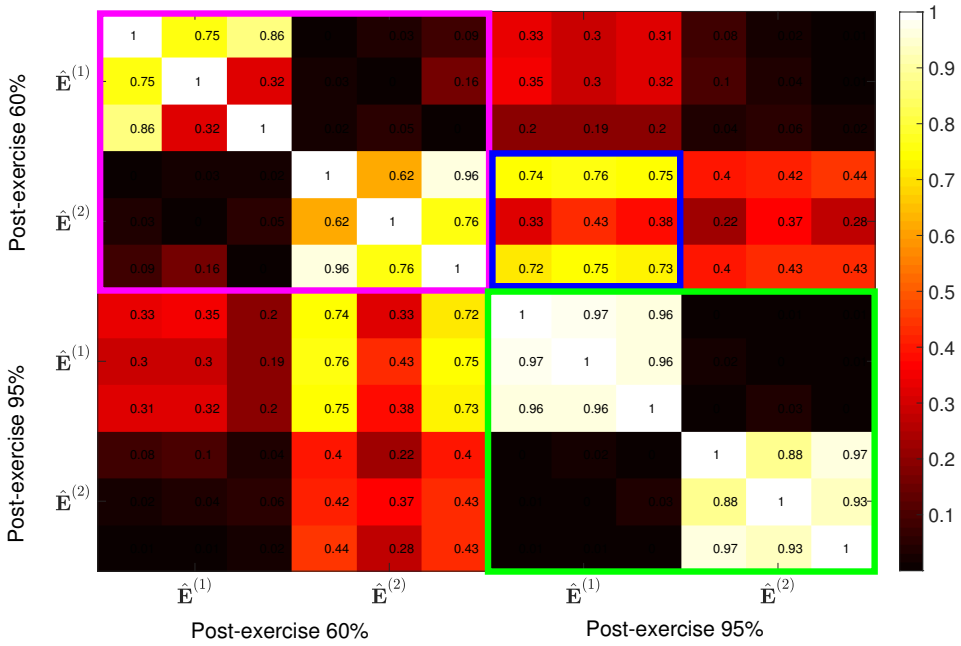


Figure 8.4.: Illustration of the absolute correlation coefficient values within post-exercise measures.

contrast to to the pre-95% intensity (shown in Figure 8.5a), where the components are only pairwise correlated. Finally, the first and the second SCVs at 60% and 95% intensities, respectively, have higher correlation coefficients between them (average 0.6) than pre-exercise SCVs. This can be seen in the blue-framed rectangle in Figure 8.4 and it indicates that these are related components.

8.2.4. Discussion

The aim of the current study was to elucidate changes between ANS subsystems related to the physical exercise. To gain insights into those exercise-induced changes, we analyzed time series data of HR, EDA, and Temp during pre- and post-exercise. We proposed analysis tools to reveal bimodal and multimodal interactions and presented the results. Bimodal analysis showed high variations in pre-exercise correlations, while in post-exercise the correlations were higher especially after high-intensity exercise. The multimodal results indicate that in measures taken before the exercise, cross-modality interactions exist but do not seem to follow a specific pattern, while in post-exercise measures the cross-modality interactions increase and show similarities between the different intensity tests, indicating an exercise-specific organization of the ANS modalities.

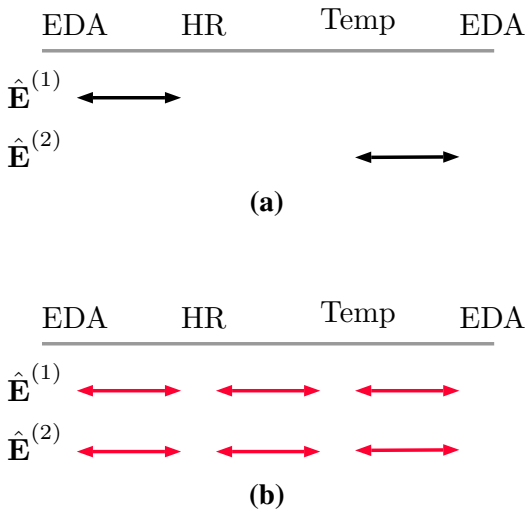


Figure 8.5.: Estimated correlation structure among the SCVs for 95% intensity a) pre-exercise b) post-exercise.

Thus, in this study we confirmed that physical exercise has an impact on several subsystems of the ANS. Bimodal analysis confirms the day-to-day variability in a sense that the number of correlated components differ strongly in the pre-exercise measures. In the post-exercise measures, the number of correlated components is similar across modality pairs. However, the strength of correlation differs between modality pairs. Especially EDA and Temp show strong correlations, which might be the result of common or similar anatomical and functional pathways as well as their contribution to thermoregulation. However, similar correlations between HR and EDA, as well as HR and Temp indicate that post-exercise measures cannot solely be due to thermoregulatory effects. The results also show intensity effects. In the post-exercise, the number of correlated components and their strengths are higher for the high than for the moderate intensity. For the high intensity, all pairwise correlations show an increase from pre- to post-exercise, while the picture is mixed for the moderate intensity. Based on the idea of a centrally interconnected regulation of ANS subsystems, a multimodal approach offers a possibility to characterize the central integration of multiple modalities. On a descriptive level, the multimodal analysis detects the components correlated across all three modalities with average correlations higher than 0.8. In the high intensity condition, the post-exercise correlations increase strongly and are higher than for the moderate intensity. Moreover, these components are correlated across all modalities as seen from the estimated correlation structure. This provides a first hint towards the assumed integration of the subsystems in the central autonomic network (CAN) [120], [121]. The finding that the post-exercise components correlate across intensities offers an interesting starting point for further investigations to describe exercise-specific organizations of the ANS.

8.3. Ultramarathon-induced interactions in the ANS

8.3.1. Participants and dataset

15 male ultramarathon runners participated in this study. All runners completed a 65km ultramarathon. Immediately before and after the ultramarathon, resting measurements to determine the ANS function were performed with the Empatica E4 device. The sensor was placed at the participants' left wrist. Successive time series data of HR, EDA and Temp was acquired during 5 minutes in resting, supine position. More details about the experiment can be found in [74]. From each measurement, three minutes were manually selected to control for data quality and avoid movement artefacts. We excluded two participants from further analysis as the sensor was unable to measure their heart rate.

8.3.2. Bimodal and multimodal interactions

Proposed analysis- The proposed analysis for this study is the same as explained in Sections 8.2.2 and 8.2.3. Since the number of subjects in this study is 13, the maximum PCA rank for the PCA-CCA detector was set to five.

Results- The GLRT-based PCA-CCA detector detected one significant component in pre- and post-exercise as shown in Table 8.2. The correlation strength of the detected component increased from pre- to post-exercise for all modality pairs.

Modality pair	Pre-exercise		Post-exercise	
	Number of correlated components	Estimated canonical correlations	Number of correlated components	Estimated canonical correlations
EDA-HR	1	0.78	1	0.87
HR-Temp	1	0.59	1	0.94
Temp-EDA	1	0.83	1	0.96

Table 8.2.: Pairwise PCA-CCA results for pre- and post-exercise measures.

For the multimodal analysis, the absolute values of the correlation coefficients among the first two estimated SCVs $\hat{\mathbf{E}}^{(1)}$ and $\hat{\mathbf{E}}^{(2)}$ obtained from mCCA are shown in Figure 8.6 for pre-exercise and in Figure 8.7 for post-exercise. Figure 8.6 reveals that in pre-exercise, high correlations among the extracted SCVs occur mainly pairwise, i.e., there are high correlations between EDA-Temp in $\hat{\mathbf{E}}^{(1)}$ and EDA-HR in $\hat{\mathbf{E}}^{(2)}$. This can also be seen in Figure 8.8a,

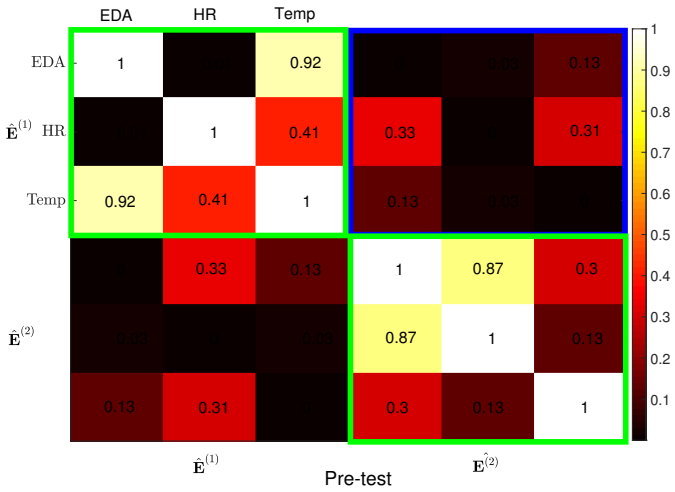


Figure 8.6.: Illustration of the absolute correlation coefficients within pre-exercise measures. On the horizontal and vertical axes the first 2 SCVs $\hat{\mathbf{E}}^{(1)}$, $\hat{\mathbf{E}}^{(2)}$ are depicted. The green-framed blocks show the correlation within an SCV. The blue-framed blocks show the correlations among the SCVs.

