

UNDERSTANDING EPILEPSY SEIZURE STRUCTURE USING TENSOR ANALYSIS

By

Evrin Acar Ataman

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
Major Subject: Computer Science

Approved by the
Examining Committee:

Bülent Yener, Thesis Adviser

Rasmus Bro, Member

Kristin Bennett, Member

Petros Drineas, Member

Mukkai Krishnamoorthy, Member

Mohammed Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2008
(For Graduation May 2008)

© Copyright 2008
by
Evrin Acar Ataman
All Rights Reserved

CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	viii
ACKNOWLEDGMENT	xiii
ABSTRACT	xiv
1. INTRODUCTION	1
1.1 Epilepsy	1
1.2 Electroencephalogram (EEG)	2
1.2.1 EEG Acquisition	2
1.3 Our Contributions and Organization	5
2. RELATED WORK AND BACKGROUND	7
2.1 Definitions and Notations	8
2.1.1 Multiway Arrays (Tensors)	8
2.1.2 Rank-One Tensor	9
2.1.3 Tensor Frobenius Norm	10
2.1.4 Matricization	10
2.1.5 Tensor-Matrix Multiplication	10
2.1.6 Matrix Operations	12
2.2 Factor Models	12
2.3 Multiway Models	15
2.3.1 PARAFAC-family	16
2.3.1.1 PARAFAC	16
2.3.1.2 Extensions of PARAFAC	19
2.3.2 Tucker-family	21
2.3.2.1 Tucker3	21
2.3.2.2 Extensions of Tucker3	23
2.3.3 Tucker vs. SVD vs. PARAFAC	24
2.3.4 Other Models	26
2.4 Algorithms	27
2.4.1 Tucker Model	27

2.4.2	PARAFAC Model	30
2.4.3	Preprocessing	31
2.5	Applications	32
2.5.1	Chemistry	32
2.5.2	Social Network Analysis/Text-mining	34
2.5.3	Computer Vision	35
2.5.4	Process Analysis	36
2.6	Software	37
2.7	Summary	38
2.8	Discussions	39
3.	EPILEPTIC SEIZURE LOCALIZATION	41
3.1	Related Work	41
3.1.1	Multiway Models in Computational Neuroscience	43
3.1.2	Multiway Models in Epilepsy	44
3.2	Epilepsy Tensor Construction	45
3.3	Artifact Extraction, Removal and Seizure Localization	47
3.3.1	Dataset	48
3.3.2	Artifact Extraction	50
3.3.3	Seizure Origin Localization	51
3.3.4	Artifact Removal	53
3.3.5	Parameter Selection	57
3.4	Summary and Discussions	60
4.	EPILEPTIC SEIZURE RECOGNITION	62
4.1	Related Work	62
4.2	Epilepsy Feature Tensor Construction	64
4.2.1	Time domain features	65
4.2.1.1	Hjorth parameters	65
4.2.1.2	Mean Absolute Slope	66
4.2.1.3	Spatial Information	66
4.2.2	Frequency domain features	67
4.2.2.1	Frequency Spectrum	67
4.2.2.2	Relative Energy	68
4.2.2.3	Spectral Entropy	68

4.3	Multilinear Regression	69
4.3.1	Partial Least Squares	69
4.3.2	Multilinear Partial Least Squares	70
4.3.3	Feature Selection	72
4.4	Patient-Specific Seizure Recognition	74
4.4.1	Dataset	74
4.4.2	Results and Interpretations	77
4.5	Patient Non-Specific Seizure Recognition	81
4.5.1	Dataset	82
4.5.1.1	Additional Features	82
4.5.1.2	Handling inter-patient differences	83
4.5.2	Results and Interpretations	83
4.6	Parameter Selection	88
4.7	Summary and Discussions	89
5.	CONCLUSIONS AND FUTURE WORK	93
5.1	Epileptic Seizure Localization	93
5.2	Epileptic Seizure Recognition	94
5.3	Epileptic Seizure Prediction	96
	LITERATURE CITED	97

LIST OF TABLES

2.1	Selection of models from PARAFAC family	19
2.2	Selection of models from Tucker family	24
3.1	Dataset of multi-channel ictal EEG. 100 scales corresponding to frequencies in the frequency band of 0.5-50Hz are used.	49
3.2	Core Consistency for different component numbers	59
4.1	EEG Dataset for Patient-Specific Seizure Recognition. The last column shows the size of the epilepsy feature tensor with modes: time epochs, features and channels. Each tensor contains a seizure as well as data before and after that seizure. The number of epochs (first mode) in each tensor differs depending on the duration of a seizure.	76
4.2	Seizure vs. Non-seizure. Performance of three-way (NPLS-based) and two-way (SVM-based) approaches in terms of geometric mean of sensitivity and specificity of the model. The row corresponding to NPLS + LDA shows the results without feature selection while the row corresponding to NPLS + LDA (FS) demonstrates the results of the model with feature selection.	80
4.3	Pre-seizure vs. Post-seizure (binary classification within non-seizure epochs). Each entry shows the performance of the model when it is trained on non-seizure epochs before/after some seizures of a patient and tested on non-seizure epochs before/after another seizure of that particular patient.	80
4.4	Subsets of features used in the patient-specific seizure recognition model of each patient. Patient 1, 2, 7 and 8 have right temporal seizures. Patient 3 suffers from left frontal while Patient 4 and 9 suffer from left temporal seizures. Patient 5 is bilateral central frontal and Patient 6 is bilateral occipital. While subsets of features tend to be similar based on seizure origins, it is not possible to make generalizations on a small set of patients.	81
4.5	EEG Dataset for Patient Non-Specific Seizure Recognition. Each epilepsy feature tensor contains a seizure as well as data before and after that seizure. The number of epochs (first mode) in each tensor differs depending on the duration of a seizure.	84

4.6 Seizure vs. Non-seizure. Performance of three-way (NPLS-based) approach in terms of the geometric mean of sensitivity and specificity of the model. The different columns correspond to different preprocessing techniques explained in the text. 86

LIST OF FIGURES

1.1	The standard 10-20 international electrode placement system. The figure is taken from [118].	3
1.2	The EEG signal recorded at channel $T_6 - Cz$ during a seizure. The period between two vertical red lines is identified visually as the seizure period by the neurologists.	4
2.1	(A) Columns, (B) Rows, (C) Tubes. The figure is taken from [15]. . . .	9
2.2	(A) Horizontal Slices, (B) Vertical Slices, (C) Frontal Slices.	9
2.3	Matricization of a three-way array in the first mode. A three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is unfolded in the first mode and a matrix of size $I \times JK$, denoted by $\mathbf{X}_{(1)}$ is formed. The subscript i in $\mathbf{X}_{(i)}$ indicates the mode of matricization. The figures at the top and the bottom illustrate the matricization in the first mode as defined in [68] and [40], respectively. The definition in [68] is commonly used.	11
2.4	The categorization of multiway models briefly explained in this chapter. We study multiway models under three categories: PARAFAC family, Tucker family and alternative approaches.	16
2.5	Illustration of a PARAFAC model. A 2-component PARAFAC model, where a three-way array $\underline{\mathbf{X}}$ is expressed as the sum of two rank-one tensors. \mathbf{a}_i , \mathbf{b}_i and \mathbf{c}_i are the i^{th} components in the first, second and third modes, respectively. $\underline{\mathbf{E}}$ is a three-way array containing the residual terms.	17
2.6	(A) A full core array of size $P \times Q \times R$ with maximum PQR nonzero elements, (B) A super-diagonal core array of size $R \times R \times R$ with maximum R nonzero elements.	18
2.7	Illustration of a Tucker3 model. A (P, Q, R) -component Tucker3 model, where a three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is modeled with component matrices $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$ in the first, second and third modes, respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ contains the error terms.	22
2.8	Modeling of a fluorescence dataset using a 3-component PARAFAC model. \mathbf{a}_i , \mathbf{b}_i and \mathbf{c}_i correspond to the i^{th} component in samples, emission and excitation modes. We also illustrate the vector outer product of \mathbf{b}_i and \mathbf{c}_i , which shows the fluorescence landscape of each analyte used in the preparation of the samples.	33

3.1	Two-way Data Construction.	45
3.2	Continuous wavelet transformation (CWT) of a signal from a single electrode (data corresponding to the single column of the matrix on the left) forms the frontal slice corresponding to a particular electrode. When the signals from all electrodes are represented in both time and frequency domains through CWT, a third-order tensor with modes time samples, scales (frequency) and electrodes is constructed.	46
3.3	Epilepsy Tensor. $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ represents the multi-channel ictal EEG data transformed by continuous wavelet transformation using a Mexican-hat wavelet and arranged as a three-way array. Each entry of $\underline{\mathbf{X}}$, x_{ijk} , corresponds to the square of the absolute value of a wavelet coefficient at i^{th} time sample, j^{th} scale and k^{th} electrode.	47
3.4	A 2-component PARAFAC model on an epilepsy tensor $\underline{\mathbf{X}}$ for a particular seizure. We demonstrate the modeling of an epilepsy tensor by a 2-component PARAFAC model, where the first component corresponds to an eye-artifact while the second component represents a seizure. Top: Temporal (\mathbf{a}_1), spectral (\mathbf{b}_1) and spatial (\mathbf{c}_1) signatures of an eye-artifact. \mathbf{a}_1 represents the coefficients of time samples, \mathbf{b}_1 represents the coefficients of scales. Since there is a peak in higher scales on the plot of \mathbf{b}_1 , it indicates that this artifact takes place at lower frequencies. \mathbf{c}_1 contains the coefficients of electrodes. These coefficients are demonstrated on a colormap using EEGLab [42]. Bottom: Temporal (\mathbf{a}_2), spectral (\mathbf{b}_2) and spatial (\mathbf{c}_2) signatures of a seizure. Similar to the first component, \mathbf{a}_2 represents the coefficients of time samples, \mathbf{b}_2 represents the coefficients of scales. There is a peak in lower scales on the figure corresponding to \mathbf{b}_2 , which indicates that the seizure takes place at higher frequencies. Finally, \mathbf{c}_2 contains the coefficients of electrodes, which are used to localize the seizure around electrodes T_4 and T_6	52
3.5	Multiway analysis of multi-channel ictal EEG. After the collection of multi-channel EEG data from epilepsy patients, we normalize the data and construct a three-way $\underline{\mathbf{X}}$ called an <i>Epilepsy Tensor</i> through wavelet transformation. $\underline{\mathbf{X}}$ is then downsampled and scaled in scales mode before multiway analysis. Preprocessed three-way array is modeled using a PARAFAC model for artifact extraction and localization of epileptic focus. Finally, PARAFAC components are compared with clinical findings of epilepsy patients. For the cases when a PARAFAC model cannot capture the seizure, we apply an artifact removal method by modeling the preprocessed three-way using a Tucker3 model to detect potential artifacts. The tensor formed after artifact removal is modeled using a PARAFAC model to extract the signatures of an artifact and a seizure.	55

3.6	The spatial signatures extracted by a PARAFAC model after applying different artifact removal approaches on seizure 6 from Table 3.1. Method I refers to using Tucker3 as the artifact removal approach; method II refers to the artifact removal process based on Tucker1 and method III removes artifacts on raw data using SVD. For method I and II, even though the spatial signatures are not exactly the same, they are very similar for both models and the seizure localization is the same. The first component in method I and the second component in method II localizes the seizure (the components are given in the order of the variation they explain but the explained variation by each component is so close that the ordering flips in different methods). The spatial signatures captured by the PARAFAC model after using SVD-based artifact removal approach on raw data cannot localize the seizure. Color scales in the figures are the same as the scale in Figure 3.4.	57
3.7	The spatial signatures extracted by a PARAFAC model after applying different artifact removal approaches on seizure 7 from Table 3.1. Method I refers to using Tucker3 as the artifact removal approach; method II refers to the artifact removal process based on Tucker1 and method III removes artifacts on raw data using SVD. The second component in method I localizes the seizure origin whereas no other component extracted by other methods can localize the seizure. Color scales in the figures are the same as the scale in Figure 3.4.	58
3.8	Some illustrative examples of artifact extraction and seizure origin localization. We present our results corresponding to the electrode mode for four of the seizures when they are modeled using a PARAFAC model. Color scales in the figures are the same as the scale in Figure 3.4. \mathbf{c}_1 , \mathbf{c}_2 and \mathbf{c}_3 stand for the first, second and third components in the electrodes mode, respectively. (1) Seizure 1. First component represents an eye-artifact while the second component localizes the seizure. (2) Seizure 8. First component shows the seizure origin and the second component corresponds to an artifact, which has a low frequency signature. The third component cannot be visually identified. (3) Seizure 7. This is one of the examples where artifact removal is applied. The components are the PARAFAC components extracted after artifact removal. While the first and third components are the artifacts, the second component represents the seizure. (4) Seizure 10. The first component localizes the seizure around F_7 and C_3 while the second component corresponds to an artifact.	61
4.1	(A) Multiple features extracted from single-channel EEG data are arranged as a matrix, (B) When multiple channels are taken into consideration, the data form a three-way array with modes: <i>time epochs</i> , <i>features</i> , <i>channels</i>	63

4.2	Epilepsy Feature Tensor. $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ represents the multi-channel EEG data, which are transformed into the feature space by computing certain measures characterizing seizure dynamics. Each entry of $\underline{\mathbf{X}}$, x_{ijk} , corresponds to the value of j^{th} feature of i^{th} time epoch at k^{th} channel.	65
4.3	Mean Absolute Slope of epochs from all channels for the fifth seizure of the second patient in Table 4.1. Epochs marked with blue and red belong to non-seizure and seizure periods, respectively. Green epochs are the transition epochs from pre-seizure to seizure or seizure to post-seizure periods.	66
4.4	Spatial Information of epochs from all channels for the second seizure of the second patient in Table 4.1. Epochs marked with blue and red belong to non-seizure and seizure periods, respectively. We observe a clear increase in similarity between neighboring channels during a seizure period. Green epochs are the transition epochs from pre-seizure to seizure or seizure to post-seizure periods.	67
4.5	Patient-Specific Seizure Recognition Model. Multi-channel EEG signals corresponding to the data before, during and after each seizure of a patient are arranged as a third-order epilepsy feature tensor. Then training and test sets are constructed by leaving out one seizure (together with data before and after that seizure period) at a time. The model built on the training set is used to predict the labels of the time epochs in the test set using N-PLS and LDA. Final step is performance evaluation using the average performance of the model on test sets. . . .	79
4.6	The figures show how one of the features, i.e., activity, behaves during seizures of some patients from Table 4.5. The epochs represented in blue and red correspond to the non-seizure and seizure periods while the epochs in green contain time samples from both seizure and non-seizure periods. We observe that there is significant order of magnitude difference between seizures of different patients.	85

4.7 Performance of the patient non-specific seizure recognition model with different preprocessing approaches for the fourth patient in Table 4.5. Blue and red dots correspond to true and predicted classes of time epochs, respectively. The figure at the top shows that we can only partially detect the seizure; in other words we can only classify the epochs at the end of the seizure as seizure when we preprocess the test set according to the whole training set containing seizures of different patients (the second column of Table 4.6). On the other hand, if we preprocess each patient separately, then we observe that we can detect almost the whole seizure period with a short delay at the beginning of the seizure (the third column of Table 4.6). 87

4.8 The figure illustrates how differently mean absolute slope behaves in pre-seizure, seizure and post-seizure periods for one of the seizures of Patient 1 from Table 4.1. This feature, therefore, can differentiate between seizure and non-seizure as well as pre-seizure and post-seizure epochs. The epochs represented in blue and red correspond to the non-seizure and seizure periods while the epochs in green contain time samples from both seizure and non-seizure periods. 91

ACKNOWLEDGMENT

I would like to thank my thesis adviser Prof. Bülent Yener for introducing me to a very exciting research area. Even though I started my Ph.D. in a completely different field in 2003, through his visionary ideas on multi-modal analysis of chatroom communications data, I had the opportunity to be introduced to the field of multi-way data analysis. I would like to also thank him for believing in me and supporting me through my studies in RPI.

I would like to express my deepest gratitude to Prof. Rasmus Bro for giving me the chance to work with him in the Chemometrics Group in University of Copenhagen in Summer of 2006, which was the start of a great research experience. Although we have communicated remotely, he has been involved in every step of my thesis through his guidance in data analysis. He has also initiated motivating projects, which gave me the opportunity to work with researchers worldwide. Many thanks for always supporting me with your positive attitude. I would like to thank both Rasmus Bro and Tamara Kolda for not only answering my endless questions but also helping me to know the multiway community better and giving me the opportunity to improve myself in the field.

Special thanks to Prof. Petros Drineas, Prof. Kristin Bennett, Prof. Mohammed Zaki and Prof. Mukkai Krishnamoorthy. I would like to acknowledge Prof. Canan A. Bingöl and Prof. Anthony L. Ritaccio for providing us epileptic EEG signals and for their guidance on epilepsy and EEG.

Finally, I would like to thank my officemates, particularly special thanks to Fikret Sivrikaya and Seyit A. Çamtepe, who have answered so many of my technical questions and made my life much easier in Troy.

ABSTRACT

Data in many disciplines are arranged as two-way datasets; in other words, matrices. However, matrices may not be enough to fully represent the information content of the data and two-way analysis techniques may fail in terms of capturing and interpreting the underlying structure in a dataset. Tensors, on the other hand, represent datasets by preserving their multi-modal structures and tensor decomposition methods, which are mostly based on generalizations of two-way factor models to higher-order datasets, can extract the true underlying structures of the data. In this thesis, we introduce mathematical models based on multi-modal data construction and analysis with a goal of understanding epilepsy seizure dynamics and developing automated and objective approaches for the analysis of large amounts of scalp electroencephalogram (EEG) data.

In the first part of this study, we address the problem of identification of a seizure origin through an analysis of ictal EEG, which is proven to be an effective standard in epileptic focus localization. We rearrange multi-channel ictal EEG data as a third-order *Epilepsy Tensor* with modes: time samples, scales and channels, through continuous wavelet transform. Then we demonstrate that multiway analysis techniques, in particular Parallel Factor Analysis, can successfully model the complex structure of an epilepsy seizure, localize an epileptic seizure origin and extract artifacts. Furthermore, we introduce an approach for removing artifacts using multilinear subspace analysis.

In the second part, we focus on seizure recognition and aim to automatically differentiate between seizure and non-seizure periods. We represent multi-channel EEG data using a set of features. These features expected to have distinct trends during seizure and non-seizure periods include features from both time and frequency domains. First, we rearrange multi-channel EEG signals as a third-order tensor called an *Epilepsy Feature Tensor* with modes: time epochs, features and channels. Second, we model the epilepsy feature tensor using a multilinear regression model, i.e., Multilinear Partial Least Squares, which is the generalization of Partial Least

Squares regression to higher-order datasets. This two-step approach facilitates EEG data analysis from multiple channels represented by several features from different domains. We develop patient-specific as well as patient non-specific seizure detection models and obtain promising performance in both approaches.

In summary, this thesis demonstrates how multi-channel epileptic EEG signals can be rearranged as multi-modal datasets and how tensor decompositions can be used to mark the seizure period or localize the seizure origin. Nevertheless, in order for these methodologies to be clinically applicable, the performance of the proposed techniques should be tested and enhanced on large datasets containing heterogeneous epileptic patterns and patients.

CHAPTER 1

INTRODUCTION

1.1 Epilepsy

Epilepsy is a neurological disorder affecting the nervous system and characterized by recurrent seizures. While some possible causes of seizures are brain injuries, family history with epileptic seizures, cysts, tumors or infections such as meningitis, the underlying reasons for recurrent seizures may not be even known in some cases. According to World Health Organization's records, there are over 50 million people suffering from epilepsy worldwide and 2.4 million new cases occur each year [146]. Therefore, it is known to be the most common and serious brain disorder worldwide.

Fortunately, there are treatment options mainly based on antiepileptic drugs, which can prevent sixty to eighty percent of patients having seizures [45]. However, in some cases the use of medications may be inadequate or the treatments cannot succeed in stopping the seizures completely. In such cases, brain surgeries may be performed in order to help patients have seizure-free lives. During an epilepsy surgery, the goal is to remove the brain tissues or tumors causing the seizures identified through intensive clinical studies. Therefore, it is extremely important to localize the region of the brain causing the seizures precisely. The cortical region responsible for generating seizures is called the *seizure onset zone*, a terminology we will be using frequently throughout this study.

Among the diagnosis techniques used in clinical evaluations, e.g., physical and neurological exams, magnetic resonance imaging (MRI) scans, ictal single photon emission computed tomography (SPECT) and video-EEG monitoring, monitoring and analysis of EEG signals is the "gold standard" used to detect seizures and identify seizure onset zones [103]. For instance, even though a tumor identified by MRI can be a possible cause of epileptic seizures, it is not necessarily the seizure onset zone or not all epilepsy patients have tumors. Consequently, analysis of EEG signals is essential for localizing the epileptic focus with confidence.

1.2 Electroencephalogram (EEG)

EEG is a measure of the electrical activity in the brain. It is recorded using electrodes with high conductance placed at particular locations of the head. One approach is to collect the measurements by placing electrodes on the surface of the head, which produces the recordings known as *surface* or *scalp EEG*. Another technique is to place electrodes within the scalp through a surgery and collect recordings using these depth electrodes. The recordings collected using these electrodes are called *intracranial EEG recordings (iEEG)*. iEEG provides more precise localization of an epileptogenic focus compared to scalp EEG recordings. In addition to better spatial resolution, other advantages of iEEG are increased sensitivity and being artifact-free, e.g., the absence of muscle artifacts. On the other hand, the placement of depth electrodes requires a surgery and it is an invasive process with possible complications. Since there are other non-invasive clinical evaluation tools, monitoring intracranial recordings may no longer be needed in all patients. Recently, types of epilepsy patients, who would benefit most from pre-surgical evaluations based on iEEG, have been reported in [116].

Considering the disadvantages of iEEG monitoring, it is important to understand the role of scalp EEG recordings better and learn as much as we can using only the surface electrodes. Scalp EEG recordings have proved to be useful in the literature, e.g., [138] and their localizing value and the necessity in the decision making process before epilepsy surgeries have been reviewed in [106]. Compared to iEEG, scalp EEG recordings have some limitations such as low spatial resolution and being contaminated with artifacts. However, research in this field, e.g., [4, 38, 41, 128, 144], has been focusing on improving the analysis of scalp EEG recordings using computational approaches based on artifact removal methods and source localization techniques that would identify the seizure location more precisely.

1.2.1 EEG Acquisition

In this study, we use scalp EEG recordings collected from epilepsy patients treated at different epilepsy centers: Marmara Epilepsy Center (Istanbul, Turkey), Yeditepe Epilepsy Center (Istanbul, Turkey) and Albany Medical College (New

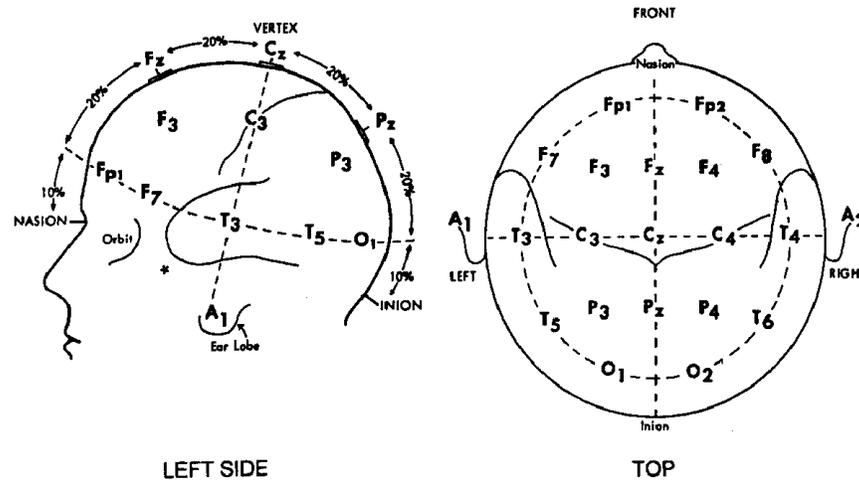


Figure 1.1: The standard 10-20 international electrode placement system. The figure is taken from [118].

York, US). We focus on scalp EEG recordings corresponding to the ictal period for developing models to localize epileptic seizure origins and collect ictal data from 7 patients (discussed in detail in Chapter 3). In the second part of the study for seizure recognition, we use data from not only ictal period but also pre-ictal and post-ictal periods. Our criterion for patient selection for developing models for patient-specific seizure recognition is that a patient should have at least three recorded seizures. We analyze scalp EEG recordings of 9 epilepsy patients and mark their seizures using a patient-specific seizure recognition model. Finally, we construct a dataset containing 9 epilepsy patients suffering from right or left temporal seizures in order to build and test our patient non-specific seizure recognition model. In total, we have collected scalp EEG recordings from 17 epilepsy patients.

The electrode placement (montage) follows the standard international 10-20 electrode placement system given in Figure 1.1. This standard provides the basis for electrode placement according to an international standard. The data from Marmara Epilepsy Center are recorded using 17 electrodes (Unused electrodes: A_1 , A_2 , P_3) while the data from Yeditepe Epilepsy Center and Albany Medical College have recordings from 18 electrodes (Unused electrodes: A_1 , A_2). The montage chosen for computational analysis is the referential montage with Cz reference. Figure 1.2

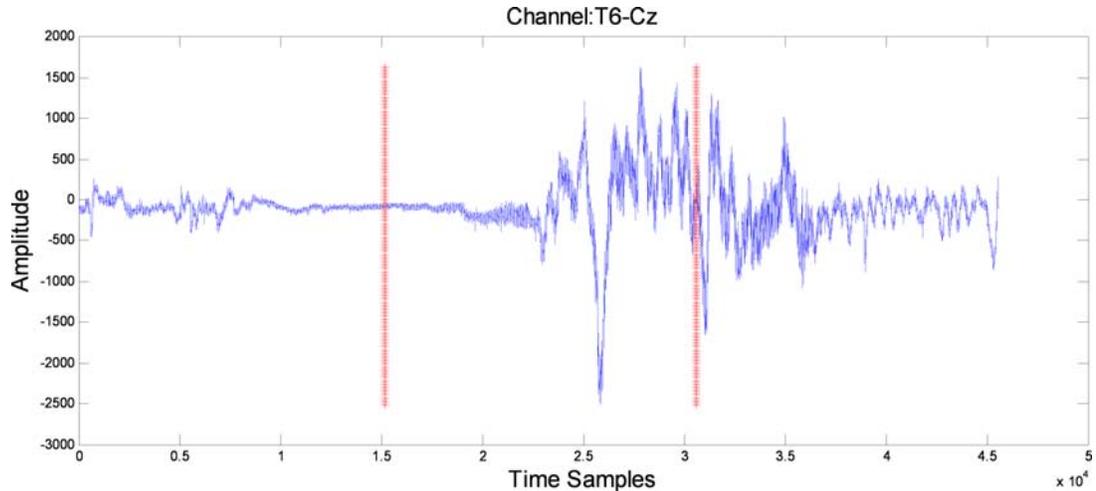


Figure 1.2: The EEG signal recorded at channel $T_6 - Cz$ during a seizure. The period between two vertical red lines is identified visually as the seizure period by the neurologists.

shows an EEG signal recorded at a single channel, i.e., $T_6 - Cz$, during a seizure of one of the patients used in this study. The period in-between vertical red lines is the seizure period. Such recordings are collected from all channels and we obtain multi-channel EEG data for each seizure of a patient. As already mentioned above, scalp EEG recordings have low spatial resolution and in order to obtain more precise localization, extra electrodes may be utilized. Optimal number of electrodes or the distances between electrodes are not exactly known and different studies suggest different numbers and placement criteria ([106] and references therein).

Epilepsy is classified into different categories based on properties like types of seizures, the causes of seizures, the part of the brain involved in the seizures or frequency of the seizures, etc. While the final classification is far from complete, seizures are often classified into two types [84]: generalized seizures and partial (focal) seizures. The main distinction between these two types of seizures is how the brain or the parts of the brain are involved in epileptic activities. Generalized seizures start with electrical discharges affecting the whole brain. Partial seizures, on the other hand, affect only a portion of the brain and epilepsy is focused in a particular region of the brain. All the patients we have included in our study suffer from partial seizures.

1.3 Our Contributions and Organization

Multiple days of multi-channel EEG recordings are often visually reviewed by EEG technologists and neurologists in their entirety to annotate epileptiform abnormalities and ictal events with a goal of understanding *when* and *where* the seizures start. However, there are mainly two disadvantages of visual analysis of EEG signals. First one is that it is a massive consumption of man-hours and the second one is being highly subjective and error-prone. Therefore, automation of the detection of the underlying brain dynamics in EEG signals from epilepsy patients is significant in order to obtain fast, robust and objective EEG analysis.

In this thesis, we address two important problems in epilepsy treatment based on the analysis of multi-channel scalp EEG recordings. The first one is epileptic focus localization and the second one is epileptic seizure recognition. Our goal is to develop automated and objective approaches for seizure localization and recognition in order to save manpower spent on visual analysis of massive amounts of EEG data as well as to remove the subjectivity in the visual analysis process. Tensors and tensor decomposition methods form the underlying principles we employ in constructing our multi-modal datasets from multi-channel EEG data and building our mathematical models. Therefore, we provide a review of these concepts together with some basics about multiway data analysis in Chapter 2 [6]. Our contributions are as follows:

- **Epileptic Focus Localization:** We propose a novel approach based on multiway models to study epilepsy seizure structure. We construct an *Epilepsy Tensor* with three modes, i.e., time samples, scales and electrodes, through wavelet analysis of multi-channel ictal EEG [2, 4]. We then demonstrate that multiway analysis techniques, in particular Parallel Factor Analysis (PARAFAC), provide promising results in modeling the complex structure of an epilepsy seizure, localizing a seizure origin and extracting artifacts. Furthermore, we introduce an approach for removing artifacts using multilinear subspace analysis. Seizure localization, artifact extraction and removal are discussed in detail in Chapter 3.

- **Epileptic Seizure Recognition:** With a goal of differentiating between seizure and non-seizure periods, we extract various features from both time and frequency domains to represent scalp EEG recordings. We rearrange multi-channel EEG recordings as a third-order tensor called an *Epilepsy Feature Tensor* with modes: time epochs, features and channels [5, 8]. We then model the epilepsy feature tensor using a multilinear discriminant analysis based on Multilinear Partial Least Squares (N-PLS), which is the generalization of Partial Least Squares regression to tensors. This two-step approach facilitates the analysis of EEG data from multiple channels represented by several features from different domains. In Chapter 4, we build a supervised seizure recognition model, which is trained on some seizures of a patient and then tested on other seizures of that particular patient (*patient-specific*) or trained and tested on seizures of different patients (*patient non-specific*). We demonstrate that multi-modal data construction and analysis approach provides promising performance in terms of marking the seizure period automatically.

