APPLICATIONS AND STATISTICAL MODELING OF ELECTROENCEPHALOGRAMS USING IDENTITY VECTORS

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

by
Christian Radcliffe Ward
May 2019

Examining Committee Members:

Dr. Iyad Obeid, Advisor, Dept. of Electrial and Computer Engineering
Dr. Zoran Obradovic, Dept. of Computer and Information Sciences
Dr. Joseph Picone, Dept. of Electrical and Computer Engineering
Dr. Yimin Zhang, Dept. of Electrical and Computer Engineering
Dr. Kai Zhang, External, Dept. of Computer and Information Sciences

ABSTRACT

In recent years, electroencephalograms (EEGs) have been the subject of intense signal processing research. The ability of software to group, cluster, or identify trends in EEG data has applications that range from clinical support tools for neurologists to brain-computer interfaces. However, a persistent limitation in the development of EEG classification algorithms has been a lack of clinician labeled data which is necessary to train the supervised neural networks and deep learning systems. This work addresses this issue by presenting an unsupervised technique for classifying EEGs and elucidating common data modes that do not depend on labeled data.

Specifically, this work introduces the application of Identity Vectors (I-Vectors) to EEG signals. I-Vectors were originally developed in the speech processing community to parse multiple facets of speaker data (speaker, language, accent, age, etc). The similarities between EEG and speech data suggest that I-Vectors are a strong candidate for developing data models that can differentiate between subjects, channels, and medical conditions. I-Vectors work by building a Universal Background Model (UBM) of signal features that is based on weighted Gaussian clusters. This UBM is then projected into a lower dimensional space through a Total Variability Matrix which seeks to maximize the differences between the UBM and a group of "enrollment" signals. Optionally, further dimensionality reduction can typically be achieved through linear discriminant analysis (LDA) before generating the final I-Vectors.

This work develops the application of I-Vectors to EEGs by addressing three key research aims. First: can the I-Vector technique be used to classify EEG data with equivalent performance to other machine learning classifiers. Secondly: how should I-Vector parameters be tuned to optimize performance on EEG data. And thirdly:

What properties of EEG data do I-Vectors take advantage of, and can this knowledge be used to inform the EEG classification process.

I-Vector performance was rigorously evaluated using larger and more diverse data sets than have been used in comparable published literature, specifically various blends of the PhysioNet Motor Movement Database and the Temple University Hospital EEG Corpus. Benchmark comparisons were made against well-known classifiers in the EEG domain, namely the Mahalanobis Distance and Gaussian Mixture Model-Universal Background Model (GMMUBM) classifiers. Performance was also evaluated using three different EEG feature sets as system inputs, namely Power Spectral Density, Spectral Coherence, and Cepstral Coefficients.

Ultimately, the I-Vectors exceeded the performance of the MD classifier and reported an equal error rate 5% higher higher than the GMMUBMs. This was achieved using I-Vectors that were one to two orders of magnitude smaller than those in the GMMUBM classifier and half the size of the MD classifier. These results Indicated the technique was robust and has the potential to scale for use on large datasets such as the Temple University Hospital EEG Corpus.

For my parents,

Who endured years of hows and whys,

and managed to retain their sanity while nurturing my curiosity.

Thank you.

ACKNOWLEDGEMENTS

When I decided to leave my aerospace career behind and pursue a PhD I was fairly certain I was making the right decision for myself. As has been the case with many of my major life choices, the reasoning *a priori* tends to differ greatly from the reasoning *a posteriori*. Coming in I thought this was my chance to finally develop all the habits and skills I jealously saw in my peers and found lacking in myself. Heading out I have realized I possessed those same skills and more, but had been looking from the wrong perspective. I say this not to diminish the improvements made to my engineering, math, science, or writing skills, but to elevate the importance of believing in oneself and trusting your success has indeed been attributed to you. Well mostly you, as there are a number of amazing people without which this self reflective paragraph would not be possible.

First and foremost, I must thank and will forever be grateful to my advisor Doctor Iyad Obeid. Coming across his TEDx Philly talk drew me to his research and his personality. My first visit, prompted by an email from the Graduate school asking if I had a sponsoring advisor, was when I knew a PhD was actually within reach. I had met someone strikingly similar to myself who has been remarkable in helping me grow as a researcher and a person. He provided me with the space and time necessary to formulate my research and explore my interests over my six year tenure. Outside of a gentle nudge from him and Doctor Joseph Picone on the potential of Identity Vectors, I was given exceptional freedom to carry out my work. When times of need arose for personal and professional issues he was insightful and supportive. In the moment I was oblivious to most of this, but now it is clear that PhD students all need some amount of personal incubation. Mine may have taken a bit longer than most, but his patience and guidance are qualities I will forever pay-forward in my personal and professional lives.

My initial years at Temple were also shaped by taking courses taught by Doctor Picone. I frequently engaged him in conversations about the curriculum, the research and applications of these tools in the real world. These were fantastic and necessary discussions as they helped assure me I made the correct choice in returning to school and provided a critical external opinion int he infancy of my research. Thanks must also be extended to Doctor Dennis Silage who became a sounding board for conversations on academia and career development.

While a PhD eventually turns into an individual journey, Dr Obeid populated his Neural Instrumentation Lab with all the right people to walk among. While all of those I met were wonderful, a few deserve special recognition starting with Doctor Alessandro Napoli. After Dr Obeid, Alessandro provided the strongest guide of what it meant to be a researcher and a person through how he carried himself during his final PhD year and his return as a Post Doc. Our ability to bond on a personal and professional level helped make some of the more turbulent days and weeks of my research.

There were also a number of less permanent members of the lab and fellow graduate students who contributed to my academic and personal growth. It is therefore necessary to thank them all, in order of appearance, Bogdan Niemoczynski, Elliot Franz, Andrew Powell, Silvia López de Diego, Vira Oleksyuk, Vinit Shaw, Dr Stephen Michael Glass, Elliot Krome, and Victor Espinoza. I may have been able to do it without you all, but it would not have been nearly as fun or worthwhile.

Not to be be forgotten, my circle of friends from high school and beyond did an amazing job of tolerating my lack of presence over the past six years. Despite moving far closer to the majority of them, and in some cases to the same city. I think I saw them less than when I lived in Maryland. However, this did not dissuade them from inviting me out and checking up on me from time to time. Sometimes, the little things have the biggest impacts and for that I am thankful for Dr. James Kelly, Dr. David Marchand, Christopher Hallenbeck, Micah Frost, Kathryn Fox, Dr. Suzanne MacFarland, Noemí Fernández, Dr. Merritt Edlind, and Andrew and Michelle Sirianni.

Lastly, I must thank my family. It is no easy task reintegrating into your hometown ten years after you had left. More so when you also slyly reintegrate back into the same room you occupied as a teenager. The acceptance and understanding offered to me from my parents, Brian and Nancy, are two of their many gifts I will forever be paying forward. It also helped to have my Aunt Donna and Nonna Rina in the neighborhood to reminded me of the importance of family, and that no matter who you are or what you do they will always be there to support you and love you unconditionally. While not a resident of the Philadelphia area any longer, my sister Emily and her husband Kellen were unassailable supports of my work from afar. Although they frequently offered up their home as getaway and a reminder of the fruits of my labors. Going on to separately thank each of you would never be enough considering everything we have shared in our lives has been instrumental in making my dream of a PhD possible. Please know that I finally understand how difficult it must be to watch a son, brother, nephew, or grandson transform right before your eyes from a precocious child to a loud and inquisitive man. I am eternally grateful for all of you and the five that watch over me from afar.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF FIGURES

Figure Page

LIST OF TABLES

LIST OF ABBREVIATIONS

AAC	American Academy of Clinicians
ABPN	American Board of Psychiatry and Neurology, Inc.
ACNS	American Clinical Neurophysiology Society
AD	Alzheimer's Disease
ADHD	attention-deficit/hyperactivity disorder
ADJUST	Automatic EEG artifact Detection based on the Joint Use of Spatial and Temporal features
ANN	Artifical Neural Network
ApEn	Approximate Entropy
BW	Baum-Welch
BCI	brain-computer interface
BSS	blind source separation
CHB	Children's Hopsital of Boston Massachusettes Institute of Technology Scalp EEG Database
CD	Cosine Distance
CEP	Cepstral Coefficient
COH	spectral coherence
CRR	Correct Recognition Rate
CRIM	Centre de Recherche d'Informatique de Montreal
CSP	common spatial pattern
DBN	Deep Belief Network
DET	Detection Error Tradeoff

DNN	deep neural network
DP	Dircihlet Processes
DT	Decision Tree
EC	Eyes Closed
ECG	electrocardiogram
ED	Euclidean Distance
EDF	European Data Format
EEG	electroencephalogram
EER	equal error rate
EM	expectation maximiation
EMD	emperical mode decomposition
EMG	electromyography
EO	Eyes Opened
EOG	electrooculography
ERP	evoked response potential
ET	epileptiform transient
FA	factor analysis
FAR	false acceptance rate
FASTER	Fully Automated Statistical Thresholding for EEG artifact Rejection
FRR	false rejection rate
FSC	Fuzzy Sugeno Classifier
FFT	Fast Fourier Transform
GFP	global field potential
GMM	Gaussian Mixture Model
GPED	generalized periodic epileptiform discharge

GMM-UBM	Gaussian Mixture Model-Universal Background Model
GMMHMM	Gaussian Mixture Model based Hidden Markov Model
HDP	Heirarchical Dirichlet Process
HMM	Hidden Markov Model
HTER	half total error rate
HTK	Hidden Markov Toolkit
ICA	independent component analysis
ICU	Intensive Care Unit
iEEG	intracranial electroencephalogram
IMF	intrinsic mode function
I-Vector	Identity Vector
JFA	joint factor analysis
KNN	K-Nearest Neighbor
LDA	Linear Discriminate Analysis
LS-SVM	Least Squares Support Vector Machine
LSTMNN	Long Short-Term Memory Neural Network
LMBPNN	Levenberg-Marquardt Backpropagation Neural Network
LOOCV	leave one out cross validation
MAP	maximum a priori
MCI	mild cognitive impairment
mCRR	mean Correct Recognition Rate
MD	Mahalanobis Distance
mEER	mean Equal Error Rate
ML	Machine Learning
MLE	maximum likelihood estimation
MLPNN	multilayer perceptron neural network

MFCC	Mel Frequency Cepstral Coefficient
MSR	Microsoft Research
mMSE	modified multiscale sample entropy
NBC	Naive Bayes Classifier
NEDC	Neural Engineering Data Consortium
NN	Neural Network
PCA	principal component analysis
PD	periodic discharge
PLDA	probabilistic linear discriminant analysis
PLED	periodic lateralized epileptiform discharge
PLI	phase lag index
PMean	pooled mean
PhysioNet Database	PhysioNet EEG Motor Movement/Imagery Database
PNN	Probabilistic Neural Network
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
RA1	Research Aim 1
RA2	Research Aim 2
RA3	Research Aim 3
RBFNN	Radial Basis Functional Neural Network
REM	random eye movement
RF	Random Forest
RMS	Root Mean Squared
ROC	receiver operating characteristic
SampEn	Sample Entropy
SOM	self organizing map

SPMD	Single Program Multiple Data
SVM	Support Vector Machine
TBR	theta beta ratio
TCP	Trans-Cranial Parasagittal
TUH	Temple University Hopsital
TVM	total variability matrix
TUH-EEG	Temple University EEG Corpus
UBM	Universial Background Model
VEP	Visually Evoked Potential
WPD	wavelet packet decomposition

Chapter 1

INTRODUCTION

The ability to communicate underlies the major functions of the brain. Given the array of tools at our disposal (voice, facial expressions, hands, feet and eyes) our ability to communicate is limited only by our inventiveness. However, this system of communication limits our brain by forcing it to indirectly communicate through these tools. When we wish to study the brain itself problems arise because the majority of measurements come through indirect means. This is further complicated as the ideas to be expressed become more complex either in terms of emotional context or severity, such as pain and illness.

Presently, electroencephalography is the principle method of directly communicating with the brain. While the communication is one directional, in that we can only listen, it affords opportunities not available through our human faculties. electroencephalograms (EEGs) may be used to discern the incidence of epilepsy and stroke [1], study neural responses to stimuli [2], or even neural control feedback [3]. Recently, the advent of inexpensive commodity-grade EEG headsets [4] has expanded the field to include areas such as gaming, neuro-modulation, and mindfulness training [5].

These advances allow for direct and more timely interpretation of EEGs via the creation of digital signal processing tools that can identify or predict neural activity [6]. In clinical settings this technology assists neurologists in reviewing long recordings [7], communicating with patients [2], and processing artifacts [8]. These tools leverage multidimensional statistical models [9, 10, 11] to enhance our understanding of EEGs. In research settings, this technology has facilitated advances in brain-computer interfaces (BCIs) [12] and seizure prediction [13].

Historically, computer-based EEG interpretation has been only moderately effective despite large quantities of research [14, 15]. One key problem is that brain function (and by extension an EEG recording) is highly variable, requiring very large sample sizes in order to create robust statistical models [16]. The most powerful statistical methods generally require even larger samples sizes to assure convergence [17]. Until recently it has been difficult to collect, store, and process such large EEG datasets.

Modern digital data collection methods, in both clinical and research settings, have made 'big neural data' feasible [18]. However, these datasets must be *annotated* prior to being useful for training statistical models. Annotated data is produced when an expert reviews the recordings by marking which segments of the recordings correspond to known phenomena [19]. These annotations can be at the macro scale (such as 'seizure') or the micro scale (such as 'sharp spike wave'). Not surprisingly, EEG annotation is manually intensive making it rarely cost effective to ask clinicians to perform it at a fine-grained level [20]

There are communication problems between even well trained clinicians on how and what to annotate on recordings. This is evident by moderate consensus agreement when annotating simple events such as variations of spike waveforms [21, 22, 23, 24]. Conflicting annotations make it difficult to produce 'gold standards' of annotations used for training new clinicians and for leveraging the power of *supervised* Machine Learning (ML) techniques.

Supervised ML techniques rely on this annotated data, more commonly called *labeled* data within the ML community, to produce sufficient models of known classifications. By using prior knowledge of the data, models can be trained to classify previously unseen data in classes such as background, seizure, and sleep. However, a common problem with these techniques is a lack of strong consensus for each class [6]. Thus the system is inherently limited by the quality and quantity of its prior knowledge.

The difficulty increases when building *unsupervised* ML techniques that operate on unlabeled data [20, 25]. Now the techniques are tasked with first determining how to partition the data into classes and then performing classification on those self-generated labels. This typically requires additional data beyond a supervised approach, but removes the stipulation of prior knowledge.