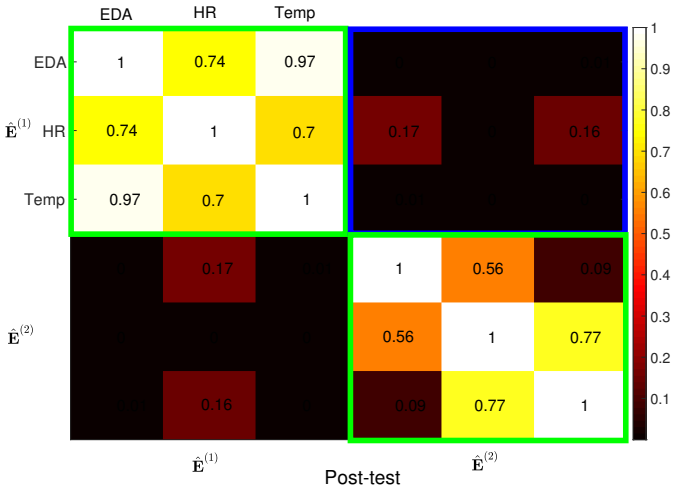


Figure 8.7.: Illustration of the absolute correlation coefficients within post-exercise measures.

which shows the estimated correlation structure. These results indicate that the interactions in the pre-exercise scenario are mainly limited to two modalities and among different components. However in post-exercise, the correlation among the first SCV of all three modalities is high indicating that the components of all three modalities are strongly interacting with each other. This can also be seen in Figure 8.8b, where the correlation structure is more dense compared to the pre-exercise and all three modalities contribute to the correlation between the first group of components as they are correlated between modality pairs EDA-HR and EDA-Temp.

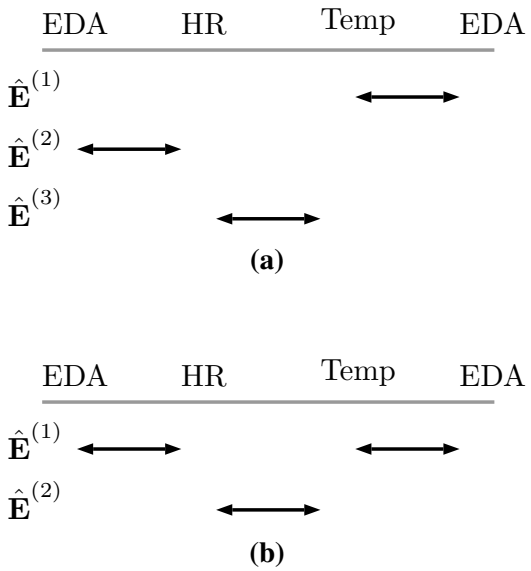


Figure 8.8.: Estimated correlation structure among the SCVs for a) pre-exercise and b) post-exercise.

8.3.3. Discussion

In the present study, we described changes across the ANS modalities induced by running an ultramarathon. We focused on the cardiac, the electrodermal and the thermoregulatory subsystems. The interaction between the modalities increased for all modality pairs and across all modalities, possibly pointing towards a reorganization of central ANS control to restore the dynamic balance after this intense physical load. We hypothesized that the changes in the ANS are interconnected since the CAN is the main driver and control mechanism of the ANS activity across modalities. Consequently, the analysis of interactions between ANS modalities offers an approach to derive information on CAN activities. Therefore, we proposed bimodal and multimodal analyses as tools to analyze the interactions between different modalities. Our results confirmed that interactions between the modalities changed after an ultramarathon. In line with running at high intensities, as in Section 8.2, the strength of correlations within all modality pairs increased.

8.4. Summary

These studies provide insights into how the ANS holistically responds to intense physical stressors. In sum, our results indicate that performing a sports activity (an exercise on the treadmill or running an ultramarathon) affects the ANS activity through the interactions be-

tween its subsystems. As a response to the physical load, measures of the ANS correlate more strongly between modalities, which might be indicative for an integrative and centrally controlled regulation to maintain the internal and dynamic balance in the human body.

Our results suggest that physical activity seems to be a holistic stimulus that alters the overall interrelation of the subsystems. The effect of physical exercise depends on the intensity. The analysis methods could be applied to analyze data in the context of training control and to potentially detect ANS states related to intense stress and overtraining. This might be of future interest to provide information on what kind of intensity is the best to achieve a certain ANS state, e.g., for the last training before a race or a match. Furthermore, it would be of interest to see how sports and physical exercise affect the ANS stress response and if this is transferable to other stressors.

We would also like to point out the limitations of our studies. By choosing one single device for the measurement, we were limited in the modality selection. Future studies might add respiratory rate or blood pressure changes to the analysis. Proving the central origin of the multimodal changes in peripheral ANS channels might be methodologically challenging but will be of interest for future studies. Moreover, the number of subjects in both studies was small. Further studies with larger number of subjects should be performed.

9. Epileptic seizure-induced changes of interrelations within the autonomic nervous system

In this chapter, we apply the techniques proposed in this thesis to analyze the multimodal interactions of the ANS in response to an epileptic seizure. Continuous EDA, HR, skin temperature and respiratory rate (RR) were measured by Empatica E4 device in two groups of children. One group of children had epileptic seizures while the other group had no seizures during the monitoring. In the group having epileptic seizures, significant changes in the correlation strength and correlation structure between the extracted components were found right before and after the seizure. This offers an interesting avenue for a potential biomarker for seizure detection and seizure prediction [1].

9.1. Introduction

Epilepsy is one of the most common neurological disorders which affects around 50 million people in the world [129]. It is associated with epileptic seizures caused by sudden excessive electrical discharges from the brain. The side effects of these seizures are both physical such as frequent injuries, broken bones, and also mental illnesses such as anxiety and depression

[1] This chapter is based on the paper: "Seizure-induced changes of interrelations within the autonomic nervous system, S. Vieluf, T. Hasija, P. J. Schreier, R. El Atrache, S. Hammond, F. M. Touserkani, T. Loddenkemper, and C. Reinsberger, *Submitted for review*, 2020". It is a result of a joint collaboration between the Signal and System Theory group and the Institute of Sports Medicine at Paderborn University, and the Division of Epilepsy and Clinical Neurophysiology at Boston Children's Hospital, Harvard Medical School. The data was recorded and preprocessed at Boston Children's Hospital. I specifically implemented and presented the bimodal and multimodal data analyses and generated all the figures used in this chapter. Sections introducing the study, interpreting and discussing the results have contributions from all authors.

[130], [131]. Moreover, the uncertainty of when an epileptic seizure will occur is one of the greatest stresses for the epileptic patients. Therefore, a significant improvement in the quality of life of patients can be made by developing an accurate system for detection and prediction of epileptic seizures [132].

EEG is the gold standard in epilepsy detection and diagnostics. However, continuous recording of EEG data on a daily or an hourly basis for seizure detection is far from convenient. Especially in children in their everyday life, it is essential that the system is easy to use and is non-invasive. Wearables and wrist-worn sensors that record several modalities of the activity in the ANS offer one possibility for developing a convenient seizure detection/prediction system [133].

Epileptic seizures are known to alter the ANS activity in several modalities [4]. These changes have been observed both preictally, i.e., before a seizure occurs and postictally, i.e., after the seizure has occurred. For example, an elevated heart rate in preictal, ictal and postictal periods was reported in [134], and an increased sweat production resulting in an EDA peak in postictal periods was reported in [7], [134]. These effects have been most prominent in patients with generalized tonic-clonic seizures (GTCS) [134]. On the other hand, the ANS states might also relate to the likelihood of an epileptic seizure [4]. Therefore, it is crucial to identify markers that provide improved monitoring and characterization of ANS functions related to an epileptic seizure.

Multimodal approaches can to some degree account for and analyze the large inter- and intra-individual variability and have been shown to contribute to a better seizure detection compared to unimodal approaches [7], [135]. Furthermore, studies examining the interactions of the respiratory and the cardiac system on a signal level showed that the information transmission between systems was altered in relation to occurring seizures [8], [136]. Currently, there is a paucity of studies focusing on multimodal signal analysis involving other systems than cardiac or respiratory functions. As cardiac changes have been shown to occur preictally, multimodal changes might also become apparent prior to seizures. In this study, we aim to evaluate interactions and changes between multimodal peripheral ANS measures prior to and after seizures, compared to the baseline and also to a group of children having no epileptic seizures. We assume that seizure-induced ANS changes have a common source, and therefore show up in several modalities and in their interaction pattern.