CHAPTER 2

RELATED WORK AND BACKGROUND

Mathematical models we develop for the analysis of ictal EEG patterns, epileptic seizure localization and recognition are based on the construction of multi-modal datasets and the analysis of those datasets using multiway analysis techniques. In this chapter, we first introduce some definitions from linear and multilinear algebra as well as the notation for multiway datasets that will be used throughout this study. We then review significant contributions in the literature on multiway models, algorithms as well as their applications in diverse disciplines including chemometrics, social network analysis, text mining and computer vision. We study the applications of multiway analysis techniques in computational neuroscience in depth in the next chapters.

Multiway data analysis, dating back to 1920s to the studies of tensor decompositions by Hitchcock [59,60], is the extension of two-way data analysis to higher-order datasets. Multiway analysis is often used for extracting hidden structures and capturing underlying correlations between variables in a multiway array. For example, multiway analysis of multi-channel EEG data enables us to capture the correlation between the channels by representing the signals both in time and frequency domains. Multi-channel EEG recordings are commonly represented as an $I \times J$ matrix containing signals recorded for I time samples at J channels. In order to capture the underlying brain dynamics, often frequency content of the signals, for instance signal power at K particular frequencies, also needs to be considered. In that case, EEG data can be arranged as an $I \times J \times K$ three-way dataset [90]. Multiway analysis of a three-way EEG array can then be used to extract the signatures of brain dynamics in time, frequency and electrode domains.

It has been shown in numerous research areas including social networks [1], neuroscience [46], process analysis [52] and text-mining [34] that underlying information content of the data may not be captured accurately or identified uniquely by two-way analysis methods. Two-way analysis methods, by which we refer to

those based on factor models here, suffer from rotational freedom unless specific constraints such as statistical independence or orthogonality are enforced. On the other hand, these constraints requiring prior knowledge or unrealistic assumptions are not often necessary for multiway models such as Parallel Factor Analysis (PARAFAC) [56] since multiway models may be uniquely defined such that there is no alternative solution, which fits the data exactly the same as the fitted model. For example, in fluorescence spectroscopy, PARAFAC can uniquely identify the pure spectra of chemicals from measurements of mixtures of chemicals. Consequently, multiway analysis with advantages over two-way analysis in terms of uniqueness as well as robustness to noise and ease of interpretation has been a popular exploratory analysis tool in a variety of application areas, which we discuss throughout this chapter.

2.1 Definitions and Notations

Before introducing multilinear models, we introduce the terminology for multiway arrays as well as some definitions from linear and multilinear algebra that will be used in the rest of the study.

2.1.1 Multiway Arrays (Tensors)

Multiway arrays, often referred to as tensors, are higher-order generalizations of vectors and matrices. Higher-order arrays are represented as $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$, where the order of $\underline{\mathbf{X}}$ is N ($N > 2$) while a vector and a matrix are arrays of order 1 and 2, respectively. Higher-order arrays have a different terminology compared to two-way datasets. Each dimension of a multiway array is called a *mode* introduced in [124,125] (or *way*) and the number of variables in each mode is used to indicate the dimensionality of a mode. For instance, $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is a multiway array with N modes (called an N -way array or an N^{th} -order tensor) with I_1, I_2, \dots, I_N dimensions in the first, second, ..., N^{th} mode, respectively. Each entry of $\underline{\mathbf{X}}$ is denoted by $x_{i_1 i_2 \dots i_N}$. For a special case, where $N = 3$, let $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a three-way array. Then $x_{i_1 i_2 i_3}$ denotes the entry in the i_1^{th} row, i_2^{th} column and i_3^{th} tube of $\underline{\mathbf{X}}$ (Figure 2.1). When an index is fixed in one of the modes and the indices vary in the two other modes,

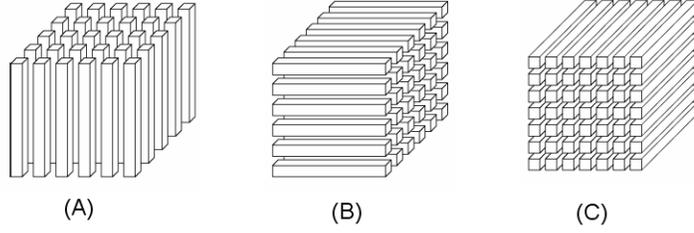


Figure 2.1: (A) Columns, (B) Rows, (C) Tubes. The figure is taken from [15].

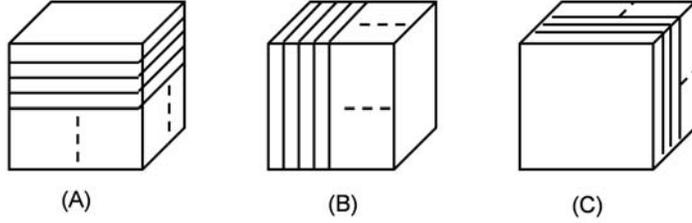


Figure 2.2: (A) Horizontal Slices, (B) Vertical Slices, (C) Frontal Slices.

this data partition is called a *slice (or a slab)*. For example, when the i_1^{th} row of $\underline{\mathbf{X}}$ is fixed, then it is a horizontal slice of size $I_2 \times I_3$ (Figure 2.2-A) or similarly, if the i_2^{th} column of $\underline{\mathbf{X}}$ is fixed, it is a vertical slice of size $I_1 \times I_3$, etc. (Figure 2.2-B).

We denote higher-order arrays using underlined boldface letters, e.g., $\underline{\mathbf{X}}$, following the standard notation in [68]. Matrices and vectors are represented by boldface capital, e.g., \mathbf{X} , and boldface lowercase letters, e.g., \mathbf{x} , respectively. Scalars are denoted by lowercase or uppercase italic letters, e.g., x or X . Matrix and tensor entries are represented by lowercase letters with subscripts, e.g., x_{ij} or x_{ijk} .

2.1.2 Rank-One Tensor

An N^{th} -order rank-one tensor is a tensor that can be written as the outer product of N vectors. Vector outer product is defined as follows. Let \mathbf{a} , \mathbf{b} and \mathbf{c} be column vectors of size $I \times 1$ and $J \times 1$ and $K \times 1$ and $\underline{\mathbf{Y}}$ is a tensor of size $I \times J \times K$, then

$$\underline{\mathbf{Y}} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \text{ if and only if } y_{ijk} = a_i b_j c_k \quad (2.1)$$

We will see later in this chapter that a PARAFAC model is based on the

representation of a higher-order data as a sum of rank-one tensors. In Chapter 3 we model multi-channel EEG data represented in both time and frequency domains using a PARAFAC model as a sum of rank-one tensors and use each vector forming each one of the rank-one tensors as signatures of artifacts or seizures.

2.1.3 Tensor Frobenius Norm

The Frobenius norm of a matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ denoted by $\|\mathbf{X}\|_F$ is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^I \sum_{j=1}^J x_{ij}^2}$. Similar to the matrix Frobenius norm, the Frobenius norm of a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is also defined as in Equation 2.2.

$$\|\underline{\mathbf{X}}\| = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \dots \sum_{i_N=1}^{I_N} x_{i_1 i_2 \dots i_N}^2} \quad (2.2)$$

2.1.4 Matricization

Matricization (or unfolding, flattening) means transforming a third or higher-order array into a two-way dataset. Let $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ be a three-way array. Then matricization of $\underline{\mathbf{X}}$ in the i^{th} mode is denoted by $\mathbf{X}_{(i)}$. It has multiple definitions in the literature [40, 68], e.g., two different definitions of unfolding in the first mode are illustrated in Figure 2.3.

Matricization is commonly used in our study. Apart from its use in the implementation of PARAFAC, Tucker3 and Multilinear PLS (N-PLS) algorithms, we often employ matricization (the version given in [68]) in order to perform operations on tensors. For instance, we unfold the epilepsy tensor with modes: time samples, scales and electrodes, in the electrodes mode in order to remove artifacts in Chapter 3. In Chapter 4, epilepsy feature tensor is unfolded in order to compare the proposed multiway approach with a two-analysis technique.

2.1.5 Tensor-Matrix Multiplication

Matrix multiplication is generalized to tensor-matrix multiplication through matricization of tensors. An N^{th} -order tensor, $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_n \times \dots \times I_N}$, can be multiplied by a matrix, $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, by matricizing the tensor in the n^{th} mode and computing the matrix product $\mathbf{Y} = \mathbf{U}\mathbf{X}_{(n)}$. The matrix product, \mathbf{Y} , is then re-

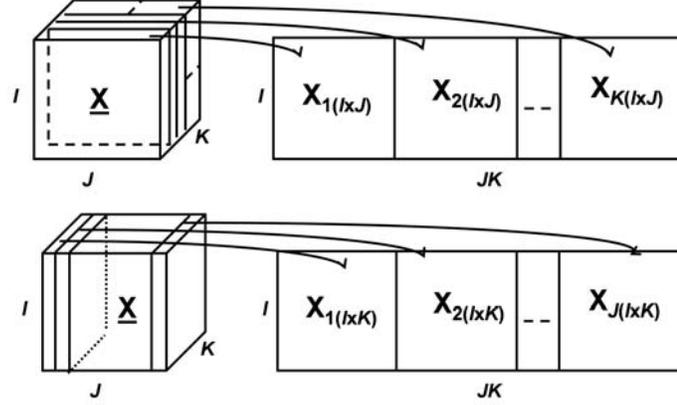


Figure 2.3: Matricization of a three-way array in the first mode. A three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is unfolded in the first mode and a matrix of size $I \times JK$, denoted by $\mathbf{X}_{(1)}$ is formed. The subscript i in $\mathbf{X}_{(i)}$ indicates the mode of matricization. The figures at the top and the bottom illustrate the matricization in the first mode as defined in [68] and [40], respectively. The definition in [68] is commonly used.

shaped as a tensor of size $I_1 \times I_2 \dots \times I_{n-1} \times J_n \times I_{n+1} \dots \times I_N$. More formally, the n -mode product of a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \dots \times I_n \times \dots \times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$ is denoted by $\underline{\mathbf{X}} \times_n \mathbf{U}$, and defined as in Equation 2.3 [40]:

$$(\underline{\mathbf{X}} \times_n \mathbf{U})_{i_1 i_2 \dots i_{n-1} j_n i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 \dots i_{n-1} i_n i_{n+1} \dots i_N} u_{j_n i_n} \quad (2.3)$$

where $x_{i_1 i_2 \dots i_N}$ and $u_{j_n i_n}$ represent the entries of an N^{th} -order tensor and a matrix, respectively. For more operations on tensors such as tensor–tensor, tensor–vector multiplications, the reader is referred to [15]. We use the tensor–matrix product every time we unfold the tensor and compute a matrix–matrix product. For instance, in the definition of the artifact removal process in Chapter 3, we unfold the tensor in the third mode and then compute the matrix–matrix product of the unfolded tensor with the projection matrix. In the computation of the approach proposed for feature selection, tensor–matrix product in the features mode is used. The notation for tensor–matrix product is \times_i but we often denote this operation clearly by unfolding the tensor in the i^{th} mode and computing the matrix–matrix product.

2.1.6 Matrix Operations

The Kronecker product of two matrices, $\mathbf{M} \in \mathbb{R}^{I \times J}$ and $\mathbf{N} \in \mathbb{R}^{K \times L}$, is denoted by $\mathbf{M} \otimes \mathbf{N}$ and defined as in Equation 2.4.

$$\mathbf{M} \otimes \mathbf{N} = \begin{bmatrix} m_{11}\mathbf{N} & m_{12}\mathbf{N} & \dots & m_{1J}\mathbf{N} \\ m_{21}\mathbf{N} & m_{22}\mathbf{N} & \dots & m_{2J}\mathbf{N} \\ \dots & \dots & \dots & \dots \\ m_{I1}\mathbf{N} & m_{I2}\mathbf{N} & \dots & m_{IJ}\mathbf{N} \end{bmatrix} \quad (2.4)$$

The columnwise Kronecker product of matrices is called the Khatri-Rao product, which is defined as follows: Let \mathbf{m}_k and \mathbf{n}_k represent the k^{th} column of $\mathbf{M} \in \mathbb{R}^{I \times K}$ and $\mathbf{N} \in \mathbb{R}^{J \times K}$, respectively for $k = 1, 2, \dots, K$. Then Khatri-Rao product denoted by $\mathbf{M} \odot \mathbf{N}$ is given as $\mathbf{M} \odot \mathbf{N} = [\mathbf{m}_1 \otimes \mathbf{n}_1 \ \mathbf{m}_2 \otimes \mathbf{n}_2 \ \dots \ \mathbf{m}_K \otimes \mathbf{n}_K]$. Some properties of these matrix operations are listed in [115] for interested readers. These matrix operations are important to understand the structural formulas and the implementations of the algorithms for PARAFAC and Tucker in this section and for N-PLS in Chapter 4.

2.2 Factor Models

In general, multiway data analysis methods are extensions of two-way analysis techniques based on the idea of linear factor models. In this section, we briefly introduce factor models in the context of bilinear factor models and in the next section, we introduce the formulations of multilinear models. Let matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ represent the original dataset. We can model \mathbf{X} using an R -component bilinear model as in Equation 2.5.

$$x_{ij} = \sum_{r=1}^R a_{ir}b_{jr} + e_{ij} \quad (2.5)$$

where $\sum_{r=1}^R a_{ir}b_{jr}$ is the structural part of the model consisting of matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$. Each column of \mathbf{B} corresponds to a factor and each row of \mathbf{A} contains the scores corresponding to R factors. Matrix $\mathbf{E} \in \mathbb{R}^{I \times J}$ contains the residuals. Analysis of residuals quantifies how well a model fits the data. More

formally, the sum of squares of residuals accounts for the unexplained variation in a least squares sense. Model fit is then defined as the ratio of explained variation in the structural part to the total variation in data. Using bilinear or multilinear models, *factors (or components, loadings)*, which are linear combinations of variables, are extracted. Since the extracted factors summarize the data, they are often used to interpret the underlying information content of the data. For example, in two-way EEG analysis, where EEG data recorded at different time samples at various channels are analyzed, R factors in the time mode can be considered as the signatures of R different brain activities. The scores in the channels mode, then, represent the contribution of an activity, e.g. an eye blink, a muscle artifact or an epileptic seizure, etc., to the signal recorded at a particular channel. Similarly, in three-way analysis of EEG data arranged as a third-order tensor with modes: time samples, frequency, channels, R factors in the time and frequency modes are considered as the temporal and spectral signatures of R different brain activities. The scores in the channels mode are then used to interpret the contributions of these activities with the extracted temporal and spectral signatures to the signal recorded at a particular channel.

By imposing orthogonality constraints on the factors, we can reformulate a bilinear factor model as the Singular Value Decomposition (SVD) [51]. For matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$, the singular value decomposition theorem states that there exist orthogonal matrices $\mathbf{U} \in \mathbb{R}^{I \times I}$ and $\mathbf{V} \in \mathbb{R}^{J \times J}$ such that

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T \quad (2.6)$$

where Σ is a diagonal matrix with $\sigma_1, \sigma_2, \dots, \sigma_R$ on the diagonal and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R$, $R = \min(I, J)$. The columns of matrices \mathbf{U} and \mathbf{V} are the *left* and *right singular vectors*, respectively and the diagonal entries of Σ are the *singular values*. SVD is commonly applied in many disciplines as a rank reduction or a noise removal method since the truncated form of SVD, where only first K ($K < R$) singular values and vectors are used, gives the best rank- K approximation of the data.

As in two-way factor models, tensor analysis aims to explore the relationships between the variables used to represent the data and find a summarization of the

data. While most multiway analysis techniques preserve the multiway nature of the data, some techniques such as Tucker1 [76] are based on matricization of a multiway array. Once a three-way array is flattened and arranged as a two-way dataset, two-way analysis methods, e.g., SVD and other factor models, can be employed in understanding the structure in data. Rearranging multiway arrays as two-way datasets and analyzing them with two-way methods, though, may result in information loss and misinterpretation especially if the data are noisy. For instance, we represent multi-channel EEG data in both time and frequency domains using a third-order tensor with modes: time samples, frequency and electrodes. When we model this tensor using a PARAFAC model, we assume that a brain activity is defined by certain signatures in time and frequency domains. The signal recorded at each electrode is then a certain mixture of these brain activities. On the other hand, if we unfold the tensor in the electrodes mode and apply a two-way factor model, we extract components from electrodes and also from the mode, which is the combination of frequency and time. It is then difficult to interpret the components in the frequency-time mode and understand the brain dynamics corresponding to those components. Besides, since we do not have the assumption of a brain activity with certain spectral and temporal signatures anymore, a two-way factor model may extract as many factors as possible to explain the variation in the data. Extra variation captured by a two-way factor model may actually explain noise rather than a specific structure. Thus, multiway models are more advantageous in terms of interpretation and accuracy compared to two-way models. In addition to the ease of interpretation and accuracy, as we have already mentioned, some multiway models such as PARAFAC are also unique under mild conditions given by the well-known result of Kruskal [78, 79, 113], in contrast to the two-way factor models suffering from rotational ambiguity. Finally, it is always desirable to choose the simplest model for the data and multiway models may be the simplest possible model. For instance, let $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a three-way array. If we model the unfolded data $\mathbf{X}_{(1)}$ (of size $I_1 \times I_2 I_3$) using an R -component two-way factor model, the model will have $(I_1 + I_2 I_3)R$ parameters. On the other hand, if the three-way array is modeled using an R -component PARAFAC model, we need to determine only $(I_1 + I_2 + I_3)R$

parameters. In that sense, a multiway model can be considered statistically simpler than a two-way model [20, 94].

Multilinear models, i.e., PARAFAC [56], Tucker [124–126] and their derivatives, capture the multilinear structure in data. Multilinearity of the model denotes that the model is linear in each mode and factors extracted from each mode are linear combinations of the variables in that mode. A component matrix, whose columns are the factors determined by the model, is then constructed to summarize the structure in each mode. These models have been applied on various datasets shown to contain multilinear structure, e.g., three-way fluorescence spectroscopic datasets with modes: *samples* \times *emission* \times *excitation* [10], wavelet-transformed multi-channel EEG arranged as a three-way array with modes: *frequency* \times *time samples* \times *channels* [4, 41, 90] and on many more data types described briefly in this chapter.

2.3 Multiway Models

The most well-known and commonly applied multiway models in the literature are Tucker models and the PARAFAC model, which is also called CANDECOMP (Canonical Decomposition) [29]¹. We will briefly describe these models as well as recent models built on the principles of PARAFAC and Tucker under three categories (Figure 2.4). First category describes PARAFAC as well as the models, which have relaxed the restrictions enforced by a PARAFAC model to capture data-specific structures. Second category contains the models that belong to the Tucker family and the extensions of Tucker models. Last category includes the models, which fall under neither the first nor the second category but still address the problem of analyzing multiway arrays. In spite of the categorization, models in different families are closely related to each other, e.g., PARALIND can be considered as a constrained version of a Tucker3 model. This categorization is primarily for the ease of presentation and understanding of the models.

In the rest of the chapter, we discuss these models in the context of three-way arrays but most of these models (e.g., the ones in Table 2.1 and Table 2.2) have

¹CANDECOMP was proposed independently but considered equivalent to PARAFAC.

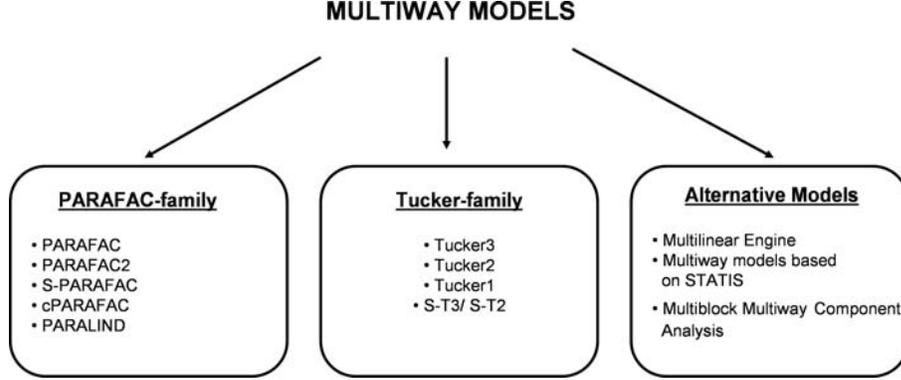


Figure 2.4: The categorization of multiway models briefly explained in this chapter. We study multiway models under three categories: PARAFAC family, Tucker family and alternative approaches.

already been extended to N -way arrays .

2.3.1 PARAFAC-family

2.3.1.1 PARAFAC

PARAFAC [56], which has been originally introduced as the polyadic form of a tensor in [59], is an extension of bilinear factor models to multilinear data. PARAFAC is based on Cattell’s principle of Parallel Proportional Profiles [30]. The idea behind Parallel Proportional Profiles is that if the same factors are present in two samples under different conditions, then each factor in the first sample is expected to have the same pattern in the second sample but these patterns will be scaled depending on the conditions. Mathematically, a PARAFAC model can be represented as the decomposition of a tensor as a linear combination of rank-one tensors. Let $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ be a three-way array. Then an R -component PARAFAC model can be expressed as in Equation 2.7, where \mathbf{a}_i , \mathbf{b}_i and \mathbf{c}_i indicate the i^{th} column of component matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$, respectively. $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ is a three-way array containing the residuals.

$$\underline{\mathbf{X}} = \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \underline{\mathbf{E}} \quad (2.7)$$

The symbol \circ denotes the vector outer product. Illustration of a 2-component

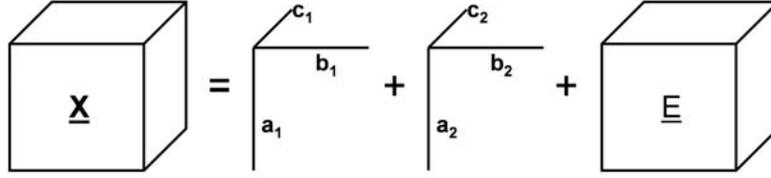


Figure 2.5: Illustration of a PARAFAC model. A 2-component PARAFAC model, where a three-way array $\underline{\mathbf{X}}$ is expressed as the sum of two rank-one tensors. \mathbf{a}_i , \mathbf{b}_i and \mathbf{c}_i are the i^{th} components in the first, second and third modes, respectively. $\underline{\mathbf{E}}$ is a three-way array containing the residual terms.

PARAFAC model on a three-way dataset is given in Figure 2.5.

The motivation behind PARAFAC is to obtain a unique solution such that component matrices are determined uniquely up to a permutation, i.e., rank-one tensors can be arbitrarily reordered, and scaling of columns. It is this uniqueness property that makes PARAFAC a popular technique in various fields. For example in fluorescence spectroscopic data analysis [10], a unique PARAFAC model allows us to find physically and chemically meaningful factors directly from measurements of mixtures of chemicals. Uniqueness is achieved by the restrictions imposed by the model. The most significant restriction is that factors in different modes can only interact factorwise. The interaction between factors in different modes are represented by a core array in multiway models. For example, for a three-way model, the core array is a third-order tensor, $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$, where g_{pqr} represents the interaction of the p^{th} factor in the first, q^{th} factor in the second and r^{th} factor in the third mode (Figure 2.6 (A)). In an R -component three-way PARAFAC model, the core array is restricted to be a super-diagonal core array, $\underline{\mathbf{G}} \in \mathbb{R}^{R \times R \times R}$ where $g_{ijk} \neq 0$ if $i = j = k$, otherwise $g_{ijk} = 0$ (Figure 2.6 (B)). In other words, a super-diagonal core indicates that i^{th} factor in the first mode (\mathbf{a}_i) can only interact with i^{th} factor in the second (\mathbf{b}_i) and the third modes (\mathbf{c}_i). Consequently, the super-diagonal core is a vector of coefficients. These are the coefficients used in the linear combination of rank-one tensors. In Figure 2.5, we assume that the components in one of the modes are scaled by those coefficients and therefore, do not explicitly show the core array.

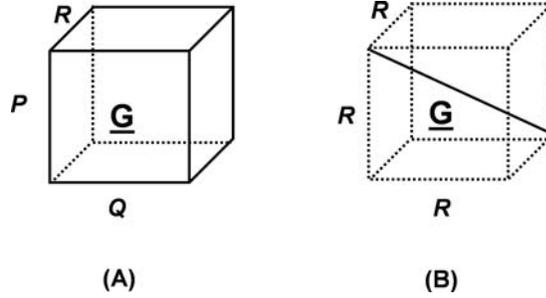


Figure 2.6: (A) A full core array of size $P \times Q \times R$ with maximum PQR nonzero elements, (B) A super-diagonal core array of size $R \times R \times R$ with maximum R nonzero elements.

As a consequence of superdiagonality, the same number of factors should be extracted in each mode. There are several techniques for determining the number of factors in a PARAFAC model, e.g., residual analysis, visual appearance of loadings, number of iterations of the algorithm and core consistency [25]. Among these techniques, core consistency diagnostic quantifies the resemblance between a Tucker3 core and a super-diagonal PARAFAC core and suggests whether a PARAFAC model is a valid model for the data. We demonstrate the use of core consistency in detail in Chapter 3. Core consistency diagnostic has been commonly applied in the literature [4, 10, 46, 90]. However, we should mention that there is no bulletproof way to determine the optimal number of factors (optimal in terms of interpretation) for real data. Therefore, it is often suggested that several diagnostic tools are used together rather than a single method [10, 25].

There are two approaches for fitting a PARAFAC model: *direct fitting* and *indirect fitting*. Direct fitting is defined as fitting a PARAFAC model to the raw data such as a three-way array with modes: $objects \times variables_1 \times variables_2$. Equation 2.8 demonstrates the direct fitting approach using an alternative formulation of a PARAFAC model in matrix notation.

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T + \mathbf{E}_k \quad (2.8)$$

where \mathbf{X}_k represents the k^{th} frontal slice of a three-way array; \mathbf{A} and \mathbf{B} are the component matrices in the first and second mode, respectively. \mathbf{D}_k is a diagonal

Table 2.1: Selection of models from PARAFAC family

Model	Mathematical Formulation	Handles Rank-deficiency
PARAFAC	$x_{ijk} = \sum_{r=1}^R a_{ir} b_{jr} c_{kr} + e_{ijk}$	×
PARAFAC2	$\mathbf{X}_k = \mathbf{A}_k \mathbf{D}_k \mathbf{B}^T + \mathbf{E}_k$	×
S-PARAFAC ¹	$x_{ijk} = \sum_{r=1}^R a_{(i+s_{jr})r} b_{jr} c_{kr} + e_{ijk}$	×
cPARAFAC ²	$x_{ijk} = \sum_{r=1}^R a_{ir} b_{(j-\theta)r} c_{kr} + e_{ijk}$	×
PARALIND ³	$\mathbf{X}_k = \mathbf{A} \mathbf{H} \mathbf{D}_k \mathbf{B}^T + \mathbf{E}_k$	✓

¹ s_{jr} represents the shift at column j for the r^{th} factor.

² θ is used to capture the shifts in the log-frequency spectrogram.

³ \mathbf{H} represents the dependency matrix.

matrix, whose diagonal elements correspond to the k^{th} row of the third component matrix \mathbf{C} . Finally, \mathbf{E}_k contains the error terms corresponding to the entries in the k^{th} frontal slice. While direct fitting is applied on the raw data, indirect fitting is applied on covariance matrices of data slices [20]. For indirect fitting, raw data are rearranged as a three-way dataset of covariance matrices, for instance in the form of *objects* \times *objects* \times *variables*₁ or *objects* \times *objects* \times *variables*₂ assuming one is particularly interested in the object mode. The relationship between direct and indirect fitting approaches is similar to the one between SVD on a data matrix and eigenvalue decomposition on a covariance matrix.

2.3.1.2 Extensions of PARAFAC

Some of the extensions of PARAFAC are PARAFAC2, Shifted PARAFAC (S-PARAFAC), Convolutional PARAFAC (cPARAFAC) and Parallel Factors with Linear Dependency (PARALIND). We discuss these models and their similarities briefly. The mathematical formulations for these models are given in Table 2.1.

- **PARAFAC2** [57]: This model is introduced as a less restrictive model compared to a PARAFAC model. PARAFAC2 relaxes a PARAFAC model by requiring the invariance of the matrix multiplication of a component matrix with its transpose in one mode rather than the invariance of the components themselves.

$$\mathbf{X}_k = \mathbf{A}_k \mathbf{D}_k \mathbf{B}^T + \mathbf{E}_k \quad (2.9)$$

$$\text{s.t. } \mathbf{A}_k^T \mathbf{A}_k = \Phi \quad k = 1, \dots, K$$

where \mathbf{A}_k is the component matrix in the first mode corresponding to the k^{th} frontal slice. Φ , which is the matrix product of \mathbf{A}_k with its transpose, is required to be invariant for all slices $k = 1, \dots, K$. In Equation 2.9, we observe that unlike in a PARAFAC model, the component matrix in the first mode (or one of the modes) can vary across slices in a PARAFAC2 model. This relaxation enables the use of multiway models in the cases, where a PARAFAC model cannot fully recover the underlying structure, e.g., modeling chromatographic data with retention time shifts [23]. Furthermore, PARAFAC2 solves the problem of modeling three-way arrays with slices of different dimensionality (if the dimensionality differs only in one mode). An example of such a multiway array is an environmental dataset that contains the concentrations of some chemical compounds measured at certain time periods across several sampling sites (*sampling sites* \times *parameters* \times *time*) [117]. It is quite common to have measurements from sampling sites for varying time periods, which would result in a three-way array with different dimensionality in one of the modes (e.g., time mode in this case). A PARAFAC2 model using an indirect fitting approach can also handle different dimensionality across slices. Nevertheless, directly fitting PARAFAC2 on raw data has more advantages than indirect fitting in terms of imposing constraints, handling missing data and generalization of the model to N -way arrays [69].

- **S-PARAFAC** [58]: S-PARAFAC has been introduced in order to deal with shifting factors in sequential data such as time series or spectral data. While PARAFAC restricts the data to have the same factor in various proportions in all samples based on Cattell’s idea, S-PARAFAC relaxes this restriction by incorporating shifting information into the model and capturing the factors even if they are available in shifted positions in different samples. On the other hand, one limitation of an S-PARAFAC model is that it only considers one-dimensional shifts such as time shifts and does not handle multi-dimensional shifts that may be encountered in image sequences like brain scans.

When compared to PARAFAC2, S-PARAFAC and PARAFAC2 are quite similar. Both models are less-constrained versions of a PARAFAC model and can model data with shifting factors. In fact, both models have been used in the analysis of

chromatographic data with retention time shifts [23,62]. However, they also have their differences since PARAFAC2 can only capture shifts that maintain the inner product of the factors (i.e., the constraint in Equation 2.9) while S-PARAFAC can model independent shifts at each factor.