Despite the majority of research focusing on supervised ML, an unsupervised ML method may best suited for interpretation of EEGs. Unsupervised approaches are decoupled from clinicians because there is no need for labeled data. Clinicians are capable annotators, but even in their area of expertise they have biases which manifest in poor inter-rater agreement when aggregating annotations [26]. Furthermore, as the use cases of EEGs grow they advance beyond what clinicians typically annotate, meaning it is impossible to provide a documented ground truth. Given such constraints, this work introduces Identity Vectors (I-Vectors) as an unsupervised machine learning method for EEGs with the aim of supplanting the reliance on clinician annotations.

1.1 The Landscape of Electroencephalograms

Before outlining the aims of this work, a brief background is provided to a shared understanding of the relationships between EEGs, algorithms and clinicians. Chiefly among these relationships is the way in which algorithms and clinicians are trained and perform annotations. Specific attention is paid to how clinicians, as individuals and groups, produce the annotations used for algorithm development. The performance of these algorithms is outlined to contrast with the scope and performance of their human counterparts. Attention is focused on the algorithms areas of application and performance.

1.1.1 Clinician Development

Clinicians undergo extensive training, often culminating in a fellowship to specialize in the treatment of epilepsy, sleep disorders, or intensive care. These specializations require the ability to interpret EEG recordings[1] for which the clinician can be certified through the American Board of Psychiatry and Neurology, Inc. (ABPN). The American Academy of Clinicians (AAC) works with the ABPN to ensure clinicians are adequately trained, but cautions that "[N]ot all hospital credentialing boards require sub-specialty training to allow individuals to interpret EEGs"[2]. Sub-specialty certifications are limited to topics such as brain injury, neuromuscular issues, and epilepsy.

Beyond this, clinicians refine their skill on the patients they encounter through their practice of medicine. Principle among these skills is their ability to accurately annotate EEGs recordings. Annotations focus on documenting the activity of the brain via signals recorded from strategically placed electrodes extracranially (on the scalp) or intracranially (on the surface of the brain) [27]. The methodology of annotating and interpreting EEG recordings is part of the certification process, but

[1] Taken from: `https://medicine.yale.edu/neurology/education/fellowships/epilepsy_eeg/`

[2] Taken from: `https://www.aan.com/uploadedFiles/Website_Library_Assets/Documents/4.CME_and_Training/2.Training/3.Fellowship_Resources/3.How_to_Apply_for_a_Fellowship/Epilepsy\%20Fellowship\%20FAQ.pdf`

the Epilepsy Foundation contends that "EEG training for clinicians is inadequate"[3]. In spite of this, clinical annotations remain the best tool for assessing the behavior and state of a brain [28].

Even with all their training and successful treatment of the myriad of brain disorders, clinicians are not without their inconsistencies as they are human [21]. Firstly, their ability to annotate accurately is often surpassed by the amount of data produced from tests. This leads to annotation consuming a disproportionate amount of their work hours. Secondly, their formal education ensures they are in agreement on terminology and its manifestation [22]. However, performance in consensus-bases studies suggests there are disagreements over which waveforms are of interest to each clinician [21, 23, 24].

Thus it is clear that clinicians are capable interpreters because they readily determine the correct diagnosis from a EEG recording. However, their reasoning for these assessments have the potential to be disparate. This behavior is not unique to a specific subset of conditions as it is readily apparent in the lack of annotation consensus in sleep [24], seizure [21] and cardiac [29] EEG recordings.

Even when presented with data common to their expertise, pairwise clinician similarity (Cohen's κ statistic[4]) is moderate (0.41-0.60) at best [21] and group performance varies from slight (0.0-0.20) to almost perfect (0.81-1.00) [29]. This suggests clinicians identify different, but valid, indicators of disorders. Ultimately this produces multiple divergent, but correct, sets of annotation from one dataset. While not problematic for diagnosing disorders, it makes it difficult to develop ML algorithms when there are multiple 'right' answers.

[3] Taken from: `http://www.epilepsy.com/article/2014/12/eeg-training-clinicians-inadequate`

[4] The statistic is not perfect [30], but does appear to be among the most common reported in studies assessing neurologist performance.

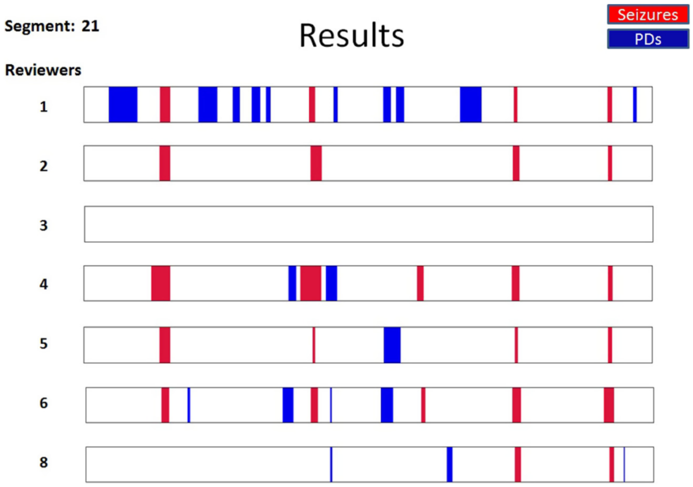


Figure 1.1. Example of EEG. In Halford et al. [23], seven reviewers were asked to annotate for seizures and PDs. The annotation results of the hour long recording, Segment 21, show that six reviewers labeled seizure events, five labeled PDs, and one labeled nothing. The quantity of annotation varies as does the spatial alignment between between reviewers.

1.1.2 Clinical Annotations

The ability to produce correct annotations is a fundamental component of EEG based research. In order to validate the performance of algorithms, clinicians must provided annotated data. These datasets are annotated through the lenses of the clinician's specialization and the patient's condition or diagnosis. As discussed previously, even when annotating the same data, clinicians struggle to come to consensus about its contents. Figure Figure 1.1 shows the results of seven clinicians annotating an hour long segment for seizures and periodic discharges (PDs). Nearly all the clinicians annotate abnormal events, save one, but the diversity and quantity of annotations are inconsistent.

Further complicating matters is that investigators often produce their own datasets, specifically for a given study. This occurs because existing datasets lack annotations, subject information, recording parameters, or protocols necessary to address their specific research questions. This makes it difficult to reuse previously annotated data because there is nothing is standardized. While one study annotates the other two do not, and then all three present with different sampling rates, recording durations, and electrode layouts.

While these decisions are practical with respect to specific studies, this behavior prevents supervised ML techniques for being applied across datasets. Without consistent sampling rates, the datasets may need to be interpolated to produce consistent windows of data. Mismatches in electrodes, inconsistent annotations, and artifacts are often manually resolved via the experiment team's limited knowledge or by possibly requiring the assistance of yet another clinician. While algorithms may overcome noise inherent in the data, this is only possible if there is a plethora of well annotated data from which to learn.

Annotations start, as shown in Figure Figure 1.2, as waveforms whose variations conform to similar behaviors. Not all annotations are related to medical conditions, as eye blinks and background are often considered to be noise. Differentiating such noise from waveforms of interest, like generalized periodic epileptiform discharges (GPEDs), periodic lateralized epileptiform discharges (PLEDs), spike and sharp wave complexes, and triphasic waves, is a critical step in reading an EEG. The American Clinical Neurophysiology Society (ACNS) defines an exhaustive list of EEG terms, including background characteristics, which are outlined in [29]. Clinicians are well versed in the terminology, but struggle in their ability to accurately match waveforms with appropriate labels [31].

The waveform examples from Wulsin et al. [14] are drawn from a seizure dataset. However, the waveforms are not unique to seizure recordings and could also be found in any of the other active EEG research fields such as attention/workload measurement [32], biometric identification [33], BCIs [5], evoked response

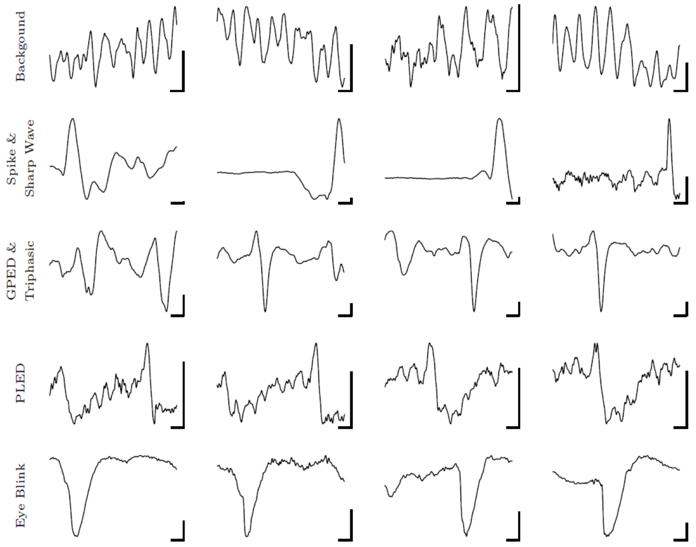


Figure 1.2. Annotation example. Annotations used for the work of Wulsin et al in [14]. Notice the placement of the spike does not need to precede or succeed the sharp wave. GPED and PLED typically occur over a range of channels making them context dependent.

potentials (ERPs) [34], and sleep stage classification [35]. Each field focuses on different facets of an EEG recording and may have distinct waveforms. Other sources for distinct waveforms include subject related traits, such as their age [24, 36] and genetics [37].

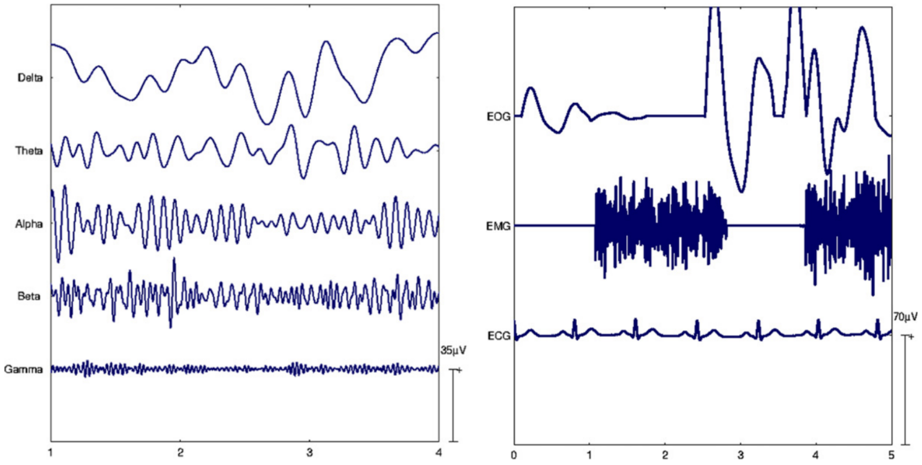
In summary, the fundamental technical challenge of training robust algorithms for automatic EEG interpretation is the diversity of annotated data. Seizure data differs from ERP data which differs from sleep data, making it difficult, if not impossible, to find clinicians capable of accurately annotating all of it. The lack of large diverse sets of thoroughly annotated data encumbers the advancement of algorithm based annotators/classifiers. This is exemplified by the struggle to develop ML algorithms capable of meeting performance levels deemed acceptable by clinicians and the inability to produce consistent universal ML classifiers.

1.1.3 Algorithm Applications

While major research avenues align with clinical applications (sleep, seizure, and various brain disorders), the use of ML provides avenues for novel applications as research progresses such as BCI, biometric verification, ERPs, and brain state workloads. Despite the variety of unique classification tasks, they all face similar fundamental performance hurdles. Chiefly among these are the necessary steps of pre-processing to address artifacts and production of acceptable feature sets. Within a given EEG recording it can be necessary to address the background waveforms that comprise the majority of the datasets.

EEG artifacts are often hard to classify because they appear as waveforms that resemble, Figure Figure 1.2, the more critical spikes and sharp waves of seizures [26] or the natural brain frequency rhythms [38], Figure Figure 1.3. While artifacts impact clinicians and algorithms, selection of an optimal feature set is unique to the algorithms. This is because feature sets are often paired with the type of EEG data being classified. The result is a wide range of potentially useful features consisting of but not limited to Power Spectral Density (PSD) features [39], spatial and temporal features [40], Cepstral Coefficient (CEP) feautres [41], auto-regressive parameters [42], and normalized raw data [43].

Despite focusing on waveforms of interesting via artifact correction and feature selection, the majority of EEG often consists of background signals[14, 44, 38]. This is frequently a problem for rare events like seizures, but is a boon to subject verification tasks and the biometric community [45]. Additionally, there are many less studied conditions that manifest throughout a recording, such as alcoholism [46], emotional state [47], pain [9], and mental focus/workload [48].



(a) Brain Rhythms

(b) Artifacts

Figure 1.3. Artifact Example. Example of artifacts (b) similarity to the natural rhythms of the human brain (a) used for the work of Uriguen et al in [38].

To motivate the implication of these areas of research a brief review of six common EEG classification fields is presented. The use of algorithms for seizure, sleep, BCIs, ERPs, and mental/workload classification are readily associated with clinician driven research, while EEG based biometrics branch out beyond their well defined knowledge base.

Seizures A substantial portion of work in this field focuses on correctly identifying and locating seizures [49, 50, 51, 52]. By isolating seizure events, researchers can focus on the properties of the seizure for the purposes of classification and waveform modeling [14, 53, 54]. The knowledge gained in this process makes it possible to predict seizures in real-time [6, 13]. Seizure events are typically high energy and frequency wavefroms with synchronization across channels [23].

Sleep Studies Sleep state classification labels the transition from wakefulness to random eye movement (REM) sleep. Sleeping EEG recordings are often cleaner due to lack of movement artifacts which improves their clarity for clinicians and reduces

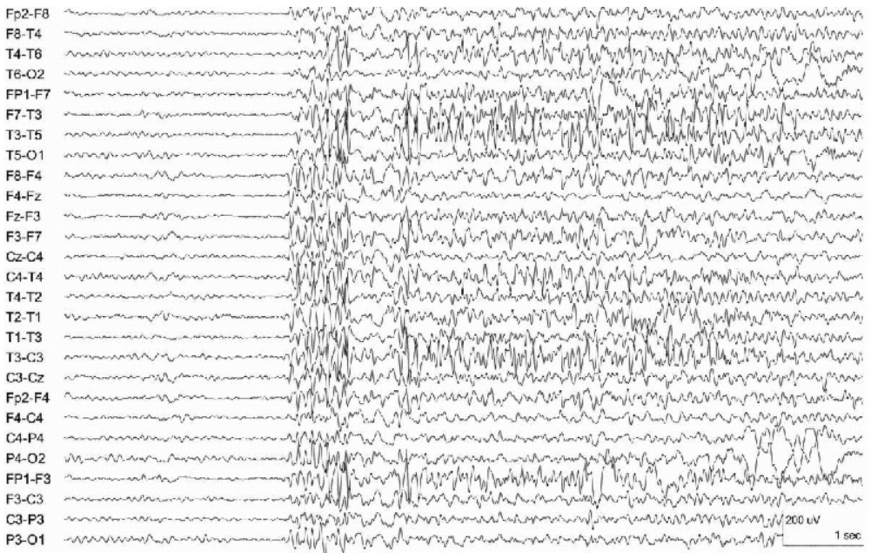


Figure 1.4. Example of a Generalized Seizure EEG. A segment of an EEG recording taken from a subject at the onset of a generalized seizure. Note that after the seizure starts, activity is not uniform across all channels. Image sourced from Tatum [55].

pre-processing for algorithms [56]. Despite this and a closed set of distinct stages, sleep stage classification suffers from inter-rater agreement problems [24]. Sleep spindles and K-Complexes serve as the main indicators of sleep along with pronounced changes in band Power Spectral Density [57]. While seizures often manifest during sleep, other issues can also be addressed such as sleep apnea [35] and overall brain functionality/health [58].