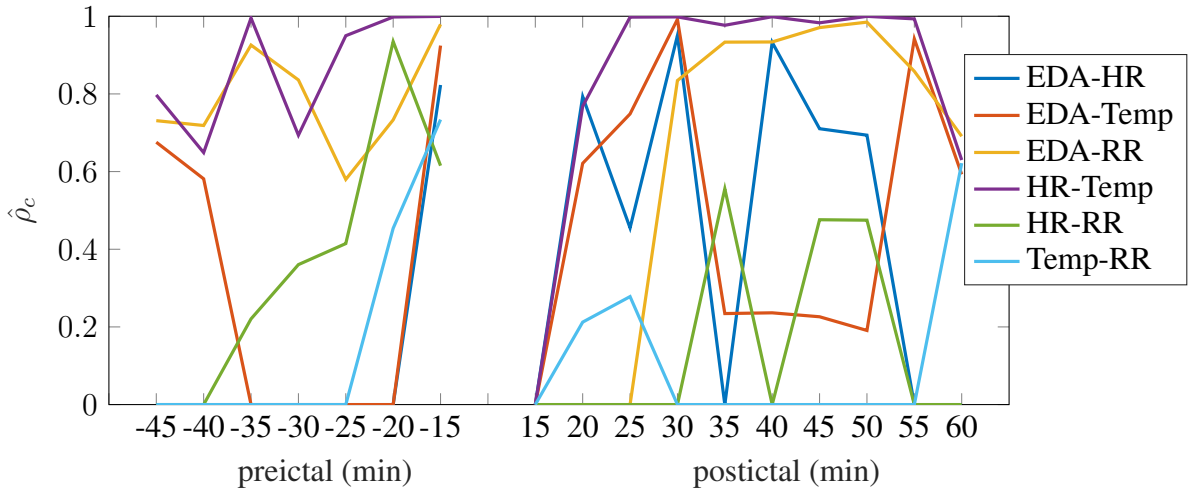
9.2. Patients and dataset

Continuous EDA, blood volume pressure (BVP), HR and Temp data was collected from 21 patients who had at least one GTCS seizure in Boston Children's Hospital. The seizure onset and offset was marked by two board-certified clinical epileptologists. The data was recorded by Empatica E4 wireless multisensor device. The RR data was calculated from the BVP. The sampling rate of EDA and Temp is 4Hz and that of HR and RR is 1Hz. We cut data from 45 minutes before the seizure onset to 60 minutes after the seizure offset. Patients with or without epilepsy who were admitted for EEG monitoring and were enrolled in the study, but had normal EEG data without seizures during their hospital stay were considered as controls. Same daytimes were selected for the controls as their corresponding matched patients. More details about the subjects and the dataset can be found in [75].

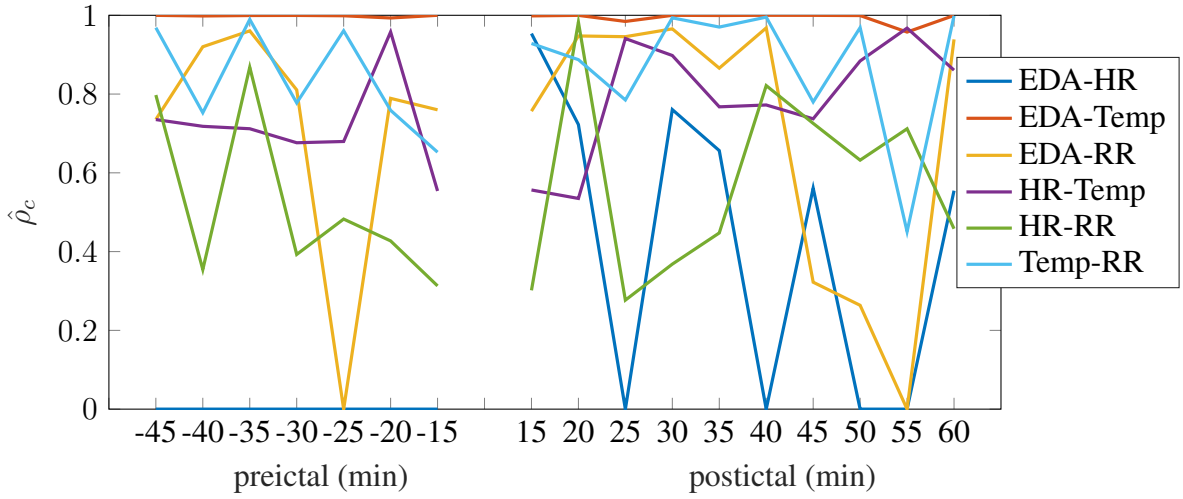
9.3. Bimodal interactions

Proposed analysis- To analyze the bimodal relationships, the joint PCA-CCA technique of Section 3.5 was applied. The preictal and postictal data were divided into consecutive windows of 15 minutes with an overlap of 10 minutes. This was done to increase the number of time points to better reflect the course of changes in the bimodal interactions. For each time window and each modality, a data matrix was generated such that the recorded time series from each subject forms a column of the data matrix. Thus, the size of the data matrix for each modality is the number of time points times the number of subjects. Matrices were generated for the patient group and the control group independently, so that the results are reported on a group level and also allow for the inter-group comparison. In this study, each subject was regarded as an observation and the number of recorded time points corresponds to the dimension of the data set. As the number of subjects is much smaller than the number of time points, the data matrices for all modalities are high-dimensional. Thus, the GLRT-based PCA-CCA detector of Section 3.5 was applied to each pair of modalities. This detector performs dimension reduction and estimates the strength of association between two modalities in a joint way [28]. As in Chapter 8, we chose this detector to have an adjustable probability of false alarm P_{fa} as the number of subjects is comparably low in our study. The strength of association between any two modalities was then estimated with an overall correlation ρ_c computed as in (8.1). We first estimate ρ_c for all six modality pairs and then average them to report the total strength of association. The maximum PCA dimension was set to

seven and the P_{fa} was set to 0.05.



(a)



(b)

Figure 9.1.: $\hat{\rho}_c$ value for each modality pair for each 15 minutes block per group a) for patients and b) for controls.

Results- Bimodal results for all modality pairs show fluctuations in $\hat{\rho}_c$ values for both patients and controls as shown in Figure 9.1. This indicates a high variability in the different ANS modality pairs. Figure 9.1a shows the pairwise correlations for different modality pairs for the patients and Figure 9.1b for the controls. The six $\hat{\rho}_c$ values are averaged to measure the total strength of correlation among all modalities and is shown in Figure 9.2. The mean $\hat{\rho}_c$ for the controls fluctuates within a certain limit compared to that of the patients. The patients demonstrate lower mean values preictally, then increasing values shortly before the seizure, followed by a postictal drop.

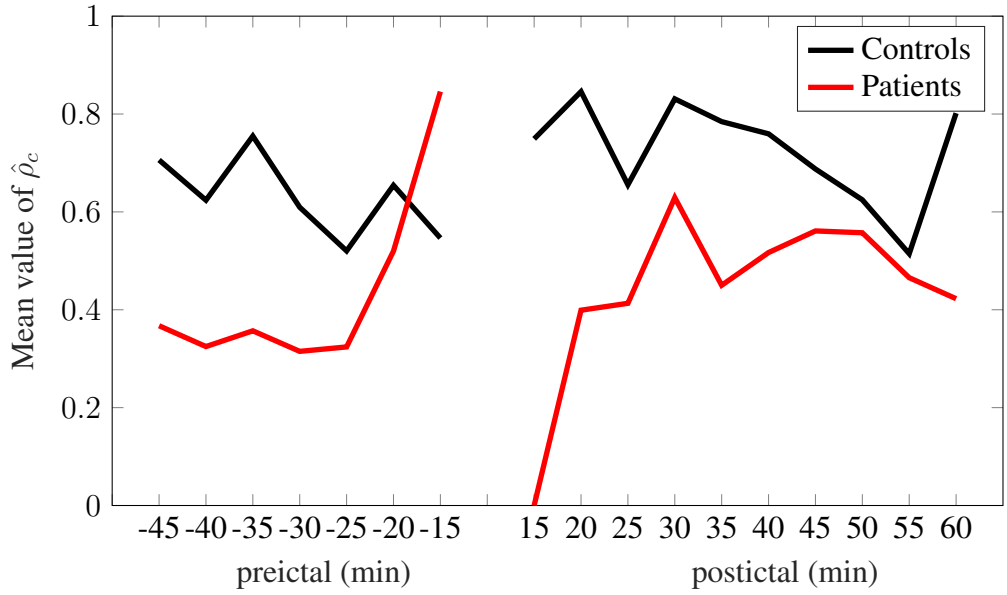


Figure 9.2.: Mean value of $\hat{\rho}_c$ across all modality pairs for each 15 minutes block per group.