- **cPARAFAC** [95]: Another extension of PARAFAC is cPARAFAC, which is a generalization of Non-negative Matrix Factor Deconvolution (NMF D) to multiway spectral data. cPARAFAC, closely related to S-PARAFAC, has been proposed for multi-channel spectral data analysis in order to model convolutive mixtures. Convolution basically means generating a mixture by sending the sources through a filter. When convolution filter is sparse, cPARAFAC becomes equivalent to S-PARAFAC.
- **PARALIND** [27]: A common problem that arises in real data analysis is that ranks of the component matrices may not be the same (called *rank deficiency*). That would require extracting different number of factors in different modes. In that case, fitting a PARAFAC model would give rank deficient solutions and would not guarantee meaningful uniqueness. PARALIND is proposed as an approach for modeling such cases. This model introduces dependency (or interaction) matrices among component matrices to enable the modeling of the data with component matrices with different ranks and capture the dependency between components. Besides, via dependency matrices, prior knowledge about the data and constraints can also be incorporated into the model.

2.3.2 Tucker-family

The models in PARAFAC family can be considered as constrained versions of less restricted multiway models, i.e., Tucker models, which are also called three-mode factor analysis [126] for three-way arrays or N -mode component analysis [66] for higher-order generalizations.

2.3.2.1 Tucker3

Similar to PARAFAC, Tucker3 is an extension of bilinear factor analysis to higher-order datasets. Equation 2.10 shows the formulation of a Tucker3 model on

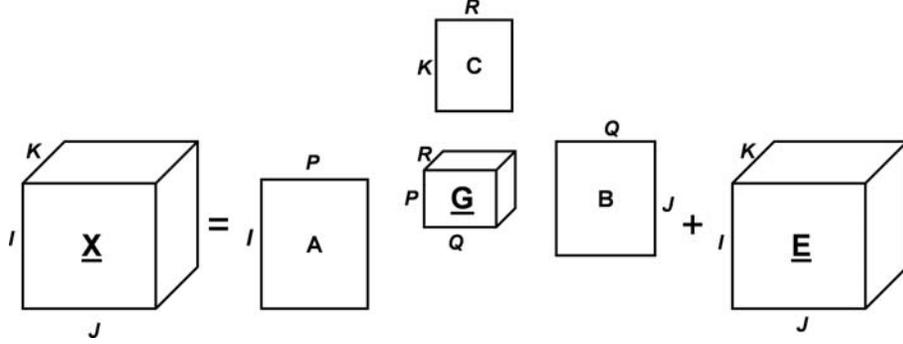


Figure 2.7: Illustration of a Tucker3 model. $\underline{\mathbf{A}}$ (P, Q, R)-component Tucker3 model, where a three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is modeled with component matrices $\underline{\mathbf{A}} \in \mathbb{R}^{I \times P}$, $\underline{\mathbf{B}} \in \mathbb{R}^{J \times Q}$ and $\underline{\mathbf{C}} \in \mathbb{R}^{K \times R}$ in the first, second and third modes, respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ contains the error terms.

a three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$.

$$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk} \quad (2.10)$$

where $\underline{\mathbf{A}} \in \mathbb{R}^{I \times P}$, $\underline{\mathbf{B}} \in \mathbb{R}^{J \times Q}$ and $\underline{\mathbf{C}} \in \mathbb{R}^{K \times R}$ are the component matrices corresponding to the first, second and third modes, respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ contains the residuals. Illustration of a Tucker3 model on a three-way array is given in Figure 2.7.

Compared to PARAFAC, a Tucker3 model is a more flexible model. This flexibility is due to the core array, $\underline{\mathbf{G}}$, which allows an interaction between a factor with any factor in other modes. While the core array enables us to explore the underlying structure of a multiway dataset much better than a restricted PARAFAC model, the full-core array structure in Tucker3 has some drawbacks. First, this property is the reason for rotational indeterminacy in Tucker3 models. Unlike PARAFAC, a Tucker3 model cannot determine component matrices uniquely. When a component matrix is rotated by a rotation matrix, it is possible to apply the inverse of the rotation matrix to the core and still obtain the same model fit. Therefore, a Tucker3 model can determine component matrices only up to a rotation. The second drawback is that the interpretation of Tucker3 models is much more difficult than that of PARAFAC models.

Originally, Tucker family contains Tucker1, Tucker2 and Tucker3 models (Table 2.2). Tucker1 is based on the simple idea of rearranging multiway data as a matrix and decomposing the unfolded data using SVD. Tucker2 and Tucker3 models allow rank reduction in more than one mode and are named after the number of modes rank reduction is applied [76]. Desired rank reduction in each mode are user-specified model parameters and determining these parameters in Tucker models is a tedious task. While using ranks indicated by SVD on unfolded data in each mode is a practical option, systematic methods, e.g., cross validation, DIFFIT [121], have also been developed. DIFFIT (Difference in Fit) enumerates all possible models and uses the differences between model fits to determine the number of components. However, high computational complexity of DIFFIT makes it inefficient. Therefore, it has later been improved by comparing approximate model fit values rather than exact model fits [70]. The most recent work in finding the number of components is based on searching for the convex hull on the plot of model fit values vs. number of free parameters [31]. This approach is more general than previously-proposed methods and helps in determining the model parameters in not only Tucker3 but also Tucker1, Tucker2 and PARAFAC models. Even though empirical comparison of DIFFIT and the convex hull approach on simulation data suggests that the convex hull approach gives promising results and outperforms previous methods, there is no straightforward way to find the optimal number of components [31]. Similar to the case of determining component numbers in a PARAFAC model, several diagnostics should be used to have a true understanding of the structure of a multiway dataset.

2.3.2.2 Extensions of Tucker3

In order to capture shifting factors, similar extensions as in PARAFAC models have also been studied for Tucker models. Shifted Tucker3 (S-T3) and Shifted Tucker2 (S-T2) introduced by [58] are the combinations of Shifted Factor Analysis with Tucker3 and Tucker2 models, respectively. Although it is not proven formally, it has been discussed in [58] that incorporating shifting information in S-T3 suggests the uniqueness of a S-T3 model. Some models that are considered to be in Tucker-family are given in Table 2.2.

Table 2.2: Selection of models from Tucker family

Model	Mathematical Formulation	Handles Rank-deficiency
Tucker1	$x_{ijk} = \sum_{p=1}^P g_{pjk} a_{ip} + e_{ijk}$	✓
Tucker2	$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q g_{pqk} a_{ip} b_{jq} + e_{ijk}$	✓
Tucker3	$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} a_{ip} b_{jq} c_{kr} + e_{ijk}$	✓
S-Tucker3 ¹	$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} a_{(i+s_{jp})p} b_{jq} c_{kr} + e_{ijk}$	✓

¹ In Shifted Tucker3, s_{jp} indicates the shift at j^{th} column for p^{th} factor. Shifted Tucker2 is formulated similarly.

2.3.3 Tucker vs. SVD vs. PARAFAC

Tucker can be considered as a generalization of SVD to higher-order tensors. The link between Tucker and SVD and how singular values and singular vectors generalize to those of higher-order datasets have been extensively studied in [40]. This is a significant milestone in multiway literature since it links multilinear algebra with the models that have been commonly used in psychometrics and chemometrics. Later, computation of singular values and singular vectors of tensors has been discussed in depth in [86], which complements the theoretical background of generalization of singular value decomposition to higher-order datasets.

The Tucker model with orthogonality constraints on the components has been also named as Higher-Order Singular Value Decomposition (HOSVD) [40](or N -mode SVD in [131, 132]). HOSVD is simply computed by flattening the tensor in each mode and calculating the left singular vectors corresponding to that mode, which are also called n -mode singular vectors. Given the component matrices formed by the left singular vectors corresponding to each mode, a core tensor can be computed. This approach is the original idea of Tucker described in [124, 125] and outlined as Method I in [126]. Unlike SVD, HOSVD does not provide the best rank- (R_1, R_2, \dots, R_N) approximation of a tensor [39], where R_i is the rank of a tensor in i^{th} mode. The rank of a tensor in n^{th} mode is called n -rank and it is the dimension of the vector space spanned by the columns of the matrix obtained by flattening the tensor in n^{th} mode. Nevertheless, it does give a good approximation of the data as shown in many applications, e.g., face recognition on an image dataset, where images are affected by several factors such as viewpoints, facial expressions,

lighting conditions, etc. [131,132]. Similarly, in our study we use a Tucker model to explore the spatial signatures of brain dynamics and identify the potential artifacts in Chapter 3. The model we fit may not be the best rank- (P, Q, R) approximation of the data but still helps us identify the dynamics with certain spatial signatures.

Compared to SVD, which is a decomposition that represents a matrix as a sum of rank-one matrices, HOSVD does not decompose a tensor as a sum of rank-one tensors. In that sense, PARAFAC is considered to be another generalization of SVD to higher-order arrays because PARAFAC decomposes a tensor as the sum of rank-one tensors. However, orthogonality constraints on all the component matrices of a PARAFAC model, in general, cannot be satisfied. In order to be able to decompose a tensor with a PARAFAC model, which will give component matrices with orthogonal columns, a tensor should be diagonalizable and in general they are not [71].

SVD has been quite popular in every field of data analysis from signal processing to social network analysis, from chemometrics to image compression because it enables noise filtering through dimensionality reduction. The first R significant singular vectors may represent the data very well and SVD provides the best rank- R approximation for a matrix. Besides, if the best rank- $(R + 1)$ approximation is sought, then the first R singular vectors are kept the same and only one more singular vector is computed. This property has played an important role in the development of online SVD algorithms, which compute SVD of a data stream by updating singular vectors rather than computing SVD of the whole dataset every time the dataset is updated [83]. However, the best rank- R approximation through SVD in matrices cannot be generalized to tensors [71,72]. For instance, in practice when we model multi-channel EEG data represented in both time and frequency domains using a PARAFAC model as a sum of rank-one tensors, if we fit an $(R + 1)$ -component PARAFAC model, we may not get the same first R rank-one tensors that we get using an R -component PARAFAC. In [72], through a counterexample it is demonstrated that the best rank- $(R + 1)$ approximation of a tensor does not necessarily contain the components present in the best rank- R approximation. By best rank- R approximation of a tensor, [72] refers to the orthogonal rank decomposition of a tensor, where a tensor $\underline{\mathbf{X}}$ is expressed as the weighted sum of rank-one

tensors as in Equation 2.11.

$$\underline{\mathbf{X}} = \sum_{r=1}^R \sigma_r \underline{\mathbf{U}}_r \quad (2.11)$$

where $\underline{\mathbf{U}}_i \perp \underline{\mathbf{U}}_j$ for all $i \neq j$. The symbol \perp is defined as follows: Let $\underline{\mathbf{U}}_i$ and $\underline{\mathbf{U}}_j$ be third-order rank-one tensors so that they can be written as the outer product of vectors, e.g., $\underline{\mathbf{U}}_i = \mathbf{u}_i^{(1)} \circ \mathbf{u}_i^{(2)} \circ \mathbf{u}_i^{(3)}$. Then $\underline{\mathbf{U}}_i$ and $\underline{\mathbf{U}}_j$ are orthogonal ($\underline{\mathbf{U}}_i \perp \underline{\mathbf{U}}_j$) if $\prod_{r=1}^3 \mathbf{u}_i^{(r)} \cdot \mathbf{u}_j^{(r)} = 0$. The minimal number of rank-one tensors (minimal R) needed to express $\underline{\mathbf{X}}$ in the form given in Equation 2.11 is called the *orthogonal rank* of $\underline{\mathbf{X}}$. For a detailed discussion on rank decompositions and algorithms proposed for computing the best rank approximations, the reader is referred to [39], [71] and [143].

2.3.4 Other Models

There are several other models based on approaches other than PARAFAC and Tucker models for multiway data analysis. In this section, we briefly introduce some of these models: Multilinear Engine (ME) [100], multiway models based on STATIS [117] and multiblock multiway models [114].

ME is a program that is capable of fitting different structural models including PARAFAC and PARAFAC2 on multiway arrays using a general-purpose optimization/curve fitting approach. Although models mentioned so far are only capable of modeling multilinearity in data, structure tables created by specified variables and functions enable ME to fit multilinear as well as quasi-multilinear models. Multilinear models are based on mathematical expressions, which are linear with respect to each set of variables corresponding to different modes whereas quasi-multilinear models contain nonlinearity in the sense of polynomials. Therefore, the multilinear engine can explore a wider range of structures in data compared to PARAFAC, Tucker3, etc.

Another model focusing on three-way data analysis is STATIS [117] originally studied in [28]. When compared to N -way analysis methods, which explore each mode simultaneously, the STATIS-based model in [117] explores each mode separately. It considers each observation/sample as a slice of a three-way array and computes the covariance matrix corresponding to that slice. The basic principle in

the model is to apply Principal Component Analysis (PCA) on a global covariance matrix formed as a linear combination of covariance matrices corresponding to individual slices. Similar to indirect fitting approach, it is possible to analyze three-way arrays with slices of different sizes using STATIS. One disadvantage of STATIS is that it cannot be generalized to N -way arrays.

Methods, referenced so far, focus on the analysis of a single multiway array. On the other hand, multiblock multiway arrays are also encountered in various studies such as control of batch processes, where more than two blocks of multiway arrays need to be analyzed simultaneously. One approach to deal with multiblock multiway component problems is to analyze each multiway array using a certain structural model such as a Tucker3 or a PARAFAC model and then combine summaries of information from different multiway arrays in a single matrix [114]. The matrix containing summaries from different arrays can then be analyzed using bilinear factor models. This approach can be considered as a generalized version of Collective PCA [67] to higher-order datasets.

2.4 Algorithms

Algorithms for fitting multiway models are, in general, iterative algorithms and based on Alternating Least Squares (ALS). In this section we briefly discuss the algorithms used for fitting, in particular, Tucker and PARAFAC models.

2.4.1 Tucker Model

The original algorithm for computing a Tucker3 model was described by Tucker in [126] as his 'Method I'. As we have already mentioned in the previous section, this method unfolds the data in each mode, e.g., $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$ for a three-way tensor $\underline{\mathbf{X}}$, and then computes the eigenvectors of $\mathbf{X}_{(i)}\mathbf{X}_{(i)}^T$; in other words, the left singular vectors of $\mathbf{X}_{(i)}$ to construct the component matrix for mode i . Finally, component matrices and the original data are used to compute the core tensor. Algorithm 1 describes these steps for modeling a third-order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ using a (P, Q, R) -component Tucker3 model. The number of components extracted from each mode are user-specified parameters and we have already discussed the tech-

Algorithm 1 Tucker3($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$, P , Q , R)

- 1: \mathbf{A} = First P left singular vectors of $\mathbf{X}_{(1)}$
 - 2: \mathbf{B} = First Q left singular vectors of $\mathbf{X}_{(2)}$
 - 3: \mathbf{C} = First R left singular vectors of $\mathbf{X}_{(3)}$
 - 4: $\underline{\mathbf{G}} = \underline{\mathbf{X}} \times_1 \mathbf{A}^T \times_2 \mathbf{B}^T \times_3 \mathbf{C}^T$
-

niques proposed for finding the component numbers in Section 2.3.2.1. Algorithm 1 has been named differently in various studies, e.g., HOSVD in [40], N -mode SVD in [131,132] or CubeSVD in [120].

Since Algorithm 1 does not provide the optimal solution in least squares sense ([40] provides an error bound), this approach has been extended by alternating least squares in [76] for three-way arrays and called TUCKALS3. It has later been formulated also for N -way arrays [66]. TUCKALS3 finds the best approximation for $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ by minimizing the least squares error function given in Equation 2.12.

$$\|\underline{\mathbf{X}} - \tilde{\underline{\mathbf{X}}}\|^2 = \|\underline{\mathbf{X}} - \underline{\mathbf{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}\|^2 \quad (2.12)$$

where \mathbf{A} , \mathbf{B} , \mathbf{C} are columnwise orthogonal component matrices in the first, second and third modes, respectively and $\underline{\mathbf{G}}$ is the core tensor. ALS algorithm is often initialized by the component matrices obtained in Algorithm 1 and these component matrices are later estimated one at a time keeping the estimates for other component matrices fixed. Estimation of component matrices is repeated until a convergence criterion, e.g., no change in model fit; in other words no change in the error given in Equation 2.12 or fixed number of iterations, is satisfied. Algorithm 2 describes the ALS-based approach for computing a (P, Q, R) -component Tucker3 model on $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$. This algorithm, or more specifically its N -mode generalization, has been studied under different names in the literature, e.g., Higher-Order Orthogonal Iteration in [39] and N -mode Orthogonal Iteration in [133].

By replacing one of the component matrices with an identity matrix, Algorithm 2 can be modified to model a third-order tensor using a Tucker2 model. The algorithm for fitting a Tucker2 model has been called TUCKALS2 [76]. As we have already mentioned, Tucker1 model, which assumes that two of the component matrices are identity matrices and performs rank reduction in only one mode, can be

Algorithm 2 Tucker3-ALS($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$, P , Q , R)

- 1: Initialize $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$
 - 2: **while** *the convergence criterion not satisfied* **do**
 - 3: $\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_2 \mathbf{B}^T \times_3 \mathbf{C}^T$
 - 4: $\mathbf{A} =$ the first P left singular vectors of $\mathbf{Y}_{(1)}$
 - 5: $\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{A}^T \times_3 \mathbf{C}^T$
 - 6: $\mathbf{B} =$ the first Q left singular vectors of $\mathbf{Y}_{(2)}$
 - 7: $\underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{A}^T \times_2 \mathbf{B}^T$
 - 8: $\mathbf{C} =$ the first R left singular vectors of $\mathbf{Y}_{(3)}$
 - 9: $\underline{\mathbf{G}} = \underline{\mathbf{X}} \times_1 \mathbf{A}^T \times_2 \mathbf{B}^T \times_3 \mathbf{C}^T$
-

simply computed by decomposing the unfolded data using SVD.

Time complexity of these algorithms depends on the dimensionality of the data in each mode as well as the rank reduction in each mode. For instance, for an $N \times N \times N$ third-order tensor, if we compute the full HOSVD by extracting N components in each mode, then time complexity is $O(N^4)$ (assuming constant number of iterations in Algorithm 2). However, we often do not need the full HOSVD in practice and compute the first R components in each mode, where we assume that the number of components extracted from each mode equals R and $R \ll N$. In that case, the complexity becomes much less, i.e., $O(N^3R)$.

In addition to Algorithm 1 and Algorithm 2, recently other algorithms such as Slice Projection [137] and Multislice Projection [127] for computing a Tucker3 model have been proposed. These approaches are based on the idea of projecting individual slices onto component matrices, where component matrices are estimated iteratively using alternating least squares. These algorithms have been compared with Algorithm 1 and Algorithm 2 in [127] based on several metrics including space complexity and the model fit. The results demonstrate that while Algorithm 2 performs best in terms of approximating the original data, the algorithms based on the slice projection approach have advantages in terms of handling very large tensors in the case of insufficient memory.

Algorithm 3 PARAFAC-ALS($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$, R)

- 1: Initialize \mathbf{A} , \mathbf{B} and \mathbf{C} , where $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$
 - 2: **while** *the convergence criterion not satisfied* **do**
 - 3: $\mathbf{Z} = \mathbf{C} \odot \mathbf{B}$
 $\mathbf{A} = \mathbf{X}_{(1)} \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1}$
 - 4: $\mathbf{Z} = \mathbf{C} \odot \mathbf{A}$
 $\mathbf{B} = \mathbf{X}_{(2)} \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1}$
 - 5: $\mathbf{Z} = \mathbf{B} \odot \mathbf{A}$
 $\mathbf{C} = \mathbf{X}_{(3)} \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1}$
-

2.4.2 PARAFAC Model

The original algorithms proposed for computing a PARAFAC model are based on the alternating least squares approach [29, 56]. As in Tucker ALS algorithm, estimates for each component matrices are computed one at a time. In order to understand how the estimate for each component is determined, we rewrite the PARAFAC model on a three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ given in Equation 2.7 and Equation 2.8 using the Khatri-Rao product as in Equation 2.13.

$$\mathbf{X}_{(1)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_{(1)} \quad (2.13)$$

where $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$ are the component matrices corresponding to the first, second and third modes, respectively. We can then write the least squares objective function for a PARAFAC model as the minimization of:

$$\|\underline{\mathbf{X}} - \tilde{\underline{\mathbf{X}}}\|^2 = \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F^2 \quad (2.14)$$

If we minimize this function with respect to \mathbf{A} , we compute the least squares estimate for \mathbf{A} as $\mathbf{X}_{(1)}((\mathbf{C} \odot \mathbf{B})^T)^\dagger$, where \dagger denotes the pseudo-inverse of a matrix defined as follows for matrix \mathbf{Y} , $\mathbf{Y}^\dagger = (\mathbf{Y}^T \mathbf{Y})^{-1} \mathbf{Y}^T$. The estimates for other modes can also be computed similarly. The ALS-based algorithm iteratively updating these estimates until a convergence criterion, e.g., no change in model fit or fixed number of iterations, is satisfied is given in Algorithm 3. The time complexity is $O(N^3 R)$ per iteration when an R -component PARAFAC model is applied on an $N \times N \times N$ tensor.

Although Algorithm 3 is the most commonly used approach to fit a PARAFAC model, it does not guarantee convergence to the global optimum. Different initializations of the algorithm may converge to different local optima. The common approaches for initializing a PARAFAC model are either using random starting points or choosing the initial values based on generalized eigenvalue decompositions [20]. Furthermore, ALS suffers from slow convergence rate. Alternative algorithms have been proposed for fitting a PARAFAC model, in particular to third-order tensors, with the objective to improve PARAFAC-ALS in terms of convergence rate and robustness to overfactoring. [47] compares the performance of some of these algorithms such as Alternating Slice-wise Diagonalization (ASD) [64], Self Weighted Alternating Trilinear Diagonalization (SWATLD) [33] and many more in terms of speed, model fit, sensitivity to overfactoring and predictive ability of the models on real datasets. It has been shown that ASD may be a good alternative to ALS particularly for the cases when the slowness of ALS is a concern. Another comparative study on PARAFAC algorithms is [123], which compares some other algorithms in addition to PARAFAC-ALS, SWATLD, ASD from [47]. [123] highlights the advantages of SWATLD algorithm in the case of overfactoring and also suggests PMF3 [99] and Levenberg-Marquadt algorithm [99, 122] as alternative algorithms with better convergence properties than ALS for fitting a PARAFAC model.

As a generalization of incrementally computing the rank-one decomposition of a matrix, a greedy algorithm has also been given in [75] to compute rank-one tensors incrementally to compute an R -component PARAFAC model. Even though this greedy approach does not give the optimal rank- R approximation of a tensor, incremental rank-one approximation can be preferred due to its simplicity [143].

2.4.3 Preprocessing

Preprocessing is not often mentioned as a separate step in the algorithms discussed above. However, it is a crucial step in data analysis in general. Similar to preprocessing in two-way data analysis, centering and scaling both generalize to multiway arrays [26]. Centering across one mode of a three-way dataset is performed by simply matricizing the data in the desired mode and applying two-way centering.

Scaling, on the other hand, requires for instance, in the three-way case, scaling each slice of a third-order tensor corresponding to each variable rather than scaling columns as in two-way analysis.

2.5 Applications

As already mentioned throughout this chapter, multiway models are employed in numerous disciplines addressing the problem of finding the multilinear structure in multiway datasets. There are many applications in various fields and this chapter offers some representative examples from different research areas.

2.5.1 Chemistry

We start with one of the most popular applications of a PARAFAC model: modeling a fluorescence excitation-emission dataset, which is a commonly used data type in chemistry, medicine and food science. Such data typically consist of samples containing different concentrations of several chemical compounds. The goal of PARAFAC analysis on this data type is to determine the compounds found in each sample as well as the relative concentrations of compounds. Fluorescence spectroscopy enables the generation of three-way datasets with modes: *samples* \times *emission* \times *excitation*. Among many other applications of PARAFAC, modeling of fluorescence spectroscopy is the one, which demonstrates the modeling power and interpretation of factors of a PARAFAC model most clearly. An example of a PARAFAC model on a fluorescence spectroscopic dataset is given as an in-depth study on a Fish dataset and data with known fluorophores in [10]. This study is an important resource demonstrating the underlying idea of the structural model of PARAFAC, its benefits and limitations.

Here we include a small example to demonstrate the modeling power of PARAFAC on fluorescence data. The sample dataset contains five samples with different amounts of three amino acids (tyrosine, tryptophan and phenylalanine) (This dataset was originally measured by Claus A. Andersson and published in [21]). We have a three-way array of size $5 \times 201 \times 61$, where the first, second and third modes correspond to the samples, emission and excitation wavelengths, respectively. We

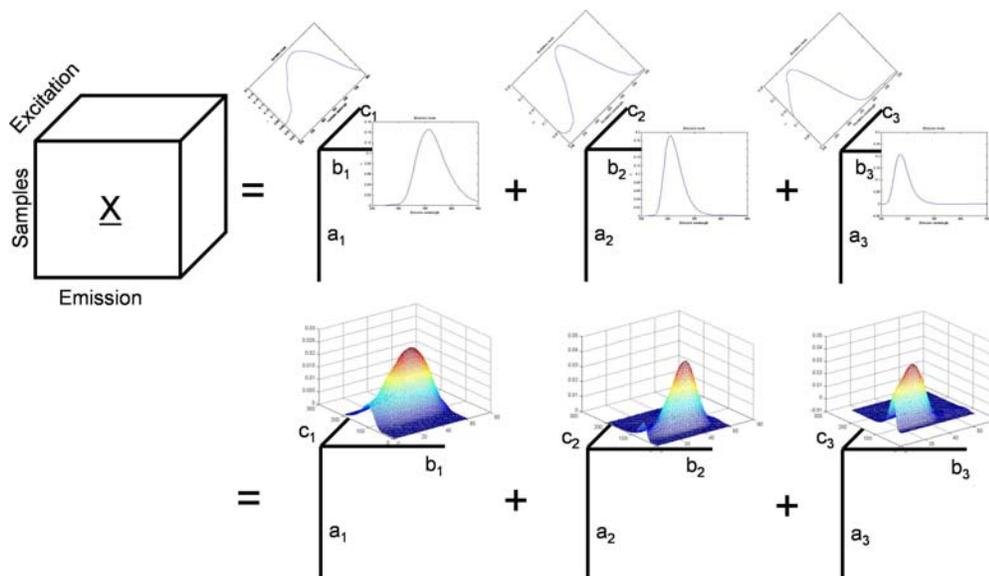


Figure 2.8: Modeling of a fluorescence dataset using a 3-component PARAFAC model. a_i , b_i and c_i correspond to the i^{th} component in samples, emission and excitation modes. We also illustrate the vector outer product of b_i and c_i , which shows the fluorescence landscape of each analyte used in the preparation of the samples.

model the data using a three-component PARAFAC model and the components in the excitation and emission modes are illustrated in Figure 2.8. Fluorescence landscapes formed using these component matrices are also given in the second line in Figure 2.8. Each fluorescence landscape indeed corresponds to one of the amino acids. The coefficients in component \mathbf{a}_i then indicates how much the amino acid whose fluorescence landscape is given by \mathbf{b}_i and \mathbf{c}_i contributes to each sample. Each sample, therefore, is a mixture of different amounts of several amino acids; in other words, original samples corresponding to horizontal slices of \mathbf{X} are linear combinations of the fluorescence landscapes extracted by a PARAFAC model. The reason why a PARAFAC model is appropriate for modeling fluorescence data is that each one of the fluorescence landscapes is a rank-one matrix and a particular analyte has specific signatures in emission and excitation modes. There is no need to model such a dataset using a Tucker3 model since the components in each mode only interact with components with the same id in other modes, indicating a super-diagonal core structure as in the case of a PARAFAC model.

Modeling power and limitations of a PARAFAC model have also been addressed in another study modeling chromatographic data [23]. Even though initial challenge of uniquely identifying the components in chromatographic data can be solved by a PARAFAC model, PARAFAC falls short in extracting the components when elution profiles of the components follow a shifting pattern throughout the experiments. On the other hand, PARAFAC2 succeeds in modeling the shifting factors and recovering the underlying components in chromatographic data with retention time shifts.

Recently, multiway analysis techniques combined with a clustering and visualization approach have been applied on chromatographic measurements of a metabolite profiling dataset in order to explore the chemical compounds accounting for the differences between different commercial extracts [7]. In addition to these studies, a recent review on multiway analysis in chemistry lists a broader range of applications on chemical datasets [22].

2.5.2 Social Network Analysis/Text-mining

Multiway data analysis has often been employed in extracting relationships in social networks. The aim of social networks analysis is to study and discover hidden structures in social networks, for instance, extracting communication patterns among people or within organizations. In [1], chatroom communications data have been arranged as a three-way array with modes: *users* \times *keywords* \times *time windows* and the performance of multiway models in capturing the underlying user group structure has been compared with that of two-way models. Another study [3] assesses the performance of collective and centralized tensor analysis approaches again on chatroom communications data. Not only chatroom but also email communications have been analyzed using multiway models [17].

In the context of web link analysis, [73] and [75] combine hyperlink and anchor text information and rearrange web graph data as a sparse three-way tensor with modes: *webpages* \times *webpages* \times *anchor text*. The web graph is then analyzed using an algorithm improved to fit a PARAFAC model to large and sparse datasets efficiently in order to capture the groupings of webpages and identify the main topics.

Furthermore, with a goal of improving personalized web searches, click-through data have also been analyzed using a multiway analysis method called CubeSVD [120], which is indeed the same as HOSVD. In this study, click-through data are arranged as a three-way array with modes: $users \times queries \times webpages$ and CubeSVD is compared with two-way methods such as Latent Semantic Indexing (LSI) and shown to outperform the two-way approaches.

Recently, [34] has made use of a PARAFAC2 model to cluster similar documents in different languages by arranging the data as a third-order tensor with modes: $terms \times documents \times languages$. It has been demonstrated that while LSI performs well for language-specific scenarios, it tends to cluster documents in the same language regardless of their similarity in terms of topic when it is applied on multilingual documents. On the other hand, a PARAFAC2 model is proposed as a better alternative in clustering documents with similar topics in different languages.

2.5.3 Computer Vision

Approximations of tensors have proved to be useful in computer vision applications such as image compression, representation and recognition. Images have two dimensional nature where x and y -coordinates being the two modes. When data have one more mode, e.g., temporal dimension, different illuminations or viewpoints etc., it forms a higher-order tensor. Most image coding techniques consider images as vectors by reshaping the data as a vector. In order to prevent vectorization of image coordinates, [141] introduces two-dimensional PCA (2DPCA) approach based on an image covariance matrix computed by preserving the two-dimensional nature of the images and proposes to use principal components of the covariance matrix for image representation. This approach is later extended to two-dimensional SVD (2DSVD) [44] by first computing both column-column and row-row covariance matrices and then finding the low-rank approximation of each image using the eigenvectors of these covariance matrices based on an approximation method called GLRAM introduced in [142]. These studies preserve the 2D nature of images instead of representing each image as a vector. Similarly, [136] preserves the 2D nature of the images, for instance, of a video sequence by forming a third-order tensor

with modes x -coordinate, y -coordinate and video frames. It has been shown that when video sequences are represented as tensors and iteratively obtained rank-one approximations of tensors are used to compress the video sequence [136], the error between the original and reconstructed images is less than the error obtained when PCA is used for compression. In [137], tensors are also used to compactly represent images and it has been shown that rank- R approximation of tensors (rank- R approximation of a tensor here refers to extracting R components in each mode; in other words, rank- (R_1, R_2, R_3) approximation of a tensor, where $R_1=R_2=R_3=R$) outperforms PCA, GLRAM and rank-one approximations of tensors in terms of the reconstruction error of a video sequence.