Biometrics Multiple studies have focused on the use of EEGs to identify and verify subjects, irrespective of any associated disease and disorder [59]. The results of such work suggest that individuals have distinct EEG fingerprints [60, 61, 62] which may relate to potential inheritable characteristics [61, 63]. A major theme in biometrics is understanding how different brain states impact these fingerprints. The work of Rocca et al. showcases brain distinctiveness when using a common testing state of resting eyes closed [33], spectral coherence as discrimination

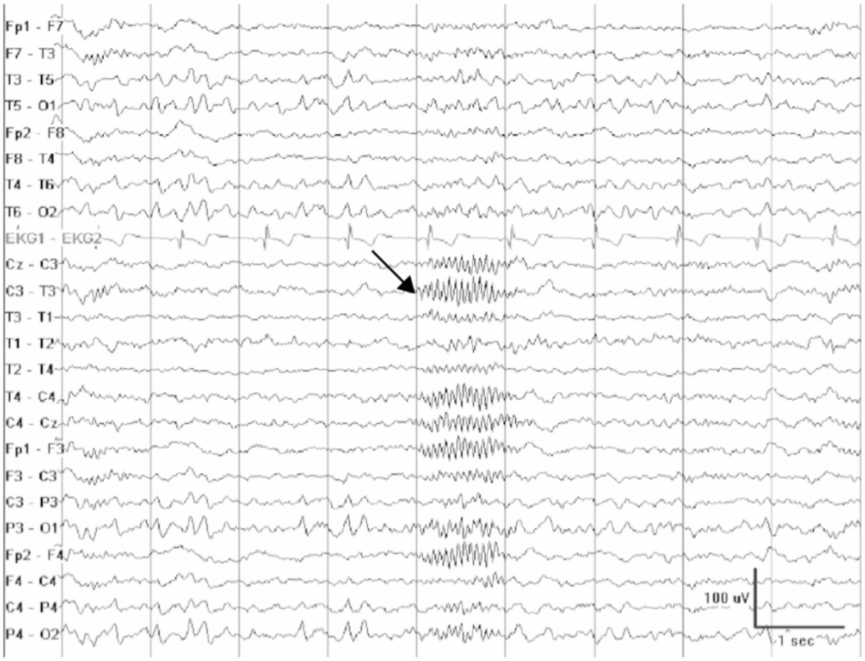


Figure 1.5. Example of Sleep EEG with Sleep Spindles. A segment of an EEG recording taken from a subject in the second of phase sleep. Note the present of sleep spindes, black arrow, across multiple channels. Image sourced from Tatum and Tatum[55].

feature [64], and techniques to reduce the feature set into sparse mappings [65]. Some approaches overlap with other applications by invoking response potentials [66], focusing on specific brains states of sleep [67], or restful states with eyes open and closed [68]. Even the longitudinal stability of biometric EEGs is tested [69] to determine viability for long term applications.

Brain Computer Interfaces BCI technology finds ways to get information into and out of a brain. The most advanced applications of this are restoring functionality to those unable to use their body [70, 71]. This requires algorithms robust to changes in subjects, but sensitive to spatial and temporal facets of EEG recordings [72, 73]. Development of subject invariant algorithms has led to disparate training protocols with transfer learning using multi-subject models [74] and zero-calibration training being subject specific [75]. This leads to a similar problem as sleep, where the waveforms are well understood, but their manifestation across populations complicates their performance.

Evoked Response Potentials ERPs are a stimulus response and not a voluntary action. A well documented case of ERP is the P300 signal that triggers in the pariatal/occiptal region 300 milliseconds after seeing an image of interest [2]. This signal is commonly used to enable subjects to communicate via P300 spellers. These spellers flash the alphabet before a subject waiting for a letter of interest to trigger an ERP, which allows them to build words [76]. This approach allows a brain to communicate without the need of a body, but also has applications for testing processing time of visual and auditory stimulus response [77].

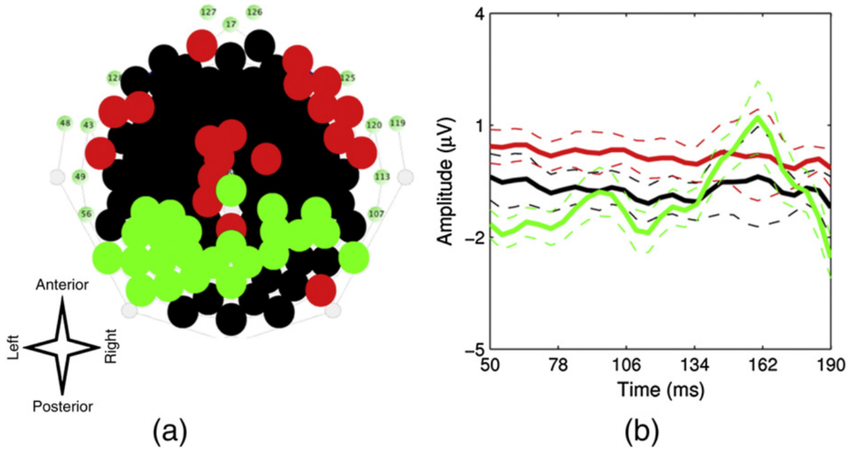


Figure 1.6. Example of an ERP. A 2D mapping of the electrodes and their group averaged waveform (solid lines). The standard deviations of the channel averaged are given as the dashed lines. Image sourced from Karamzadeh et al.[77].

Brain State/Workload Analysis of involuntary conditions address the state of a person's brain which can refer to the emotional state, disease state, or attention/workload state. Those afflicted with Alzheimer's [78], alcoholism [46], and mental disorders such as attention-deficit/hyperactivity disorder (ADHD) and Bi-Polar disorder [79] present with distinct EEG features. Knowing these conditions can manifest in the EEG recordings provides context for the how the known underlying biological changes alter a subject's EEGs. This is exemplified by studies measuring how stress impacts cognitive function [80] and a brain's workload during attention dependent tasks [48].

1.1.4 Algorithm Development

The development of ML techniques for EEG tends to focus on areas well understood by clinicians, detecting seizures [13, 14], identifying the stages of sleep [35, 81], capturing ERPs [34, 82], or processing BCI signals. Minimal focus has been given to a generalized classifier for interpreting multiple types of EEGs [83]. The approach closest to this goal is the use of EEGs for biometrics given that subject verification works on variety datasets with similar results [84, 85, 86]. While conditional classification techniques (seizure detection, sleep classification, BCIs, and subject verification) are capable, they fail to increase our overall understanding of EEGs.

Despite the lack of a generalized classifier, the data specific classifiers rely on some amount of data pre-processing. This is necessary to address recording artifacts [11, 87, 88], optimize the available channel data [64], or generate an acceptable feature set [89]. In carrying out one or more of these pre-processing steps a preliminary amount of dimensionality reduction is introduced which becomes more pronounced as the data is windowed into epochs for a given algorithm [90, 91, 92].

Unfortunately all these steps are often unique to the type of EEG being classified which means there is no well defined protocol of feature set that applies universally. For example, seizure algorithms typically process data in windows on the order of 10s of seconds [93]. Biometric algorithms utilize channel subsets to verify a subject [45]. BCIs use spatial filters to target the regions of the motor cortex [94]. ERPs focus on the occipital region where recognition of stimulus is triggered [72]. Things are further complicated by the varying performance within a dataset based upon subject or recording variation seen in BCI tasks [90, 73, 95], seizure recordings[6, 14, 96], and even biometric protocols [97, 98]. Due to this a comprehensive feature set remains elusive, but data specific feature sets have shown promise when paired with various algorithms.

These approaches leverage knowledge gained from the study of EEGs which makes them *domain knowledge*. Unfortunately domain knowledge comes from clinicians which means, as outlined previously, there are limits to its impact. It is critical in understanding artifacts and background (Figure Figure 1.2), seizures (Figure Figure 1.4), and sleep patterns (Figure Figure 1.5) [99, 57], but clinicians have minimal knowledge specific to biometrics [84]. Thus some approaches are bootstrapped by domain knowledge, but it furthers a Catch-22. Algorithms are made dependent clinician supplied insights when the algorithms task is to provide annotations to assist those same clinicians.

Within this loop of clinician annotations driving the development progress of algorithms, is the closed set of available EEG datasets. Aside from the PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) [100] and BCI competition [101] databases, much of the research is conducted on specific single use datasets [14, 92, 102]. Furthermore, when the PhysioNet Database database is used it is often the only dataset [86, 103, 68, 104]. There are studies that combine datasets, but that they tend to focus on biometric applications [105].

This lack of a robust data landscape manifests as variable algorithm performance based on the dataset[11, 91] or, within the context of BCI applications, as subjects being unable to use the system making them *'illiterate'* [106, 107]. This suggests there are intrinsic problems within processing EEG data. Algorithm performance can they be viewed as being dependent on the type of data (BCI, sleep, seizure, etc), but also the individual subjects themselves. This makes it difficult to tout any way forward given the lack of common performance benchmarks.

To address these data issues, algorithms must either be generalized, or built to tackle specific problems using domain knowledge. These two paths pair well with unsupervised (generalized) and supervised (specific) ML algorithms [34, 108, 109, 110]. The use of domain knowledge with supervised deep learning approaches have shown promise in BCI [111], sleep [112] and seizure [14, 13] classification. While their unsupervised counterparts do not require clinician support they do lag in comparative performance when addressing similar classification tasks [83, 113, 114, 110].

Efforts to address this data dependency in the BCI field produced four styles of classification schemes: *adaptive classifiers*, *matrices and tensors*, *transfer learning*, and *deep learning*[91]. In a ten year review of 2008 to 2018 by Lotte et al. [110], it was noted the similar schemes still worked, but a focus should be placed on tailoring BCI algorithms specifically to the end user to improve performance. These reviews covered supervised and unsupervised algorithms indicating that no resolute classifier has been found for the BCI community. Seizure classification has progressed, but the results are often using small esoteric datasets with 5 subjects [43] and 17 subjects [13]. Biometric techniques continue to perform well, but have also failed to expand their datasets [115, 59, 86].

The potential within the EEG ML community is vast. Presently performance varies based upon dataset, feature set, algorithm, and often subject quality. While some applications have become robust, seizure, ERP, and BCI classification show promise, this field is still maturing. Given the nascent status of EEG processing borrowing techniques from an established domain may be helpful. In general much of the technology currently deployed for ML pursuits comes from the realm of speech recognition, such as with Hidden Markov Models (HMMs) [116]. Following this trend, the development of the unsupervised learning technique called I-Vectors could offer growth of performance and understanding for EEG classification tasks [117]. I-Vectors are able to learn decision surfaces for the accent, age, content, gender, and language of a speaker [118]. Through a series of data modeling utilizing Gaussian Mixture Models (GMMs) [119] that produce a Universial Background Model (UBM) [120] capturing the variability of the training data in a total variability matrix (TVM), it is possible to reduce the dimensionality of various sized segments of data into robust discrimination vectors, I-Vectors [121].

1.2 Research Proposal

A clinician's primary focus is to treat their patients. Asking clinicians to produce perfectly annotated recordings to support algorithm research is not in the best interest of their patients or their overall productivity. There is little sense in asking clinicians for help to build datasets for algorithms who's goal is to reduce the time clinicians spend reading EEGs. This is clearly a Catch-22: The people that algorithms can help must first help to train the algorithms. However, clinicians do not have the time or group consensus to meet the needs of the algorithms.

The most direct solution is to find a way to annotate recordings without involving clinicians. As discussed ML-based solutions exist, but the field is diverse and lacks an apex technique. Despite the success of these techniques, fundamental problems continue to exist which must be overcome by all algorithms. These include variations in the quality of the recordings, the presence of adequate (in quality and quantity) annotated data, an acceptable feature set, and consistent channel layout across recordings. At its core the issue is identifying what characteristics of the EEG are relevant for a given classification task. In most instances, annotated data and prior knowledge is leveraged in order to reduce the dimensionality, and thus the uncertainty, in the algorithm's classification. This approach reinforces a reliance on annotators, which is not ideal given the disparate quality and consensus of annotations.

Annotation-based techniques are presently the dominant ML approach to classifying data. This means that clinicians effectively control the algorithms' performance which makes them an external source of error. To alleviate this constraint, unsupervised ML algorithms can be developed to match the capabilities of their supervised counterparts. The benefits of equivalent performance would be significant, as unsupervised ML enables training on large diverse datasets without the need of clinicians. Countless hours of data in need of annotation could thus be labeled, producing a steady supply of data for training supervised ML algorithms and clinicians. By using I-Vectors for this process it may also be possible to uncover novel phenomena in the data similar to their use on speech signals.

1.2.1 The Research Aims

The goal of this work was to lay the foundation for an unsupervised ML system that classified and clusterd EEG recordings. The preliminary pre-dissertation work indicated it was possible for I-Vectors to perform subject verification and to sort data by similarity[5]. While promising, these results had to be expanded to determine whether I-Vectors could overcome the annotation advantage. This primarily relied on the constrained modeling processing carried out in the generation of I-Vectors. Once

[5] See chapter 4's preliminary experiment results.

mastered, the process was largely transparent in its approach making it possible to study the decision surfaces used for the proscribed classification and clustering tasks.

In addition to understanding how the proposed system operated on EEGs, it was necessary to prove that I-Vectors could offer comparable performance to existing standard methods, including both ML algorithms and clinicians. However, given the advancement of ML algorithms, the ability to cluster and verify subjects is related only to algorithms. Clinicians do perform similar tasks, but they use resources beyond EEG recordings to make their assessments such as medical reports. Thus the performance of I-Vectors was evaluated against other well documented ML of varying complexity to highlight the tradeoffs between performance, dimensionality reduction, and algorithm complexity.

From these areas of interest, three research questions were posed:

Research Aim 1: Can an I-Vector-based classification perform as well as, or better than, other applicable ML techniques?

Research Aim 2: Under what conditions does an I-Vector based system perform best?