9.4. Multimodal interactions

Proposed analysis- The bimodal analysis revealed significant changes in the average correlation across all modality pairs just before the seizure onset and right after the seizure. However, the limitation of the bimodal analysis is that it is not straightforward to interpret whether these detected changes are due to several distinct components interacting among different modality pairs or are associated with a common component (or components) interacting across all modalities. This can be analyzed by estimating the complete correlation structure. The estimated correlation structure for the preictal data in patients using the rank-reduced version of the mCCA-HT technique of Section 6.3 is shown in Figure 9.3. This is similar to the correlation map explained in Figure 6.7, where a white block represents a nonzero correlation coefficient among a modality pair and the black block represents a zero correlation coefficient. Since the data sets generated for all modalities are high-dimensional, PCA is applied to each data set before estimating the correlation structure. The PCA dimensions are estimated using Algorithm 4 on page 101.

Results- It can be seen from Figure 9.3a,b that some time before the seizure, the components are correlated mainly across two modalities. However, just before the seizure as seen in Figure 9.3c, not only more components are interacting among different modalities as seen from an increase in the number of correlated components, but also the first extracted component is correlated across all pairs of modalities.

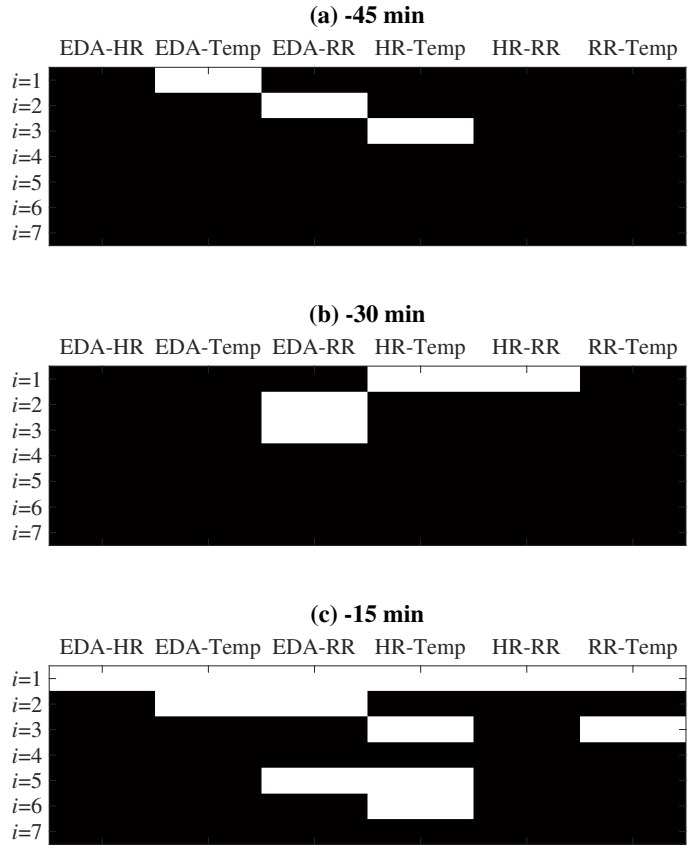


Figure 9.3.: Estimated correlation structure between the extracted omponents of EDA, HR, Temp and RR in patients for preictal data illustrated using a correlation map. The white blocks represent nonzero correlation coefficients and the black blocks represent zero correlation coefficients.

Similarly, the estimated correlation structure for the postictal data in patients is shown in Figure 9.4. For the first postictal segment in Figure 9.4a, no correlated component was detected and therefore, the correlation map is completely black. For the later segments as seen in Figure 9.4b,c,d, although components are interacting more among different modalities compared to Figure 9.4a, these interactions are limited to a few pairs of modalities and no component was found which is correlated across all modality pairs.

The estimated correlation structure for preictal and postictal data in controls is shown in Figure 9.5 and Figure 9.6, respectively. There is some variability in the correlation structure among different time segments. However, compared to the evolution of correlation structure in patients (in Figures 9.3 and 9.4), no such significant trend across time is observed

here.

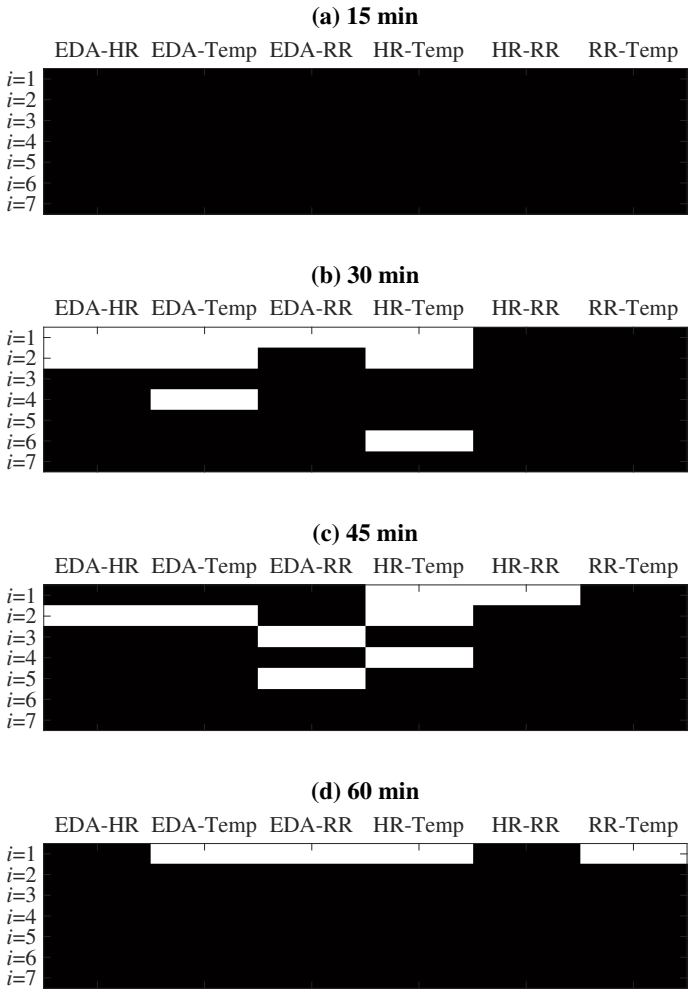


Figure 9.4.: Estimated correlation structure between the extracted components of EDA, HR, Temp and RR in patients for postictal data.

9.5. Discussion and summary

Epileptic seizures are complex stressors which challenge the homeostasis within the human body and evoke responses in the ANS that manifest themselves in multiple subsystems. To gain deeper insights into the dynamic changes of functional interactions of the ANS subsystems, interactions among HR, RR, EDA and Temp were analyzed. Our sample consisted of a

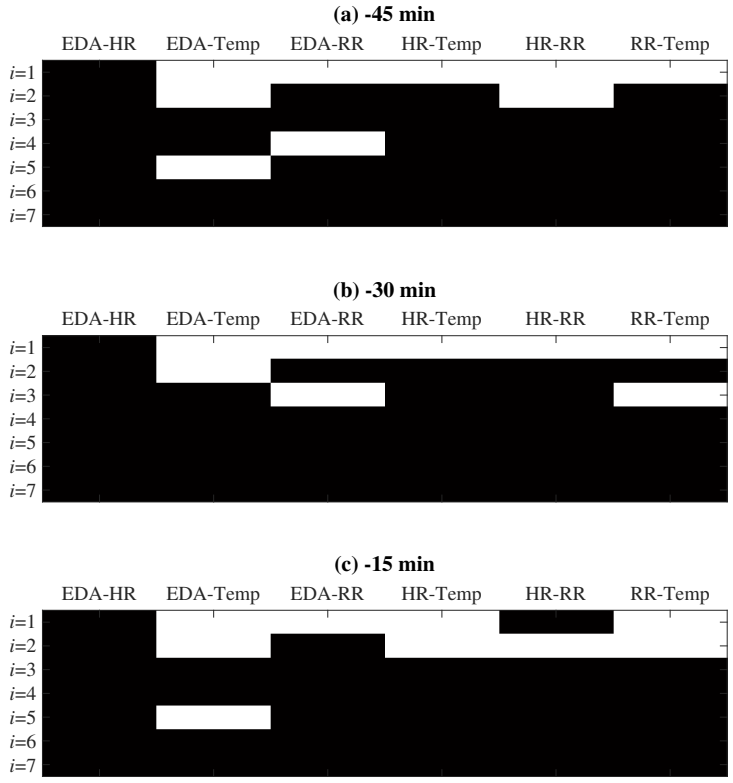


Figure 9.5.: Estimated correlation structure between the extracted components of EDA, HR, Temp and RR in controls for pretictal data.

group of patients with GTCS and a group of controls with normal EEG and no seizures.