Another application of multiway models in computer vision is face recognition. For instance, a set of face images are arranged as a fifth-order tensor, that represents face images using not only pixel information but also illuminations, expressions, viewpoints and person identities in [131]. Although face images are treated as vectors, varying conditions of the images are used to construct a multi-modal dataset. HOSVD is then used to decompose this tensor and extract the basis vectors called TensorFaces [131–133]. Component matrices extracted from each mode are used to construct person-specific, viewpoint-specific, illumination-specific or expression-specific TensorFaces, which improve the understanding of the underlying factors in an image dataset. Apart from these applications, tensors have also been employed in several other fields in computer vision, e.g., textured surface rendering [134] and handwritten digit recognition [107].

2.5.4 Process Analysis

Last research area we mention here is process monitoring. Real-time batch process monitoring is a challenging task since the complete data are needed for the analysis in general and that would require waiting till the completion of a batch. However, in-filling of missing future data and modeling using a PARAFAC model have been shown to overcome this challenge and PARAFAC has been demonstrated to be an applicable approach in controlling batch processes arranged as a three-way array with modes: $batches \times variables \times time\ samples$ [89].

Similarly, STATIS has been applied in monitoring batch processes on datasets that come from various areas such as pharmacology, spectroscopy and yeast bakery production. These datasets are arranged as three-way arrays and have different number of dimensions in the time mode [52]. STATIS can handle the different dimensionality problem easily since it is applied on covariance matrices. On these datasets, STATIS-based approach has been compared with unfolded PCA (in other words Tucker1). The three-way array is unfolded in variables mode and a matrix of *variables* \times *batches* - *time samples* is constructed. It has been observed that detecting bad batches is much easier with STATIS than it is with unfolded PCA.

2.6 Software

As multiway data analysis is spreading from chemometrics and psychometrics to other fields, software tools have also been developed and improved. Some available software for multiway data analysis are the Nway Toolbox [13], Tensor Toolbox [15, 145], PLS.Toolbox (by Eigenvector Research Inc.) and CuBatch [53], which all run under MATLAB. The Nway toolbox is the original toolbox, which has combined multiway analysis techniques such as PARAFAC and Tucker models in a software package and enabled the application of these models in different fields. The Tensor Toolbox has been initially introduced as a TensorClass, which handles mathematical operations on multiway arrays such as tensor–matrix multiplications, matricization of tensors and many other algebraic operations. It has later been extended to manipulate efficiently not only multiway arrays but also sparse tensors, where only small fraction of the elements are nonzero. CuBatch is another software package recently introduced as a multiway analysis toolbox with a user-friendly interface. It has been originally built for analyzing batch process data but it is applicable on multiway datasets in general. Some of the available models in the toolbox are PCA, PARAFAC, PARAFAC2, and Tucker models. Preprocessing techniques such as centering and scaling and different techniques for identifying outliers are also included in this toolbox. CuBatch contains the Nway Toolbox functions and it is a more developed version of the initial toolbox. Apart from these freely available toolboxes, there is also a commercial toolbox called PLS.Toolbox, which enables

the analysis of multiway arrays with numerous multiway models providing visual analysis tools. An efficient approach for analyzing multiway arrays would be to combine the Tensor Toolbox with one of the other toolboxes to have modeling, algorithmic and visualization power as a readily-available package.

In addition to software running under MATLAB, there is another software package called Multilinear Engine [100] implemented in FORTRAN. There are also other software packages for manipulating multiway arrays but they do not particularly focus on multiway data analysis or multiway models. For more information on these software packages, interested users are referred to [16] and references therein.

2.7 Summary

Multiway data analysis has recently attracted attention in many fields. That is mostly due to the nature of the datasets, which cannot be truly captured by traditional two-way analysis techniques. As datasets started to be rearranged as multiway arrays rather than matrices, multilinear models, which have been mostly used in psychometrics and chemometrics, have become more popular. These models have been followed by extended versions of original models and techniques, e.g., Shifted factor models, PARALIND, etc. in an effort to capture data-specific structures in multiway datasets. Theoretical aspects of these models such as model uniqueness and rank properties of multiway arrays have been studied in depth. New algorithms, e.g., ASD, SWATLD, etc. as alternatives to ALS have been developed. Multilinear algebra, a less-known field compared to linear algebra, has been explored to perform operations on multiway arrays and develop computationally efficient algorithms. Enhanced software tools, e.g., Nway Toolbox, CuBatch, etc. enabling multiway data analysis and mathematical operations on multiway arrays, e.g., Tensor Toolbox, have been implemented.

In this chapter, we have mainly focused on multiway models and briefly described the algorithms employed in fitting these models to multiway datasets. We have also given representative applications of multiway analysis from a variety of disciplines to illustrate the diversity of the fields making use of multiway data analysis. However, we have not mentioned some important aspects of multiway analysis.

There are concepts such as uniqueness properties of models and uniqueness conditions for models, e.g., well-known Kruskal’s condition for PARAFAC [78, 79], as well as different definitions of rank for tensors. [74] explains these concepts in depth pointing to important references in the literature. Furthermore, most of the models given in Section 2.3 can be employed for both unsupervised and supervised learning. In addition to those, there are also multiway models especially developed for supervised learning, e.g., Multilinear Partial Least Squares [19]. In this chapter, we have rather focused on only unsupervised multiway analysis but in Chapter 4, we will use supervised approaches based on N-PLS and study the generalization of regression models to higher-order datasets.

2.8 Discussions

Recent studies show that multiway models have many application areas in computer science such as social network analysis, web link analysis and a variety of other problems in data mining as well as computer vision. Besides, recent theoretical studies focusing on multiway models improve the understanding of the models commonly used and developed in chemistry and psychometrics and make them more popular in handling computer science problems.

Even though current algorithms and models are applicable on numerous datasets, there is still further progress needed in several fields. First area we want to emphasize is the summarization and analysis of data streams. Techniques discussed so far are based on offline and centralized dimensionality reduction models. On the other hand, there is also a demand for online methods to analyze data streams efficiently, especially in communication networks. For instance, a recent study tackle the problem of analyzing data streams using an approach called dynamic tensor analysis [119]. Similarly, developing distributed versions of these methods would enable efficient analysis of massive datasets.

In addition, concepts of multilinearity and nonlinearity should be studied further so that the limitations of multilinear models in capturing the structures in multiway datasets are better understood. Factors extracted by common multilinear models, i.e., PARAFAC-family and Tucker-family, are linear combinations of vari-

ables in each mode. On the other hand, these models will fail to discover nonlinear relationships between variables. Nonlinearity has been a recent topic of interest in data mining community especially since kernel methods [111] became popular. Most two-way analysis techniques, e.g., PCA, Canonical Correlation Analysis (CCA), are combined with kernels in order to capture the underlying nonlinear structure among variables [109]. Similar to a recent study in computer vision community, which has combined HOSVD with kernel methods for face recognition to capture the nonlinear relationship between image features [85], embedding kernels into multiway models should be explored more, especially from a theoretical perspective.

Finally, performance of multiway data analysis in terms of space and computational complexity should be analyzed further. Most studies applying multiway data analysis demonstrate how multiway methods improve the interpretation and accuracy of the underlying structure captured when multiway models are used instead of two-way methods. However, computational and space complexity of multiway models and algorithms have not often been discussed in the literature except for few studies including [11] and [12], where the speed of algorithms for a Tucker3 model is compared and Tucker3 is suggested as a compression method for speeding up multiway analysis; [143], which compares different algorithms for computing rank-one approximation of a tensor; and [16], which discusses efficient tensor storage schemes. Recently, for efficient analysis of massive multiway datasets, e.g., recommendation systems, hyperspectral image datasets, Tensor-CUR decomposition has been proposed [87]. Unlike multiway models discussed in this chapter, Tensor-CUR does not employ an approach based on factor models. Instead, this algorithm relies on sampling subtensors, which consist of the original data elements, based on a given probability distribution and approximating the data using the sampled subtensors. While sampling reduces the complexity of the problem, how well it captures the multilinear structure in datasets in general is another open problem.

CHAPTER 3

EPILEPTIC SEIZURE LOCALIZATION

Using some of the principles of multi-modal data construction and multiway data analysis reviewed in the previous chapter, we develop mathematical models for epileptic seizure localization in this chapter. It is extremely important to localize a seizure origin precisely since the success of an epilepsy surgery strongly depends on the correct and precise identification of a seizure onset zone. Epilepsy is defined as recurrent seizures caused by abnormally synchronous neuronal activity. The electrical symptoms of this abnormal activity are believed to uniquely define and reveal the mechanisms of the underlying abnormal neural function and structure. Localization of the initial seizure discharge gives clues about the cortical region that generates epileptic seizures, which is called a *seizure onset zone* (*epileptic focus* or *seizure origin*). Therefore, the analysis of ictal EEG (scalp or intracranial) is an effective standard for localization of an epileptic focus.

We address the problem of localization of a seizure origin through an analysis of ictal scalp EEG recordings with a goal of developing an automated and objective approach. Ictal periods are identified by neurologists visually based on EEG seizure onsets rather than clinical seizure onsets.

3.1 Related Work

The majority of the research devoted to automated detection of epileptic events concentrates around spike detection techniques. Although most of these techniques are based on single channel data, in [50] spatial information from 16-channel EEG data has been incorporated in building a detection system for epileptic sharp waves. Sharp wave source localization on multi-channel EEG data has also been applied in [49] to determine the areas of interest with epileptic activity.

The main challenge in focus localization using scalp EEG recordings is the contamination of EEG with artifacts. *Ictal EEG*, or in other words, the EEG data recorded during the seizure period of an epileptic patient, are often contaminated

with signals caused by eye blinks, eye movements and/or muscle artifacts. These artifacts undermine the efforts to localize epileptic foci and understand the characteristics of a seizure. Commonly used approaches for artifact removal are simple filtering techniques and statistical methods such as Independent Component Analysis (ICA) [37] and, lately, Canonical Correlation Analysis (CCA) [63]. The filtering methods eliminating EEG activity within certain frequencies may result in loss of significant information about the seizure structure in the cases where epileptic signals and artifacts overlap in the frequency domain. Therefore, statistical approaches, especially methods based on ICA, are quite common in artifact removal literature. Some of these statistical techniques rely on the widely-accepted assumption of independence between artifacts and epileptic brain signals. Considering EEG recorded at each electrode as a linear mixture of signals originating from independent sources, independent components are extracted using ICA [43, 54, 128, 144]. The components corresponding to artifacts are later identified by visual inspection [128, 144] or a semi-automated/automated artifact identification techniques based on high-order statistics, i.e., kurtosis, entropy [43, 54]. As an alternative to ICA, a CCA-based artifact removal approach has been recently proposed [38]. This technique is similar to ICA-based methods except for the independence assumption. The underlying idea is rather the mutual non-correlation between artifacts and epileptic signals.

Artifact removal approaches mentioned so far focus on multi-channel EEG data arranged as a two-way dataset of recordings collected at several electrodes at different time samples. Two-way analysis methods on multi-channel EEG data, however, only allow us to capture temporal and spatial signatures, such as the ones identified by ICA and CCA-based techniques. In order to capture frequency domain information in addition to temporal and spatial signatures, these methods require one more step, e.g., feature extraction as in [82], where several features based on the spectral information content of a component are extracted or the transformation of the original data into frequency domain and the application of two-way analysis methods also in the frequency domain. On the other hand, constructing multi-modal datasets from multi-channel EEG signals by representing an EEG recording in both time and frequency domains enable us to inspect the information content of EEG

signals in time, frequency and electrode domains simultaneously using multilinear component models. Before discussing how to construct such multi-modal datasets from EEG recordings, we first summarize the applications of multiway analysis techniques in neuroscience.

3.1.1 Multiway Models in Computational Neuroscience

In computational neuroscience multiway models have been applied in a variety of problems, e.g., analysis of EEG/ERP signals and fMRI (functional MRI) data. As one of the very first applications of multiway models in neuroscience, to our knowledge, [36] makes use of a PARAFAC model to identify which tasks relate to which parts of the brain and frequency bands based on the analysis of EEG signals. Another application of a PARAFAC model was introduced in the late 1980s as the decomposition of event-related potentials (ERP) [94]. Due to the problems such as rotational ambiguity and orthogonality constraints of PCA/SVD in analyzing and interpreting brain signals, [94] introduces the Topographic Components (TC) model, which is essentially the same as a PARAFAC model but rather developed in the context of ERP analysis. [94] arranges multi-channel ERP signals from multiple subjects as a third-order tensor with modes: subjects, electrodes and time samples and discusses the model from a perspective of ERP signal analysis. Following the introduction of the TC model, [48] applies a TC/PARAFAC model on evoked potential (EP) signals and demonstrates that modeling of this tensor using a TC/PARAFAC model extracts, what is called in the paper as, "temporal components", "spatial components" and "subject scores" when the data are arranged in the same way as introduced in [94].

PARAFAC in neuroscience has become even more popular with the introduction of decomposing EEG data into space, time and frequency components [90]. [90] applies continuous wavelet transformation (CWT) on each signal recorded at each channel so that the signal from a single channel is represented in both time and frequency domains. Then the wavelet-transformed data are arranged as a three-way array with modes $time\ samples \times frequency \times electrodes$ and analyzed using a PARAFAC model. This study demonstrates that factors in the first, second and

third component matrices represent the temporal, spectral and spatial signatures of the EEG data, respectively. PARAFAC models with nonnegativity constraints have later been used in another study on ERP to find the underlying structure of brain dynamics [96]. Recently, a toolbox called ERPWAVELAB [97] running under MATLAB has been released for multi-channel time-frequency analysis of brain activity using EEG data. Not only PARAFAC but also the other most well-known multiway model, i.e., Tucker3, has been showed to perform well in EEG analysis. [46] uses a Tucker3 model to study the effect of a new drug on brain activities by arranging EEG data and data collected through experiments with different doses of a drug over several patients under certain conditions as a six-way array with modes: EEG, patients, doses, conditions, time and channels. The results demonstrate that significant information is successfully extracted from a complex drug dataset by a Tucker3 model rather than two-way models such as PCA.

In addition to these studies on EEG/ERP analysis, multiway methods have also been used in the analysis of fMRI data. A third-order tensor with modes $voxels \times time\ points \times runs$ is constructed in [9] from different runs of fMRI from a patient. This tensor is then analyzed by a PARAFAC model in order to capture the spatial and temporal profiles of brain functions such as an activation triggered by finger tipping. So far we have mentioned the applications of multilinear component models. Apart from component models, multilinear regression models have also been employed in neuroscience, e.g., in [88] for extracting the connection between EEG recordings and fMRI data.

3.1.2 Multiway Models in Epilepsy

These studies have motivated the application of multiway models for understanding the structure of epileptic seizures [2, 4, 41]. Similar to the three-way array constructed in [90], multi-channel ictal EEG data are arranged as a third-order tensor called an epilepsy tensor with modes $time\ samples \times frequency \times electrodes$ using the power of wavelet coefficients in [2] and [4] and using pure wavelet coefficients in [41]. [2], to our knowledge, has been the first study focusing on the analysis of epileptic EEG data using multiway models. In [2], once epilepsy tensors

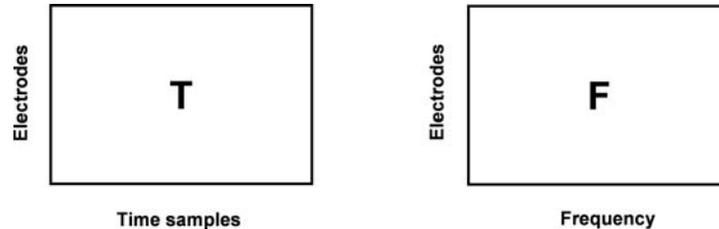


Figure 3.1: Two-way Data Construction.

are constructed, we first explore how linear, multilinear and nonlinear statistical tools perform on the EEG recordings of epilepsy patients in order to have a broader understanding of the structure of epileptic events. In [4], we extend our previous study and develop multiway models to extract and remove artifacts as well as localize epileptic foci. We focus on not only localizing an epileptic focus but also understanding epilepsy seizure structure, its signatures in time, frequency and electrode domains together with those of artifacts. We discuss our findings in these studies in the next sections.

3.2 Epilepsy Tensor Construction

Multi-channel EEG data originally form a matrix of time samples by electrodes. We may analyze the EEG recordings in the time domain, e.g., using the raw data in the time domain or computing the instantaneous signal power for each time sample at each electrode. Another approach would be to use the information in the frequency domain, for instance by computing the signal power at different frequencies through the use of Fourier Transformation (FT)(Figure 3.1). However, neither of these techniques can represent the data in both frequency and time domains.

Even though FT is a widely used technique for frequency spectrum analysis, it is not sufficient to represent the information content of non-stationary signals, e.g., EEG. The Fourier Transform assumes that all frequencies identified in frequency spectrum are available during the whole time duration. However, it is not the case for non-stationary signals. Therefore, we apply continuous wavelet transformation on the signal recorded at each electrode in order to identify the frequency component available at each time sample. As a mother wavelet, we use a Mexican-hat wavelet.

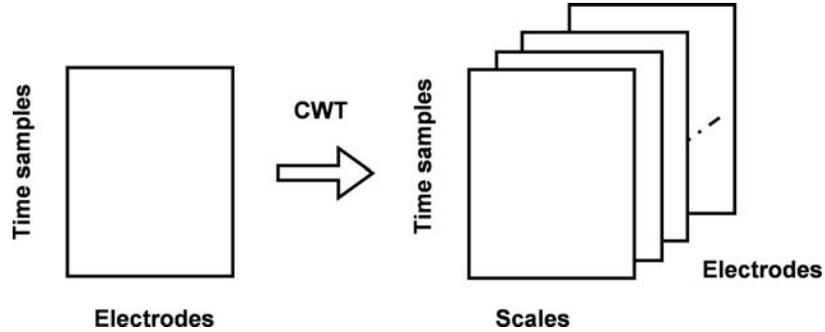


Figure 3.2: Continuous wavelet transformation (CWT) of a signal from a single electrode (data corresponding to the single column of the matrix on the left) forms the frontal slice corresponding to a particular electrode. When the signals from all electrodes are represented in both time and frequency domains through CWT, a third-order tensor with modes time samples, scales (frequency) and electrodes is constructed.

Our selection of the mother wavelet is based on a previous work [81], which shows that a Mexican-hat wavelet captures epileptic events well. It has been shown in a recent study [41] that biorthogonal wavelets also perform well in the localization of epileptic foci.

Wavelet transformation of a signal from a single electrode forms the frontal slice corresponding to a particular electrode (Figure 3.2). Similar to the way described in [90], we rearrange our data, i.e., each seizure of a patient, as a three-way array, $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ with modes: time samples, scales and electrodes (Figure 3.3). Each entry of $\underline{\mathbf{X}}$ denoted by x_{ijk} represents the square of the absolute value of the wavelet coefficient at i^{th} time sample for j^{th} scale at k^{th} electrode, or in other words the power of the wavelet transformed data. Since we use the referential montage by taking Cz as the reference electrode and analyze the electrical potential difference between two electrodes, the third mode actually does not correspond to individual electrodes but rather contains the channels. However, we use the terms electrodes and channels interchangeably throughout this study. Another clarification is regarding to scales and frequencies. Scales and frequencies are also often used interchangeably here. However, it is important to keep in mind that scales are inversely proportional to frequencies based on the relationship given in Equation

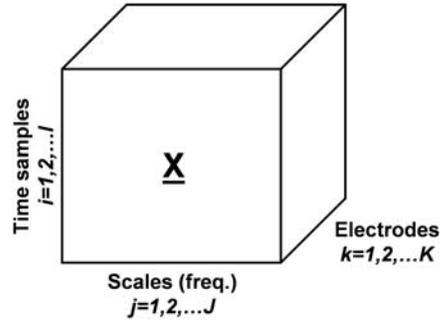


Figure 3.3: Epilepsy Tensor. $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ represents the multi-channel ictal EEG data transformed by continuous wavelet transformation using a Mexican-hat wavelet and arranged as a three-way array. Each entry of $\underline{\mathbf{X}}$, x_{ijk} , corresponds to the square of the absolute value of a wavelet coefficient at i^{th} time sample, j^{th} scale and k^{th} electrode.

3.1.

$$f = \frac{f_C}{a \times \Delta t} \quad (3.1)$$

where f_C is the center frequency of the wavelet, a is the scale, f is the frequency corresponding to the a^{th} scale and Δt is the sampling period.

We name the third-order tensor constructed using multi-channel ictal EEG recordings an *Epilepsy Tensor* and use this structure for further analysis on understanding epilepsy seizure structure in the next sections.

3.3 Artifact Extraction, Removal and Seizure Localization

Once multi-channel ictal EEG recordings are rearranged as an epilepsy tensor, we explore the characteristics of epileptic seizures by identifying individual brain dynamics that take place during seizures. While we are primarily interested in the localization of epileptic foci, in addition to that, our goal is to identify spatial, spectral and temporal signatures of an epileptic seizure as well as those of an artifact. With a goal of analyzing an epileptic seizure in these three domains, i.e., time, frequency and electrode, our contributions are as follows:

- **Epileptic focus localization:** We model epilepsy tensors using a PARAFAC model and use PARAFAC components in time, frequency and electrode domains

to define a seizure. We localize a seizure origin based on the spatial signature of a seizure extracted by a PARAFAC model and identified by a neurologist.

- **Artifact Extraction:** We extract artifacts using a PARAFAC model and use PARAFAC components as spectral, spatial and temporal signatures of an artifact in order to define an artifact.
- **Artifact Removal:** Through multilinear subspace analysis, we remove artifacts such as eye movements so that the remaining data do not contain any activity correlated with the artifact.

As a preliminary study, we have also applied a Tucker3 model with orthogonality constraints on the component matrices [2]. However, the justification of orthogonality constraints meaning that a neural activity is orthogonal to another neural activity is unclear in neuroscience. Furthermore, Tucker3 has rotational ambiguity and those two properties: rotational ambiguity and unnecessary orthogonality constraints, which did not make much sense in terms of the interpretation of brain dynamics, were the underlying reasons of the introduction of the TC model in [94] at the first place. In addition to that, the interpretation of a Tucker3 model is much harder than that of a PARAFAC model. The intuition behind modeling an epilepsy tensor using a PARAFAC model, on the other hand, is quite straightforward. We assume that when the wavelet-transformed EEG data are modeled using a PARAFAC model, or in other words as a sum of rank-one tensors, each rank-one tensor corresponds to either a seizure or an artifact. Besides, the vectors, which form each rank-one tensor, reveal the signatures of an artifact or a seizure in the time, frequency and electrode domains.

3.3.1 Dataset

We study scalp EEG recordings of 10 seizures from 7 patients with different pathology substrates. Ictal EEG recordings are done with long term video EEG monitoring using scalp electrodes in the epilepsy monitoring units of Yeditepe University Hospital and Marmara University.

Table 3.1: Dataset of multi-channel ictal EEG. 100 scales corresponding to frequencies in the frequency band of 0.5-50Hz are used.

SID ¹	PID ¹	Samp. Freq. (Hz.)	Duration (sec.)	Downsamp. Factor	Number of Electrodes	Size of Epilepsy Tensor
1	1	200	47	10	17	$940 \times 100 \times 17$
2	2	200	100	10	17	$2000 \times 100 \times 17$
3	3	200	61	10	18	$1220 \times 100 \times 18$
4	4	200	60	10	18	$1200 \times 100 \times 18$
5	4	200	74	10	18	$1480 \times 100 \times 18$
6 ²	4	200	63	10	18	$1260 \times 100 \times 18$
7 ²	4	200	76	10	18	$1520 \times 100 \times 18$
8	5	400	86	20	18	$1720 \times 100 \times 18$
9	6	200	17	1	18	$3400 \times 100 \times 18$
10	7	400	77	10	18	$3080 \times 100 \times 18$

¹ PID and SID stand for patient id and seizure id, respectively.

² These are the seizures for which artifact removal is applied.

The duration of ictal EEG corresponding to each seizure, sampling frequencies and the number of electrodes are given in Table 3.1. The raw data (*time samples* \times *electrodes* matrix) are first centered across time samples mode and scaled within the electrodes mode before constructing the tensor using continuous wavelet transform. Once the data corresponding to each seizure are rearranged as a third-order tensor, the data are then scaled in scales mode in order to capture the activity in all frequencies rather than only at low frequencies with relatively much higher energy than higher frequencies. Since the duration of an epileptic seizure ranges between 17 sec. and 100 sec. for the seizures given in Table 3.1 and sampling rates are either 200 Hz or 400 Hz, the dimensionality of each epilepsy tensor in the first mode differs from one seizure to another. In order to reduce the complexity of the analysis, after wavelet transformation, we downsample the wavelet coefficients in the time samples mode. The downsampling factor is 10 for most of the seizures. In the case of seizure 8, since both the sampling rate is high and the seizure is long, we pick 20 as the downsampling factor. On the other hand, for seizure 9, we do not downsample the data at all since the seizure is already quite short. Even though we use different downsampling factors for different seizures, we can still localize the seizure. Nevertheless, dependence of the downsampling factor and how we downsample can be further explored in future studies. During our analysis, we use MATLAB’s Wavelet

Toolbox for continuous wavelet transform, PLS_Toolbox for multiway models and EEGLab [42] for topographic maps across the scalp.

3.3.2 Artifact Extraction

Once the three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ with modes: time samples, scales and electrodes, is constructed and preprocessed, we model $\underline{\mathbf{X}}$ using an R -component PARAFAC model as in Equation 2.7. Here we go back to the underlying principle of a PARAFAC model, i.e., parallel proportional profiles [30]. The idea behind parallel proportional profiles is that if the same factors are present in two samples under different conditions, then each factor in the first sample is expected to have the same pattern in the second sample but profiles of the factors will be scaled depending on the conditions. When we take a closer look at the idea of parallel proportional profiles, we can observe that a signal from an electrode can be referred to as a sample. These samples are generated by certain underlying sources with spectral, spatial as well as temporal signatures specific to the sources. Each electrode, thus, has a coefficient representing the contribution of the source to the signal (or sample) recorded at that particular electrode. Our aim is to identify the sources, such as an eye artifact, a muscle artifact or an epileptic activity generating a seizure, based on these signatures and relative coefficients of electrodes.

An R -component PARAFAC model on $\underline{\mathbf{X}}$ extracts the components \mathbf{a}_i , \mathbf{b}_i and \mathbf{c}_i , for $i = 1, 2, \dots, R$, where these components indicate the signatures of sources in time, frequency and electrode domains, respectively as shown in Figure 3.4. Consequently, a PARAFAC model can serve as an artifact extraction method by identifying patterns indicative of artifacts. In Figure 3.4, the signatures captured by the first component characterize an eye-artifact. \mathbf{a}_1 indicates the time points the artifact takes place. \mathbf{b}_1 shows that the eye-artifact observed at the specified times (those with high coefficients in \mathbf{a}_1) has a high-scale signature indicating a low-frequency content (1.25-2.5Hz). Finally, \mathbf{c}_1 localizes the artifact around electrodes F_{P1} and F_{P2} . Based on the visual analysis of EEG recordings, neurologists identify the time and the location of an artifact. We observe that the signatures extracted by the model match with the clinically identified time and location of an artifact. The

model also gives us information about the spectral properties of an artifact. In our study, we observe that most artifacts have low frequency content. Some other examples of extracted artifact signatures are also given in Figure 3.8.

3.3.3 Seizure Origin Localization

By pursuing the same discussion on Cattell’s idea, when one of the underlying sources in the signals recorded by the electrodes is an epileptic seizure, we can argue that one or more of PARAFAC components can model a seizure in the same way an artifact is modeled. Similar to an artifact, a seizure also has a signature in time, frequency and electrode domains. Once these signatures are extracted using a PARAFAC model, the signature of a seizure in the electrode domain can be used to localize the seizure origin.

Therefore, we can also employ PARAFAC as a model for localizing a seizure origin. We observe in Figure 3.4 that the second component in time samples mode, \mathbf{a}_2 , shows an ongoing activity in an ictal period. When the second component in the second mode, \mathbf{b}_2 , is examined, we detect that this ongoing activity in the ictal period takes place in low-scales indicating a rather high-frequency content (12.5-25Hz) compared to that of the first component. Eventually, \mathbf{c}_2 suggests that the activity with described characteristics takes place particularly around electrodes T_4 and T_6 . In fact, this activity is a seizure and the component of a PARAFAC model in the electrodes mode localizes the seizure origin. These conclusions are also drawn based on the clinically identified seizure onset. Since the seizure origin is identified by neurologists as T_4 and T_6 , we expect to observe high coefficients corresponding to these electrodes in the spatial signature of the seizure (We illustrate more examples of seizure origin localization in Figure 3.8). Furthermore, the temporal signature corresponding to this activity should have an ongoing activity characterized by high-coefficients all through the seizure period. The components extracted by the model have these characteristics and therefore, they are considered as the signatures of a seizure. Seizures are also often observed to have relatively higher frequency content compared to the artifacts we have observed.