Research Aim 3: What characteristics of EEG data do I-Vectors take advantage of in their discrimination? Is this process inherently well suited for addressing EEG classification?

By answering these questions, insight into the nature of I-Vectors and EEGs was gained. This was possible because similar I-Vectors work in the speech recognition community produced strong results related to subject verification [122], language classification [123], accent detection [118], and speaker age estimation [124]. The underlying hypothesis was that EEGs had a bounded mathematical space similar to speech signals. This space can be exploited by the constraints of the TVM which shapes the I-Vectors producing nuanced classification similar to those seen in speech.

1.2.2 The Research Experiments

The Aims of this work were addressed in three experiments: *Parameter Sweeps*, *Algorithm Benchmarks*, and *UBM-TVM Relationship.* Upon completing the experiments, the process of producing I-Vectors from EEG data was understood along with which properties of EEG and I-Vector made this approach viable for producing annotations in an unsupervised manner.

Parameter Sweeps The purpose of the *Parameter Sweeps* was to determine optimal operating parameters for applying I-Vectors to EEGs. This addressed Research Aim 2 by measuring the significance of specific features, channels, UBM mixture sizes, and the TVM training process. Testing each parameter over a range of values produced trends for a best practice approach to baseline I-Vector systems. The statistical decomposition of each dataset (abnormal, normal, motion trials, and seizure) and I-Vector development process provided background and baseline results enabling comparisons against the other published results where the data is not publicly available.

Algorithm Benchmarks In order to validate I-Vectors as an option for classification and clustering of EEG data their performance was compared against a suite of ML algorithms. The algorithms were evaluated through their sensitivity and specificity and, when applicable, their ability to cluster. These experiments addressed Research Aim 1 through a series of leave one out cross validation (LOOCV) experiments based on subject and channel classifications.

UBM-TVM Relationship The relationships between UBMs and TVMs was deconstructed to examine the trade-offs made during optimization of the TVM. Using the reported performance of Gaussian Mixture Model-Universal Background Model (GMM-UBM) and I-Vector classifications, the influence of the mixture weighting were traced throughout the entire modeling process. This manifested as comparative feature and mixture mappings for each classification test. These mappings unlocked the fundamental statistical properties used to differentiate subjects which can then be compared across data sets as they are bounded by a common feature set. Ultimately this protocol turned I-Vectors into a powerful multi-modal signal analysis technique.

Chapter 2

BACKGROUND

Scarecrow:

The sum of the square roots of any two sides of an isosceles triangle is
equal to the square root of the remaining side. Oh joy! Rapture! I got a
brain! How can I ever thank you enough?

The Wizard of Oz:

You can't.

This chapter introduces the nature and use of EEGs in clinical and research settings. Clinical EEGs are used by clinicians to make diagnostic decisions in accordance with their education and training. In research settings algorithms strive to replicate the performance of clinicians through statistical modeling guided by clinician annotated data. Together these two groups are increasing our ability to discern the meaning of EEG signals.

This dissertation will examine the suitability of I-Vectors as a mathematical tool for allowing researchers to replicate clinician performance on EEGs. I-Vectors have shown promise with respect to classification and clustering of speech signals in terms of accent, age, context, gender, and language via its feature transformation process. This type of discrimination would be beneficial to understanding the phenomena that produce EEG waveforms. The I-Vector technique is introduced in depth along with the necessary criteria to evaluate it against other algorithm based discrimination techniques.

2.1 Electroencephalograms

An EEG records the electrical activity of the brain. The captured voltage signals represent the firing of neurons involved with all aspects of a brain's functionality. Through the use of EEGs we can see how the brain functions on an operational level [47], interprets stimuli [62], and changes due to diseases and disorders [46]. The applications of EEGs are primarily limited by the ability to link recorded activity to the underlying physiological condition.

A clinician's ability to annotate EEG recordings utilizes their knowledge of the relationship between waveforms and physiological conditions. An accurate diagnosis cannot be made from waveforms only as the clinician must consider the subject's history and the recording conditions of the EEG. In many cases spatial and temporal properties must be considered when assess for specific conditions related to different regions of the brain and similarities between waveforms.

Depending on application, EEG signals require radically different signal processing techniques for separating or decoding them. For example, whereas seizure and sleep waveforms are distinct and easily separable [1], EEG signals in BCI applications are typically subtle and require custom spatial and/or temporal filters [34]. This changes the discrimination techniques when dealing with BCI to spatial and temporal features [73, 125]. Auditory and visual stimulus response [2] have distinct spatial patterns as well adding to the diversity of BCI waveform morphology [75].

To distinguish spatial and temporal features, EEGs are partitioned via channels and epochs. As discussed previously, the channels are a representation of the electrodes, shaped by filtering and montages. Epochs segment the data as a function of time, typically on the order of seconds. Clinician and algorithm based approaches both rely on these techniques, but in different ways. Clinicians will review EEGs using epochs on the order of tens of seconds [24, 126], while algorithms operate on epochs of seconds [14, 75].

One of the main diagnostic applications of EEGs is the classification of seizures [14]. Seizures represent excessive electrical activity within a region of the brain which manifest as high energy waveforms. The study of sleep is also an active research area given the occurrence of seizures during sleep and sleep's impact on brain health [102]. When recording for seizure and sleep activity a substantial amount of background activity is also captured. This enables enables an analysis of overall brain function, like the presence of ADHD in children[127]. Adult EEGs also provide insight into numerous conditions such as alcoholism [46], Alzheimer's Disease [78], brain development [128], emotion [47], and stress [80].

In a research setting, BCIs promote a deeper understanding of brain functionality by allowing those with disabilities to communicate [2] and regain functionality [70]. BCIs highlight the ability of algorithms to classify waveforms beyond the capabilities of clinicians. These computer-driven methods enabled the development of novel applications in clinical monitoring, video games [5], and bio-metrics [129]. All of these use real-time classification which is not in the purview of clinicians. Specifically, bio-metrics provide the ability to dissect the facets of EEG that differentiate one person from another. This is a level of discrimination that clinicians cannot attain and serves needs far beyond clinical settings in hospitals.

Moving EEGs outside of hospitals has expanded the potential applications of EEGs[4]. It is easier to produce EEG datasets for experiments, but even with these advances there are few publicly available datasets. Those datasets available having varying levels of documentation and labeling related to conditions, subjects, and tasks. In addition, the sampling rates and number of channels have no definitive standards which furthers the disparate nature of the recordings. Recording in nonclincial environments often increases the likeliehood of artifacts, but even under ideal clinical conditions artifacts are still present requiring pre-processing[87, 8].

The following sections focus on the process and techniques of collecting EEG signals from a brain. Electrode configuration and montages are two important tools clinicians use when making a diagnosis from a recording. They provide flexibility to the clinician, but hamper the ability of algorithms to validate themselves on similar data. The experimental datasets are also introduced to highlight the difficulties of working with publicly available data.

2.1.1 Properties of Electroencephalograms

An EEG is comprised of multiple surface/scalp electrode channels capturing the continuous signals generated by the brain. These signals represent the aggregated neuronal activity of the cortical neurons in immediate proximity to each electrode. Each channel maps to a specific electrode that is placed on the scalp, extracranially, or in the case of intracranial electroencephalograms (iEEGs) directly on the brain's surface. Electrode placement for extra-cranial recordings follows a standardized layout, Figure 2.1, based upon relative distances [130]. Intra-cranial electrodes are high density electrode grids that are placed directly on the brain region of interest. This increases the complexity of the electrode and the data collected which excludes them from this work, but there is no theoretical reason I-Vectors could not operate on such signals.

The electrode configuration dictates the number of channels in the recording. To visual these signals clinicians view them indirectly as *montages*, a differential electrode configuration. Montages, Section 2.1.1, can be configured to be referential to a common ground electrode, neighboring electrode, or a contralateral electrode. These configurations aid in the diagnostic process by calling attention to patterns

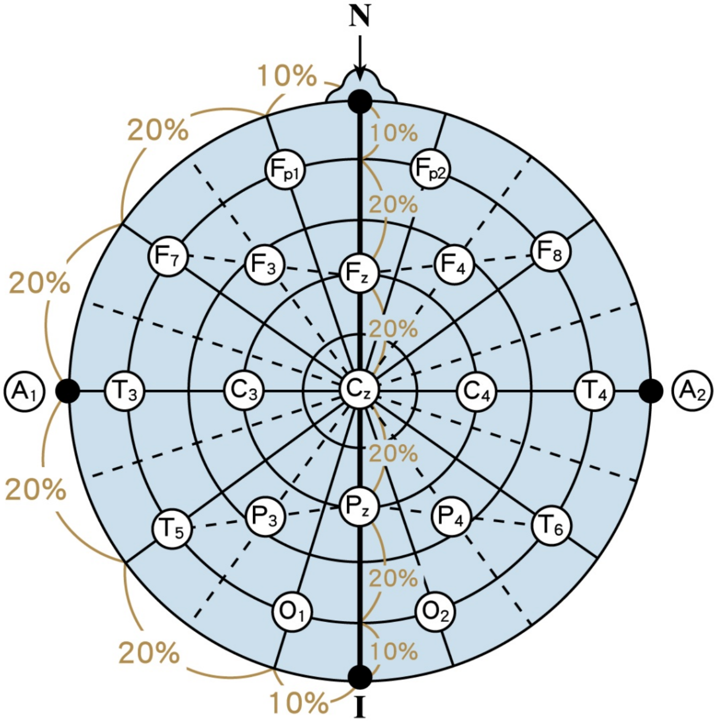


Figure 2.1. 10-20 EEG Configuration. The 10-20, 10-10, and 10-5 layouts for EEG electrodes utilize a proportional unit of measure for the distribution of electrodes. The first number represents the distance of the electrodes from the nasion and inion and the second represents the space between subsequent electrodes. With this approach adding electrodes does not change the location of the previous electrodes. Image sourced from [131].

of behavior in the recording. Below are three sets of montages for a system with eighteen channels[1].

Montages serve to improve the clarity of each channel. Theoretically they do not impact the content of the channels, but evaluating such a claim is beyond the immediate focus of this work. Filtering of the channel data, before or after inclusion in a montage, is necessary to separate signals into the five standard EEG frequency bands, Table 2.2. Signals between 2Hz to 80Hz represent the spectrum commonly viewed by clinicians[2]. For many conditions the frequency range of activity is critical

[1] Taken from: `https://www.acns.org/UserFiles/file/EEGGuideline3Montage.pdf`

[2] While this is the dominant spectrum of interest, research using iEEGs indicates activity at higher frequencies (>500Hz) may contain relevant discriminatory data related to seizures [109].

in signal classification. Motor activity signals dominate the alpha band [132], while the stages of sleep affect all but the gamma band [35].

Table 2.1. Common EEG Montages

Channel	Longitudinal Bipolar	Transverse Bipolar	Referential to Ground(Ear)
1	Fp1-F7	F7-Fp1	F7-A1
2	F7-T3	Fp1-Fp2	T3-A1
3	T3-T5	Fp2-F8	T5-A1
4	T5-O1	F7-F3	Fp1-A1
5	Fp1-F3	F3-Fz	F3-A1
6	F3-C3	Fz-F4	C3-A1
7	C3-P3	F4-F8	P3-A1
8	P3-O1	T3-C3	O1-A1
9	Fz-Cz	C3-Cz	Fz-A1
10	Cz-Pz	Cz-C4	Pz-A2
11	Fp2-F4	C4-T4	Fp2-A2
12	F4-C4	T5-P3	F4-A2
13	C4-P4	P3-Pz	C4-A2
14	P4-O2	Pz-P4	P4-A2
15	Fp2-F8	P4-T6	O2-A2
16	F8-T4	T5-O1	F8-A2
17	T4-T6	O1-O2	T4-A2
18	T6-O2	O2-T6	T6-A2

Table 2.2. EEG Frequency Bands

Band Name	Frequency Range (Hz)	Attributes
Delta	1-3	Brain health, deep sleep
Theta	4-7	ADHD rhythms, relaxation
Alpha*	8-12	motor activity, alertness
Beta	13-30	anxiety, focus
Gamma	31-80	REM sleep, stress

*When dealing with motor cortex signals it is common to encounter the Mu band (9-11Hz) which resides within the Alpha band.

2.1.2 Available Datasets

There are a number of publicly available EEG datasets[3]. These datasets are developed for specific studies independently of each other resulting in a wide variation of data content and format. Their data formats range across European Data Format (EDF), Matlab formatted files, and raw text files. The data content differs in terms of electrodes, sampling rates, and the studied phenomena.

This work applies to the PhysioNet Database dataset and the Temple University EEG Corpus (TUH-EEG) dataset. These datasets have been standardized to utilize the same 20 channel Trans-Cranial Parasagittal (TCP) montage. In addition the TUH-EEG dataset contains annotations from multiple sources providing robust labeling of events. This helps control for variation between the BCI focused PhysioNet Database dataset and predominantly seizure focused TUH-EEG dataset.

2.1.2.1 Temple University Hospital EEG Corpus

The TUH-EEG dataset contains over 25,000 EEG studies and their associated neurological evaluations taken from Temple University Hopsital (TUH) in Philadelphia, Pennsylvania [18]. Each patient's records present with different electrode configurations and sampling rates. The curated corpus uses a common 22 channel montage, TCP shown in Figure 2.2, for all subjects with a static sample rate of 250Hz.

The dataset contains longitudinal results of patients receiving continuing care at the hospital. These include multiple same patient sessions in a given day or sessions spaced out over a number of years. TUH treats patients of varying backgrounds (age, gender, diagnosis) providing breadth to the data. Recording profiles at TUH range

[3] The University of California San Diego maintains a website, `https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html`, indexing many of the publicly available datasets.

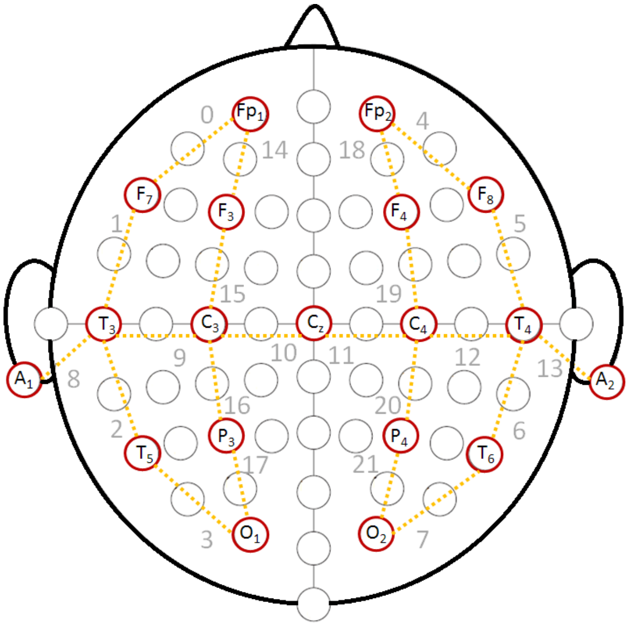


Figure 2.2. The TCP Montage Layout. The TCP Montage channels (red) used by the TUH EEG corpus is overlaid on the PhysioNet Database channel layout. Each montage link (orange) is assigned an index for storing the montage channel (gray) data in the corpus. The proper 10-20 channel names (black) are provided for the montage channels.

from 23 to 32 electrodes with sampling rates of 250Hz, 256Hz, 400Hz, or 512Hz [18]. Computerized EEG analysis is complicated by the fact that even small variations in electrode placement can hamper generalizations between subjects. This problem is exacerbated when datasets from disparate sources are combined.