Bimodal and multimodal analyses were used to infer the centrally modulated interaction of subsystems based on changes among processed signals recorded peripherally. When looking at the individual modality pairs in the bimodal analysis, day-to-day variability became evident by a large variability of correlation values across time windows. Especially when analyzing control data, the strength of correlation differs between modality pairs. We selected the same 24-hour times as the seizure times to increase the comparability of datasets in terms of influences of daytime. Taking the complexity and redundancies in the possible responses to adapt to external stimuli into account, parts of the variability within the ANS can be reduced by averaging the correlation scores across modality pairs. Our results indicate a trend after averaging, where the patient group shows moderate values in the early preictal period followed by an increase before the seizure, and finally reaching the peak just before the seizure onset. Right after the seizure, the interactions between all modalities are diminished. As the control group's average correlations remain in a certain range of variability,

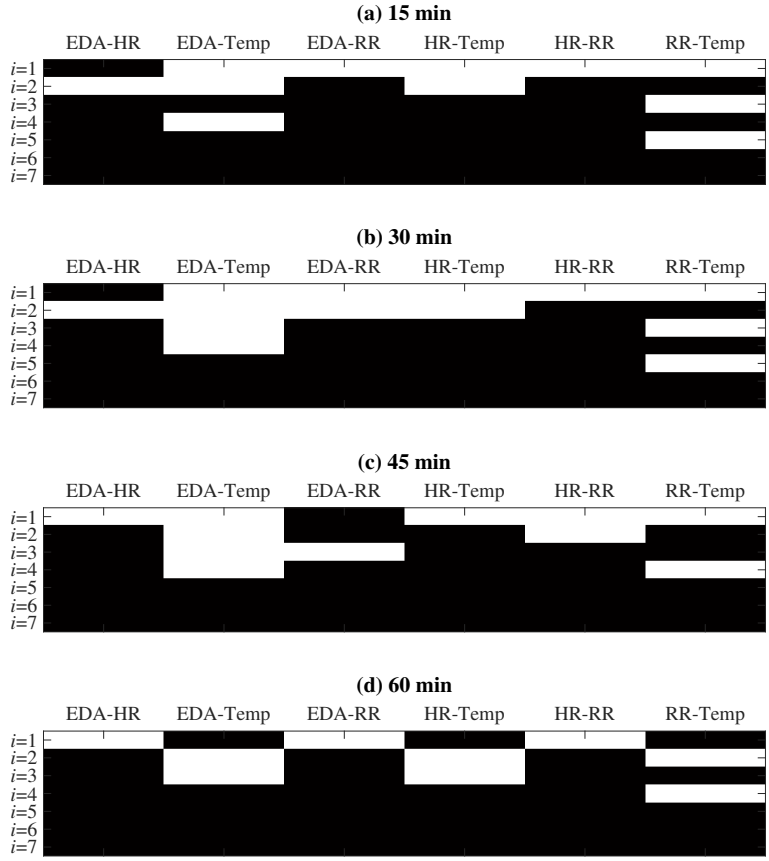


Figure 9.6.: Estimated correlation structure between the extracted components of EDA, HR, Temp and RR in controls for posttictal data.

this pattern might be indicative for seizures and would have a predictive potential.

A similar trend was seen in the multimodal analysis, where both the number of correlated components and the modalities across which they are correlated increase right before the seizure onset and decrease right after the seizure. Moreover, right before the seizure, the analysis revealed a component correlated across all pairs of modalities; thus, indicating that the interactions among the modalities are centrally driven in the ANS. This points towards a central integration of all subsystems and a possible ANS state change right before the seizure. These changes are of great interest and may offer potential biomarkers that could contribute to seizure semiology, detection, reporting and prediction [137].

However, there are also some limitations in our study. One of the challenges is the sample selection. Since the ANS changes were expected to be most prominent in GTCS rather than

in other seizure types, only patients with GTCS were analyzed, and thus the sample size was rather small. It would further be of great interest to utilize a healthy control group. Transferability and generalizability to real life situations outside the hospital setting, where different stimuli could lead to additional responses, requires further testing. Finally regarding the analyses, the selected approaches to jointly analyze ANS signals were so far used in an explorative way, and therefore offer several possibilities for modifications and expansions for future applications.

10. Summary

10.1. Conclusions

In this thesis, novel statistical techniques for completely characterizing the linear association among multiple data sets were developed and applied in various real-world applications. Each data set was represented as a set of observations (or samples) of a random vector, and correlation among different data sets was used as the statistical measure of linear association. The correlation structure for two data sets can be reduced without loss of generality to the pairwise correlation structure between the latent components. This means that the individual components are either correlated or uncorrelated between the data sets. In this case, estimating the model order that identifies the number of correlated components provides a complete summary of the joint-correlation information among the two sets. This, however, is not possible for more than two sets. Some components can be correlated across all pairs of data sets, some across different subsets of data sets and some uncorrelated among all data sets; thus, several possibilities of characterizing the joint-correlation information among the data sets exist.

A model-order selection problem was posed in Chapter 5 and two techniques to estimate the number of components that are correlated across all the data sets were developed. If it can be assumed that the components are either correlated across all data sets or are completely uncorrelated, the model order provides a complete summary of the linear association between multiple data sets. The first technique in Chapter 5 estimated the model order by assuming this special correlation structure. The advantage of such an approach is that the GLRT and its distribution under the null hypothesis were derived in closed form. This enabled to develop a technique for data sets with small number of samples compared to their dimensions, where the model order and the PCA rank applied to each data set were jointly estimated. Such a technique offers a substantial benefit over a two-step approach where first a PCA-

preprocessing is applied to all data sets followed by estimating the model order. Of course there is no free lunch. The PCA rank required to include all the correlated components is assumed to be small compared to the number of samples. However, this is a more reasonable assumption than assuming that the dimensions of the data sets are small compared to the the number of samples. The second technique in Chapter 5 is able to tackle arbitrary correlation structure among the components. In this case, the model order only characterizes the joint-correlation information which exists in all the data sets. Since it is challenging to derive the ML function with respect to the model order, a completely different approach was followed. It was derived under reasonably high SNR that the model order corresponds to the number of non-zero eigenvalues of the product of coherence matrices of all pairs of data sets. Both proposed techniques outperform the state-of-the-art competitors in various numerical scenarios.

When the correlation structure is not known apriori, assuming a special correlation structure is far from optimal as it restricts the degrees of freedom in the model. Moreover, in this general case, determining only the model order is insufficient. A model-selection problem was formulated in Chapter 6, and two novel but complementary techniques were proposed for complete characterization of the second-order association across more than two data sets. This was done by estimating the complete correlation structure, i.e., the number of correlated components and the data sets across which they are correlated. The mCCA-HT technique combined multiset CCA with pairwise model-order estimates, and thus leveraged the recent results for model-order estimation with two data sets in sample-poor regime. The joint-EVD technique used the eigenvalues and eigenvectors of the composite coherence matrix and relied solely on joint information from all of the data sets provided by them. The necessary and sufficient conditions under which the correlation structure can be identified were theoretically derived using the results from graph theory. Furthermore, extensions of both techniques for small-sample support were also introduced. It was later numerically shown that the joint-EVD technique demonstrates superior accuracy in estimating the correlation structure of components correlated across more data sets, while the mCCA-HT technique performs better when the components are sparsely correlated. To the best of my knowledge, no competing technique exist in the literature that estimates the complete correlation structure in multiple data sets without imposing strict assumptions on the correlation structure.

Throughout this thesis, minimal assumptions, which were necessary to identify the correlated components and their correlation structure, were made. Therefore, the proposed techniques can be applied to arbitrary number of data sets with different dimensions and unknown cor-

relation structure. This flexibility made it possible to employ these techniques in a variety of applications as follows. In Chapter 4, a detector for the number of improper signals in complex-valued data was developed and applied to estimate the number of sources impinging on a sensor array with very few observations. In Chapter 7, a new method for source enumeration and multi-speaker voice activity detection in WASN was developed which significantly outperforms the existing standard techniques. Later in Chapter 8, the proposed techniques were applied in the field of sports science and it was shown that physical exercise affects the interactions among various subsystems of the ANS, and that these interactions change depending on the strength of the physical load. Finally, in the field of epilepsy, specific seizure-induced changes in the interactions of four different ANS modalities were identified in Chapter 9. These results are promising and will be analyzed in the future to identify potential biomarkers for seizure detection, and more importantly for the open and challenging problem of seizure prediction.

The techniques developed in this thesis are not limited to the specific data sets and the applications shown here. Depending on whether or not an apriori correlation structure seems a reasonable assumption for an application, one of the proposed techniques can be applied. For instance, if a common underlying source vector is assumed to be observed by all the data sets, the GLRT technique of Chapter 5 can be readily employed, whereas the mCCA-HT and joint-EVD techniques of Chapter 6 can be used when such an assumption is too restrictive for the application under concern.

10.2. Future work

The techniques developed in this thesis can be directly used, modified or act as a motivation for developing new methods in the future. I list a few possible directions as follows.