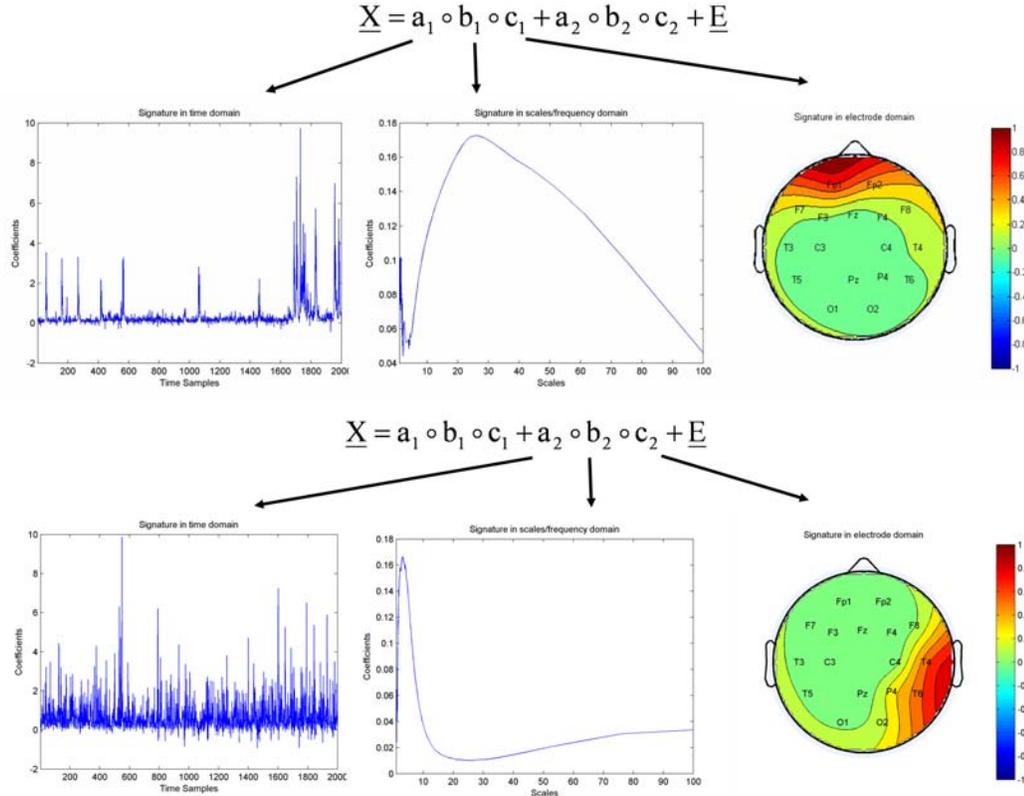


Figure 3.4: A 2-component PARAFAC model on an epilepsy tensor \underline{X} for a particular seizure. We demonstrate the modeling of an epilepsy tensor by a 2-component PARAFAC model, where the first component corresponds to an eye-artifact while the second component represents a seizure. Top: Temporal (\mathbf{a}_1), spectral (\mathbf{b}_1) and spatial (\mathbf{c}_1) signatures of an eye-artifact. \mathbf{a}_1 represents the coefficients of time samples, \mathbf{b}_1 represents the coefficients of scales. Since there is a peak in higher scales on the plot of \mathbf{b}_1 , it indicates that this artifact takes place at lower frequencies. \mathbf{c}_1 contains the coefficients of electrodes. These coefficients are demonstrated on a colormap using EEGLab [42]. Bottom: Temporal (\mathbf{a}_2), spectral (\mathbf{b}_2) and spatial (\mathbf{c}_2) signatures of a seizure. Similar to the first component, \mathbf{a}_2 represents the coefficients of time samples, \mathbf{b}_2 represents the coefficients of scales. There is a peak in lower scales on the figure corresponding to \mathbf{b}_2 , which indicates that the seizure takes place at higher frequencies. Finally, \mathbf{c}_2 contains the coefficients of electrodes, which are used to localize the seizure around electrodes T_4 and T_6 .

3.3.4 Artifact Removal

When we model the epilepsy tensor using a PARAFAC model, in some cases we cannot capture the seizure but rather observe only artifacts. In those cases, we suggest that the variation due to artifacts are removed from the data and a new tensor, $\underline{\mathbf{X}}_{NEW}$, representing the data contaminated with less artifacts is modeled using a PARAFAC model in order to localize seizures.

In order to understand the underlying structure of the data, we model $\underline{\mathbf{X}}$ using a Tucker3 model because a Tucker3 model, unlike the PARAFAC model, is known to reflect the main subspace variation in each mode assuming a multilinear structure. We fit a Tucker3 model as in Equation 3 with large number of components in each mode such that we extract enough components to capture most of the variation in the data (around 75%). Using a Tucker3 model with orthonormality constraints in each mode, we model the data with component matrices \mathbf{A} , \mathbf{B} and \mathbf{C} corresponding to time samples, scales and electrode modes, respectively and having orthonormal columns. Components in all modes are extracted in decreasing order of captured variance just like in SVD on matrices. Then based on visual inspection of the components in the electrodes mode, first N components with characteristics of a potential artifact are identified. Our goal is to remove the activity associated with these potential artifacts. Similar to the underlying idea in interference subtraction based on subspace analysis in [101], we make use of multilinear subspace analysis to remove the artifacts. We project the data onto the space orthogonal to the space spanned by the components characterizing an artifact. The steps of the artifact removal method are described more formally as follows:

1. Fit a Tucker3 model to $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ with component numbers large enough to capture most of the variation in data and extract the component matrices with orthonormal columns. (Suppose that modes of $\underline{\mathbf{X}}$ are as given in Figure 3.3).
2. Pick N components, which are identified as potential artifacts by visual inspection, in electrodes mode.
3. Form matrix $\mathbf{Q} \in \mathbb{R}^{K \times N}$ with N columns using the N components picked in

Step 2.

4. Construct an orthogonal projector, \mathbf{P} , using the matrix \mathbf{Q} .

$$\mathbf{P} = \mathbf{I} - \mathbf{Q}\mathbf{Q}^T \quad (3.2)$$

where \mathbf{I} denotes the identity matrix.

5. Compute $\underline{\mathbf{X}}_{NEW} \in \mathbb{R}^{I \times J \times K}$, which is the projection of $\underline{\mathbf{X}}$ onto the space orthogonal to the range of matrix \mathbf{Q} :

$$\underline{\mathbf{X}}_{NEW} = \underline{\mathbf{X}} \times_3 \mathbf{P} \quad (3.3)$$

where \times_3 denotes the product of tensor $\underline{\mathbf{X}}$ with matrix \mathbf{P} in the electrodes mode.

This artifact removal scheme takes out the effect of an artifact across all frequencies during an ictal period from $\underline{\mathbf{X}}$. After removing the artifacts, we remodel $\underline{\mathbf{X}}_{NEW}$ using a PARAFAC model and use PARAFAC components to identify the seizure origin and inspect spatial, spectral and temporal signatures of the remaining artifacts and seizure. While the artifact removal process enables the localization of seizures after artifact removal, signatures of seizures in scales mode suggest that seizures have very low frequency component. On the other hand, we consistently observe seizure activities at high frequencies (12.5-50Hz) for the seizures where artifact removal is not needed. The effect of the artifact removal process on modes other than the electrodes mode should be further explored in order to fully understand its side effects. We summarize the whole process of multiway analysis of multi-channel ictal EEG in Figure 3.5.

We compare our artifact removal approach based on Tucker3 analysis with two other artifact removal methods. The first approach constructs the epilepsy tensor but then unfolds the tensor in the electrodes mode and decomposes the unfolded data using SVD; in other words, applies a Tucker1 model instead of Tucker3. The same number of left singular vectors as in the case of a Tucker3 model is then used to construct the projection matrix. Except for using the components extracted by a

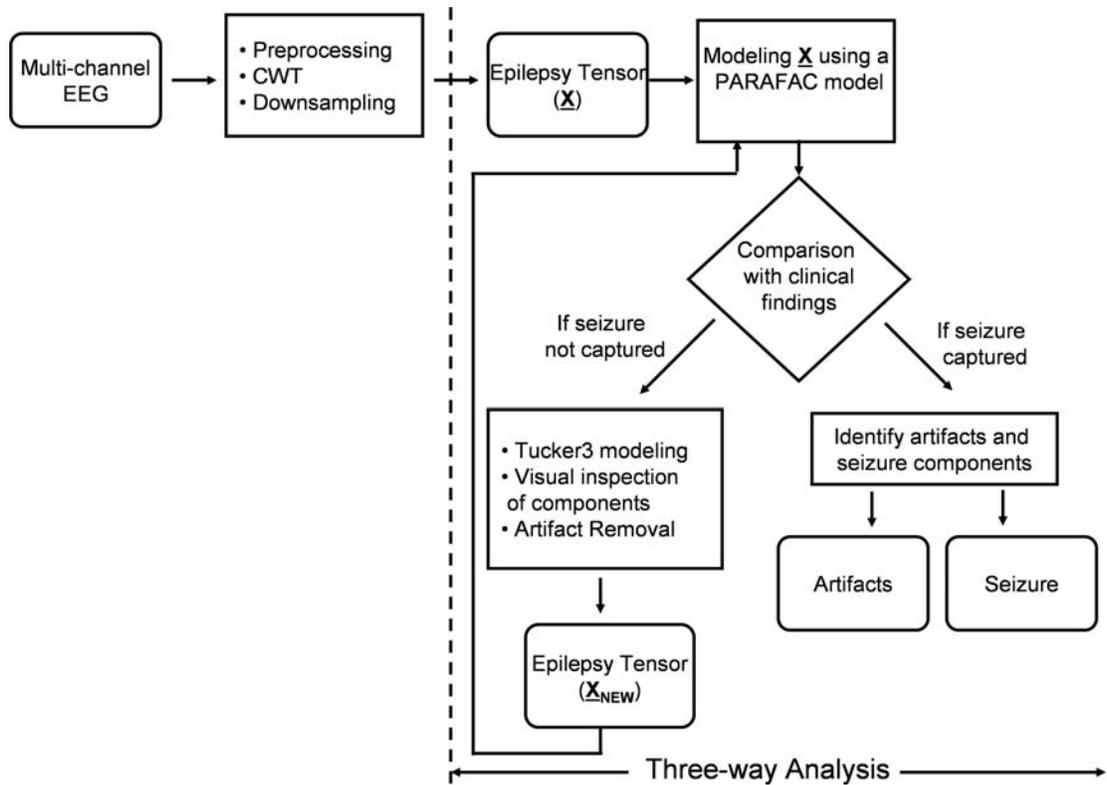


Figure 3.5: Multiway analysis of multi-channel ictal EEG. After the collection of multi-channel EEG data from epilepsy patients, we normalize the data and construct a three-way \underline{X} called an *Epilepsy Tensor* through wavelet transformation. \underline{X} is then downsampled and scaled in scales mode before multiway analysis. Preprocessed three-way array is modeled using a PARAFAC model for artifact extraction and localization of epileptic focus. Finally, PARAFAC components are compared with clinical findings of epilepsy patients. For the cases when a PARAFAC model cannot capture the seizure, we apply an artifact removal method by modeling the preprocessed three-way using a Tucker3 model to detect potential artifacts. The tensor formed after artifact removal is modeled using a PARAFAC model to extract the signatures of an artifact and a seizure.

Tucker1 model, the rest of the artifact removal scheme is the same. After removing the artifacts, we rearrange the data as a tensor and model the new tensor using a PARAFAC model. Another artifact removal approach is to remove the artifacts on the raw multi-channel EEG data. We model the original data matrix of type time samples by electrodes using SVD and construct the projection matrix using the right singular vectors. We extract as many singular vectors as needed to explain 75% of the variation and among those singular vectors, we identify the potential artifacts using visual analysis. The new matrix obtained after removing artifacts is used to construct the epilepsy tensor through continuous wavelet decomposition. The final step is the analysis of the epilepsy tensor using a PARAFAC model.

We compare the results of different artifact removal approaches on two seizures: seizures 6 and 7 from Table 3.1. These are the seizures, on which we need to perform an artifact removal method in order to capture seizures. For seizure 6, Tucker1-based artifact removal approach performs equally well as the Tucker3-based approach in terms of localizing the seizure origin as shown in Figure 3.6. On the other hand, the second seizure, i.e., seizure 7, cannot be localized using the Tucker1-based technique (Figure 3.7). These results support our discussions in Chapter 2 regarding to Tucker1 being less robust and prone to overfitting and modeling noise rather than the exact structure. Furthermore, Tucker3 model would help us more in identifying potential artifacts by providing the signatures of the artifacts in time and frequency domains explicitly. Even though Tucker3 is more complex in terms of interpretation compared to a PARAFAC model, it is still much easier than a Tucker1 model since the components in time and frequency modes are separately extracted by a Tucker3 model. We also demonstrate the results of the approach based on artifact removal from the raw dataset using SVD as the third method in Figure 3.6 and Figure 3.7. For seizure 6, since a 3-component PARAFAC model is more appropriate for the data (with high core consistency, which we discuss next in parameter selection section), we fit a 3-component model and demonstrate the three spatial signatures. None of these correlates with the seizure origin. Similarly, we observe that none of the spatial signatures extracted by a PARAFAC model after the third artifact removal approach corresponds to the spatial signature of seizure

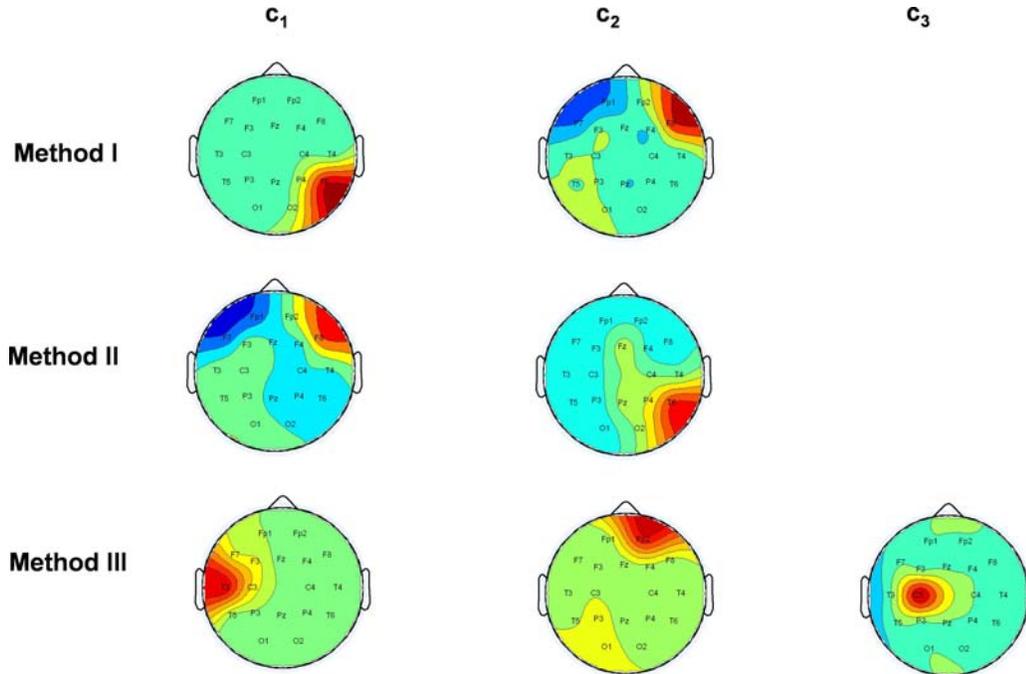


Figure 3.6: The spatial signatures extracted by a PARAFAC model after applying different artifact removal approaches on seizure 6 from Table 3.1. Method I refers to using Tucker3 as the artifact removal approach; method II refers to the artifact removal process based on Tucker1 and method III removes artifacts on raw data using SVD. For method I and II, even though the spatial signatures are not exactly the same, they are very similar for both models and the seizure localization is the same. The first component in method I and the second component in method II localizes the seizure (the components are given in the order of the variation they explain but the explained variation by each component is so close that the ordering flips in different methods). The spatial signatures captured by the PARAFAC model after using SVD-based artifact removal approach on raw data cannot localize the seizure. Color scales in the figures are the same as the scale in Figure 3.4.

7 in Figure 3.7.

3.3.5 Parameter Selection

It is important to extract the right number of components in a multilinear model in order to capture the true underlying structure in data. As we have already mentioned in Chapter 2, there are several techniques for determining the number

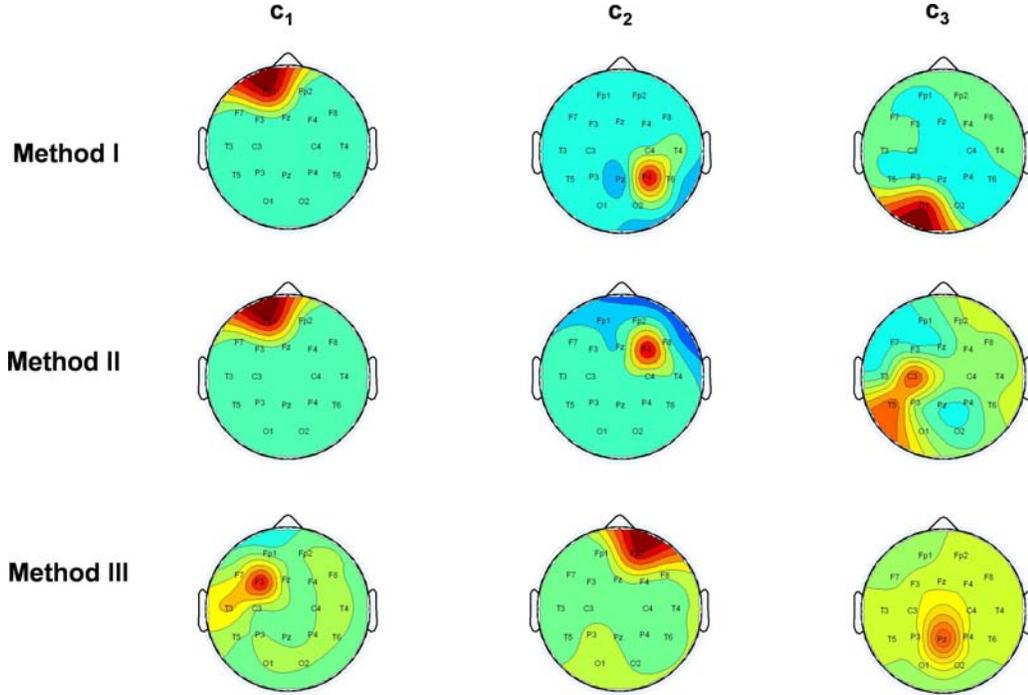


Figure 3.7: The spatial signatures extracted by a PARAFAC model after applying different artifact removal approaches on seizure 7 from Table 3.1. Method I refers to using Tucker3 as the artifact removal approach; method II refers to the artifact removal process based on Tucker1 and method III removes artifacts on raw data using SVD. The second component in method I localizes the seizure origin whereas no other component extracted by other methods can localize the seizure. Color scales in the figures are the same as the scale in Figure 3.4.

of components, e.g., residual analysis, visual appearance of loadings, the number of iterations of the algorithm and core consistency. Among these techniques, in this study, we mostly rely on the core consistency diagnostic [25] for finding the number of components of a PARAFAC model.

The core consistency quantifies the resemblance between a PARAFAC core and a Tucker3 core built based on the PARAFAC component matrices. This diagnostic suggests whether a PARAFAC model with the specified number of components is a valid model for the data. Let $\mathbf{T} \in \mathbb{R}^{R \times R \times R}$ be a super-diagonal PARAFAC core such that $t_{ijk} = 1$ if $i = j = k$, otherwise $t_{ijk} = 0$. Let $\mathbf{G} \in \mathbb{R}^{R \times R \times R}$ be a Tucker3 core

Table 3.2: Core Consistency for different component numbers

Seizure ID	$R = 2$	$R = 3$	$R = 4$	$R = 5$	$R_{selected}$
1	100%	Neg.	-	-	2
2	100%	97%	93%	Neg.	4
3	100%	96%	Neg.	-	3
4	99%	98%	Neg.	-	3
5	100%	92%	83%	64%	4
6	99%	Neg.	-	-	2
7	95%	89%	Neg.	-	3
8	100%	82%	55%	-	3
9	100%	98%	75%	-	3
10	100%	Neg.	-	-	2

*Neg. denotes negative core consistency values and the sign - is used for the component numbers, which are not considered. $R_{selected}$ shows the number of components used in our analysis.

determined using the PARAFAC component matrices, where g_{ijk} can be nonzero for all i, j, k . Then core consistency diagnostic is defined as follows:

$$\text{Core Consistency} = \left(1 - \frac{\sum_{i=1}^R \sum_{j=1}^R \sum_{k=1}^R (g_{ijk} - t_{ijk})^2}{\sum_{i=1}^R \sum_{j=1}^R \sum_{k=1}^R t_{ijk}^2}\right) \times 100$$

In [25], a general rule based on core consistency is introduced. Core consistency above 90% is often used as an indication of the trilinear structure in data and suggests that a PARAFAC model with the specified number of components would be an appropriate model for the data. A core consistency value close to or lower than 50%, on the other hand, demonstrates that a PARAFAC-like model would not be appropriate. This diagnostic has been commonly applied in the neuroscience-multiway literature [4, 46, 90] often together with other diagnostic tools in order to determine the component number.

In this section, we inspect whether core consistency diagnostic serves as a reliable tool for determining the optimal number of components (optimal in terms of interpretation of the data). We model the epilepsy tensor for each seizure with a PARAFAC model using R components, where $R = 2, 3, \dots$, until the core consistency drops considerably. Table 3.2 demonstrates the core consistency values corresponding to different number of components. We often extract as many components as

possible until the core consistency drops because additional components help us capture more variation in data (as long as the captured variation with additional components is significant). For instance, for seizure 10, we can fit a 2-component PARAFAC model with confidence since core consistency indicates that more than two components would not be suitable. On the other hand, for seizure 2, we can choose three or four components. In that case, as we mentioned above, the increase in the explained variation by extracting more components may be used in order to determine the component number. We use all four components and observe that the last two components correspond to artifacts. Table 3.2 shows that core consistency gives fairly good indications of the right number of components. However, it is still important to take into account other diagnostics, e.g., residual analysis, visual appearance of the loadings or any prior knowledge, in order to determine the component number.

3.4 Summary and Discussions

The development of an automated system capable of localizing an epileptic focus would strongly affect the outcome of epilepsy surgeries. Removing or extracting artifacts and exploring the underlying brain dynamics in a seizure are also as crucial as seizure origin localization. They would not only provide accurate focus localization but also improve the understanding of the complex structure of epilepsy, which has not yet been fully discovered.

Ictal scalp EEG recordings are frequently contaminated with movement and muscle artifacts that complicate the analysis of seizure localization. Although scalp EEG recordings have limitations in detection and localization of seizure onset, our multilinear approach based on a multiway model gives us promising results in analyzing seizures and defining seizures and artifacts in time, frequency and electrode domains. These definitions are formed by the spectral, spatial and temporal signatures extracted by multiway analysis of multi-channel EEG data arranged as a three-way array. Not only the detection of the artifacts but also the localization of all seizures are correlated with the clinical findings. Future research directions proposed to improve this study further are discussed in Chapter 5.

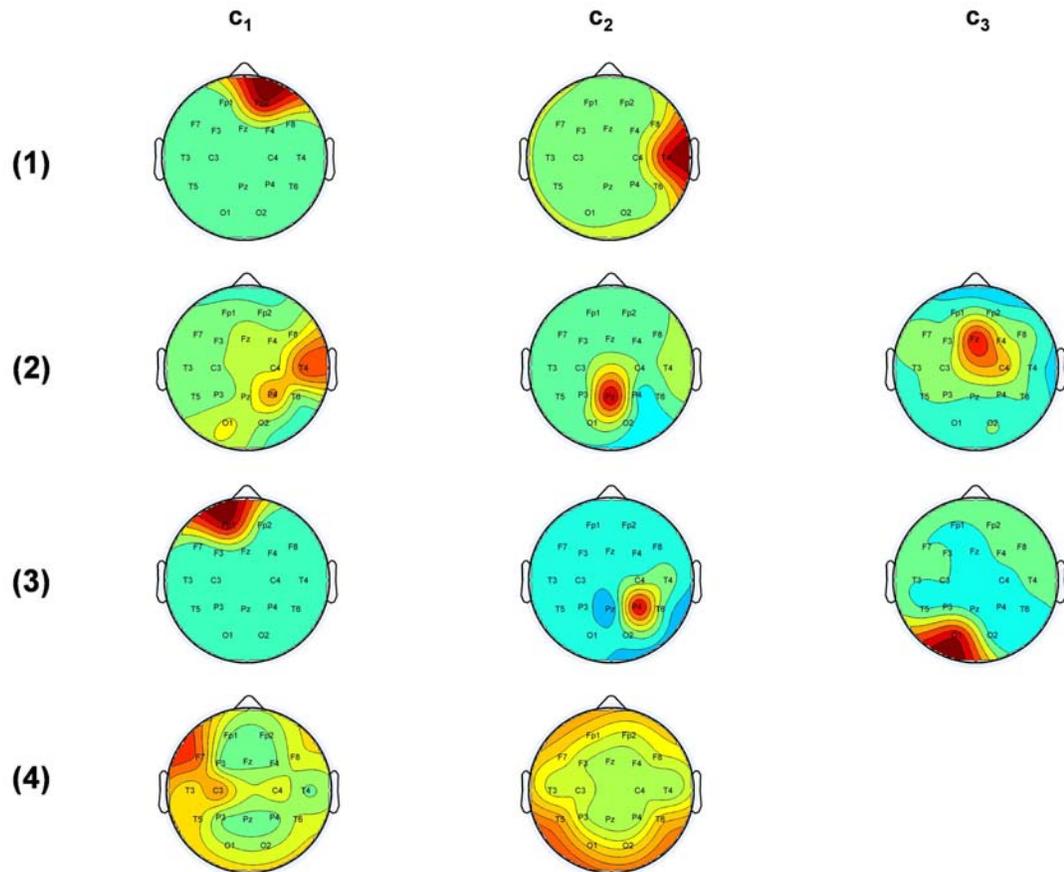


Figure 3.8: Some illustrative examples of artifact extraction and seizure origin localization. We present our results corresponding to the electrode mode for four of the seizures when they are modeled using a PARAFAC model. Color scales in the figures are the same as the scale in Figure 3.4. c_1 , c_2 and c_3 stand for the first, second and third components in the electrodes mode, respectively. (1) Seizure 1. First component represents an eye-artifact while the second component localizes the seizure. (2) Seizure 8. First component shows the seizure origin and the second component corresponds to an artifact, which has a low frequency signature. The third component cannot be visually identified. (3) Seizure 7. This is one of the examples where artifact removal is applied. The components are the PARAFAC components extracted after artifact removal. While the first and third components are the artifacts, the second component represents the seizure. (4) Seizure 10. The first component localizes the seizure around F_7 and C_3 while the second component corresponds to an artifact.

CHAPTER 4

EPILEPTIC SEIZURE RECOGNITION

In the previous chapter, we assume that the time when the seizure starts and ends is already marked by neurologists before the analysis. However, the ideal approach would be to recognize a seizure given any EEG recording. One of the common methods used in clinical evaluations is to visually analyze EEG signals and identify the seizure period. However, visual analysis of EEG signals has some drawbacks. Often EEG signals recorded for several days are scanned visually in their entirety and this is a time-consuming task. Furthermore, it is also subjective and error-prone due to fatigue, etc. Therefore, automation of the detection of the underlying brain dynamics in EEG signals is significant in order to save manpower and obtain robust and objective EEG analysis.

4.1 Related Work

A common approach in seizure recognition/detection and also in prediction is to extract information; in other words, features that can characterize seizure morphologies, from EEG recordings [93,102,105,112,129]. The procedure for feature extraction from multi-channel EEG data is often as follows: First, an EEG signal from a channel is divided into I time epochs (overlapping or non-overlapping) and then J features are extracted from each epoch. Consequently, a signal from a single channel can be represented as a matrix of size $I \times J$ (Figure 4.1-A). A great deal of effort from different disciplines has been invested in exploring the features in order to define the signature of a seizure. These features include statistical complexity measures (e.g., fractal dimension, approximate entropy, Lyapunov exponents, etc.) as well as other features from time (e.g., higher-order statistics of the signal in time domain, Hjorth parameters, etc.) and frequency domains (e.g., spectral skewness, spectral entropy, etc.). A list of features used in characterization of epileptic seizure dynamics can be found in recent studies [93,102,129].

In the literature, studies use either multiple features from a single channel or

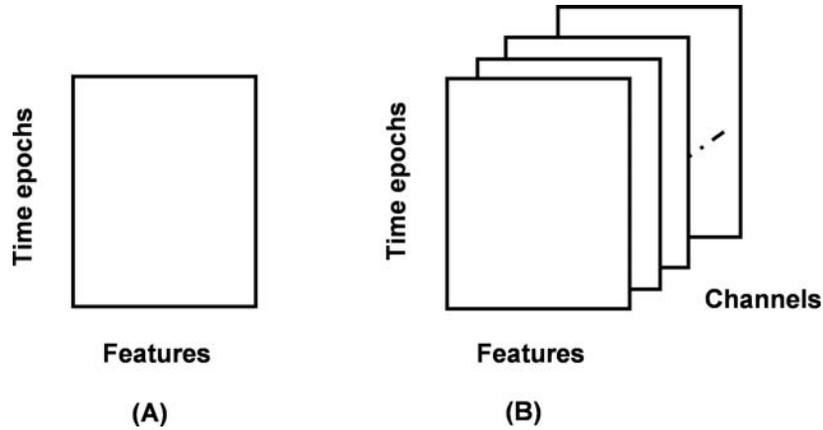


Figure 4.1: (A) Multiple features extracted from single-channel EEG data are arranged as a matrix, (B) When multiple channels are taken into consideration, the data form a three-way array with modes: *time epochs*, *features*, *channels*.

a single feature from multiple channels since data construction and data analysis techniques are often restricted to two dimensions. For instance, in [102], seizure dynamics are analyzed solely on a specific recording, which represents the characteristics of a seizure well. Then the performance of various features from different domains on that particular signal is analyzed simultaneously. On the other hand, [93] analyzes multi-channel EEG data but assesses the performance of each feature one at a time. Furthermore, different studies extract different features and employ different algorithms to distinguish between seizure and non-seizure periods (e.g., [91] and references therein), which makes it difficult to compare the performance of features. An approach capable of simultaneously analyzing features would enable the performance comparison of the features on the same data using the same classifier. Simultaneous analysis of features is also important because it may consider linear or non-linear combinations of features. While a single feature may not be very effective in discriminating between epileptic periods, combinations of several features may well be [55]. Taking into consideration the challenges addressed in the literature, we introduce a multi-modal data construction and analysis approach, which rearranges signals from K channels as a third-order tensor of size $I \times J \times K$ as shown in Figure 4.1-B. We then model the third-order tensor using multilinear discriminant

analysis by facilitating simultaneous analysis of EEG data from multiple channels based on several features from different domains.

In this study, we are particularly interested in distinguishing a seizure (ictal) period from a pre-seizure (pre-ictal) and a post-seizure (post-ictal) period. Moreover, we want to be able to characterize seizures of patients using a subset of features and understand the differences between seizures of different patients. Our ultimate goal is to mark the seizure period but not to predict an upcoming seizure or to detect the seizure onset with minimum delay. This study, therefore, differs from the related work on seizure detection and prediction, e.g., [32, 93, 112] or see [84] for a recent review on seizure prediction. They either focus on the identification of features distinguishing between inter-ictal and pre-ictal periods or aim to detect an epileptic seizure with minimum possible delay using features from a particular domain. Nevertheless, multiway data construction and analysis approaches introduced here can be easily extended to seizure prediction and detection (See Chapter 5 for more discussion).

4.2 Epilepsy Feature Tensor Construction

We introduce a novel approach, which combines the recognition power of several features from different domains by rearranging multi-channel EEG data as a third-order tensor, namely *Epilepsy Feature Tensor* [5, 8], with modes: *time epochs*, *features*, *channels*.

An EEG recording from a single channel is a sequence of time samples. We first divide the signal from each channel into time epochs and inspect whether there are certain underlying dynamics in a particular epoch. This could be achieved by extracting measures that characterize those dynamics. Then each epoch can be represented using a set of measures called *features*. Let $s(j)$ denote the time sample at time j and $\mathbf{s} = \{s(1), s(2), \dots, s(N)\}$ be the time sequence for a particular epoch of length N . We represent each feature as $f_i(\mathbf{s})$, which denotes the i^{th} feature computed on time epoch \mathbf{s} .