2.1.2.2 PhysioNet EEG Motor Movement/Imagery Database

The PhysioNet Database data contains 109 subjects following computer prompted motion/motion imagery trials at the New York State Department of Health's Wadsworth Center [94]. The recordings present 64 electrodes following a 10-20 layout sampled at 160Hz. From this base layout, the data is converted to the same 22 channel TCP montage used by the TUH-EEG.

Each subject performs two calibration trials (resting eyes open and resting eyes closed) and twelve task driven trials. The four tasks consist of opening/clenching the (1) left or (2) right first and opening/clenching both (3) fists or (4) feet as a physical and imaginary movement. A trial consists of 30 tasks that alternates between rest and motor tasks. The calibration trials last for one minute and the motor trials last for two minutes, providing 26 total minutes of subject data. The data is publicly available through the PhysioNet Database website [100].

There are 12 total motion tasks representing three groups. These groups consist of 4 repeated trials creating natural cohorts of grouped trials: {3, 7, 11; 4, 8, 12; 5, 9, 13; 6, 10, 14}. Figure 2.3 shows the layout of tasks within each trial and their associated grouping. The major experiments utilize these trial level cohorts and the unique 109 subjects to develop I-Vectors for discrimination on the trial and subject level.

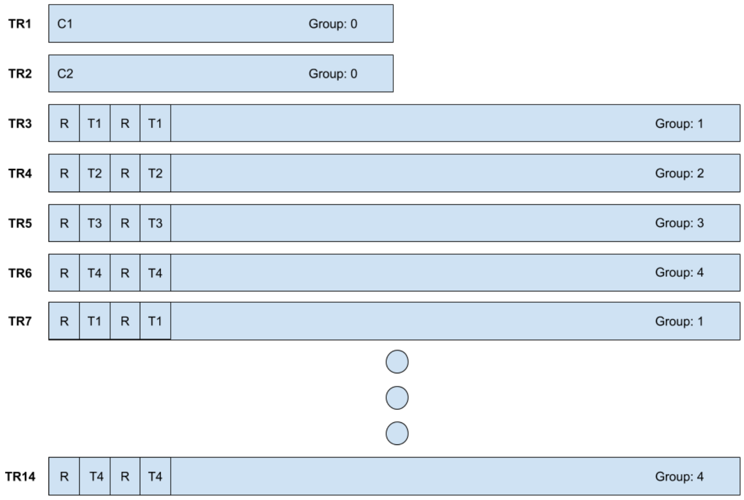


Figure 2.3. PhysioNet Trial Composition. Each subject from the PhysioNet data set completed 14 trials. Two of these trials (TR1 and TR2) are one minute calibrations trials of resting eyes open and resting eyes closed. The remaining 12 trials are two minute recordings of a predefined sequence consisting of a task state and resting state. With four tasks states, each task is repeated three times producing four groups of task related trials. These trial groups provide the basis for cohort retrieval on the trial level.

2.2 Applications and Classification of Electroencephalograms

The techniques used by algorithms and clinicians to classify and cluster EEG data are unique. An algorithm's foundation is informed by the knowledge of clinicians via their annotated data. A clinician's knowledge comes from their experience treating patients and their formal education. The algorithms are dependent on the clinicians' annotations to build their knowledge base, making them susceptible to clinician bias. Clinicians are skeptical of algorithm performance because it does not match clinical performance. As algorithms attempt to improve their classification they are competing against experts in a field that is still being understood. Progress is slow because it is difficult for algorithms and clinicians to be confidant in the reasoning of their classifications. This makes it difficult to produce accurate testing datasets given the competing views on what are accurate annotations.

Clinicians annotate EEGs recordings to diagnose their patient. Typical clinical recordings are 20 minutes or more depending on the nature of the assessment. Each recording is accompanied by a detailed EEG report [27]. These reports must document the subject, the testing carried out, and address the *clinical questions*[4]. The interpretation of an EEG recording is the main criteria when affirming a diagnosis, but must be supported by evidence indicating the recording is normal or abnormal[19].

This annotation and reporting process relies on the clinician's ability to review segments of the full recording for waveforms relevant to the clinical questions. A clinically relevant interpretation of the patient's condition may not be forthcoming

[4] Clinical questions are posed prior to testing by the clinician. They serve to inform the clinician about the patient, their condition, and what outcomes are possible. As an example, if a patient has seizures while sleeping it would be necessary to determine the location of these seizures, their severity, and how such seizures compare to other patient populations. These would all be questions answered through EEG recordings.

without reviewing the reports of other tests and/or subjects [27]. This meta-analysis across subjects is a clustering process informed by medical records and annotations. However, the EEG reports focus on determining if the results inform the clinical questions or not [19]. This does not require all relevant phenomena to be annotated, as only enough data must be collected to affirm a position. As such a clinician's ability to cluster could be hampered by their ability to annotate, which is suggested by tracking a clinician's ability to reproduce classifications [26].

In contrast, an algorithm's approach to annotation is much more broad. Depending on the desired outcome, algorithms can perform a normal/abnormal classification [7], annotate specific epochs [14] or combine these approaches to classify EEG recordings [35]. Each of these classification techniques is a subset of the classification approach used by clinicians. Performance of these algorithms is measured against gold standards generated from training data annotated by clinicians [14, 24]. The goal is develop algorithms capable of mirroring clinical performance which limits the strength of the algorithms to the strength of the clinicians.

Depending on the output of these algorithms, they are capable of clustering EEG recordings in a way clinicians cannot replicate. The ability to infer similarity of waveforms, epochs, and entire recordings across subjects is important in the development of robust BCI [74] and bio-metric applications[45]. In this area algorithms exceed the ability of clinicians by shifting how EEG recordings are evaluated through novel channel and feature selection [65, 66, 68].

Specifically, bio-metric algorithms can determine the similarity of one subject to another [45, 63]. This makes bio-metric subject verification the closest analog to I-Vectors, but they are not limited to subject comparisons. Instead they offer the ability to discriminate on multiple facets of the data without needing the same extent of bio-metric pre-processing [121]. This makes their application to EEG recordings interesting as I-Vectors may be capable of bridging classification between algorithms and clinicians.

2.2.1 Clinician Classification

For clinically annotated EEG recordings it is important that common terminology was used when describing the waveforms. Without a shared vocabulary EEG reports would be ineffectual for diagnostics and documentation [27]. Gaspard [22] tested 49 clinicians' agreement on terminology by asking them 409 questions about 37 pre-selected EEG waveforms. Their protocol removed the need of the clinician to find the epochs, enabling them to focus on each clinician's ability to describe the contents of each pre-selected epoch.

Each clinician's background varied in terms of experience (2-15+ years) and training (adult or pediatric neurology). While the epochs were sourced from only critical care patients exhibiting PLEDs, GPEDs, seizures, and other rhythmic activity. The epochs were presented using a modified biploar montage with a bandpass filter spanning 1Hz-70Hz. From these epochs, clinicians made *categorical assessments* based upon the presence of a seizure and dominant morphologies and *ordinal assessments* based upon the physical properties on the signals (sharpness, amplitude, frequency, etc). The overall and inter-rater agreement of the clinicians is presented in Table 2.3.

In 12 of the 15 categories, the clinicians' exceeded an agreement of 70% and 7 of the 15 showed near- perfect (0.81-1.00) κ statistics. The categories with the lowest agreement and weakest κ statistics were categorical classifications. With only 3 morphologies reporting κ below substantial (0.61-0.80), the results suggest the clinicians perform well as a group. Yet, those three categories indicated a universal

Table 2.3. EEG Terminology Agreement

Terminology Item	Agreement (%)	κ statistic (95% CI)
Categorical		
Seizure	93.3	91.1 (90.6-91.6)
Main Term 1	91.3	89.3 (89.1-89.6)
Main Term 2	85.2	80.3 (79.4-81.2)
Triphasic Morphology	72.9	58.2 (56.1-60.2)
Plus + Modifier	49.6	33.7 (32.4-35.1)
Any +	59.3	19.2 (17.5-20.9)
+ Fast Activity	71.9	65.5 (64.4-66.7)
+ Rhythmic Activity	76.5	67.4 (66.5-68.3)
+ Spike or Sharply Contoured	83.9	81.8 (81.2-82.5)
Ordinal		
Sharpness	91.5	84.8 (84.3-85.2)
Absolute Amplitude	96.5	94.0 (93.8-94.2)
Relative Amplitude	71.8	66.4 (65.3-67.4)
Frequency	97.8	95.1 (94.9-95.2)
Phases	89.9	83.0 (82.6-83.4)
Evolution	65.6	21.0 (19.7-22.2)

Each terminology item, aside from Seizure, could be classified with multiple responses. Fast Activity could be yes, no, or no applicable while Phases were 1, 2, 3, >3, not applicable forcing the clinicians to articulate their classifications. Agreement specifies the percentage of waveforms classified correctly. The κ score indicates the amount of inter-rater agreement, see Section A.1.1.

blind spot that would be passed on to an algorithm built from this annotated data. Since the contents of epochs were known, this showed how difficult it was for clinicians to agree on labeling of wavefroms.

These biases likely existed because clinicians were evaluated on their annotations indirectly. Their diagnoses were not solely based on a single event in the EEG, but rather the sum of the recordings in conjunction with the patient's medical history. In Halford et al. [26] the importance of detecting epileptiform transients (ETs) was found to be critical for diagnosing epilepsy. Failing to annotate some of the ETs does not change the diagnosis because the clinicians were primed to make a decision about epilepsy. Individually the 18 tested clinicians were unable to produce a Gwet agreement coefficient[5] over 0.50 with the rest of the group. This indicated a weak agreement among the clinicians. Despite varying levels of certification and years of practice, there were no distinct indicators of what characteristics represented a better annotator.

The difficulty in producing accurate annotations with respect to others existed at the intersection of finding the waveforms and then correctly labeling them. These problems were documented to various degrees when clinicians' annotation skills were tested on critically ill patients [31], patients exhibiting seizures [21, 23], comatose cardiac patients [29], and sleeping subjects [126]. The results of such studies highlighted problems with clinician inter-rater and intra-rater agreement as a function of the type of EEG data.

[5] The Gwet's AC2 is an alternative to κ statistics for quantifying inter-rater similarity, but is bounded over the same range [133].

2.2.1.1 Clinician Inter-rater Agreement

The previous section discussed this broadly and with the benefit of the waveforms being pre-selected. However, when clinicians were asked to annotate longer epochs the discrepancies shift from clinical knowledge to issues of annotation style. Their inter-rater agreement was the ability of one clinician's classification to agree with one or more other clinicians.

A pedantic instance of this was seen in Figure 2.4 where two clinicians labeled seizure events [23]. In the highlighted section, Rater B identified two discrete events while Rater A labels them as one event. Each of them notices at least 4 other seizure events, but their agreement was weakened because of their three misidentified events. Behavior such as this further complicated how to quantify agreement and disagree based upon duration of said annotations.

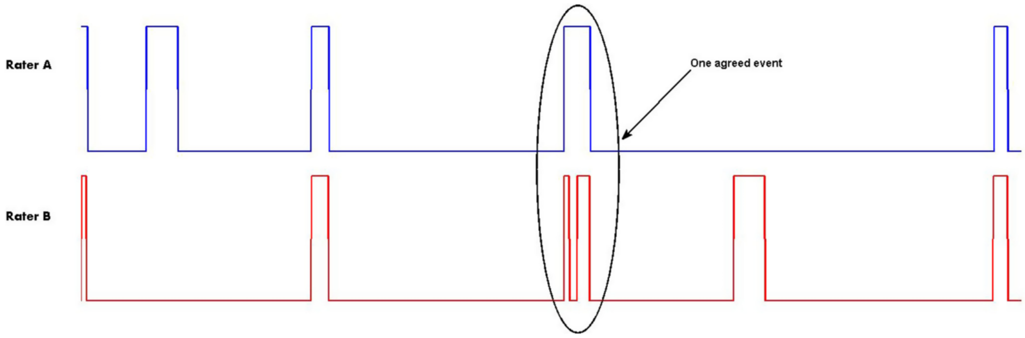


Figure 2.4. Inter-rater annotation matching. An example of how open ended annotation styles lead to inconsistencies in evaluating the accuracy of inter-rater agreements.

This example came from Halford et al.[23] where the agreement of 8 clinicians was test on 30 one hour Intensive Care Unit (ICU) EEG recordings from 20 seizure patients. Each clinician was asked to label PDs events, a strong indicator of a seizure, and true seizure events. The resultant κ statistics for the group were 0.58, moderate, for seizures and 0.38, fair, for PD. These results highlighted the difficulty in finding consensus by suggesting it surpassed their background and experience. There was a clear issue in how clinicians selected waveforms in the recordings, which resulted in less data being included in any gold standard.

Gerber et al.[31] conducted a study with a more expansive classification list than Halford et al.'s by expanding the available labels and varying the amount of available data. Two data sets, split into epochs of 10 seconds and epochs >20 minutes, were built from 11 subjects with convulsive seizures, status epilepticus[6]. The results, Table 2.4, showed the clinicians' consensus was stronger on the shorter epochs (0.04-0.68) than the longer epochs (0.07-0.44).

Table 2.4. Classification Performance of Long and Short Segments

Term	10s Epoch Kappa	20min Epoch Kappa	20min Epoch Agreement (%)
Rhythmic/periodic vs. excluded	0.68	0.44	82
Localization	0.49	0.42	66
Morphology	0.39	0.37	69
Frequency	0.34	0.27	78
"Quasi" vs. Not	0.04	0.07	57
"Frontally Predominant" vs. Not	0.40	0.08	68
+ vs. Not	0.12	0.08	62

Results of classification using segments of 10 seconds and > 20 minutes in length. Five clinicians annotated the shorter epochs and all seven clinicians annotated the longer epochs. The κ statistics for both datasets are reported along with the raw agreement percent for the 20min epoch dataset.