1. **Non-linearity and sample-to-sample dependence**: In this work, correlation among different data sets was analyzed. Although correlation is a reliable measure for linear dependence, it is does not account for non-linear dependencies, which are common in many applications [138], [139]. One way to analyze non-linear dependencies is through higher-order moments. IVA is an extension of ICA for multiple data sets and considers second and higher-order statistics to measure dependencies among different sets [18]. Another advantage of using an IVA framework is that it allows to exploit the sample-to-sample dependencies when the samples are not i.i.d. [42]. This is particularly useful for applications involving time-series or images where samples

are typically not i.i.d. However, model selection in IVA is an open problem even in the sample-rich regime. Model-selection techniques developed in this thesis can be adapted for IVA to analyze the *dependency structure*, i.e., which components in different data sets are dependent with each other. In this context, IVA-G, the IVA model for Gaussian distributed data sets [140], can act as a good starting point. A similar approach as in Section 6.4.1 can be followed by examining the eigenvalues and eigenvectors of the component covariance matrix extracted using IVA.

2. **Correlation vs causation**: The fact that correlation does not imply causation limits the applicability of components that are correlated across multiple data sets [141]. Since correlation is a bidirectional relationship, it cannot be said whether there is a cause and effect relationship between the correlated components or if there exists another confounding variable driving them. Defining true causality is a challenging task since one cannot always account for all the confounding variables. However, in many fields including econometrics, Granger causality (GC) is commonly used to test for causality. A time series X Granger-causes time series Y if the past values of X provide statistically significant information about the future values of Y [142]. When $\mathbf{X}$ and $\mathbf{Y}$ contain multiple time series, one way to test and measure GC is using partial CCA. In partial CCA, the correlation between two data sets is maximized in a low-dimensional space after eliminating the effect of the third data set [143]. Under the assumption that $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian distributed, GC can be written as a function of the nonzero partial canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$ after eliminating the effect of the past values of $\mathbf{Y}$ [144]. However, there has not been much work on GC when the data sets are sample poor. The techniques developed in this thesis can be modified for partial CCA to detect and quantitatively measure the causality in Granger-sense in the sample-poor regime.

3. **Correlation structure based joint blind source separation**: JBSS tools for multiple data sets like mCCA [21], IVA [18], group ICA [145] rely on correlation or dependency between the latent (or source) components of different data sets to jointly estimate them. However, these tools are not designed to incorporate any knowledge about the correlation structure among the underlying sources when estimating them. This is suboptimal specially when the available number of samples is not large. Typically not all sources are correlated across all data sets. If some data sets do not include any correlated sources, including them in the analysis will lead to an inaccurate source separation. This is because for a limited number of samples, including these data sets would lead to estimating more parameters, which causes an inaccurate estimation of

the covariance matrices. A possible solution is to estimate the sources in a deflationary approach, where the correlation structure for the first set of sources is estimated using the techniques developed in this thesis, and only the data sets with correlated sources are used for JBSS. For estimating the next and subsequent sets of sources, the data sets are recomputed after projecting them on the orthogonal complement of the already estimated set of sources and the previous step is repeated. This unified approach where the correlation structure is determined jointly with the estimation of the correlated sources could lead to significant improvements, and thus offers an interesting avenue for applications where the data is limited.

List of publications

Journal publications

Y. Song, P. J. Schreier, D. Ramirez, and **T. Hasija**, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Processing, 128*, pp.449-458, 2016.

T. Hasija, C. Lameiro, and P. J. Schreier, "Determining the dimension of the improper signal subspace in complex-valued data", *IEEE Signal Processing Letters, 24(11)*, pp.1606-1610, 2017.

S. Vieluf, **T. Hasija**, R. Jakobsmeyer, P. J. Schreier, and C. Reinsberger, "Exercise-induced changes of multimodal interactions within the autonomic nervous network," *Frontiers in Physiology, 10*, p.240, 2019.

S. Vieluf, V. Scheer, **T. Hasija**, P. J. Schreier, and C. Reinsberger, "Multimodal approach towards understanding the changes in the autonomic nervous system induced by an ultramarathon," *Research in Sports Medicine, 28(2)*, pp.231-240, 2020.

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, 176*, p.107613, 2020.

S. Vieluf, **T. Hasija**, P. J. Schreier, R. El Atrache, S. Hammond, F. M. Touserkani, T. Loddenkemper, and C. Reinsberger, "Seizure-induced changes of interrelations within the autonomic nervous system," *Submitted for review*, 2020.

Conference publications

T. Hasija, Y. Song, P. J. Schreier, and D. Ramirez, "Detecting the dimension of the subspace correlated across multiple data sets in the sample poor regime," *in Proceedings of the IEEE Workshop on Statistical Signal Processing, Palma de Mallorca, Spain*, pp.1-5, 2016.

Y. Song, **T. Hasija**, P. J. Schreier, and D. Ramirez, "Determining the number of signals correlated across multiple data sets for small sample support," *in Proceedings of the European Signal Processing Conference, Budapest, Hungary*, pp.1528-1532, 2016.

T. Hasija, Y. Song, P. J. Schreier, and D. Ramirez, "Bootstrap-based detection of the number of signals correlated across multiple data sets," *in Proceedings of the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, U.S.A.*, pp.610-614, 2016.

T. Marrinan, **T. Hasija**, C. Lameiro, and, P. J. Schreier, "Complete model selection in multiset canonical correlation analysis," *in Proceedings of the European Signal Processing Conference, Rome, Italy*, pp.1082-1086, 2018.

C. Lameiro, **T. Hasija**, T. Marrinan, and, P. J. Schreier, "Estimating the number of correlated components based on random projections," *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Brighton, U.K.*, pp.5152-5156, 2019.

T. Hasija, M. Gölz, M. Muma, P. J. Schreier, and A. M. Zoubir, "Source enumeration and robust voice activity detection in wireless acoustic sensor networks," *in Proceedings of the Asilomar Conference on Signals, Systems and Computers, Pacific Grove, U.S.A.*, pp.1257-1261, 2019.

References

[1] J. Z. Varghese and R. G. Boone, "Overview of autonomous vehicle sensors and systems", in *Proceedings of the International Conference on Operations Excellence and Service Engineering*, 2015, pp. 178–191.

[2] X. Chen, Z. J. Wang, and M. McKeown, "Joint blind source separation for neurophysiological data analysis: Multiset and multimodal methods", *IEEE Signal Processing Magazine*, vol. 33, no. 3, pp. 86–107, 2016.

[3] B. Müngen, M. S. Berilgen, and A. Arıkanoğlu, "Autonomic nervous system functions in interictal and postictal periods of nonepileptic psychogenic seizures and its comparison with epileptic seizures", *Seizure*, vol. 19, no. 5, pp. 269–273, 2010.

[4] Y. Nagai, "Modulation of autonomic activity in neurological conditions: Epilepsy and Tourette syndrome", *Frontiers in Neuroscience*, vol. 9, p. 278, 2015.

[5] W. Jänig and H. J. Häbler, "Specificity in the organization of the autonomic nervous system: A basis for precise neural regulation of homeostatic and protective body functions.", *Progress in Brain Research*, vol. 122, pp. 351–367, 2000.

[6] N. M. Correa, Y. O. Li, T. Adalı, and V. D. Calhoun, "Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia", *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 998–1007, 2008.

[7] D. Cogan, J. Birjandtalab, M. Nourani, J. Harvey, and V. Nagaraddi, "Multi-biosignal analysis for epileptic seizure monitoring", *International Journal of Neural Systems*, vol. 27, no. 01, p. 1 650 031, 2017.

[8] C. Varon, K. Jansen, L. Lagae, L. Faes, and S. Van Huffel, "Transient behavior of cardiorespiratory interactions towards the onset of epileptic seizures", in *Proceedings of the Computing in Cardiology 2014*, IEEE, 2014, pp. 917–920.

[9] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[10] M. K. Tippett, T. DelSole, S. J. Mason, and A. G. Barnston, "Regression-based methods for finding coupled patterns", *Journal of Climate*, vol. 21, no. 17, pp. 4384–4398, 2008.

[11] Y. Yamanishi, J. P. Vert, A. Nakaya, and M. Kanehisa, "Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis", *Bioinformatics*, vol. 19, no. suppl 1, pp. 323–330, 2003.

[12] T. Hasija, M. Gölz, M. Muma, P. J. Schreier, and A. M. Zoubir, "Source enumeration and robust voice activity detection in wireless acoustic sensor networks", in *Proceedings of the 53rd Asilomar Conference on Signals, Systems and Computers*, 2019, pp. 1257–1261.

[13] N. Asendorf and R. R. Nadakuditi, "Improving multiset canonical correlation analysis in high dimensional sample deficient settings", in *Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 112–116.

[14] V. D. Calhoun and T. Adalı, "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.

[15] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion", *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.

[16] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning", in *Proceedings of the International conference on machine learning*, 2015, pp. 1083–1092.

[17] H. Hotelling, "Relations between two sets of variates", *Biometrika*, pp. 321–377, 1936.

[18] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components", in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*, Springer, 2006, pp. 165–172.

[19] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition", *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[20] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Application to medical imaging", *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.

[21] J. R. Kettenring, "Canonical analysis of several sets of variables", *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[22] 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*, vol. 3, 1968, pp. 227–228.

[23] Y. O. Li, T. Adalı, 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.

[24] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules", *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[25] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise", *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1652–1663, 1993.