Once several features from both the time and frequency domains are extracted from each epoch, we can then represent a signal using a set of feature vectors.

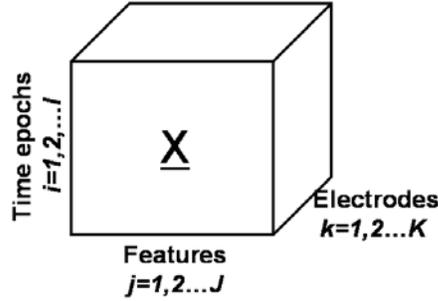


Figure 4.2: Epilepsy Feature Tensor. $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ represents the multi-channel EEG data, which are transformed into the feature space by computing certain measures characterizing seizure dynamics. Each entry of $\underline{\mathbf{X}}$, x_{ijk} , corresponds to the value of j^{th} feature of i^{th} time epoch at k^{th} channel.

Consequently, the signal from a single channel is represented as a matrix, which is one of the frontal slices of a third-order tensor. When we use the data from all channels, we construct a third-order tensor as in Figure 4.2. We do not make any assumptions about the seizure origin but rather construct the dataset using the signals from all channels. In this section, we briefly explain the features we use in constructing epilepsy feature tensors.

4.2.1 Time domain features

4.2.1.1 Hjorth parameters

Hjorth parameters [61] including activity, mobility and complexity are computed as defined in [102] as follows:

$$\begin{aligned}
 \text{Activity} : \quad & f_1(s) = \sigma_s^2 \\
 \text{Mobility} : \quad & f_2(s) = \sigma_{s'}/\sigma_s \\
 \text{Complexity} : \quad & f_3(s) = (\sigma_{s''}/\sigma_{s'})/(\sigma_{s'}/\sigma_s)
 \end{aligned}$$

where σ_s stands for the standard deviation of a time sequence \mathbf{s} ; \mathbf{s}' and \mathbf{s}'' denote the first and second difference of a time series \mathbf{s} , respectively. The d^{th} difference of a time series can be denoted as $(1 - \mathbf{B})^d s(t)$, where \mathbf{B} is the backshift operator. The backshift operator applied to a time sample can be represented as $\mathbf{B}^j s(t) = s(t - j)$ [139].

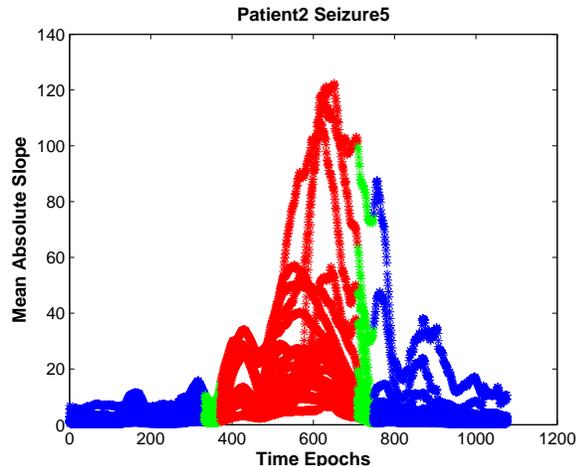


Figure 4.3: Mean Absolute Slope of epochs from all channels for the fifth seizure of the second patient in Table 4.1. Epochs marked with blue and red belong to non-seizure and seizure periods, respectively. Green epochs are the transition epochs from pre-seizure to seizure or seizure to post-seizure periods.

4.2.1.2 Mean Absolute Slope

Absolute slope is calculated using the consecutive differences between time samples in a time sequence: $AS(t) = |s(t+1) - s(t)|$ for each time sample $s(t)$ in a time epoch s [108]. In addition to its simplicity and efficiency, absolute slope can capture both high-amplitude slow and low-amplitude fast activities. We extract the mean of absolute slopes computed for each time sample in a time epoch as the fourth feature, $f_4(s)$ (Figure 4.3). We believe that this feature would be a more reliable feature for intracranial EEG recordings, which are not contaminated with artifacts and a less reliable feature, in our case, for scalp EEG recordings often contaminated with artifacts. However, we have observed that this feature contributes to seizure recognition in almost half of the patients in our dataset [Table 4.4].

4.2.1.3 Spatial Information

During visual analysis, neurologists take into consideration not only the signal from a single channel but also the activity in other channels, especially in the neighboring channels and expect to observe synchronization. Therefore, in order to quantify the similarity between neighboring channels in each time epoch, we first

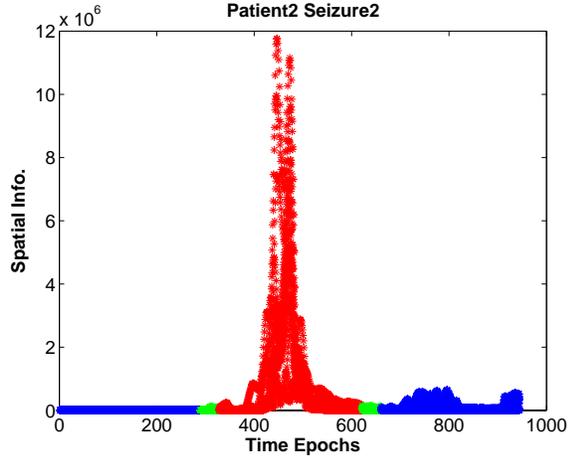


Figure 4.4: Spatial Information of epochs from all channels for the second seizure of the second patient in Table 4.1. Epochs marked with blue and red belong to non-seizure and seizure periods, respectively. We observe a clear increase in similarity between neighboring channels during a seizure period. Green epochs are the transition epochs from pre-seizure to seizure or seizure to post-seizure periods.

define neighbors for each channel and then use the covariance between neighboring channels as a feature (Figure 4.4). Let \mathbf{X} be a matrix of type: time samples by channels, for a particular time epoch \mathbf{s} . We define spatial information, the fifth feature extracted from an epoch \mathbf{s} , for channel i as $f_5(\mathbf{s}, i) = \sum_{j \in NEIGH_i} |\mathbf{C}_{ij}|$, where $NEIGH_i$ contains the neighbors of channel i and \mathbf{C} is the covariance matrix corresponding to the channels in \mathbf{X} .

4.2.2 Frequency domain features

4.2.2.1 Frequency Spectrum

We reduce the time series at least to a mean-stationary time series by taking the first difference of the signal before computing the amplitude spectrum. Then given a time series \mathbf{s} corresponding to a particular epoch, we use a Fast Fourier Transform (FFT) to obtain the Fourier coefficients, c_k , where $c_k = \frac{1}{N} \sum_{t=1}^N s(t) e^{-i \frac{2\pi k}{N} t}$. Based on the Fourier coefficients, we construct the amplitude spectrum using $|c_k|$. The amplitude spectrum is used to extract the sixth feature ($f_6(\mathbf{s})$), which is the median frequency.

4.2.2.2 Relative Energy

We extract a set of features from the energy spread of the signals across different EEG frequency bands. Five frequency bands in accordance with the traditional EEG frequency bands ([104] and references therein) are chosen: δ (0.5 - 3.5Hz), θ (3.5 - 7.5Hz), α (7.5 - 12.5Hz), β (12.5 - 30Hz) and γ (> 30 Hz). We apply continuous wavelet transform between 0.5-50Hz using a Mexican-hat wavelet as the mother wavelet on each epoch. Wavelet coefficients are later used to observe the energy spread across these five frequency bands in each epoch. Let E_f be the estimate of the energy in frequency band f and E_T be the estimate for the total energy in all frequency bands computed as follows:

$$E_f = \sum_{i=1}^N \sum_{j=1}^S |c_{ij}|^2$$

$$E_T = \sum_{f=1}^5 E_f$$

where c_{ij} denotes the wavelet coefficient corresponding to the i^{th} time sample in an epoch and j^{th} scale. N is the length of an epoch and S is the number of scales. The relative energy in each frequency band is then defined as the ratio of the energy in each frequency band to the total energy ($\frac{E_f}{E_T}$).

4.2.2.3 Spectral Entropy

Based on the relative energies computed above, we also compute spectral entropy as a measure used to quantify the uncertainty in the frequency domain. We compute spectral entropy, H , using Shannon's entropy measure [110] as follows:

$$H = - \sum_{f=1}^5 \frac{E_f}{E_T} \log\left(\frac{E_f}{E_T}\right)$$

The list of these features can be easily extended by adding vertical slices to the three-way dataset given in Figure 4.2. Once the tensor corresponding to each seizure of a patient (together with the recordings before and after that particular seizure) is constructed, we build a seizure recognition method using a multilinear regression model called Multilinear PLS, which we explore in detail in the next section.

4.3 Multilinear Regression

Regression models, e.g., multiple linear regression, Partial Least Squares (PLS) and Principal Component Regression (PCR), are commonly applied in prediction and classification problems in diverse disciplines. While these models are employed on datasets of order no higher than two (vectors or matrices), the independent variable in this study, i.e., an epilepsy feature tensor, is a third-order tensor (Figure 4.2). This section briefly describes PLS first and then introduces the regression model, i.e., Multilinear Partial Least Squares (N-PLS), developed for higher-order data analysis.

4.3.1 Partial Least Squares

Partial Least Squares regression, similar to multiple linear regression, aims to build a model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, where $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{y} \in \mathbb{R}^I$ are the independent and dependent variables, respectively. $\mathbf{b} \in \mathbb{R}^J$ contains the regression coefficients and $\mathbf{e} \in \mathbb{R}^I$ is the error term. Here we formulate the model considering that the dependent (or response) variable is a single variable (univariate), which is the case throughout this chapter.

Unlike multiple linear regression, PLS maps the data to a lower-dimensional space and constructs matrix \mathbf{T} whose columns are the extracted factors from original data \mathbf{X} . The model is built using this low-rank approximation of the data. Therefore, the model can be denoted by $\mathbf{y} = \mathbf{T}\mathbf{b} + \mathbf{e}$, where $\mathbf{T} \in \mathbb{R}^{I \times N}$ and $N < J$. In this respect, PLS resembles Principal Component Regression, which maps the data to a lower dimensional space before computing the regression coefficients. However, PCR formulates the mapping without taking into account the dependent variable whereas PLS uses the response variable while extracting the factors from \mathbf{X} . This property makes PLS a more suitable and successful method for classification problems. The underlying objective while extracting the factors is to maximize the covariance between the factors and the dependent variable, or in other words, $\max_w \text{cov}(\mathbf{X}\mathbf{w}, \mathbf{y})$. The steps for extracting the factors satisfying this objective function is given in Algorithm 4. For an in-depth discussion about the algorithm and the formulations of PLS and PCR from an optimization perspective, the reader is referred to [18].

Algorithm 4 PLS($\mathbf{X}, \mathbf{y}, N$)

- 1: $\mathbf{y}_0 = \mathbf{y}, \mathbf{X}_0 = \mathbf{X}$
- 2: **for** $i = 1$ to N **do**
- 3: $\mathbf{W}(:, i) = \mathbf{X}_{i-1}^T \mathbf{y}_{i-1}$
- 4: $\mathbf{T}(:, i) = \mathbf{X}_{i-1} \mathbf{W}(:, i) / \|\mathbf{X}_{i-1} \mathbf{W}(:, i)\|$
- 5: $\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{T}(:, i) \mathbf{T}(:, i)^T \mathbf{X}_{i-1}$
- 6: $\mathbf{y}_i = \mathbf{y}_{i-1} - \mathbf{T}(:, i) \mathbf{T}(:, i)^T \mathbf{y}_{i-1}$
- 7: $\mathbf{y}_i = \mathbf{y}_i / \|\mathbf{y}_i\|$
- 8: $\mathbf{b} = \mathbf{W}(\mathbf{T}^T \mathbf{X} \mathbf{W})^{-1} \mathbf{T}^T \mathbf{y}$

* $\mathbf{A}(:, j)$ represents the j^{th} column of matrix \mathbf{A} . The subscript i in \mathbf{X}_i and \mathbf{y}_i indicates the iteration number.

4.3.2 Multilinear Partial Least Squares

Multilinear PLS is introduced as a generalization of PLS to multiway datasets [19]. This method can handle the situations where dependent and/or independent variables are multiway arrays. In this study, we confine our attention to the case where the independent variable, $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$, is a three-way array of type epilepsy feature tensor and the dependent variable, $\mathbf{y} \in \mathbb{R}^I$, is a vector containing the class assignments of time epochs. Multilinear PLS models the dataset $\underline{\mathbf{X}}$ by extracting a component, $\mathbf{t} \in \mathbb{R}^I$, from the first mode such that $\text{cov}(\mathbf{t}, \mathbf{y})$ is maximized. A pre-defined number of components, N , is extracted iteratively and the matrix $\mathbf{T} \in \mathbb{R}^{I \times N}$, whose columns are the extracted components (\mathbf{t} 's), is constructed. In addition to \mathbf{T} , component matrices, \mathbf{W}^J and \mathbf{W}^K , corresponding to the second and third modes, respectively are also formed. The notation used for the component matrices indicates that a component matrix corresponds to the mode with the dimensionality denoted by the superscript.

The steps of the algorithm are summarized in Algorithm 5. In order to make a comparison between the algorithms for PLS and N-PLS easily, here we include the original N-PLS algorithm introduced in [19]. On the other hand, the tri-linear model fit to the independent data in Step 7 has been later replaced with a Tucker3-like model. This modification does not have any implications in terms of the regression part but rather modifies the way the independent variable is modeled [24]. During our analysis, we use the implementation of N-PLS in PLS_Toolbox (by Eigenvector

Algorithm 5 Multilinear PLS($\underline{\mathbf{X}}, \mathbf{y}, N$)

- 1: $\mathbf{y}_0 = \mathbf{y}, \mathbf{X}_0 = \mathbf{X}_{(1)}$
- 2: **for** $i = 1$ to N **do**
- 3: $\mathbf{z} = \mathbf{y}_{i-1}^T \mathbf{X}_{i-1}$
 Reshape \mathbf{z} as a matrix $\mathbf{Z} \in \mathbb{R}^{J \times K}$ such that $z_{mn} = \mathbf{z}(m + J * (n - 1))$
- 4: {Compute singular value decomposition of matrix \mathbf{Z} }
 $\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}^T$
- 5: $\mathbf{w}^J = \mathbf{U}(:, 1), \mathbf{w}^K = \mathbf{V}(:, 1)$
 $\mathbf{W}^J(:, i) = \mathbf{w}^J, \mathbf{W}^K(:, i) = \mathbf{w}^K$
- 6: $\mathbf{T}(:, i) = \mathbf{X}_{i-1}(\mathbf{w}^K \otimes \mathbf{w}^J)$
- 7: $\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{T}(:, i)(\mathbf{w}^K \otimes \mathbf{w}^J)'$
- 8: $\mathbf{b}_i = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{y}_{i-1} = \mathbf{T}^\dagger \mathbf{y}_{i-1}$
- 9: {Regression and Deflation}
 $\mathbf{y}_i = \mathbf{y}_{i-1} - \mathbf{T}\mathbf{b}_i = (\mathbf{I} - \mathbf{T}\mathbf{T}^\dagger)\mathbf{y}_{i-1}$

* $\mathbf{X}_{(1)}$ stands for the tensor $\underline{\mathbf{X}}$ matricized in the first mode. \mathbf{X}_i indicates matricized data in the first mode updated/deflated by the computation of i components. $\mathbf{A}(:, j)$ represents the j^{th} column of matrix \mathbf{A} . \mathbf{W}^J and \mathbf{W}^K correspond to the component matrices in the second and third mode, respectively. \mathbf{T}^\dagger stands for pseudo-inverse defined as $\mathbf{T}^\dagger = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T$.

Research Inc.), which uses the modified version of the algorithm. The underlying reasons for this modification as well as its advantages are discussed in detail in [24].

Since N-PLS is a regression method and we need a binary classifier to classify time epochs as seizure and non-seizure, we combine N-PLS with linear discriminant analysis (LDA). The procedure for combining N-PLS with LDA is as follows: When we model the training set, $\underline{\mathbf{X}}_{train} \in \mathbb{R}^{I \times J \times K}$, using N-PLS, we extract the component matrices corresponding to each mode of a three-way array. Let $\mathbf{T}_{train} \in \mathbb{R}^{I \times N}$, $\mathbf{W}^J \in \mathbb{R}^{J \times N}$ and $\mathbf{W}^K \in \mathbb{R}^{K \times N}$ be the component matrices corresponding to the first, second and third modes, respectively. We can use this model to predict the labels of the time epochs in other EEG recordings; in other words the labels of the time epochs in our test set, which contains the left-out seizure of a patient in patient-specific seizure recognition (Figure 4.5) or the seizures of the left-out patients in patient non-specific seizure recognition. Let $\underline{\mathbf{X}}_{test} \in \mathbb{R}^{R \times J \times K}$ be a third-order tensor representing the time epochs in our test set. We compute $\mathbf{T}_{test} \in \mathbb{R}^{R \times N}$ using the component matrices \mathbf{W}^J and \mathbf{W}^K extracted from the training set based

on the general formula in Equation 4.1 derived in [115]:

$$\begin{aligned} \mathbf{R} &= [\mathbf{w}_1 (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{w}_2 \dots (\prod_{n=1}^{N-1} (\mathbf{I} - \mathbf{w}_n \mathbf{w}_n^T)) \mathbf{w}_N] \\ \mathbf{T}_{test} &= \mathbf{X}_{test(1)} \mathbf{R} \end{aligned} \quad (4.1)$$

where $\mathbf{X}_{test(1)}$ is the matrix formed by unfolding $\underline{\mathbf{X}}_{test}$ in the first mode and vector \mathbf{w}_i equals to the Kronecker product of i^{th} column of matrices \mathbf{W}^K and \mathbf{W}^J : $\mathbf{w}_i = \mathbf{w}_i^K \otimes \mathbf{w}_i^J$. Once we obtain the t-scores for the epochs in the test set, we can then determine the class (seizure or non-seizure) of each time epoch by comparing \mathbf{T}_{test} with \mathbf{T}_{train} through LDA using a discriminant function as given in Equation 4.2 in [98].

$$L_g(\mathbf{t}) = \log(\pi_g) - \frac{1}{2}(\mathbf{t} - \bar{\mathbf{t}}_g)^T \mathbf{S}_{within}^{-1} (\mathbf{t} - \bar{\mathbf{t}}_g) + \log|\mathbf{S}_{within}| \quad (4.2)$$

$L_g(t)$ is the score for the g^{th} group for a time epoch represented by a column vector $\mathbf{t} \in \mathbb{R}^N$. π_g is a prior probability for each group. A time epoch is assigned to the group which has the highest score. $\bar{\mathbf{t}}_g$ indicates the mean of time epochs in group g and \mathbf{S}_{within} is the pooled covariance matrix computed as follows: $\mathbf{S}_{within} = \frac{1}{I-G} \sum_{g=1}^G \sum_{j=1}^{I_g} (\mathbf{t}_{gj} - \bar{\mathbf{t}}_g)(\mathbf{t}_{gj} - \bar{\mathbf{t}}_g)^T$. I and G represent the number of time epochs and number of groups, respectively, e.g., $G = 2$ for seizure/non-seizure classification. I_g is the number of time epochs in group g and \mathbf{t}_{gj} is the vector corresponding to the j^{th} epoch in group g . The prior probabilities are assumed to be the same. Several other assumptions, e.g., each group following a multivariate normal distribution with identical covariance matrices, are also made at this step to come up with a simpler discriminant function as explained in [77].

4.3.3 Feature Selection

Not every feature in our feature set may be a powerful discriminator between seizure and non-seizure dynamics. Therefore, we identify the significant features for seizure recognition using a variable selection approach.

Our variable selection method is an extension of Variable Importance in Projection (VIP) [140] to three-way datasets. VIP is used in two-way regression analysis and based on the idea of factor models. In linear factor models, several components

summarizing the data are extracted either to explain the variance in the data, e.g., as in PCA, or to capture the correlation between two datasets, e.g., as in PLS or Canonical Correlation Analysis. The components extracted in these linear factor models are linear combinations of the variables in the data. The variable selection method, VIP, computes a VIP-score for each variable in order to quantify a variable's importance by using the coefficient of a variable in each component together with each component's significance in regression. Variables with a VIP-score under a certain threshold are then removed from the data since they are considered insignificant. Let $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{y} \in \mathbb{R}^I$ be the independent and dependent variables. $\mathbf{T} \in \mathbb{R}^{I \times N}$ represents the lower dimensional space \mathbf{X} is mapped to and $\mathbf{b} \in \mathbb{R}^N$ contains the regression coefficients such that we can write $\mathbf{y} = \mathbf{T}\mathbf{b} + \mathbf{e}$ and $\mathbf{X} = \mathbf{T}\mathbf{W} + \mathbf{E}$, where \mathbf{e} and \mathbf{E} contain the residuals. The VIP-score of the i^{th} variable is then calculated as follows [35]:

$$VIP_i = \sqrt{I \times \frac{\sum_{n=1}^N b_n^2 \mathbf{t}_n^T \mathbf{t}_n (w_{in}/|\mathbf{w}_n|)^2}{\sum_{n=1}^N b_n^2 \mathbf{t}_n^T \mathbf{t}_n}}$$

where \mathbf{w}_n and \mathbf{t}_n correspond to the n^{th} column of matrix \mathbf{W} and \mathbf{T} , respectively and w_{in} is the entry in the i^{th} row of the n^{th} column of matrix \mathbf{W} . b_n is the regression coefficient for the n^{th} component; in other words, the n^{th} entry of vector \mathbf{b} .

Similarly, in N-PLS we extract component matrices corresponding to each mode of a higher-order dataset. Each column of a component matrix contains the coefficients corresponding to the variables in a specific mode and represents a component, which is a linear combination of the variables. Let the independent and dependent variables be $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ and $\mathbf{y} \in \mathbb{R}^I$, respectively and let $\mathbf{T} \in \mathbb{R}^{I \times N}$, $\mathbf{W}^J \in \mathbb{R}^{J \times N}$ and $\mathbf{W}^K \in \mathbb{R}^{K \times N}$ be the component matrices corresponding to the first (time epochs), second (features) and third (channels) modes. In the computation of VIP scores for variables in one mode of a three-way array, we replace matrix \mathbf{W} with the component matrix in the mode where we select variables, in our case with \mathbf{W}^J corresponding to the features mode. In addition, we compute matrix \mathbf{F} , where

$\mathbf{F} = \mathbf{X}_{(2)}^T \mathbf{W}^J$ and use the columns of matrix \mathbf{F} , i.e., \mathbf{f}_n , instead of t-scores.

$$VIP_i = \sqrt{I \times \frac{\sum_{n=1}^N b_n^2 \mathbf{f}_n^T \mathbf{f}_n (w_{in}^J / |\mathbf{w}_n^J|)^2}{\sum_{n=1}^N b_n^2 \mathbf{f}_n^T \mathbf{f}_n}}$$

Since the average of squared VIP scores equals 1, a general criterion for variable selection is to select the variables with VIP score greater than 1. On the other hand, we just want to remove insignificant variables and include most of the variables contributing to seizure recognition in our analysis. Therefore, we lower the threshold to 0.7 and set the threshold to the same value for all patients.

When we analyze epilepsy feature tensors with N-PLS, we have the chance to select features independent of the channels because N-PLS models the data by constructing different component matrices for each mode. On the other hand, if we matricized an epilepsy feature tensor, then we would obtain a matrix of *time epochs* by *features – channels*. In that case, we would not be able to select only features but we would rather need to select a feature from a particular channel since each variable would be a combination of features and channels.

4.4 Patient-Specific Seizure Recognition

We build our patient-specific seizure recognition model on a training set constructed using all but one seizure of a patient. Once the training set is formed, the training set is centered across the time epochs mode and scaled within the features mode before the analysis since features have different ranges of magnitudes (See Figure 4.3 and Figure 4.4). Centering/scaling a three-way array is different than centering/scaling two-way datasets as we have already mentioned in Chapter 2. Before the analysis, the dependent data are also centered. We then regress the data for all the seizures in the training set onto the \mathbf{y} -vector using Multilinear PLS regression and build a model based on Algorithm 5.

4.4.1 Dataset

Our dataset contains multi-channel scalp EEG recordings of 32 seizures from 9 patients suffering from focal epileptic seizures. Multi-channel scalp EEG signals

from epilepsy patients with at least three recorded seizures are included in our analysis. The EEG data have been collected via scalp electrodes in the epilepsy monitoring units of Yeditepe University Hospital and Albany Medical College. The recording of EEG with referential electrode Cz is used for computational analysis. The number of seizures per patient as well as sizes of epilepsy feature tensors with modes: time epochs, features and channels, are given in Table 4.1. EEG recordings are not preprocessed to remove artifacts. The data for one of the patients are sampled at 200Hz and the data for other patients are sampled at 400Hz. Before the analysis, we apply a bandstop filter at 50 Hz and its harmonics (for the data from Yeditepe University) and at 60Hz and its harmonics (for the data from Albany Medical College) to remove the artifacts from the power source.

The data corresponding to a seizure of a patient contain a certain amount of data before the seizure, the seizure period and a certain amount of data after the seizure period. Each signal is divided into epochs of 10 seconds. Consequently, each epoch typically contains 2000 or 4000 samples depending on the sampling frequency. The epochs are formed using a sliding window approach such that consecutive epochs differ only in 100 samples. For each epoch we compute the features given in Table 4.4 and a matrix of size *number of time epochs* \times 7 is created for the signal from a single channel. When all channels are included in the analysis, this forms a three-way array of *number of time epochs* \times 7 \times 18 for each seizure (Figure 4.5).

The ictal period is visually identified by neurologists based on EEG seizure onset for each seizure of a patient. In accordance with the markings, the epochs are divided into two classes: epochs that belong to the seizure period and the ones outside the seizure period. The dependent variable, i.e., \mathbf{y} -vector in Algorithm 5, corresponding to the time epochs mode of an epilepsy feature tensor is then constructed such that: $y_i = 1$ if i^{th} epoch is outside the seizure period and $y_i = 2$ if i^{th} epoch belongs to the seizure period. Since epochs are formed using a sliding window approach, some epochs contain samples from both pre-seizure and seizure periods or both seizure and post-seizure periods. These epochs are excluded from training and test sets so that the performance of the model is not affected by epochs containing the characteristics of different seizure dynamics.

Table 4.1: EEG Dataset for Patient-Specific Seizure Recognition. The last column shows the size of the epilepsy feature tensor with modes: time epochs, features and channels. Each tensor contains a seizure as well as data before and after that seizure. The number of epochs (first mode) in each tensor differs depending on the duration of a seizure.

Patient ID	Seizure ID	Size of Epilepsy Feature Tensor
1	1	$302 \times 7 \times 18$
	2	$386 \times 7 \times 18$
	3	$320 \times 7 \times 18$
	4	$398 \times 7 \times 18$
	5	$444 \times 7 \times 18$
2	1	$878 \times 7 \times 18$
	2	$866 \times 7 \times 18$
	3	$902 \times 7 \times 18$
	4	$986 \times 7 \times 18$
	5	$998 \times 7 \times 18$
3	1	$790 \times 7 \times 18$
	2	$746 \times 7 \times 18$
	3	$1034 \times 7 \times 18$
4	1	$1174 \times 7 \times 18$
	2	$1346 \times 7 \times 18$
	3	$1170 \times 7 \times 18$
5	1	$62 \times 7 \times 18$
	2	$74 \times 7 \times 18$
	3	$458 \times 7 \times 18$
6	1	$226 \times 7 \times 18$
	2	$186 \times 7 \times 18$
	3	$186 \times 7 \times 18$
	4	$186 \times 7 \times 18$
7	1	$638 \times 7 \times 18$
	2	$630 \times 7 \times 18$
	3	$578 \times 7 \times 18$
8	1	$866 \times 7 \times 18$
	2	$1082 \times 7 \times 18$
	3	$842 \times 7 \times 18$
9	1	$1442 \times 7 \times 18$
	2	$1814 \times 7 \times 18$
	3	$1294 \times 7 \times 18$

4.4.2 Results and Interpretations

We determine the performance of the model for a patient by computing the average performance over all seizures of that patient. We build a training set using all but one seizure of a patient. We use the training set to determine the number of components in N-PLS and also to select a subset of features. We then test the model on the left-out seizure of that particular patient.

As a performance evaluation criterion, we use the geometric mean of sensitivity and specificity, which is called g-means defined as $g = \sqrt{\text{sensitivity} \times \text{specificity}}$ [80]. Sensitivity indicates the proportion of the true-positives to the sum of true-positives and false-negatives, where true-positives are the time epochs that belong to the seizure period and are classified as seizure; false-negatives are the seizure epochs that are classified as non-seizure. Specificity, on the other hand, is the ratio of true-negatives to the sum of true-negatives and false-positives, where true-negatives are the time epochs that belong to non-seizure period and are classified as non-seizure; false-positives are the non-seizure epochs classified as seizure.

Table 4.2 demonstrates the performance of the model on nine patients, who have at least three recorded seizures. We show the average g-means for each patient both with feature selection and without feature selection. We observe that feature selection is especially useful for Patient 4, 5 and 6 to detect seizures. For instance in Patient 5, who has three seizures, first two seizures are not detected at all without feature selection and this results in very poor performance (since sensitivity is 0 for the undetected seizures). On the other hand, when we select a subset of features based on the EEG signals of the patient in the training set, we refine the model and detect all seizures of the patient with average g-means around 83%.

Table 4.4 shows the subset of features used in seizure-recognition for each patient. Since we form training sets by leaving-out one seizure at a time, different features can be selected from each training set. The features given in Table 4.4 correspond to the union of subsets of features selected from each training set. These subsets of features can be further used to understand the differences between patients. For instance, different seizure locations may result in differences in the features used for seizure recognition. Nevertheless, we should point out that feature

selection may also result in overfitting the seizures in the training sets. Therefore, in the cases where there is variation among seizures of a patient, feature selection may degrade the performance.

Furthermore, we assess the performance of the multi-modal data construction and modeling approach by comparing its performance with that of a two-way classification model. We unfold the epilepsy feature tensor in the time epochs mode as shown in Figure 2.3 and then use SVM [130] to classify epochs as seizure and non-seizure. Similarly, [112] has previously proposed a patient-specific seizure detection model by representing each time epoch with a feature vector and then classifying the time epochs using SVMs. When we unfold the epilepsy feature tensor in the time epochs mode, we have $7 \times 18 = 126$ features corresponding to each time epoch. We employ SVM² to classify the time epochs based on those 126 features. For each patient, we build a patient-specific model using all but one seizure of a patient and then test the model on the left-out seizure and recordings before and after that particular seizure. After each seizure is left-out once, we compute the average performance of the model for each patient. We use radial basis function kernel with a parameter adjusted for each patient. The parameter for each patient is determined using cross-validation on the training set similar to the way the number of components for an N-PLS model is determined (discussed in Parameter Selection in Section 4.6).