The most critical labels (rhythmic/periodic vs. excluded, localization, and morphology) exceed 65% agreement, but only rhythmic/periodic exceeds 80%. This meant that on average each clinician failed to recognize 20% to 35% of what the

[6] Status epilepticus is the categorization of a person's state when seizures occur close together or occur for a prolonged duration(>5 minutes).

other clinicians annotated. Without definitively labeled data it was impossible to determine if the 35% gap is due to false positives or false negatives. Such knowledge could be used to determine if they were over-jealous or overly-shrewd in their annotations. However, it was possible their performance was impeded by alignment issues similar to those seen in Halford et al.'s work. The results otherwise suggested that the clinicians agree at a moderate to fair level which was enough to make accurate medical decisions, but not sufficient from which to train algorithms.

Gerber et al.'s best reported inter-rater agreement was inline with Halford et al.'s. This trend persisted in the work of Grant et al.'s work [21]. Their study evaluated the agreement of 6 clinicians (adult and pediatric neurologists) classifying 7 categories (status epilepticus, seizure, epileptiform discharges w/ and w/o slowing, slowing, normal, uninterpretable) of waveforms in 150 30 minute EEG epochs. Each clinician reviewed a unique set of 150 epochs from the full dataset's 300 30-minute epochs. Over the 15 inter-rater pairs, their inter-rater κ scores ranged from 0.29 to 0.62 suggesting fair to substantial agreement among the pairs.

Westhall et al. [29] had a smaller subject pool, 4 clinicians, but asked them to evaluate EEG recordings for specific to *Prespecified EEG patterns*, *Background EEG*, or *Periodic or rhythmic patterns*. Each $>$ 20 minute recording was drawn from a pool of 103 comatose cardiac arrest patients. For the prespecified EEG patterns the κ statistics ranged from 0.42 to 0.71, Table 2.6. Meanwhile, the background and periodic patterns produced inter-rater κ statistics between -0.07 to 0.82, Table 2.7.

Just as the results of Gerber et al. showed strongest performance for critical waveforms, Westhall et al. did too. However, performance outside these critical waveforms was extremely poor in terms of classification agreement and κ statistics. This might have been caused by the increase in classification categories, compared to Gerber et al., Grant et al., or Halford et al, but more likely suggested the

Table 2.5. Inter-rater Clinician κ Agreement

Clinician	Clinician Pair					
	A	B	C	D	E	F
A	-	0.43	0.52	0.37	0.37	0.50
B	-	-	0.48	0.41	0.37	0.29
C	-	-	-	0.49	0.56	0.62
D	-	-	-	-	0.48	0.35
E	-	-	-	-	-	0.42
F	-	-	-	-	-	-

The pair averaged κ score was 0.44 which made the overall agreement moderate.

Table 2.6. Inter-rater Clinician Agreement

EEG Waveform	Agreement (%)	κ statistic
Highly Malignant	75	0.71 (0.55-0.79)
Malignant	63	0.42 (0.34-0.51)
Benign	63	0.42 (0.34-0.51)

Agreement and Kappa statistics using the ACNS classification labels for inter-rater performance on specified EEG patterns.

clinicians fundamentally disagreed over the non-prespecified EEG patterns given their previously discussed terminology consensus. Conversely, if background EEG or periodic patterns were necessary to make a diagnosis it would be difficult to resolve an understanding from the work of these clinicians.

Table 2.7. Inter- and Intra-rater Agreement

Condition	Inter-rater Agreement (%)	κ	Intra-rater Agreement (%)	κ
Background EEG				
Continuity	37	0.76	62	0.86
Voltage	47	0.65	75	0.31
Predominant Frequency	3	0.36	30	0.17
Reactivity to sound	42	0.25	82	0.76
Reactivity to pain	32	0.17	69	0.44
Periodic or rhythmic patterns				
Periodic or rhythmic discharges	50	0.56	80	0.55
Prevalence	39	0.49	70	0.58
Typical frequency	6	0.82	55	0.80
Maximum frequency	14	0.74	54	0.68
Sharpness	74	0.73	75	0.58
Absolute amplitude	44	0.42	86	0.59
Stimulus induced pattern	63	0.19	80	0.48
Evolution	13	0.19	76	0.30
Plus Modifier present	19	0.17	84	0.28
Triphasic morphology	61	-0.07	63	0.00

2.2.1.2 Clinician Intra-rater Agreement

Clinicians difficulty in producing acceptable κ statistics in inter-rater testing extended themselves via intra-rater testing as well. In most cases, intra-rater agreement addressed a clinician's ability to reproduce annotations on data they previously annotated. Gerber et al., Grant et al., and Westhall et al. ran specific intra-rater experiments to track inter-rater behavior.

Gerber et al. evaluated the ability of 5 clinicians to reproduce their results on the 10 second epochs 12 months after the original study. The same epochs were used, presented in a randomized order, and each clinician was asked to follow the classification scheme from the original study. The resultant κ statistics, Table 2.8, showed the difficulty clinicians had in agreeing with themselves. Compared against inter-rater agreement, Table 2.4, the intra-rater agreement was only marginally better.

Table 2.8. Intra-rater agreement after 12 months

Clinician	Rhythmic/ Periodic vs. Excluded	Local.	Morp.	Freq.	"Quasi" vs. Not	"Frontally Predominant" vs. Not	"Plus" vs. Not
1	0.79	0.58	0.67	0.30	0.28	0.32	-0.03
2	0.86	0.60	0.55	0.24	0.25	0.38	0.00
3	0.68	0.51	0.15	0.28	0.32	0.45	0.28
4	0.73	0.68	0.58	0.29	-0.08	0.57	0.24
5	0.76	0.46	0.40	0.19	0.28	0.67	0.00
Mean κ	0.76	0.57	0.47	0.26	0.21	0.48	0.098

The 5 clinicians in the original 10s epoch evaluations, re-evaluate the same set of data 12 months later. These results represent how well each clinician agrees with their original classifications.

The follow-on experiment in Grant occurred 4 months after the initial study. In this case, the range of intra-rater agreement (0.33 to 0.73, mean of 0.59) was better than that of the inter-rater agreement (0.29 to 0.62). However, the intra-rater results suggested clinician A was the worst performer. This conflicted with clinician A's inter-rater agreements, Table 2.5. The worst inter-rater agreements did not involve clinician A, but rather clinicians B, D, and F. These results suggested inter- and intra-rater agreement scores were poor tools for understanding a clinician's annotation ability, but confirmed their ability to generate consistent diagnoses.

The trend of intra-rater agreement, Table 2.9, scoring higher than inter-rater agreement, Table 2.6, was repeated by the clinicians Westhall et al tested as well. Repeating their original experimental protocol 6 months later produced very high intra-rater classification agreements, Table 2.9. However, the κ statistic for highly malignant, 0.64, was lower than its inter-rater counterpart, 0.71. Despite each clinician improving and/or maintaining their classification ability, they were unable to identify the same waveforms as they did previously. This again spoke to nature of clinicians ability to only need a minimum amount of insight to generate a consistent diagnosis.

Table 2.9. Intra-rater Classification

EEG Waveform	Agreement (%)	κ score
Highly Malignant	88	0.64 (0.48-0.83)
Malignant	98	0.93 (0.57-1.00)
Benign	98	0.93 (0.57-1.00)

Agreement and Kappa statistics using the ACNS classification labels for intra-rater performance.

The other features in Table 2.7 represented less discrete facets of EEG waveforms. These features required qualitative analysis which increased the difficulty of classification consensus, exemplified by the abundance of slight and poor inter-rater κ statistics. Intra-rater agreement showed minimal improvement of κ statistics, while the averaged intra-rater agreement % was better than its counterpart. This suggested clinicians were capable of reproducing their work, but were prevented from doing so by their innate biases thus limiting their κ statistics.

As a whole these intra- and inter-rater studies indicated clinicians were consistent within themselves, and their cohorts, when classifying EEG recordings. However that consistency did not appear to translate into producing data acceptable for use as a gold standard. While the results of each study offered suggestions as to why such consensus was difficult to reach, there was no single conclusive factor. The size of the epochs, the category of classification, the duration of the annotated waveform, and the clinician's training and experience all impacted the resultant κ statistics. Their inability to come to agreement did not, however, diminish their ability to diagnosis. The only shortcoming was that it limited the quality and quantity of data available on which to train ML algorithms.

2.2.2 Algorithm Classification

Despite robust waveform nomenclature, translating EEG signals into features for algorithm classification was an open field. With no feature consistency, each study was free to develop their own features such as using a unique feature set [14], borrowing from a previous study's features [134], or forgoing features and using the raw data directly [135]. Regardless of the type of features, they all segmented the recordings into *epochs* which served as the input to the algorithms.

Most epochs represented a window in time, typically on the order of seconds, that contained the data from one or more EEG channels. The duration of the epochs drove a trade off between categorizing phenomena occurring rapidly, PDs, or slowly, such as sleep states. Given the number of channels in a recording, their duration, and the sampling rate EEG recordings typically produced significant amounts of data. The use of epochs was the first step of dimensionality reduction by attempting to normalize the raw data into manageable segments across channels, subjects, sessions, and datasets.

Thus the features used for these epochs needed to excel at minimizing the amount of data while maximizing the information density relative to the data type. This was a difficult task given the depth of EEG signals which was why feature sets were frequently developed for specific use cases like seizures [13], BCIs [136], sleep [24], alcoholism [46], ADHD [42], and beyond. The combinations of features and epochs allowed each study to focus on their specific goals, but made it difficult to produce a robust universal feature set.

This problem was compounded by the EEG community's continual adaption of the newest ML algorithms in an effort to increase classification performance. This behavior was not much different from the development of speech technologies until they resolved a robust universal feature set [137] as they developed a myriad of techniques to address their classification problems, such as K-Nearest Neighbors (KNNs), Support Vector Machines (SVMs), Neural Networks (NNs), and GMMs. Often a given a combination of features and datasets performed better or worse than another depending on the algorithm and its parameters. This made it hard to determine if performance gains were due to algorithms, dataset, feature set, or something else.

The following sections reviewed algorithms that used *statistical models*, *supervised algorithms*, and *unsupervised algorithms* common to the EEG classification landscape. Statistical models formed the basis of numerous ML techniques and were frequently used to filter out artifacts via thresholding, detect ERPs, or interpret common spatial patterns (CSPs). Supervised algorithms used labeled data from clinicians and *a prior* knowledge to build classifiers focused on specific phenomena like seizures and mental states. Meanwhile, unsupervised algorithms leveraged the power of statistical models built from large unlabled datasets to classify conditions for which annotations were hard to obtain. These techniques were applied at one time or another on datasets generated from sleep, seizures, ADHD, or BCI EEGs.

2.2.2.1 Statistical Algorithms

Statistical modeling of known EEG phenomena provided a robust platform for developing basic classification algorithms. The type of modeling depended on the waveform, similar to how features were adapted, but classification was primarily based on one-versus-all evaluation. These approaches were mathematically straightforward and required minimal data relative to the defined phenomena. Their success, however, was data dependent as they required a thorough set of labeled data to operate. This made them ultimately reliant on the knowledge on clinicians.

An ERP represented an involuntary response by the brain when it perceived a targeted external stimulus. One of the most common instances of these events, the P300 response, was used to development basic BCI spellers. A P300-spellers were built to detect responses to auditory and/or visual stimulus enabling a person to spell words with their brain [76]. This phenomena was ideal for statistical modeling as brief subject specific training readily produced acceptable performances [138].

Guger et al. [138] showed that 5 minutes of training were enough to elevate the majority of the subjects to 60% or better accuracy, Figure 2.5. The training period asked the subjects to spell specific words and then used Linear Discriminate Analysis (LDA) to tune the weights of the 8 pre-selected channels. Subjects operated the speller by responding to a single character being flashed, single character speller, or by alternating flashing of rows and columns, row-column speller.

Figure 2.5. ERP Classification Performance

Classification accuracy (%)	Row-column speller % of sessions 81 subjects	Single character speller % of sessions 38 subjects
100	72.8	55.3
80-100	88.9	76.3
60-79	6.2	10.6
40-59	3.7	7.9
20-39	0.0	2.6
0-19	1.2	2.6

This approach represented a highly effective real-time communication platform that did not require excessive training data nor overly complex signal processing. The main drawback was the time required to produce a single letter, 28.8 seconds for row-column spelling and 54 seconds for single character spelling. The technique itself was very specific to ERPs which meant it did not contribute much to other EEG applications. This necessitated the development of different statistical models for addressing the detection of Alzheimer's Disease (AD) [90], ADHD [139], and seizures [140] events.

The ability to detect and classify seizures has remained a core focus of EEG research in terms of reviewing existing recordings as well as enable accurate predictions. Chu et al. [13] applied *attractor states*[7] to EEG data in an effort to improve seizure prediction and detection via statistical discrimination. The technique was tested on two datasets, the Children's Hopsital of Boston Massachusettes Institute of Technology Scalp EEG Database (CHB) and adult seizures from the Department of Neurosurgery of Seoul National University Hospital, using 50% overlapping channel independent 20s epochs. The raw epochs were converted to frequency banded Fourier coefficient features used to build seizure and non-seizure state models.

Their seizure predictions, using a 30 second horizon, averaged 90.20% sensitivity on the training data and 86.67% sensitivity on the testing data (2 subjects reported 0%). Decreases in sensitivity correlated with a drop in average false positives per hour from 0.476 on the training data to 0.367 on the testing data. The peak rate of false positives were 1.667 and sensitivity for multiple subjects was 0.0%. The results suggested a simple model can predict seizure onset, correctly predicting 39 of the 45

[7] Attractor states are stable states which the data trends towards given its natural behavior. The concept originated from the work of Scheffer et al.[140], but is beyond the scope of discussion in this work.

documented seizures across the 17 subjects. However, the failure to detect anything for 2 subjects (1 seizure each) and missing 2 seizures from another subjects indicated the technique may not be sufficient for all types of seizures nor all patients.

Understanding sleep cycles aided in understand seizures given seizures frequently occur at night [6], but first the stages of sleep needed to be classified. Warby et al.[24] compared the performance of six statistical sleep spindle, sleep stage markers, algorithms[8] against clinicians and non-experts. The dataset consisted of 32,112 25s single channel epochs from 110 healthy subjects split into training, testing, and verification data. The verification data, built from 2,000 epochs scored by 5.3 clinicians on average, serves as the gold standard.

Each of the algorithms applied different flavors of energy thresholding (Root Mean Squared (RMS), PSD, or Fast Fourier Transform (FFT)) on a bandwidth filtered (9-16Hz) portion of the epochs. The algorithms' performances, Table 2.10, were not in agreement with the gold standard (GS), but they did agree with the automated group consensus (AGC). Overall, the algorithms were the weakest classifiers while the clinicians were the strongest at classifying sleep spindles. The non-experts performed better than the algorithms which suggested these statistical based algorithms may not be an effective classification technique for this task.