[26] P. Stoica, K. M. Wong, and Q. Wu, "On a nonparametric detection method for array signal processing in correlated noise fields", *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 1030–1032, 1996.

[27] W. Chen, J. P. Reilly, and K. M. Wong, "Detection of the number of signals in noise with banded covariance matrices", *IEE Proceedings-Radar, Sonar and Navigation*, vol. 143, no. 5, pp. 289–294, 1996.

[28] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support", *Signal Processing*, vol. 128, pp. 449–458, 2016.

[29] C. Lameiro, T. Hasija, T. Marrinan, and P. J. Schreier, "Estimating the number of correlated components based on random projections", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2019, pp. 5152–5156.

[30] Y. Wu, K. W. Tam, and F. Li, "Determination of number of sources with multiple arrays in correlated noise fields", *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1257–1260, 2002.

[31] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, "Detecting the dimension of the subspace correlated across multiple data sets in the sample poor regime", in *Proceedings of the IEEE Workshop on Statistical Signal Processing*, 2016, pp. 1–5.

[32] Y. Song, T. Hasija, P. J. Schreier, and D. Ramírez, "Determining the number of signals correlated across multiple data sets for small sample support", in *Proceedings of the European Signal Processing Conference*, 2016, pp. 1528–1532.

[33] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, "Bootstrap-based detection of the number of signals correlated across multiple data sets", in *Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers*, 2016, pp. 610–614.

[34] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Estimation of common subspace order across multiple datasets: Application to multi-subject fMRI data", in *Proceedings of the 51st Annual Conference on Information Sciences and Systems*, 2017, pp. 1–5.

[35] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces", *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers*, pp. 994–997, 2004.

[36] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis", *Machine Learning*, vol. 83, no. 3, pp. 331–353, 2011.

[37] R. R. Nadakuditi, "Fundamental finite-sample limit of canonical correlation analysis based detection of correlated high-dimensional signals in white noise", in *Proceedings of the Statistical Signal Processing Workshop*, IEEE, 2011, pp. 397–400.

[38] N. Asendorf and R. R. Nadakuditi, "Improved detection of correlated signals in low-rank-plus-noise type data sets using informative canonical correlation analysis", *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3451–3467, 2017.

[39] C. Lameiro and P. J. Schreier, "A sparse CCA algorithm with application to model-order selection for small sample support", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2017, pp. 4721–4725.

[40] V. Tsatsishvili, F. Cong, P. Toiviainen, and T. Ristaniemi, "Combining PCA and multiset CCA for dimension reduction when group ICA is applied to decompose naturalistic fMRI data", in *Proceedings of the International Joint Conference on Neural Networks*, 2015, pp. 1–6.

[41] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate, and capacity", *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 824–831, 2000.

[42] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds", *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, 2014.

[43] J. Vía, M. Anderson, X.-L. Li, and T. Adalı, "A maximum likelihood approach for independent vector analysis of Gaussian data sets", in *Proceedings of the International Workshop on Machine Learning for Signal Processing*, IEEE, 2011, pp. 1–6.

[44] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge University Press, 2010.

[45] T. Adalı and V. D. Calhoun, "Complex ICA of brain imaging data", *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 136–139, 2007.

[46] X.-L. Li, T. Adalı, and M. Anderson, "Noncircular principal component analysis and its application to model selection", *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4516–4528, 2011.

[47] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 85–88.

[48] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 1958, vol. 2.

[49] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods", *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[50] 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.

[51] 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.

[52] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[53] H. Akaike, "Information theory and an extension of the maximum likelihood principle", in *Selected Papers of Hirotugu Akaike*, Springer, 1998, pp. 199–213.

[54] G. Schwarz *et al.*, "Estimating the dimension of a model", *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[55] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[56] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal statistical society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[57] E. Aharoni and S. Rosset, "Generalized alpha-investing: Definitions, optimality results and application to public databases", *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 771–794, 2014.

[58] M. S. Bartlett, "A note on the multiplying factors for various chi-square approximations", *Journal of the Royal Statistical Society*, pp. 296–298, 1954.

[59] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[60] P. Stoica, Y. Selén, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection", *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 794–797, 2004.

[61] A. M. Zoubir and D. R. Iskander, *Bootstrap Techniques For Signal Processing*. Cambridge University Press, 2004.

[62] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[63] B. Efron and R. J. Tibshirani, *An Introduction To The Bootstrap*. CRC press, 1994.

[64] P. Hall and S. R. Wilson, "Two guidelines for bootstrap hypothesis testing", *Biometrics*, pp. 757–762, 1991.

[65] M. Hassan, S. Boudaoud, J. Terrien, B. Karlsson, and C. Marque, "Combination of canonical correlation analysis and empirical mode decomposition applied to denoising the labor electrohysterogram", *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 9, pp. 2441–2447, 2011.

[66] K. T. Sweeney, S. F. McLoone, and T. E. Ward, "The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique", *IEEE transactions on biomedical engineering*, vol. 60, no. 1, pp. 97–105, 2012.

[67] M. S. Bartlett, "The statistical significance of canonical correlations", *Biometrika*, pp. 29–37, 1941.

[68] H. D. Vinod, "Canonical ridge and econometrics of joint production", *Journal of econometrics*, vol. 4, no. 2, pp. 147–166, 1976.

[69] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical correlation analysis when the data are curves", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 55, no. 3, pp. 725–740, 1993.

[70] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis", *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[71] C. Lameiro and P. J. Schreier, "Cross-validation techniques for determining the number of correlated components between two data sets when the number of samples is very small", in *Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, 2016, pp. 601–605.

[72] Y. Levin-Schwartz, Y. Song, P. J. Schreier, V. D. Calhoun, and T. Adalı, "Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data", *NeuroImage*, vol. 134, pp. 486–493, 2016.

[73] S. Vieluf, T. Hasija, R. Jakobsmeyer, P. J. Schreier, and C. Reinsberger, "Exercise-induced changes of multimodal interactions within the autonomic nervous network", *Frontiers in physiology*, vol. 10, p. 240, 2019.

[74] S. Vieluf, V. Scheer, T. Hasija, P. J. Schreier, and C. Reinsberger, "Multimodal approach towards understanding the changes in the autonomic nervous system induced by an ultramarathon", *Research in Sports Medicine*, pp. 1–10, 2019.

[75] S. Vieluf, T. Hasija, P. J. Schreier, R El Atrache, S Hammond, F. Touserkani, T Loddenkemper, and C. Reinsberger, "Seizure-induced changes of interrelations within the autonomic nervous system", *Submitted for review*, 2020.

[76] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[77] I. Santamaria, L. L. Scharf, D. Cochran, and J. Vía, "Passive detection of rank-one signals with a multiantenna reference channel", in *Proceedings of the European Signal Processing Conference*, 2016, pp. 140–144.

[78] D. Lawley, "Tests of significance for the latent roots of covariance and correlation matrices", *Biometrika*, vol. 43, pp. 128–136, 1956.

[79] F. R. Bach and M. I. Jordan, "Kernel independent component analysis", *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[80] 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.

[81] I. M. Johnstone, "Multivariate analysis and jacobi ensembles: Largest eigenvalue, tracy–widom limits and rates of convergence", *Annals of statistics*, vol. 36, no. 6, p. 2638, 2008.

[82] R. R. Nadakuditi and J. W. Silverstein, "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples", *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 468–480, 2010.

[83] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples", *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.

[84] T. Adalı, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety", *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, 2011.

[85] C. N. Mooers, "A technique for the cross spectrum analysis of pairs of complex-valued time series, with emphasis on properties of polarized components and rotational invariants", in *Deep Sea Research and Oceanographic Abstracts*, vol. 20, 1973, pp. 1129–1141.

[86] P. Chevalier and A. Blin, "Widely linear MVDR beamformers for the reception of an unknown signal corrupted by noncircular interferences", *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5323–5336, 2007.

[87] F. Roemer and M. Haardt, "Multidimensional unitary tensor-ESPRIT for non-circular sources", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3577–3580.

[88] P. Chargé, Y. Wang, and J. Saillard, "A non-circular sources direction finding method using polynomial rooting", *Signal Processing*, vol. 81, no. 8, pp. 1765–1770, 2001.

[89] H. Li, N. M. Correa, P. A. Rodriguez, V. D. Calhoun, and T. Adalı, "Application of independent component analysis with adaptive density model to complex-valued fMRI data", *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 10, pp. 2794–2803, 2011.

[90] G. E. Box, "A general distribution theory for a class of likelihood criteria", *Biometrika*, vol. 36, no. 3/4, pp. 317–346, 1949.

[91] A. T. Walden and P. Rubin-Delanchy, "On testing for impropriety of complex-valued Gaussian vectors", *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 825–834, 2009.

[92] M. Haardt and F. Roemer, "Enhancements of unitary ESPRIT for non-circular sources", in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2004, pp. 101–104.