Table 4.2 demonstrates the performance of seizure recognition using a two-way approach for each patient. We observe that while SVM has a fairly good performance in terms of seizure detection, for the cases when it performs poorly, our multi-modal approach using feature selection improves the performance of the model. For example, in Patient 5, two-way analysis approach cannot detect one of the seizures at all and this results in low average g-means while NPLS+LDA with feature selection can capture all seizures. By preserving the multi-modality of the data, multiway data analysis keeps the model simple and makes the interpretation easier so that we can easily select features, which in turn would improve the performance resulting in some cases in much better performance than SVM.

²Implementation of support vector machines called *SVM^{light}* [65] is used in the analysis.

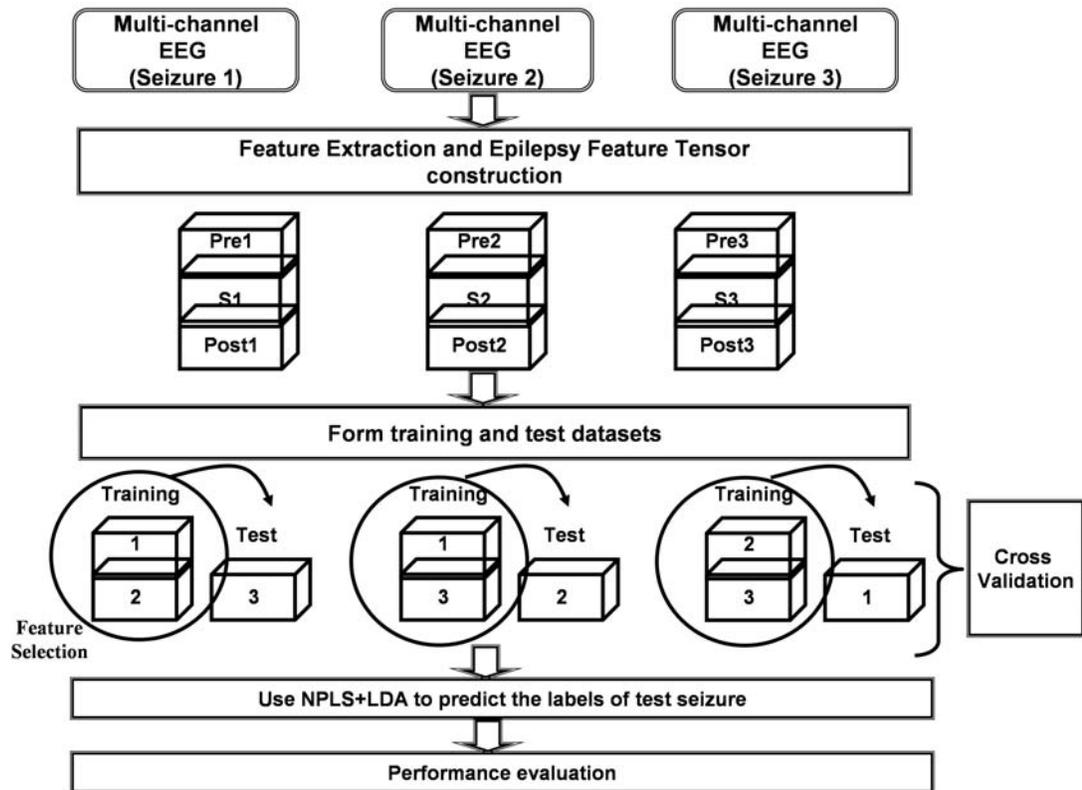


Figure 4.5: Patient-Specific Seizure Recognition Model. Multi-channel EEG signals corresponding to the data before, during and after each seizure of a patient are arranged as a third-order epilepsy feature tensor. Then training and test sets are constructed by leaving out one seizure (together with data before and after that seizure period) at a time. The model built on the training set is used to predict the labels of the time epochs in the test set using N-PLS and LDA. Final step is performance evaluation using the average performance of the model on test sets.

Table 4.2: Seizure vs. Non-seizure. Performance of three-way (NPLS-based) and two-way (SVM-based) approaches in terms of geometric mean of sensitivity and specificity of the model. The row corresponding to NPLS + LDA shows the results without feature selection while the row corresponding to NPLS + LDA (FS) demonstrates the results of the model with feature selection.

Seizure vs. Non-seizure	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6	Patient7	Patient8	Patient9
NPLS + LDA	85.3%	97.6%	91.3%	75.0%	28.6%	72.3%	97.0%	86.0%	84.5%
NPLS + LDA (FS)	86.6%	96.7%	91.1%	77.3%	83.1%	89.3%	92.1%	78.4%	77.4%
SVM	86.9%	98.6%	98.4%	76.3%	44.8%	88.3%	98.2%	94.0%	93.4%

Table 4.3: Pre-seizure vs. Post-seizure (binary classification within non-seizure epochs). Each entry shows the performance of the model when it is trained on non-seizure epochs before/after some seizures of a patient and tested on non-seizure epochs before/after another seizure of that particular patient.

Pre-seizure vs. Post-seizure	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6	Patient7	Patient8	Patient9
NPLS + LDA	93.1%	98.0%	94.2%	88.4%	8.1%	64.7%	85.9%	85.3%	75.8%

Table 4.4: Subsets of features used in the patient-specific seizure recognition model of each patient. Patient 1, 2, 7 and 8 have right temporal seizures. Patient 3 suffers from left frontal while Patient 4 and 9 suffer from left temporal seizures. Patient 5 is bilateral central frontal and Patient 6 is bilateral occipital. While subsets of features tend to be similar based on seizure origins, it is not possible to make generalizations on a small set of patients.

PID ¹	Activity	Mobility	Complexity	Mean Abs. Slope	Spatial Info	Median Freq.	Spectral Entropy
1	✓	✓	✓	✓	✓	×	✓
2	✓	✓	✓	✓	✓	×	✓
3	✓	✓	✓	×	✓	×	✓
4	✓	✓	✓	×	✓	×	✓
5	✓	✓	✓	×	×	×	✓
6	✓	✓	✓	×	✓	×	×
7	✓	✓	✓	✓	✓	×	✓
8	✓	✓	✓	✓	✓	×	×
9	✓	✓	×	×	✓	×	×

¹ PID stands for patient id.

4.5 Patient Non-Specific Seizure Recognition

In patient-specific seizure recognition, several seizures of a patient need to be recorded first in order to construct a training set. Only after we have those prior seizures, we can build a model specific to that particular patient. On the other hand, the ideal and practical approach is to build a model on seizures of some patients, which have been already recorded and then use that model to recognize seizures of new patients. This approach is called patient non-specific seizure recognition. Patient non-specific seizure recognition is more challenging than patient-specific cases considering that patients suffer from seizures with different morphologic and topographic characteristics and training on one type and testing on another may not perform well. In this section, we develop a patient non-specific seizure recognition model and emphasize the changes we make to the patient-specific seizure recognition approach to model the inter-patient variation.

4.5.1 Dataset

Our dataset contains 26 seizures from 9 patients suffering from temporal seizures. Six patients have right temporal seizures while the seizures of the remaining three patients are left temporal. The number of seizures for each patient, sizes of epilepsy feature tensors and the lateralization of the epileptic focus are given in Table 4.5. We choose only the patients with temporal seizures for our analysis and include both left and right temporal seizures in order to explore whether the location of an epileptic focus affects the performance of the model. The EEG data have been recorded in epilepsy monitoring units of Yeditepe University Hospital and Albany Medical College using the referential montage according to Cz electrode. Similar to the case in patient-specific seizure recognition in the previous section, we filter the noise from the power source by applying bandstop filters at 50Hz and its harmonics for the data from Yeditepe University Hospital and at 60 Hz and its harmonics for the data from Albany Medical College. Apart from this filtering process, no other method is used to remove artifacts such as eye blinks, eye movements or muscle artifacts.

The same approach as in patient-specific seizure recognition is used for constructing an epilepsy feature tensor for each seizure of a patient including data from pre-seizure and post-seizure periods in addition to the seizure itself. However, there are two main differences in the construction of epilepsy feature tensors.

4.5.1.1 Additional Features

The first difference in tensor construction is that we include four features quantifying the relative energy in different frequency bands in addition to the seven features used in the previous section. The frequency spectrum is divided into five bands, which are defined in Section 4.2. We include the relative energies in θ , α , β and γ bands. The first band corresponding to 0-3.5Hz interval is ignored since it has been discussed in [105] that non-seizure sleep EEG can be frequently available between 0-3 Hz. Consequently, the dimensionality of the tensors in the second mode given in Table 4.5 is 11 instead of 7.

We should point out that it is possible to use these features while building

the patient-specific seizure recognition model as well. We have observed that even though the performance of the model for some patients increases, the average performance over all patients drops when these features are included in patient-specific seizure recognition. On the other hand, these additional features improves the performance of patient non-specific seizure recognition on the sample dataset given in Table 4.5. Therefore, we use seven of the eleven features for patient-specific seizure recognition while we use all eleven features for patient non-specific seizure recognition. As we have already mentioned, the feature sets used in building these models can be easily modified by adding/removing vertical slices to/from epilepsy feature tensors.

4.5.1.2 Handling inter-patient differences

The second difference is the way the tensor is preprocessed before the analysis. One of the main challenges in developing a patient non-specific seizure recognition model is to handle the inter-patient variation. We have observed that orders of magnitude of features differ remarkably from one patient to another. When we normalize features in order not to just model the features with higher magnitudes but to give equal importance to every feature, seizures of some patients are not well-modeled and become unrecognizable unless we take into consideration the variation between different patients.

We have identified that activity and spatial information are the two features that have quite different orders of magnitude across different patients as illustrated in Figure 4.6. In order to alleviate this problem, we take the log of the features instead of using their actual values. Another way around this problem would be to redefine each feature in a relative fashion; relative in the sense we define relative energies in different frequency bands if possible or relative based on some reference signal specific to a patient [105].

4.5.2 Results and Interpretations

We evaluate the performance of our patient non-specific seizure recognition model by training the model on the seizures of all but one patient. Then we test the model on the seizures of the left-out patient and compute the average performance

Table 4.5: EEG Dataset for Patient Non-Specific Seizure Recognition. Each epilepsy feature tensor contains a seizure as well as data before and after that seizure. The number of epochs (first mode) in each tensor differs depending on the duration of a seizure.

Patient ID	Seizure ID	Size of Epilepsy Feature Tensor	Epileptic Focus
1	1	$1382 \times 11 \times 18$	RT ¹
2	1	$302 \times 11 \times 18$	RT
	2	$386 \times 11 \times 18$	
	3	$320 \times 11 \times 18$	
	4	$398 \times 11 \times 18$	
	5	$444 \times 11 \times 18$	
3	1	$878 \times 11 \times 18$	RT
	2	$866 \times 11 \times 18$	
	3	$902 \times 11 \times 18$	
	4	$986 \times 11 \times 18$	
	5	$998 \times 11 \times 18$	
4	1	$1174 \times 11 \times 18$	LT ²
	2	$1346 \times 11 \times 18$	
	3	$1170 \times 11 \times 18$	
5	1	$638 \times 11 \times 18$	RT
	2	$630 \times 11 \times 18$	
	3	$578 \times 11 \times 18$	
6	1	$866 \times 11 \times 18$	RT
	2	$1082 \times 11 \times 18$	
	3	$842 \times 11 \times 18$	
7	1	$626 \times 11 \times 18$	LT
	2	$734 \times 11 \times 18$	
8	1	$1442 \times 11 \times 18$	LT
	2	$1814 \times 11 \times 18$	
	3	$1294 \times 11 \times 18$	
9	1	$1238 \times 11 \times 18$	RT

¹ RT stands for right temporal.

² LT stands for left temporal.

of the model for that patient. We determine the number of components extracted by N-PLS using cross-validation on the training set by leaving out one seizure at a time. The component number, which gives the best overall performance in the training set, is selected to build the model. As in patient-specific seizure recognition, we preprocess the training set first and then preprocess the test set accordingly. However, unlike in patient-specific case, we should keep in mind that the training

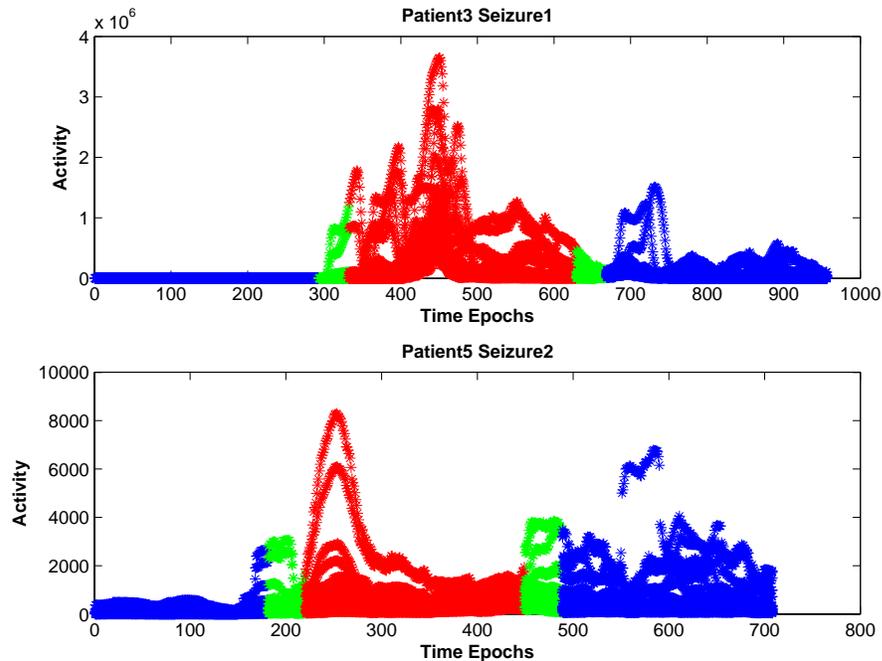


Figure 4.6: The figures show how one of the features, i.e., activity, behaves during seizures of some patients from Table 4.5. The epochs represented in blue and red correspond to the non-seizure and seizure periods while the epochs in green contain time samples from both seizure and non-seizure periods. We observe that there is significant order of magnitude difference between seizures of different patients.

set contains seizures of different patients and this may have complications as we further discuss in the next paragraph.

The geometric mean of sensitivity and specificity is used as the performance evaluation criterion. The second column of Table 4.6 shows the performance of the patient non-specific seizure recognition model. While we can successfully classify seizure and non-seizure epochs for most of the patients, we observe that the performance of the model is not good for Patient 4, 8 and 9. Our hypothesis is that these failures are due to scale differences between different patients but not due to, for instance, different lateralizations of an epileptic focus. In order to show the underlying reason for low performance, we individually preprocess each patient in such a way that the seizures of each patient are separately centered across time epochs

Table 4.6: Seizure vs. Non-seizure. Performance of three-way (NPLS-based) approach in terms of the geometric mean of sensitivity and specificity of the model. The different columns correspond to different preprocessing techniques explained in the text.

Patient ID	NPLS+LDA (v1)	NPLS+LDA (v2)	NPLS+LDA (v1+heuristic)
1	87.7%	81.4%	89.0%
2	88.8%	87.0%	90.2%
3	83.7%	88.1%	85.2%
4	41.9%	84.7%	39.8%
5	95.6%	94.9%	96.9%
6	95.6%	92.6%	95.6%
7	94.5%	88.1%	95.6%
8	69.6%	71.1%	69.3%
9	68.1%	72.8%	68.6%
MEAN	80.6%	84.5%	81.1%

mode and scaled within features mode. Consequently, seizures in the test set are not preprocessed according to the training set. When we build and test the model after such a preprocessing approach, the performances given in the third column of Table 4.6 are obtained. We observe a dramatic increase for some patients, e.g., the performance of Patient4 increases from about 42% to 85%. Figure 4.7 further illustrates the reason of this increase in performance. When we use the training set containing seizures of other patients to preprocess the seizures of a patient in the test set, we can only partially recognize a seizure as shown on the top figure in Figure 4.7. On the other hand, if we individually preprocess each patient using his/her own seizures, we then recognize most of the seizure and sensitivity of the model increases resulting in a better performance in terms of g-means. Even though we get lower performance for some patients with the first approach, we still use this approach but also make use of individual preprocessing in order to understand the underlying reason of low performances in some patients. We emphasize here that low performances are due to the order of magnitude differences among patients. Log transformation of some features improves the performance considerably but further improvements are still needed in order to fully handle inter-patient variation.

Even though g-means is used to assess the performance of a seizure recognition

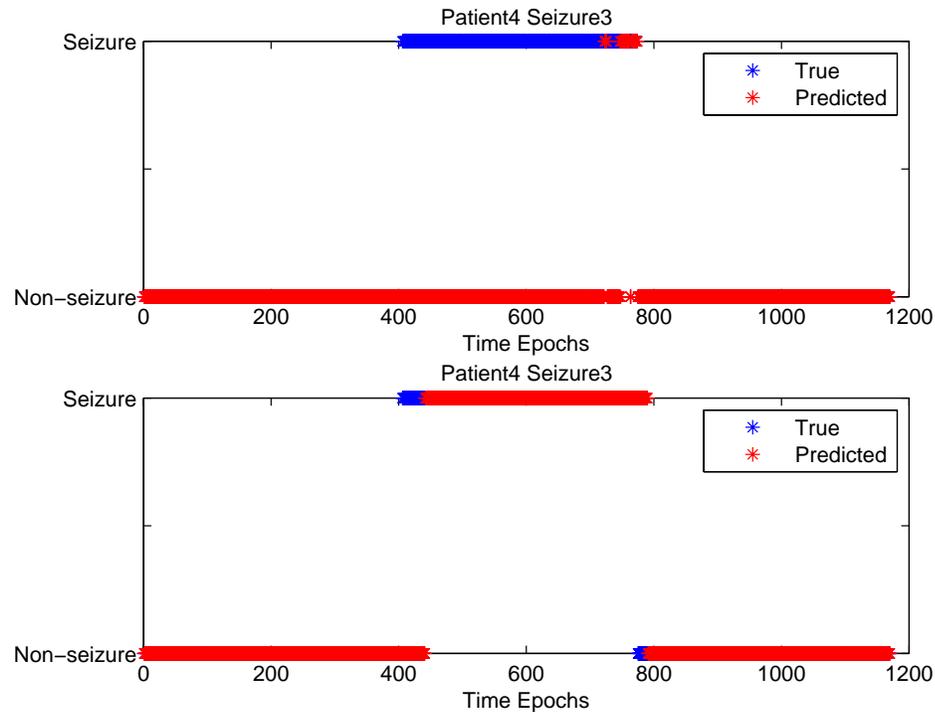


Figure 4.7: Performance of the patient non-specific seizure recognition model with different preprocessing approaches for the fourth patient in Table 4.5. Blue and red dots correspond to true and predicted classes of time epochs, respectively. The figure at the top shows that we can only partially detect the seizure; in other words we can only classify the epochs at the end of the seizure as seizure when we preprocess the test set according to the whole training set containing seizures of different patients (the second column of Table 4.6). On the other hand, if we preprocess each patient separately, then we observe that we can detect almost the whole seizure period with a short delay at the beginning of the seizure (the third column of Table 4.6).

model in this study, specificity is often the main concern. Seizure recognition models with high false-positive rates resulting in unnecessary seizure notifications would lose reliability over time and high false-positive rate is the main problem of current seizure detection and prediction techniques. When we further explore our model, we observe that the model sometimes wrongly predicts one or two isolated epochs as seizure. We can eliminate these false-positives using a heuristic, which suggests that it is not seizure until, for instance, ten consecutive epochs are classified as seizure.

In [112] a similar approach is used to handle the same problem by classifying an epoch as seizure only if three consecutive 2-second epochs are classified as seizure. We demonstrate the performance of the model in the fourth column of Table 4.6 when the original model (performance of which is given in the second column of Table 4.6) is modified using such a heuristic. The performance for most patients slightly increases since the isolated epochs wrongly-classified as seizure are converted to non-seizure epochs. Even though there is no significant performance change, it may be essential to use such a heuristic in order to increase the reliability of a model.

4.6 Parameter Selection

In both patient-specific and patient non-specific seizure recognition, our model is based on N-PLS. As seen in Algorithm 5, the number of components in N-PLS, N , is a user-defined parameter. In order to determine N , we use cross-validation on the training set. Each seizure in the training set is left out once and tested for different number of components ranging from 1 to pre-defined maximum number of components. We then compare the predictions obtained by the model for all seizures in the training set with their actual labels. The component number, which gives the best overall classification performance in terms of both sensitivity and specificity, is selected to build the model to be used on the test set.

In addition to the number of components N , there are other parameters to be determined in our analysis. For instance, we set the duration of an epoch to 10 seconds. It has been set to different values in the literature, e.g., 1 second [102], 2 seconds [105,112], 10 seconds [129] and around 20 seconds [93]. Besides, the duration of overlap between consecutive epochs, the maximum number of components in an N-PLS model and the threshold for a VIP score for feature selection are some of the other user-defined parameters. In this study, for each parameter we use the same value for each patient. In future studies, the sensitivity of the performance of the model on each patient to each one of these parameters should be explored further.

4.7 Summary and Discussions

Epileptic seizures are often detected by visually analyzing large amounts of EEG data recorded over nights and days. Therefore, a system automatically marking seizures would save manpower as well as remove the subjectivity of the visual process. We introduce a multi-modal data construction and analysis approach for seizure recognition using multi-channel scalp EEG signals. Multi-modality of the data enables us to represent EEG signals from multiple channels using various features from different domains as a third-order tensor called an Epilepsy Feature Tensor with modes: time epochs, features and channels. We analyze these multiway arrays using a multilinear discriminant analysis based on N-PLS in order to classify time epochs as seizure or non-seizure. We develop both patient-specific and patient non-specific seizure recognition models and our results demonstrate that multiway data analysis can detect seizures with promising performance. Furthermore, multiway models also have the potential to improve our understanding of different seizure structures by giving us the chance to compare seizures of patients through the features used in seizure recognition.

Our datasets contain data before and after each seizure as well as the seizure itself and we evaluate the performance of our model using out-of-sample test sets. Another good test case is to use inter-ictal data, which contain inter-ictal epileptiform waves but do not have any epileptic seizures. Neurologists visually analyze these signals to decide whether an outpatient may have an epileptic disorder. We make use of such a dataset of 4-minute long from Patient 7 in Table 4.5 as a control dataset and use our model trained on all patients except for Patient 7 to predict the labels of the epochs in this dataset. We do not get any false-positives and the label of each epoch in the test set is successfully predicted as non-seizure. Nevertheless, our inter-ictal control data are quite short and the performance of the model would be better assessed if it was thoroughly tested on continuous inter-ictal EEG [84].

We have tried to extract various features that can differentiate between seizure and non-seizure periods. However, it is important to emphasize that what we call a set of non-seizure epochs is not homogeneous and contains epochs from both pre-seizure and post-seizure periods. In other words, there are two different classes

reflecting different brain dynamics within the set of non-seizure epochs. Therefore, we also explore whether the features, which we use in representing our EEG signals, can capture the differences between pre-seizure and post-seizure periods. Table 4.3 shows that if we only analyze the data from pre-seizure and post-seizure periods, we can classify epochs into pre-seizure and post-seizure classes with very high performance for most of the patients. These results suggest that pre-seizure and post-seizure periods are not very similar in terms of the features we use in our analysis. With a goal of identifying whether the subset of features used for differentiating between seizure and non-seizure differs from the subset of features distinguishing between pre-seizure and post-seizure, we model the data for each patient using N-PLS enabling feature selection. In order to capture the features differentiating between pre-seizure and post-seizure, we exclude the seizure part of the data and just focus on modeling pre-seizure and post-seizure epochs. The results reveal that the subsets of features for these two classification problems are not distinct. Consequently, we cannot, for instance, exclude the subset of features differentiating between pre-seizure and post-seizure epochs from the feature set in order to get better classification of seizure and non-seizure epochs because same features may play an important role in both classification problems. Figure 4.8 illustrates one of the features, i.e., mean absolute slope, across different epochs: pre-seizure, seizure and post-seizure. We can see that it can discriminate between seizure and non-seizure as well as pre-seizure and post-seizure. In future studies, extracting features, which are different only in seizure period but behave quite similarly in pre-seizure and post-seizure periods, may improve the performance of the model.

In [91], the performance of different approaches in seizure detection has been summarized by presenting the classification accuracies given in the literature for the publicly available EEG dataset described in [14]. We would like to point out that comparison of our results with those would be misleading due to major differences in the type of the data. In this study, we aim to differentiate between non-seizure and seizure phases using multi-channel EEG data recorded *extracranially* (scalp EEG). We have also mentioned that non-seizure phases correspond to pre-seizure and post-seizure periods. Therefore, our goal is to mark the seizure period. On the other

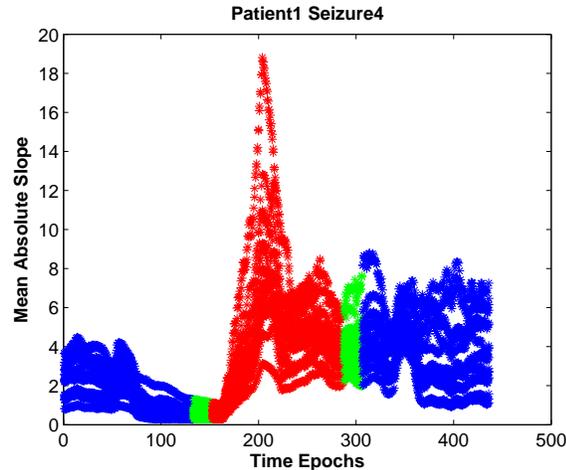


Figure 4.8: The figure illustrates how differently mean absolute slope behaves in pre-seizure, seizure and post-seizure periods for one of the seizures of Patient 1 from Table 4.1. This feature, therefore, can differentiate between seizure and non-seizure as well as pre-seizure and post-seizure epochs. The epochs represented in blue and red correspond to the non-seizure and seizure periods while the epochs in green contain time samples from both seizure and non-seizure periods.

hand, in previous work ([91] and references therein), even if the problem definition is presented as the differentiation of non-seizure and seizure periods, the concept of non-seizure is defined differently. Epochs that belong to a non-seizure period include seizure-free data from healthy patients recorded extracranially as well as seizure-free data from epilepsy patients recorded *intracranially* (iEEG). Consequently, in our case, it is more challenging to differentiate a few seconds before and after a seizure period from the seizure compared to differentiating EEG of a healthy patient from the seizure. Besides, we obtain these results using scalp EEG recordings.

Final remark is regarding to the computational complexity of the construction of epilepsy feature tensors and their analysis using multiway analysis techniques. The performance bottleneck in this approach is the construction of an epilepsy feature tensor, in particular the computation of the features, which are based on continuous wavelet transform. Spectral entropy and relative energies in different frequency bands are computed using continuous wavelet transform and its computational complexity is $O(N \log N)$ per scale or $O(N)$ per scale at best [135], where

N is the size of an epoch. This computation is repeated for each epoch from each electrode resulting in high time complexity compared to the computations of other features, which are linear in the size of an epoch. Since our main concern has not been the computational complexity but rather the performance of the model in terms of sensitivity and specificity, we have focused on a variety of features regardless of their computational complexities and regardless of the computational complexity of a specific implementation. Nevertheless, in order to apply this approach on very large datasets or in real-time, special attention should be paid to the computational complexity of each feature included in the analysis.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Previous chapters outline the underlying multi-modal approaches we propose to automate the analysis of epileptic EEG signals with a goal of detecting seizures, localizing seizure origins and understanding seizure dynamics. In this chapter, we give an outline of future research directions that should be explored further.

5.1 Epileptic Seizure Localization

We have introduced our multi-modal seizure localization approach, which rearranges multi-channel ictal scalp EEG data as a third-order Epilepsy Tensor with *time samples*, *frequency* and *channels* modes in order to automatically localize a seizure origin. We then model the third-order tensor using a PARAFAC model. After modeling the data using a PARAFAC model, we have so far relied on the clinical feedback from neurologists to identify whether a rank-one tensor corresponds to a seizure or an artifact. On the other hand, for a fully automated seizure localization system, an automated approach should be developed to identify the rank-one tensor corresponding to the seizure and then use its component in the channels mode to localize the seizure origin.

Another observation in our study is that while the seizure localization is restricted to a smaller area and the concordance with visual analysis is high in some patients, e.g., patients with tumor, lateralization is well-defined but localization is more widespread in some other patients, e.g., patients with mesial temporal sclerosis (MTS). This observation may emphasize the limitations of scalp EEG recordings in terms of spatial resolution since seizures spread to a larger area by the time they reach the scalp in MTS patients and that wider area is what scalp EEG recordings can capture. It would be interesting to look for the correlation between the accuracy of the models and types of epilepsy patients in more detail using a larger set of patients with different etiological pathologies.

Finally, we would like to point out that although multiway analysis methods

are powerful enough to capture the multilinear structure of the data, the models can only extract factors that are linear combinations of the features. Therefore, they cannot capture the nonlinear relationships among samples. Since current nonlinear methods are limited to two-way arrays, nonlinear structures in multiway arrays cannot be captured by classical nonlinear analysis techniques, either. With a goal of understanding the relationship between electrodes better, we believe that our preliminary studies on the analysis of EEG signals by combining kernel methods with a multilinear model [2] should be studied more in depth. This approach has also recently been applied in another discipline, i.e., image synthesis and recognition, by applying HOSVD in feature space [85].

5.2 Epileptic Seizure Recognition

We have constructed multi-modal datasets from multi-channel scalp EEG recordings using various features from different domains. These datasets have been analyzed using models based on N-PLS, which is a generalization of PLS to higher-order datasets in order to automatically mark an epileptic seizure by differentiating between seizure and pre-seizure/post-seizure periods. We have also compared multi-modal approaches in patient-specific seizure recognition with a two-way approach based on SVMs and discussed their performance in terms of interpretation and classification. However, available multiway models, e.g., multilinear PLS, cannot capture the nonlinear relationships between the variables unlike SVM-based techniques capable of capturing the nonlinearity in the data using kernels. Therefore, if mapping the features into a higher dimensional space improves the performance of the models, multiway models may also be extended to nonlinear multiway models, either by mapping the data to the feature space explicitly or incorporating the kernel matrix into the algorithms and the design of the models.

Besides, in this study we have combined N-PLS regression with LDA in order to classify time epochs as seizure or non-seizure. Nevertheless, this approach can be considered to be an ad-hoc approach, which initially solves a regression problem and then combines it with a discriminant analysis. An alternative approach would be to replace the least squares loss function in N-PLS with another loss function in

order to solve the classification problem in a single step generalizing the approach introduced in [92] to higher-order datasets.

Furthermore, since there are order of magnitude differences in the ranges of some features used in constructing our datasets, we have employed log transformation for handling the inter-patient variation in order to successfully recognize seizures by training the model on seizures of some patients and testing on seizures of other patients. We have showed that the performance would be much better if we do not have that kind of variation between patients at the first place by individually preprocessing each patient. Approaches other than log transform that could work as well as individually preprocessing each patient should be developed. Once this is achieved, the next step should be to understand the occurrence of false-positives better, inspect what kind of activities on raw EEG data may result in false-positives in the model and how those can be prevented. We also want to point out that we have used a specific montage in our analysis but the performance of different montages may vary and it would be interesting to assess the performance of different montages.