Huang et al. [145] aimed to detect the presence of AD in a set of 93 subjects labeled as having AD, mild cognitive impairment (MCI) or healthy controls. Classification used the 15 2s epochs from each subject which were built on their alpha (8.0-11.5Hz) and theta (4.0-7.5Hz) global field potential (GFP), a generalized EEG amplitude measurement. The algorithm reported an AD classification accuracy of 84% against control subjects. This represented an optimal feature set,

[8] The six algorithms were drawn from six unique studies cited here: {a1[141], a2[142], a3[143], a4[144], a5[99], and a6[57]}

Table 2.10. Sleep Spindle Detection F1 Scores

Algorithm	Gold Standard	Non-Expert Group	Automated Group
a1	0.28	0.22	0.28
a2	0.28	0.30	0.40
a3	0.21	0.17	0.21
a4	0.50	0.46	0.79
a5	0.52	0.49	0.84
a6	0.41	0.37	0.48

which started as epochs of FFTs decomposed into their GFP across frequency bands (delta (1-3.5Hz), theta, alpha, beta 1 (12-15.5Hz), and beta 2 (16-19.5Hz)). These features were then localized with respect to regions of the brain: antero-posterior (Loc-X), left-right (Loc-Y), and superior-inferior (Loc-Z). The resultant values of each feature permutation is shown in Figure 2.6.

ADHD was another omnipresent condition that could be detected through a subject's theta beta ratio (TBR) [146]. Lenartowicz et al. [146] reviewed multiple approaches for distinguishing ADHD patients from controls based on temporal and spatial features and the ratios of energy present in frequency bands and specific channels. The studies reported divergent performance when using TBR as a discrimination metric. Monastra et al. [139] reported an accuracy of 91% (90% sensitivity, 94% specificity) while Buyck et al. [147] reported an accuracy of 49-55%.

Detecting ADHD through EEG recordings appears possible based on the TBR, but Lenartowicz et al. conclude the technique is not reliable enough to be a diagnostic test. The work of Monastra et al. was carried out in 2001, but advancement in the field, like Buyck et al.'s 2014 work, indicate variations in ADHD morphology make TBR a poor classification metric. Despite a clear clinical utility in using EEG recordings for ADHD diagnosis [148], the research suggested the condition was not

Figure 2.6. Raw Feature Means for AD Classification

Band	Group	GFP	Loc-X	Loc-Y	Loc-Z
Delta	AD	13.4(9.3)	12.5(9.8)	1.8(4.2)	-5.6(6.0)
	C	7.3(2.3)	12.9(8.6)	0.1(4.7)	-4.4(5.8)
	MCI	10.4(5.2)	12.2(11.3)	1.8(4.6)	-6.2(6.0)
Theta	AD	15.6(14.6)	-2.6(7.6)	2.1(5.2)	-0.2(6.9)
	C	8.0(6.5)	-5.7(7.2)	1.4(5.9)	-4.0(5.0)
	MCI	10.2(10.8)	-3.6(12.3)	2.7(5.5)	-2.0(6.3)
Alpha	AD	14.1(14.5)	-12.6(11.5)	-2.1(7.1)	1.7(8.9)
	C	31.2(30.2)	-21.0(7.3)	-0.4(5.4)	-3.4(7.2)
	MCI	40.1(43.3)	-19.9(11.1)	-0.1(6.3)	-1.7(9.3)
Beta 1	AD	3.7(3.7)	-6.2(11.2)	-1.5(8.9)	5.2(9.9)
	C	3.6(1.9)	-12.1(10.1)	2.2(5.8)	1.4(9.1)
	MCI	5.2(5.2)	-13.9(12.3)	1.3(6.9)	2.5(8.8)
Beta 2	AD	2.1(1.7)	0.3(12.8)	-2.3(10.8)	8.3(10.6)
	C	2.9(1.7)	-8.2(11.8)	1.8(7.4)	4.4(8.6)
	MCI	4.2(4.6)	-8.8(13.9)	1.0(10.4)	4.8(11.0)

The table contains the mean values of the GFP for each frequency band over a given brain region. These represent the features the algorithms uses to discern AD subjects from MCI subjects and healthy controls.

yet understood to the point of being able to develop a robust statistical classification model for it.

However, Buyck et al. found that TBR did make an excellent, AUC 0.965, discriminator for age classification. This exemplified the difficulty in building a robust feature set for a given classification task as differnet sets of features could conflate multiple conditions. The best examples of this were the efforts made for detection and correction of EEG artifacts [8].

The most common artifacts (eye blink, muscle artifacts, and eye movements) were caused by the subject making them difficult to mitigate during recording. Jung et al. [149] indicated the overlap between artifacts and waveforms of interest prevents many novel artifact detection techniques from having a broader impact. Their work compared the performance of independent component analysis (ICA) to principal component analysis (PCA) on a dataset of normal and autistic subjects. Despite both techniques being capable of separating the signals from the noise, ICA offered the best performance for correcting the original recordings.

Delorme et al. [150] devised a more comprehensive experiment[9] for detecting artifacts. They applied six thresholding schemes to raw data and data processed each with ICA, building on the success of Jung et al. Their results, Figure 2.7, showed that applying ICA improved the classification performance regardless of the artifact's source. However, the use of ICA did not improve the performance of each algorithm.

[9] They compared five methods to determine how best to identify artifacts within a recording. (1) Extreme values: Artifacts detected if amplitudes exceeded a predetermined threshold. (2) Linear trends: Least squares thresholding against an average of the activity in an epoch. (3) Data improbability: Likelihood of an observations with respect to all observations from each channel. Each epoch became a product of likelihoods which should decrease if artifact events are detected. (4) Kurtosis: Measure the 'peakedness' of each epoch's distribution. (5) Spectral pattern: model scalp topology in conjunction with frequency spectrum.

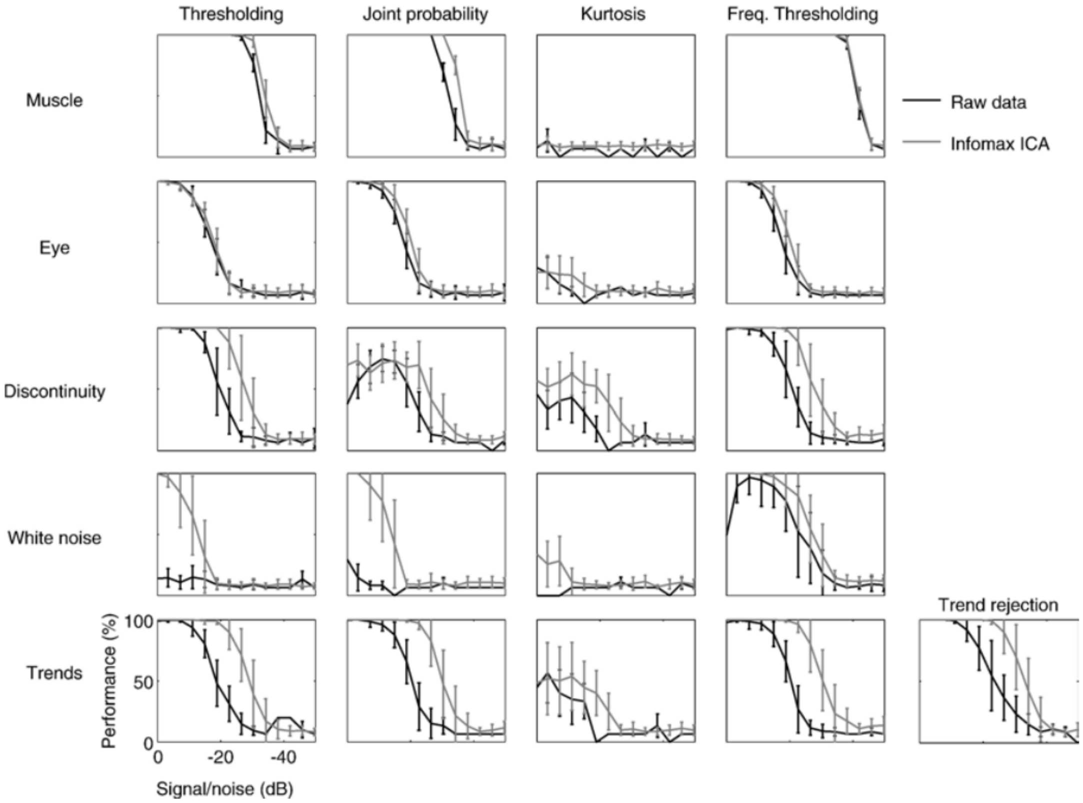


Figure 2.7. Statistical Thresholding of Artifacts. Classification performance of thresholding approaches based upon the signal to noise ratio of the artifact and the signal.

The largest changes in performance were related to the artifacts (discontinuity and white noise) and not the algorithms. This suggested that the algorithm's performance does matter, i.e. Kurtosis performed universally poor, but ICA was only able to impact performance when the noise was distinct from the signal. This, of course, is the definition of ICA, but highlighted the problem of its use on EEG data where waveforms and artifacts presented as seemingly identical signals.

Despite this limitation, these experiments were mostly a success which lead to the development of Fully Automated Statistical Thresholding for EEG artifact Rejection (FASTER) [8] and Automatic EEG artifact Detection based on the Joint Use of Spatial and Temporal features (ADJUST) [40]. These techniques provided universal artifact detection and rejection across multiple types of EEG data. FASTER relied on a parameter set consisting of variance, Hurst exponent[10], amplitude range, and channel deviation over five thresholding levels (channel, epoch, epoch ICA, channel-epochs, and channel average). ADJUST used spatial and temporal feature extraction to classify and remove artifacts from ICA filtered data. These results again highlighted how different feature sets and algorithms achieved acceptable performance making it hard to know which option is 'right'.

Table 2.11. FASTER's Artifact Detection Performance

Channels	Channel Sensitivity(%)	Channel Specificity(%)	Epoch Sensitivity(%)	Epoch Specificity(%)
128	94.47	98.96	60.24	97.53
64	97.02	98.48	61.83	97.54
32	5.88	96.81	58.64	97.49

[10] The Hurst exponent is a measure of the changes in lag observed from the auto-correlation of pairs of points in a time series.

ADJUST was more complex, but achieved 60% sensitivity and 97% specificity, Table 2.11, on 2 second channel independent epochs drawn from 47 subjects. FASTER was less complex than ADJUST and replicated the clinician's 95.2% artifact detection performance. The most significant aspect of these results were that the testing dataset was built from 10 subjects that were withheld from the 21 subject training dataset.

Artifact detection and correction continues to be an active research topic, but the reliance on ICA remained. Mahajan et al.[87] reported exceptional performance using ICA on 12 electrodes followed by modified multiscale sample entropy (mMSE) and Kurtosis and thresholding. Their eye blink detection algorithm reported 90% sensitivity and 98% specificity across four subjects.

These statistical approaches to classification are promising, but they are developed on small datasets with simple goals. Adapting them for use on larger datasets with more extensive classifications needs seemed to be beyond their capability. At the very least they showed when EEG signals were broken down to their core components it was possible to reliably discriminate among them. This suggested reiterated the idea that it was possible ML algorithms to at least match a clinician's performance.

2.2.2.2 Supervised Algorithms

Supervised ML algorithms build statistical models from datasets with labeled classes. Each class would ideally represent a subset of related data (artifacts, sleep spindles, or ETs) that the algorithm would learn to distinguish between. Given a diverse feature set, the algorithms build decision surfaces based upon the strongest statistical properties of the features unique to each known class. These decision surfaces allowed classifications to be learned instead of having to infer them directly from the dataset.

These algorithms were setup with the aim of emulating a clinician's classification performance. In doing so, they tied themselves to the performance of those provided the labeled data. This is the main limitation of supervised learning: The algorithms must be shown what to classify making their success dependent on the properties of the training data. If the test contains a new class, the algorithm will struggle to define it and it may go undetected unless additional analyses were undertaken. However, the strength of this approach is that supervised ML classification algorithms work extremely well for well known phenomena (artifacts, seizure, and sleep). This has been shown to be true even when such conditions occurred rarely or were learned from a small number of epochs [151]. This naturally meant they worked best paired with phenomena in smaller sets of clinically annotated data (BCIs, emotions, and workload) EEGs.

The classification of sleep relied on detecting waveforms known as k-complexes and sleep spindles which are unique to a sleeping brain. There is also generalized brain activity specific to the energy bands that accompany each stage of sleep[126]. Thus each stage of sleep contains a mixture of unique waveforms and shifts in the rhythms, ratios of energy in the EEG bands, and waveforms that make it distinct from other brain conditions. Such behavior is most notable in the that dominant (>50%) alpha rhythms where remain indicative of being awake. Stage 1 typically contains a split (50%\50%) of alpha and delta rhythms. Stage 2 contains sleep spindles and diminished (<20%) delta rhythms. Stage 3 sees a resurgence (20%-50%) of delta rhythms. Stage 4 and REM sleep are classified by dominant delta rhythms.

These discrete states made the adaptation of supervised ML algorithms straightforward. In Schluter et al.[35] the stages of sleep were classified with

Decision Trees (DTs) by bagging[11] on an array of physiological data[12]. The classification was performed on 33,542 30 second epochs drawn from 15 subjects, Table 2.12. On the whole separating wakefulness, REM sleep, and from the stages of sleep was excellent. However, identifying the distinct stages of sleep proved difficult especially for stage 1 and stage 3. These results incorporated data in addition to the EEG recordings, suggesting EEG alone may not be sufficient for accurate classification.

Table 2.12. Classification of the Stages of Sleep

Stages	W	S1	S2	S3	S4	REM
W	97.0	2.4	0.6	0.1	0.0	0.5
S1	9.1	58.1	20.2	0.8	0.2	11.6
S2	0.5	4.7	91.7	5.5	0.8	0.2
S3	0.0	0.1	20.2	62.8	18.2	0.1
S4	0.1	0.2	1.0	12.6	86.8	0.1
REM	0.7	2.3	3.0	0.1	0.0	96.6

Radha et al. [102] used different algorithms to classifying the stages of sleep, but produced similar results to that of Schluter et al. Their data consisted of 30s epochs of 34 features drawn from 10 health subjects. They compared two supervised algorithms, Random Forest (RF) and SVM, ability to classify the epochs into REM sleep and the 3 stages of non-REM sleep (N1,N2,N3). By using supervised algorithms it was necessary to have a clinician provide labeled training data. However, this also allowed a κ statistic to be associated with each algorithm's performance relative to

[11] Bagging, bootstrap aggregating, is a technique employed to reduce the variance of ML algorithms. The original data was re-sampled with replacement to produce multiple data sets containing redundant data.

[12] Sleep studies frequently collect electrocardiogram (ECG), EEG,electromyography (EMG), and electrooculography (EOG). In this work, aside from EEG data, EMG and EOG were used to help classify the sleep stages.

the clinician, Figure 2.8. Prior to classification the feature set was optimized for the differential montage channel (F4-A1), an epoch duration of 30s, and only 20 of the original 34 features.