[93] J.-P. Delmas, "Asymptotically minimum variance second-order estimation for non-circular signals with application to DOA estimation", *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1235–1241, 2004.

[94] M. Bhandary, "Detection of the number of signals in the presence of white noise in decentralized processing", *IEEE Transactions on Signal Processing*, vol. 46, no. 3, pp. 800–803, 1998.

[95] X. F. Gong, Q. H. Lin, F. Y. Cong, and L. De Lathauwer, "Double coupled canonical polyadic decomposition for joint blind source separation", *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3475–3490, 2018.

[96] W. W. Hager, "Updating the inverse of a matrix", *SIAM review*, vol. 31, no. 2, pp. 221–239, 1989.

[97] R. F. Brcich, A. M. Zoubir, and P. Pelin, "Detection of sources using bootstrap techniques", *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 206–215, 2002.

[98] S. Aouada, D. Traskov, N. d'Heureuse, and A. M. Zoubir, "Application of the bootstrap to source detection in nonuniform noise", *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, Philadelphia PA, USA*, 2005.

[99] Y.-O. Li, W. Wang, T. Adalı, and V. D. Calhoun, "Cca for joint blind source separation of multiple datasets with application to group fmri analysis", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1837–1840.

[100] N. M. Correa, T. Adalı, 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.

[101] G. R. Naik and D. K. Kumar, "An overview of independent component analysis and its applications", *Informatica*, vol. 35, no. 1, 2011.

[102] D. J. Struik, *A Source Book In Mathematics*. Princeton University Press, 1986, pp. 89–93.

[103] Z. B. Charles, M. Farber, C. R. Johnson, and L. Kennedy-Shaffer, "Nonpositive eigenvalues of hollow, symmetric, nonnegative matrices", *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1384–1400, 2013.

[104] R. Bellman, *Introduction to matrix analysis*. Siam, 1997, vol. 19.

[105] F. P. Ramsey, "On a problem of formal logic", in *Classic Papers in Combinatorics*, Springer, 2009, pp. 1–24.

[106] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective", in *Proceedings of the IEEE symposium on communications and vehicular technology*, IEEE, 2011, pp. 1–6.

[107] M. Cobos, F. Antonacci, A. Mouchtaris, and B. Lee, "Wireless acoustic sensor networks and applications", *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

[108] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks", *Signal Processing*, vol. 107, pp. 4–20, 2015.

[109] M. H. Bahari, L. K. Hamaidi, M. Muma, J. Plata-Chaves, M. Moonen, A. M. Zoubir, and A. Bertrand, "Distributed multi-speaker voice activity detection for wireless acoustic sensor networks", *ArXiv preprint arXiv:1703.05782*, 2017.

[110] L. K. Hamaidi, M. Muma, and A. M. Zoubir, "Multi-speaker voice activity detection by an improved multiplicative non-negative independent component analysis with sparseness constraints", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing)*, IEEE, 2017, pp. 4611–4615.

[111] ——, "Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction", in *Proceedings of the European Signal Processing Conference*, IEEE, 2017, pp. 161–165.

[112] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[113] D. Wang and J. Lim, "The unimportance of phase in speech enhancement", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[114] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[115] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement", *Speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[116] A. M. Zoubir and D. R. Iskander, *Bootstrap techniques for signal processing*. Cambridge University Press, 2004.

[117] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables", *Journal of Royal Statistical Society, Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[118] C. L. Mallows, "Some comments on c_p", *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[119] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: An integrative review of the heart's anatomy and heart rate variability", *Frontiers in psychology*, vol. 5, p. 1040, 2014.

[120] F. Beissner, K. Meissner, K. J. Bär, and V. Napadow, "The autonomic brain: An activation likelihood estimation meta-analysis for central processing of autonomic function", *Journal of Neuroscience*, vol. 33, no. 25, pp. 10 503–10 511, 2013.

[121] P. M. Macey, J. A. Ogren, R. Kumar, and R. M. Harper, "Functional imaging of autonomic regulation: Methods and key findings", *Frontiers in neuroscience*, vol. 9, p. 513, 2016.

[122] C. R. Bellenger, R. L. Thomson, E. Y. Robertson, K. Davison, M. J. Nelson, L. Karavirta, and J. D. Buckley, "The effect of functional overreaching on parameters of autonomic heart rate regulation", *European journal of applied physiology*, vol. 117, no. 3, pp. 541–550, 2017.

[123] M. P. Tulppo, A. M. Kiviniemi, A. J. Hautala, M. Kallio, T. Seppänen, S. Tiinanen, T. H. Mäkikallio, and H. V. Huikuri, "Sympatho-vagal interaction in the recovery phase of exercise", *Clinical physiology and functional imaging*, vol. 31, no. 4, pp. 272–281, 2011.

[124] S. Schulz, M. Bolz, K.-J. Bär, and A. Voss, "Central-and autonomic nervous system coupling in schizophrenia", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2067, p. 20 150 178, 2016.

[125] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[126] S. Boettger, C. Puta, V. K. Yeragani, L. Donath, H. J. Mueller, H. H. Gabriel, and K.-J. Baer, "Heart rate variability, QT variability, and electrodermal activity during exercise", *Medicine and Science in Sports and Exercise*, vol. 42, no. 3, pp. 443–8, 2010.

[127] J. Achten and A. E. Jeukendrup, "Heart rate monitoring", *Sports medicine*, vol. 33, no. 7, pp. 517–538, 2003.

[128] P. H. Venables, "Autonomic activity.", *Annals of the New York Academy of Sciences*, 1991.

[129] P. N. Banerjee, D. Filippi, and W. A. Hauser, "The descriptive epidemiology of epilepsy - a review", *Epilepsy research*, vol. 85, no. 1, pp. 31–45, 2009.

[130] D. Buck, G. A. Baker, A. Jacoby, D. F. Smith, and D. W. Chadwick, "Patients' experiences of injury as a result of epilepsy", *Epilepsia*, vol. 38, no. 4, pp. 439–444, 1997.

[131] M. Jackson and D Turkington, "Depression and anxiety in epilepsy", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 1, pp. i45–i47, 2005.

[132] D. W. Loring, K. J. Meador, and G. P. Lee, "Determinants of quality of life in epilepsy", *Epilepsy & Behavior*, vol. 5, no. 6, pp. 976–980, 2004.

[133] G. Regalia, F. Onorati, M. Lai, C. Caborni, and R. W. Picard, "Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the Empatica wristbands", *Epilepsy research*, vol. 153, pp. 79–82, 2019.

[134] M. Z. Poh, T Loddenkemper, C Reinsberger, N. Swenson, S Goyal, J. Madsen, and R. W. Picard, "Autonomic changes with seizures correlate with postictal EEG suppression", *Neurology*, vol. 78, no. 23, pp. 1868–1876, 2012.

[135] S. Vieluf, R. El Atrache, S. Hammond, F. M. Touserkani, T. Loddenkemper, and C. Reinsberger, "Peripheral multimodal monitoring of ANS changes related to epilepsy", *Epilepsy & Behavior*, vol. 96, pp. 69–79, 2019.

[136] D. M. Goldenholz, A. Kuhn, A. Austermuehle, M. Bachler, C. Mayer, S. Wassertheurer, S. K. Inati, and W. H. Theodore, "Long-term monitoring of cardiorespiratory patterns in drug-resistant epilepsy", *Epilepsia*, vol. 58, no. 1, pp. 77–84, 2017.

[137] J. Pavei, R. G. Heinzen, B. Novakova, R. Walz, A. J. Serra, M. Reuber, A. Ponnusamy, and J. L. Marques, "Early seizure detection based on cardiac autonomic regulation dynamics", *Frontiers in physiology*, vol. 8, p. 765, 2017.

[138] D. K. Chatkoff, K. J. Maier, and C. Klein, "Nonlinear associations between chronic stress and cardiovascular reactivity and recovery", *International Journal of Psychophysiology*, vol. 77, no. 2, pp. 150–156, 2010.

[139] A. Dionisio, R. Menezes, and D. A. Mendes, "Mutual information: A measure of dependency for nonlinear time series", *Physica A: Statistical Mechanics and its Applications*, vol. 344, no. 1-2, pp. 326–329, 2004.

[140] M. Anderson, X.-L. Li, and T. Adalı, "Nonorthogonal independent vector analysis using multivariate Gaussian model", in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Springer, 2010, pp. 354–361.

[141] N. Altman and M. Krzywinski, "Association, correlation and causation", *Nature Methods*, vol. 12, no. 10, pp. 899–900, 2015.

[142] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[143] B. R. Rao, "Partial canonical correlations", *Trabajos de estadistica y de investigación operativa*, vol. 20, no. 2-3, pp. 211–219, 1969.

[144] T. Shibuya, T. Harada, and Y. Kuniyoshi, "Reliable index for measuring information flow", *Physical Review E*, vol. 84, 061109, no. 6, pp. 1–7, 2011.

[145] V. D. Calhoun, T. Adalı, 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, no. 3, pp. 140–151, 2001.