Another extension of our study in Chapter 4 should be to have a larger set of patients suffering from seizures with localization other than temporal lobes. We have limited our analysis to right and left temporal seizures for patient non-specific seizure recognition and we have not identified any dependence on localization. Nevertheless, for instance, frontal seizures or seizures with different lateralizations should be included in the dataset and the performance of the model should be assessed using a larger set of patients with different characteristics.

As we have also mentioned at the end of Chapter 4, a good way to evaluate the performance of the model is to use continuous multi-channel inter-ictal EEG datasets as the control data. Consequently, this suggests that for a thorough evaluation of a model, we need an annotated data archive representing different aspects of possible epileptic seizures. It is also important to have shared databases so that the performance of different models can be compared on the very same dataset. The efforts for creating such databases have been underway by the organizers of International Workshop on Epileptic Seizure Prediction. The fourth workshop in

this series will be held in 2009 [147].

5.3 Epileptic Seizure Prediction

Our multi-modal data construction and analysis approach can also be applied for epileptic seizure prediction. Unlike seizure recognition, where our goal is to differentiate between non-seizure and seizure periods, the focus in seizure prediction is to distinguish between inter-ictal and pre-ictal periods. The aim is to predict seizures as early as possible in order to be able to develop a warning system for upcoming seizures. Together with neurologists, the pre-ictal period can be marked to a certain duration prior to the seizure onset. The model can then be trained on some inter-ictal and pre-ictal data of a patient and later used to predict other seizures of that particular patient (patient-specific seizure prediction). However, the ideal case would again be to predict the seizures of a patient by training the model on seizures of other patients (patient non-specific seizure prediction). Evaluation of a seizure prediction system is more challenging than seizure recognition since false-alarms mean unnecessary warnings of a seizure and they are more critical. Therefore, the system should have high sensitivity, very low false-alarm rate as well as a well-defined prediction horizon.

This seizure prediction approach can also be extended to real-time seizure prediction by building a seizure prediction model on a training set offline and then using this model to monitor and process EEG signals in real-time. The computationally expensive part of real-time seizure prediction is to build the model on a large set of training data. Since this part can be performed offline, the model can be efficiently used to predict the labels of the new data stream.

However, proving that this approach is better than any other approach proposed in the literature would be challenging. It should be demonstrated that the new approach can provide lower false-alarm rates and has a better temporal resolution.

LITERATURE CITED

- [1] Acar E., Çamtepe S. A., Krishnamoorthy M. and Yener B., Modeling and Multiway Analysis of Chatroom Tensors, In Proc. of IEEE International Conference on Intelligence and Security Informatics, *Lecture Notes in Computer Science*, Springer Verlag, 2005, 3495: 256-268.
- [2] Acar E., Bingöl C. A., Bingöl H. and Yener B., Computational Analysis of Epileptic Focus Localization, In Proc. of the 4th IASTED International Conference on Biomedical Engineering, 2006, pp. 317-322.
- [3] Acar E., Çamtepe S. A. and Yener B., Collective Sampling and Analysis of High Order Tensors for Chatroom Communications, In Proc. of IEEE International Conference on Intelligence and Security Informatics, *Lecture Notes In Computer Science*, Springer Verlag, 2006, 3975: 213-224.
- [4] Acar E., Bingöl C. A., Bingöl H., Bro R. and Yener B., Multiway Analysis of Epilepsy Tensors, *Bioinformatics*, 2007, 23(13): i10-i18.
- [5] Acar E., Bingöl C. A., Bingöl H., Bro R. and Yener B., Seizure Recognition on Epilepsy Feature Tensor, In Proc. of the 29th Annual Conference of IEEE Engineering in Medicine and Biology Society, 2007, pp. 4273-4276.
- [6] Acar E. and Yener B., Unsupervised Multiway Data Analysis: A Literature Survey, *RPI Technical Report 07-06*, Rensselaer Polytechnic Institute, 2007.
- [7] Acar E., Bro R. and Schmidt B., New Exploratory Clustering Tool, *Journal of Chemometrics*, 2008, 22(1): 91-100.
- [8] Acar E., Bingöl C. A., Bingöl H., Bro R., Ritaccio A. L. and Yener B., Modeling and Detection of Epileptic Seizures using Multi-modal Data Construction and Analysis, *RPI Technical Report 08-02*, Rensselaer Polytechnic Institute, 2008.
- [9] Andersen A. H. and Rayens W. S., Structure-seeking multilinear methods for the analysis of fMRI data, *Neuroimage*, 2004, 22(2): 728-739.
- [10] Andersen C. M. and Bro R., Practical Aspects of PARAFAC Modeling of Fluorescence Excitation-Emission Data, *Journal of Chemometrics*, 2003, 17(4): 200-215.
- [11] Andersson C. A. and Bro R., Improving the speed of multi-way algorithms: Part I. Tucker3, *Chemometrics and Intelligent Laboratory Systems*, 1998, 42(1-2): 93-103.

- [12] Andersson C. A. and Bro R., Improving the speed of multi-way algorithms: Part II. Compression, *Chemometrics and Intelligent Laboratory Systems*, 1998, 42(1-2): 105–113.
- [13] Andersson C. A. and Bro R., The N-way Toolbox for MATLAB, *Chemometrics and Intelligent Laboratory Systems*, 2000, 52(1): 1-4.
- [14] Andrzejak R. G., Lehnertz K., Mormann F., Rieke C., David P. and Elger C. E., Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity: Dependence on Recording Region and Brain State, *Physical Review E*, 2001, 64(6): 061907.
- [15] Bader B. W. and Kolda T. G., Algorithm 862: MATLAB Tensor Classes for Fast Algorithm Prototyping, *ACM Transactions on Mathematical Software*, 2006, 32(4): 635-653.
- [16] Bader B. W. and Kolda T. G., Efficient MATLAB computations with sparse and factored tensors, *SIAM Journal on Scientific Computing*, 2006, 30(1): 205–231.
- [17] Bader B. W., Harshman R. A. and Kolda T. G., Temporal Analysis of Semantic Graphs using ASALSAN, In Proc. of the 7th IEEE International Conference on Data Mining, 2007, pp. 33–42.
- [18] Bennett K. P. and Embrechts M. J., An Optimization Perspective on Partial Least Squares, *Advances in Learning Theory: Methods, Models and Applications*, NATO Science Series III: Computer & Systems Sciences, IOS Press, Amsterdam, 2003, 190: 227-250.
- [19] Bro R., Multiway Calibration. Multilinear PLS, *Journal of Chemometrics*, 1996, 10(1): 47-61.
- [20] Bro R., PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, 1997, 38(2): 149–171.
- [21] Bro R., *Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications*, PhD Thesis, University of Amsterdam, Amsterdam, Holland, 1998.
- [22] Bro R., Review on Multiway Analysis in Chemistry–2000-2005, *Critical Reviews in Analytical Chemistry*, 2006, 36(3–4): 279–293.
- [23] Bro R., Andersson C. A. and Kiers H. A. L., PARAFAC2 - Part II: Modeling Chromatographic Data with Retention Time Shifts, *Journal of Chemometrics*, 1999, 13(3-4): 295-309.

- [24] Bro R., Smilde A. K. and de Jong S., On the Difference Between Low-Rank and Subspace Approximation Improved Model for Multi-linear PLS regression, *Chemometrics and Intelligent Laboratory Systems*, 2001, 58: 3-13.
- [25] Bro R. and Kiers H. A. L., A New Efficient Method for Determining the Number of Components in PARAFAC models, *Journal of Chemometrics*, 2003, 17(5): 274-286.
- [26] Bro R. and Smilde A. K., Centering and Scaling in Component Analysis, *Journal of Chemometrics*, 2003, 17(1): 16-33.
- [27] Bro R., Harshman R. A. and Sidiropoulos N. D., Modeling Multi-way Data with Linearly Dependent Loadings, *Technical Report 2005-176*, KVL, 2005.
- [28] Carlier A., Lavit C., Pages M., Pernin M. and Turlot J., A Comparative Review of Methods, Which Handle a Set of Indexed Data Tables, *Multiway Data Analysis*, Elsevier, North-Holland, 1989, pp. 85-101.
- [29] Carroll J. D. and Chang J., Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of 'Eckart-Young' Decomposition, *Psychometrika*, 1970, 35(3): 218-319.
- [30] Cattell R. B., Parallel Proportional Profiles and Other Principles for Determining the Choice of Factors by Rotation, *Psychometrika*, 1944, 9(4): 267-283.
- [31] Ceulemans E. and Kiers H. A. L., Selecting among Three-mode Principal Component Models of Different Types and Complexities: A Numerical Convex-hull Based Method, *British Journal of Mathematical and Statistical Psychology*, 2006, 59(1): 133-150.
- [32] Chaovalitwongse W., Pardalos P. M. and Prokopyev O. A., Electroencephalogram (EEG) Time Series Classification: Applications in Epilepsy, *Annals of Operations Research*, 2006, 148: 227-250.
- [33] Chen Z. P., Wu H. L., Jiang J. H., Li Y. and Yu R. Q., A Novel Trilinear Decomposition Algorithm for Second-Order Linear Calibration, *Chemometrics and Intelligent Laboratory Systems*, 2000, 52(1): 75-86.
- [34] Chew P. A., Bader B. W., Kolda T. G. and Abdelali A., Cross-language Information Retrieval using PARAFAC2, In Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 143-152.
- [35] Chong I. G. and Jun C. H., Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, 2005, 78(1-2): 103-112.

- [36] Cole H. W. and Ray W. J., EEG Correlates of Emotional Tasks Related to Attentional Demands, *International Journal of Psychophysiology*, 1985, 3(1): 33-41.
- [37] Comon P., Independent Component Analysis - a New Concept?, *Signal Processing*, 1994, 36: 287-314.
- [38] de Clercq W., Vergult A., Vanrumste B., van Hees J., Palmimi A., van Paesschen W. and van Huffel S., A New Muscle Artifact Removal Technique to Improve the Interpretation of the Ictal Scalp Electroencephalogram, In Proc. of the 27th Annual Conference of the IEEE Engineering in Medicine and Biology, 2005, pp. 944-947.
- [39] de Lathauwer L., de Moor B. and Vandewalle J., On the Best Rank-1 and Rank- (R_1, R_2, \dots, R_N) Approximation of Higher-Order Tensors, *SIAM Journal on Matrix Analysis and Applications*, 2000, 21(4): 1324-1342.
- [40] de Lathauwer L., de Moor B. and Vandewalle J., A Multilinear Singular Value Decomposition, *SIAM Journal on Matrix Analysis and Applications*, 2000, 21(4): 1253-1278.
- [41] de Vos M., Vergult A., de Lathauwer L., de Clercq W., van Huffel S., Dupont P., Palmimi A., van Paesschen W., Canonical Decomposition of Ictal Scalp EEG Reliably Detects the Seizure Onset Zone, *Neuroimage*, 2007, 37(3): 844-854.
- [42] Delorme A. and Makeig S., EEGLAB: an Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis, *Journal of Neuroscience Methods*, 2004, 134(1): 9-21.
- [43] Delorme A., Makeig S. and Sejnowski T., Automatic Artifact Rejection for EEG data Using High-Order Statistics and Independent Component Analysis, In Proc. of the 3rd International Workshop on ICA, 2001, pp. 457-462.
- [44] Ding C. and Ye J., Two-Dimensional Singular Value Decomposition for 2D Maps and Images, In Proc. SIAM International Conference on Data Mining, 2005, pp. 32-43.
- [45] Engel J., Epilepsy in the World Today: Medical Point of View, *Epilepsia*, 2002, 43(6): 12-13.
- [46] Estienne F., Matthijs N., Massart D. L., Ricoux P. and Leibovici D., Multi-way Modelling of High-dimensionality Electroencephalographic data, *Chemometrics and Intelligent Laboratory Systems*, 2001, 58(1): 59-72.
- [47] Faber N. M., Bro R. and Hopke P. K., Recent Developments in CANDECOMP/PARAFAC Algorithms: A Critical Review, *Chemometrics and Intelligent Laboratory Systems*, 2003, 65(1): 119-137.

- [48] Field A. S. and Graupe D., Topographic Component (Parallel Factor) Analysis of Multichannel Evoked Potentials: Practical Issues in Trilinear Spatiotemporal Decomposition, *Brain Topography*, 1991, 3(4): 407-423.
- [49] Flanagan D., Agarwal R. and Gotman J., Computer-aided Spatial Classification of Epileptic Spikes, *Journal of Clinical Neurophysiology*, 2002, 19(2): 125-135.
- [50] Glover J. R., Raghavan N., Ktonas P.Y. and Frost J. D., Context-based Automated Detection of Epileptogenic Sharp Transients in the EEG: Elimination of False Positives, *IEEE Transactions on Biomedical Engineering*, 1989, 36(5): 519-527.
- [51] Golub G. H. and van Loan C. F., *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.
- [52] Gourvéneç S., Stanimirova I., Saby C. A., Airiau C. Y. and Massart D. L., Monitoring Batch Processes with the STATIS approach, *Journal of Chemometrics*, 2005, 19(5-7): 288-300.
- [53] Gourvéneç S., Tomasi G., Durvillec C., di Crescenzo E., Saby C. A., Massart D. L., Bro R. and Oppenheim G., CuBatch, a MATLAB Interface for N-mode Data Analysis, *Chemometrics and Intelligent Laboratory Systems*, 2005, 77(1-2): 122-130.
- [54] Greco A., Mammone N., Morabito F.C. and Versaci M., Semi-Automatic Artifact Rejection Procedure based on Kurtosis, Renyi's Entropy and Independent Component Scalp Maps, In Proc. of the 5th World Enformatica Conference, 2005.
- [55] Guyon I. and Elisseeff A., An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 2003, 3(7-8): 1157-1182.
- [56] Harshman R. A., Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multi-modal Factor Analysis, *UCLA working papers in phonetics*, 1970, 16: 1-84.
- [57] Harshman R. A., PARAFAC2: Mathematical and Technical Notes, *UCLA working papers in phonetics*, 1972, 22: 30-44.
- [58] Harshman R. A., Hong S. and Lundy M. E., Shifted Factor Analysis - Part I: Models and Properties, *Journal of Chemometrics*, 2003, 17(7): 363-378.
- [59] Hitchcock F. L., The expression of a tensor or a polyadic as a sum of products, *Journal of Mathematics and Physics*, 1927, 6(1): 164-189.
- [60] Hitchcock F. L., Multiple invariants and generalized rank of a p-way matrix or tensor, *Journal of Mathematics and Physics*, 1927, 7: 39-79.

- [61] Hjorth B., EEG analysis based on time domain properties, *Electroencephalography and Clinical Neurophysiology*, 1970, 29(3): 306-310.
- [62] Hong S. and Harshman R. A., Shifted factor analysis - Part III: N-way Generalization and Application, *Journal of Chemometrics*, 2003, 17(7): 389-399.
- [63] Hotelling H., Relations Between Two Sets of Variates, *Biometrika*, 1936, 28(3): 321-377.
- [64] Jiang J. H., Wu H. L., Li Y. and Yu R. Q., Three-way Data Resolution by Alternating Slice-wise Diagonalization (ASD) Method, *Journal of Chemometrics*, 2000, 14(1): 15-36.
- [65] Joachims T., *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, MIT-Press, B. Scholkopf and C. Burges and A. Smola, 1999.
- [66] Kapteyn A., Neudecker H. and Wansbeek T., An Approach to N-mode Components Analysis, *Psychometrika*, 1986, 51(2): 269-275.
- [67] Kargupta H., Huang W., Sivakumar K. and Johnson E., Distributed Clustering Using Collective Principal Component Analysis, *Knowledge and Information Systems Journal*, 2001, 3(4): 422-448.
- [68] Kiers H. A. L., Towards a Standardized Notation and Terminology in Multiway Analysis, *Journal of Chemometrics*, 2000, 14(3): 105-122.
- [69] Kiers H. A. L., ten Berge J. M. F. and Bro R., PARAFAC2 - Part I: A Direct Fitting Algorithm for the PARAFAC2 Model, *Journal of Chemometrics*, 1999, 13(3-4): 275-294.
- [70] Kiers H. A. L. and der Kinderen A., A Fast Method for Choosing the Numbers of Components in Tucker3 Analysis, *British Journal of Mathematical and Statistical Psychology*, 2003, 56(1): 119-125.
- [71] Kolda T. G., Orthogonal Tensor Decompositions, *SIAM Journal on Matrix Analysis and Applications*, 2001, 23(1): 243-255.
- [72] Kolda T. G., Counterexample to the Possibility of an Extension of the Eckart-Young Low-rank Approximation Theorem for Orthogonal Rank Tensor Decomposition, *SIAM Journal on Matrix Analysis and Applications*, 2003, 24(3): 762-767.
- [73] Kolda T. G. and Bader B. W., The TOPHITS Model for Higher-Order Web Link Analysis, Workshop on Link Analysis, Counterterrorism and Security, 2006.

- [74] Kolda T. G. and Bader B. W., Tensor Decompositions and Applications, *Technical Report SAND2007-6702*, Sandia National Labs., 2007.
- [75] Kolda T. G., Bader B. W. and Kenny J. P., Higher-order web link analysis using multilinear algebra, In Proc. of the 5th IEEE International Conference on Data Mining, 2005, pp. 242–249.
- [76] Kroonenberg P. M. and de Leeuw J., Principal Component Analysis of Three-mode Data by means of Alternating Least Squares Algorithms, *Psychometrika*, 1980, 45(1): 69-97.
- [77] Krzanowski W. J., *Principles of Multivariate Analysis*, Clarendon Press, Oxford, 1988.
- [78] Kruskal J. B., Three-way arrays: Rank and Uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra and its Applications*, 1977, 18(2): 95-138.
- [79] Kruskal J. B., Rank Decomposition, and Uniqueness for 3-way and N-way Arrays, *Multway Data Analysis*, Elsevier, North-Holland, 1989, pp. 8-18.
- [80] Kubat M. and Matwin S., Addressing the Curse of Imbalanced Training Datasets: One sided selection, In Proc. of International Conference on Machine Learning, 1997, 30(2-3): 179-186.
- [81] Latka M., Was Z., Kozik A. and West B. J., Wavelet Analysis of Epileptic Spikes, *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, 2003, 67(5): 105-122.
- [82] LeVan P., Urrestarazu E. and Gotman J., A System for Automatic Artifact Removal in Ictal Scalp EEG based on Independent Component Analysis and Bayesian Classification, *Clinical Neurophysiology*, 2006, 117(4): 912-927.
- [83] Levy A. and Lindenbaum M., Sequential Karhunen-Loeve Basis Extraction and Its Applications to Images, *IEEE Transactions on Image Processing*, 2000, 9(8): 1371-1374.
- [84] Lehnertz K., Mormann F., Osterhage H., Müller A., Prusseit J., Chernihovskyi A., Staniek M., Krug D., Bialonski S. and Elger C. E., State-of-the-Art of Seizure Prediction, *Journal of Clinical Neurophysiology*, 2007, 24(2): 147-153.
- [85] Li Y., Du Y. and Lin X., Kernel-based Multifactor Analysis for Image Synthesis and Recognition, In Proc. of the 3rd International Conference on Computer Vision, 2005, 1: 114-119.
- [86] Lim L. H., Singular Values and Eigenvalues of Tensors: a Variational Approach, In Proc. of IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005, pp. 129-132.

- [87] Mahoney M. W., Maggioni M. and Drineas P., Tensor-CUR decompositions for tensor-based data, In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 327–336.
- [88] Martinez-Montes E., Valdes-Sosa P. A., Miwakeichi F., Goldman R. I. and Cohen M. S., Concurrent EEG/fMRI analysis by multiway Partial Least Squares, *Neuroimage*, 2004, 22(3): 1023-1034.
- [89] Meng X., Morris A. J. and Martin E. B., On-line monitoring of batch processes using a PARAFAC representation, *Journal of Chemometrics*, 2003, 17(1): 65–81.
- [90] Miwakeichi F., Martnez-Montes E., Valds-Sosa P., Nishiyama N., Mizuhara H. and Yamaguchi Y., Decomposing EEG Data into Space-Time-Frequency Components Using Parallel Factor Analysis, *Neuroimage*, 2004, 22(3): 1035-1045.
- [91] Mohseni H. R., Maghsoudi A. and Shamsollahi M. B., Seizure Detection in EEG Signals: A Comparison of Different Approaches, In Proc. of the 28th Annual Conference of the IEEE Engineering in Medicine and Biology, Supplement, 2006, pp. 6724-6727.
- [92] Momma M. and Bennett K. P., Constructing Orthogonal Latent Features for Arbitrary Loss, *Feature Extraction, Foundations and Applications*, Springer, 2006.
- [93] Mormanna F., Kreuz T., Rieke C., Andrzejak R. G., Kraskov A., David P., Elger C. E. and Lehnertz K., On the Predictability of Epileptic Seizures, *Clinical Neurophysiology*, 2005, 116(3): 569-587.
- [94] Möcks J., Decomposing event-related potentials: A new topographic components model, *Biological Psychology*, 1988, 26(1-3): 199-215.
- [95] Mørup M. and Schmidt M. N., Sparse Non-negative Tensor 2D Deconvolution (SNTF2D) for Multichannel Time-Frequency Analysis, *Technical Report*, Technical University of Denmark, DTU, 2006.
- [96] Mørup M., Hansen L. K., Hermann C. S., Parnas J. and Arnfred S. M., Parallel Factor Analysis as an Exploratory Tool for Wavelet Transformed Event-Related EEG, *Neuroimage*, 2006, 29(3): 938-947.
- [97] Mørup M., Hansen L. K. and Arnfred S. M., ERPWAVELAB A Toolbox for Multi-Channel Analysis of Time-Frequency Transformed Event Related Potentials, *Journal of Neuroscience Methods*, 2007, 161(2): 361-368.
- [98] Nørgaard L., Bro R., Westard F. and Engelsen S. B., A modification of Canonical Variates Analysis to Handle Higly Collinear Multivariate Data, *Journal of Chemometrics*, 2006, 20(8-10): 425-435.

- [99] Paatero P., A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis, *Chemometrics and Intelligent Laboratory Systems*, 1997, 38(2): 223–242.
- [100] Paatero P., The Multilinear Engine - a Table-driven, Least Squares Program for Solving Multilinear Problems, Including the N-way Parallel Factor Analysis Model, *Journal of Computational and Graphical Statistics*, 1999, 8(4): 854-888.
- [101] Parra L. C., Spence C. D., Gerson A. D. and Sajda P., Recipes for Linear Analysis of EEG, *Neuroimage*, 2005, 28(2): 326-341.
- [102] Päivinen N., Lammi S., Pitkänen A., Nissinen J., Penttonen M. and Grönfors T., Epileptic Seizure Detection: A Nonlinear Viewpoint, *Computer Methods and Programs in Biomedicine*, 2005, 79(2): 151-159.
- [103] Rosenow F. and Lüders H., Presurgical Evaluation of Epilepsy, *Brain*, 2001, 124: 1683-1700.
- [104] Rosso O. A., Martin M. T. and Plastino A., Evidence of Self-Organization in Brain Electrical Activity Using Wavelet-Based Information Tools, *Physica A*, 2005, 347: 444-464.
- [105] Saab M. E. and Gotman J. , A System to Detect the Onset of Epileptic Seizures in Scalp EEG, *Clinical Neurophysiology*, 2005, 116(2): 427–442.
- [106] Sarco D. P., Burke J. F. and Madsen J. R., Electroencephalography in Epilepsy Surgery Planning, *Child's Nervous System*, 2006, 22: 760-765.
- [107] Savas B. and Eldén L., Handwritten digit classification using higher order singular value decomposition, *Pattern Recognition*, 2007, 40(3): 993-1003.
- [108] Schindler K., Leung H., Elger C. E. and Lehnertz K., Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG, *Brain*, 2007, 130(1): 65-77.
- [109] Schölkopf B., Smola A. and Muller K. R., Kernel Principal Component Analysis, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
- [110] Shannon C. E., A Mathematical Theory of Communication, *Bell System Technical Journal*, 1948, 27: 379-423.
- [111] Shawe-Taylor J. and Cristianini N., *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, 2004.

- [112] Shoeb A., Edwards H., Connolly J., Bourgeois B., Treves S. T. and Guttag J., Patient-Specific Seizure Onset Detection, In Proc. of the 26th Annual Conference of the IEEE Engineering in Medicine and Biology, 2004, 1: 419-422.
- [113] Sidiropoulos N. D. and Bro R., On the uniqueness of multilinear decomposition of N-way arrays, *Journal of Chemometrics*, 2000, 14(3): 229-239.
- [114] Smilde A. K., Westerhuis J. A. and Boqué R., Multiway Multiblock Component and Covariates Regression Models, *Journal of Chemometrics*, 2000, 14(3): 301-331.
- [115] Smilde A. K., Bro R. and Geladi P., *Multi-way Analysis. Applications in the Chemical Sciences*, Wiley, England, 2004.
- [116] Sordo M. P., Diosy D., Zenteno J. F. T., Sahjpaul R. and Wiebe S., Usefulness of Intracranial EEG in the Decision Process for Epilepsy Surgery, *Epilepsy Research*, 2007, 74(2-3): 176-182.
- [117] Stanimirova I., Walczak B., Massart D. L., Simeonov V., Saby C. A. and di Crescenzo E., STATIS, a Three-way Method for Data Analysis. Application to Environmental Data, *Chemometrics and Intelligent Laboratory Systems*, 2004, 73(2): 219-233.
- [118] Stern J. M. and Engel J., *Atlas of EEG Patterns*, Lippincott Williams and Wilkins, Philadelphia, PA, 2005.
- [119] Sun J., Tao D. and Faloutsos C., Beyond Streams and graphs: dynamic tensor analysis, In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 374-383.
- [120] Sun J. T., Zeng H. J., Liu H., Lu Y. and Chen Z., CubeSVD: A Novel Approach to Personalized Web Search, In Proc. of the 14th International World Wide Web Conference, 2005, pp. 382-390.
- [121] Timmerman M. E. and Kiers H. A. L., Three Mode Principal Components Analysis: Choosing the Numbers of Components and Sensitivity to Local Optima, *British Journal of Mathematical and Statistical Psychology*, 2000, 53(1): 1-16.
- [122] Tomasi G. and Bro R., PARAFAC and missing values, *Chemometrics and Intelligent Laboratory Systems*, 2005, 75(2): 163-180.
- [123] Tomasi G. and Bro R., A Comparison of Algorithms for Fitting the PARAFAC model, *Computational Statistics and Data Analysis*, 2006, 50(7): 1700-1734.

- [124] Tucker L. R., Implications of Factor Analysis to Three-Way Matrices of Measurement of Change, *Problems in Measuring Change*, 1963, The University of Wisconsin Press, Madison, pp. 122-137.
- [125] Tucker L. R., The Extension of Factor Analysis to Three-dimensional Matrices, *Contributions to Mathematical Psychology*, H. Gulliksen and N. Frederiksen (Eds.), Holt, Rinehart and Winston, New York, 1964, pp. 110-182.
- [126] Tucker L. R., Some Mathematical Notes on Three-mode Factor Analysis, *Psychometrika*, 1966, 31: 279-311.
- [127] Turney P. D., Empirical Evaluation of Four Tensor Decomposition Algorithms, *Technical Report ERB-1152*, National Research Council, Institute for Information Technology, 2007.
- [128] Urrestarazu E., Iriarte J., Alegre M., Valencia M., Viteri C. and Artieda J., Independent Component Analysis Removing Artifacts in Ictal Recordings, *Epilepsia*, 2004, 45(9): 1071-1078.
- [129] van Putten M. J. A. M, Kind T., Visser F. and Lagerburg V., Detecting Temporal Lobe Seizures from Scalp EEG Recordings: A comparison of various features, *Clinical Neurophysiology*, 2005, 116(10): 2480-2489.
- [130] Vapnik V. N., *The Nature of Statistical Learning Theory*, Springer, Berlin, 1995.
- [131] Vasilescu M. A. O. and Terzopoulos D., Multilinear Analysis of Image Ensembles: TensorFaces, In Proc. of the 7th European Conference on Computer Vision, *Lecture Notes in Computer Science*, 2002, 2350: 447-460.
- [132] Vasilescu M. A. O. and Terzopoulos D., Multilinear Image Analysis for Facial Recognition, In Proc. of International Conference on Pattern Recognition, 2002, 2: 511-514.
- [133] Vasilescu M. A. O. and Terzopoulos D., Multilinear Subspace Analysis of Image Ensembles, In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, pp. 93-99.
- [134] Vasilescu M. A. O. and Terzopoulos D., TensorTextures: Multilinear Image-Based Rendering, *ACM Transactions on Graphics*, 2004, 23(3): 336-342.
- [135] Vrhel M. J., Lee C. and Unser M., Fast Continuous Wavelet Transform: A least squares formulation, *Signal Processing*, 1997, 57(2): 103-119.
- [136] Wang H. and Ahuja N., Compact Representation of Multidimensional Data Using Tensor Rank-One Decomposition, In Proc. of International Conference on Pattern Recognition, 2004, 1: 44-47.

- [137] Wang H. and Ahuja N., A Tensor Approximation Approach to Dimensionality Reduction, *International Journal of Computer Vision*, 2008, 76(3): 217–229.
- [138] Walczak T. S., Radtke R. A. and Lewis D. V., Accuracy and Interobserver Reliability of Scalp Ictal EEG, *Neurology*, 1992, 42(12): 2279-2285.
- [139] Wei W. S., *Time Series Analysis: Univariate and Multivariate Methods (2nd edn)*, Addison Wesley, USA, 2006.
- [140] Word S., Johansson E. and Cocchi M., *3D QSAR in Drug Design: Theory, Methods and Applications*, Kluwer/Escom, The Netherlands, 1993.
- [141] Yang J., Zhang D., Frangi A. F. and Yang J., Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(1): 131–137.
- [142] Ye J., Generalized Low Rank Approximation of Matrices, *Machine Learning*, 2005, 61(1–3): 167–191.
- [143] Zhang T. and Golub G. H., Rank-One Approximation to High Order Tensors, *SIAM Journal on Matrix Analysis and Applications*, 2001, 23(2): 534-550.
- [144] Zhou W. and Gotman J., Removing Eye-movement Artifacts from the EEG during the Intracarotid Amobarbital Procedure, *Epilepsia*, 2005, 46(3): 409-414.
- [145] Bader B. W. and Kolda T. G., MATLAB Tensor Toolbox Version 2.2, <http://csmr.ca.sandia.gov/tgkolda/TensorToolbox> (last accessed April 2008).
- [146] World Health Organization, http://www.who.int/mental_health/neurology/epilepsy/en/ (last accessed April 2008).
- [147] International Workshop on Epileptic Seizure Prediction, <https://epilepsy.uni-freiburg.de/seizure-prediction-workshop-2007> (last accessed April 2008).