Figure 2.8. Single EEG Channel Sleep Scoring

Sleep Stage	SVM 1vA Precision	SVM 1vA Recall	SVM 1v1 Precision	SVM 1v1 Recall	RF Precision	RF Recall
W	0.86	0.51	0.75	0.71	0.78	0.73
N1	0.00	0.00	0.18	0.00	0.52	0.31
N2	0.86	0.83	0.85	0.88	0.85	0.91
N3	0.32	0.70	0.82	0.70	0.83	0.73
REM	0.56	0.55	0.58	0.79	0.69	0.70
Accuracy	0.69		0.77		0.80	
κ	0.46		0.61		0.66	

Precision and recall of SVM and RF classification using a single EEG channel for sleep stage classification. In this study non-REM sleep is broken into only three stages (N1, N2, N3) making it difficult to compare to the standard four non-REM stages of sleep shown in Table 2.12.

These results were comparable, Table 2.12, to Schluter et al., which was a study that used far more data. The moderate to substantial κ statistics suggested the algorithms performed as well as a clinician would have given the previously reported inter-rater agreements. However, it was possible that the feature optimization drove this performance. Sleep states were not a unique phenomena and they tended to represent major changes in brain activity, the necessity of channel and feature optimization suggests this algorithm/feature combination was only able to find the strongest indicator of sleep and may be missing out on the nuances of individual sleep stages.

Similar to sleep, seizures had frequently been categorized into distinct stages: *normal* indicative of a normal healthy state, *pre-ictal* indicative of a build up to a seizure, *ictal* indicative of an active seizure [152], and *post-ictal* indicative of the time following a seizure [13]. Accurate detection of these stages, specifically pre-ictal, could help improve the diagnosis and treatment of epilepsy [6]. Seizure classification was always a primary research focus of automated algorithms because of number of people affected by them[6]. Effort has been continually applied to improve the classification of seizures which tended to focused on developing better features than the FFT based frequency band powers [14, 54, 13] and improving algorithms [152, 25, 153]. These efforts were predicated on, and thus limited by, the availability of annotated data and the quality of the annotations.

Wulsin et al.[14] used raw data and diverse feature subsets derived from a stock listing of features [13] to compare seizure detection as a function of algorithms and features. Despite efforts to find a suitable feature subset, the strongest classification occurred when using the raw data as the input features. In addition to the feature analysis, four classification algorithms (DTs, SVMs, KNNs, and Deep Belief Networks (DBNs)) were evaluated with SVMs producing the best classifications, Figure 2.9.

Bajaj et al.[54] used emperical mode decomposition (EMD)[14] features as inputs for a Least Squares Support Vector Machine (LS-SVM) driven seizure classifier. The data was sourced from 100 23.6s channel epochs drawn from 5 subjects. EMD separated the nonlinear and non-stationary components of the EEGs into intrinsic mode functions (IMFs). The two dominant IMFs, amplitude modulation and frequency modulation, produced a peak sensitivity and specificity of 100% while averaging 94% sensitivity and specificity over the dataset by using a common supervised ML algorithm in SVMs. As the classifier was not knew, the success of this work was likely driven by the use

[13] area, normalized decay, frequency band power, line length, mean energy, average peak/valley amplitude, normalized peak number, peak variation, root mean square, wavelet energy, and zero crossings

[14] A detailed review of EMD is omitted, but if interested the work of Huang et al.[154] introduced technique and its applications.

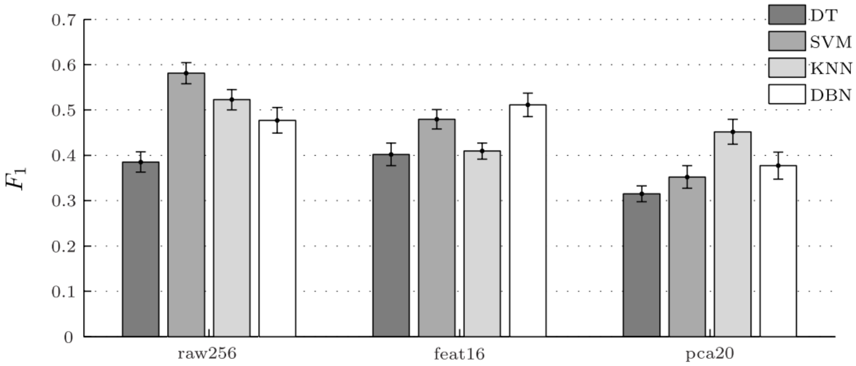


Figure 2.9. F1 Performance of Four Supervised Algorithms. Wulsin et al. evaluate algorithm peformance based upon the F_1 measure, where $F_1 = 2 * (sensitivity * precision)/(sensitivity + precision)$. The results are presented to compare the algorithms and feature sets against each other. The feature sets are comprised of: *raw256* represents the raw waveform data, *feat16* are the hand selected 16 features, and *pca20* are the 20 features chosen by PCA.

of EMD features or qualities of the 5 subject dataset. This was a recurrent problem with EEG algorithm development of unique datasets and unique features obscuring the cause of classification improvement.

The alternative to diverse features and datasets was to test range of algorithms. This is what Acharya et al.[152] did by focusing on six supervised ML algorithms: Fuzzy Sugeno Classifier (FSC), SVM, KNN, Probabilistic Neural Network (PNN), DT, and Naive Bayes Classifier (NBC), and one unsupervised ML algorithm: GMM. This was a better approach than Bajaj et al.'s as it provides multiple reference points on a constrained dataset. These six algorithms used four different types of entropy calculations as features: Approximate Entropy (ApEn)[155], Sample Entropy (SampEn)[156], and S1 entropy and S2 entropy[157]. Distinct epochs were drawn from 5 healthy and 5 epilepsy subjects that produced 200 healthy, 200 pre-ictal, and 100 ictal artifact free single channel 23.6 second epochs.

Table 2.13. Algorithm Performance Using Entropy Based Features

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
FSC	98.1	99.4	100
SVM	95.9	97.2	100
KNN	93.0	97.8	97.8
PNN	93.0	97.8	97.8
DT	88.5	98.3	91.1
GMM	95.9	98.3	95.6
NBC	88.1	94.4	97.8

Classification accuracy of entropy based feature sets for various classifiers.

The sensitivity and specificity the algorithms were similar, Table 2.13, but the best accuracy was achieved by the FSC classifier. The separability of the trained seizure states, Table 2.14, produced a p-value less than 0.0001 for each entropy. However, it was hard to assess the strength of the individual algorithms given the small size of the dataset and the natural discrimination strength of the features. The strong performance across all the algorithms suggested the results were driven by the features, but the dataset was, again, too small to have known for certain.

Table 2.14. Entropy Level Based Seizure Labeling

Class	Normal	Pre-ictal	Epileptic
ApEn	$2.2734 \pm 3.320 \times 10^{-2}$	1.8650 ± 0.331	1.9325 ± 0.215
SampEn	1.3130 ± 0.120	0.99332 ± 0.189	0.92628 ± 0.139
S1	$0.57012 \pm 7.120 \times 10^{-2}$	$0.47208 \pm 6.149 \times 10^{-2}$	0.48325 ± 1.55
S2	$0.76827 \pm 3.125 \times 10^{-2}$	$0.68072 \pm 3.790 \times 10^{-2}$	$0.73184 \pm 4.555 \times 10^{-2}$

Ghosh-Dastidar et al. [25] benchmarked a novel wavelet-chaos-neural network Levenberg-Marquardt Backpropagation Neural Network (LMBPNN) classifier against the same data as Acharya et al. They picked features (standard deviation, correlation dimension, and largest Lyapunov exponent) that were specific to each frequency band and grouped them together into various band specific sets. The epochs were evaluated by supervised techniques (Radial Basis Functional Neural Network (RBFNN) and LMBPNN), an unsupervised technique (k-means clustering), and statistical discriminant techniques (Quadratic Discriminant Analysis (QDA) and LDA using Euclidean and Mahalanobis distance metrics).

The various combination of band-specific features sets were used to resolve an optimal set for the classifiers. These tests provided an exhaustive analysis of the relationship between algorithm and feature set performance, which was frequently lacking in other research. However, the best performance resulted when using a mixed-band feature set. The impact of feature set on the performance of LMBPNN is seen in Figure 2.10.

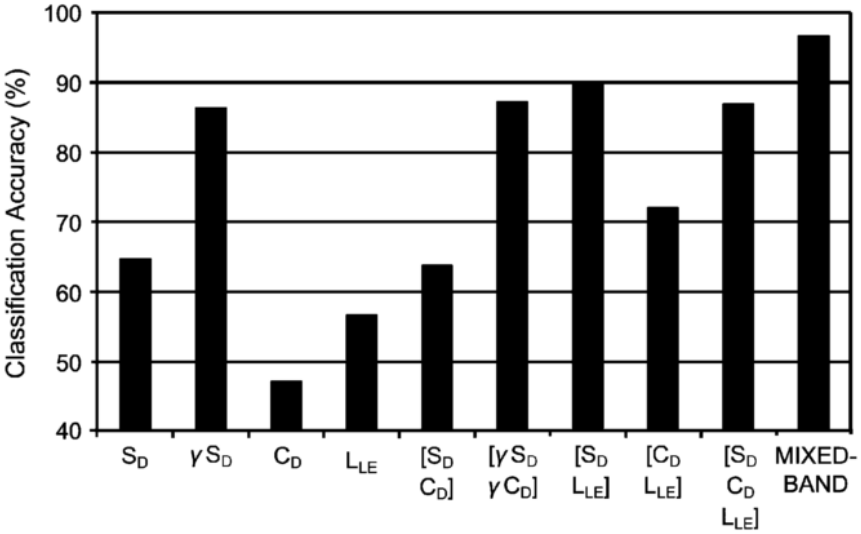


Figure 2.10. Impact of features on LMBPNN classification. When iterating over the available feature sets, the performance of LMBPNN responds differently for each combination. The Greek letters indicate the band specific features (S_D is standard deviation, C_D is correlation dimension, and L_{LE} is Lyapunov exponent). The mixed-band feature set uses the band independent S_D and C_D with $\alpha S_D C_D L_{LE}$, $\beta S_D C_D$ and $\gamma S_D C_D$.

The classification performances were similar so only the maximum accuracy was reported in Table 2.15 and even these cover a wide range. Overall, the proposed LMBPNN provided the strongest peak performance, but it relied on a large mixed-band feature set. This suggested that the features were the driving force of classification, and yet only some of the algorithms were able to adequately used them. Again, the difficulty in improving performance appeared to stem from being able to model the phenomena in a meaningful way for the chosen classifiers.

Table 2.15. Classification Accuracy of Single and Mixed-Band Features

Algorithm	Maximum Accuracy (%)
k-means	59.3
LDA w/ Euclidean	79.6
LDA w/ Mahalanobis	84.8
QDA	85.5
RBFNN	76.5
LMBPNN	89.9
QDA*	93.8
LMBPNN*	96.7

The values shown are the maximum accuracy achieved by each algorithm given on a single or (*) mixed-band feature set.

2.2.2.3 Unsupervised Algorithms

Unsupervised ML algorithms differed from supervised ML algorithms in that they do not require labeled data. This has made them a historically useful as starting point when knowledge of a domain was limited. Their decision surfaces were created directly through the data which removed any bias present in the labeling, but traded it for bias in the datasts. This also meant unsupervised approaches worked best when operating on datasets with enough data to represent each class of interest. Thus without an equitable distribution of data, classes may be ignored or poorly modeled leading to weak classification performance.

Given the need for large and diverse datasets and improvement to supervised classification techniques, the use of unsupervised classification of EEG recordings has diminished. However, unsupervised algorithms endured given their ease of use and ability to produce benchmarks for their supervised counterparts. Acharya et al.[152] showed GMM produced competitive accuracy and sensitivity, but not specificity to their tested supervised methods in Table 2.13. Alternatively, there were cases where it peformed much worse such as Ghosh-Dastidar et al.'s [25] k-means clustering of Table 2.15. As unsupervised algorithms relied on the dataset more than supervised techniques, such contrasting performances were common in the EEG literature.

Gabor et al.[158] tested a single unsupervised algorithm, a self organizing map (SOM)[15] NN, for seizure detection on 24 recordings from 22 subjects. The algorithm was trained to classify seizures from features produced by a wavelet transform using 4s epochs built from the 10 channels of each recording. A separate feature set using 8s epochs was used, but the duration was found to be too long as it masked out shorter seizures.

In total 62 seizures were captured from the 24 recordings of which the algorithm detected 56 (90%). The average false positives per hour (0.71) produced more false positives than true positives given the average recording duration of 22.02 hours. As discussed previously, unsupervised techniques are sensitive to the distribution of the

[15] A detailed review of SOMs was omitted, but if interested the work of Kohonen[159] formalized the implementation. This technique attempts to mimic the structure of the brain by parsing the data in an unsupervised fashion to create a flat, two dimensional, map linking elements of the data together.

training data which manifested in this case as poor false alarm rates. In addition, the age range (<1 to 43 years old), small training set (5 of the 24 recordings), and epoch duration were all factors working against the algorithm.

Not all unsupervised algorithms focus on classifying the data, as some are deployed for dimensionality reduction. One such approach was the use of of unsupervised LDA in areas where clinicians' skills were weaker such as BCI [83]. LDA is within the realm of factor analysis (FA), as are ICA and PCA. A detailed review of FA techniques is given in Section 2.4.1 and what follows now touches on their use in EEG applications.

Vidaurre et al. [83] used three flavors of LDA to enhance BCI performance over four datasets[16]. The experiments focused on developing an unsupervised solution to transitioning between training and feedback sessions of BCI tasks. Each version of LDA focused on a different aspect of the features: LDA-I, targeted changes in the pooled mean (PMean) between the features of the training and feedback data, LDA-II incorporated updates to the covariance matrix with PMean, and LDA-III scaled the mean and covariance using CSPs. These techniques were compared against a supervised version of LDA to determine the strength of the unsupervised techniques.

Unsupervised techniques main strength resides in their simplicity when compared to ever advancing supervised techniques. Their evaluation was frequently used to evaluate the performance gain versus increased complexity and requirements. This made head to head comparisons, Figure 2.11, critical to development of both type of algorithm. Using the first dataset, the supervised algorithms slightly out performed the unsupervised algorithms. On the second, larger, dataset the PMean based algorithms met or exceeded the performance of the

[16] The first dataset was comprised of 19 sessions recorded from 10 subjects performing motor imagery tasks. The second dataset consisted of 80 subjects performing 75 motor imagery trials with calibration. The third dataset involved 7 quadriplegics attempting to move use a BCI mouse. The final dataset was a repeat of the second dataset without any calibration for the users.

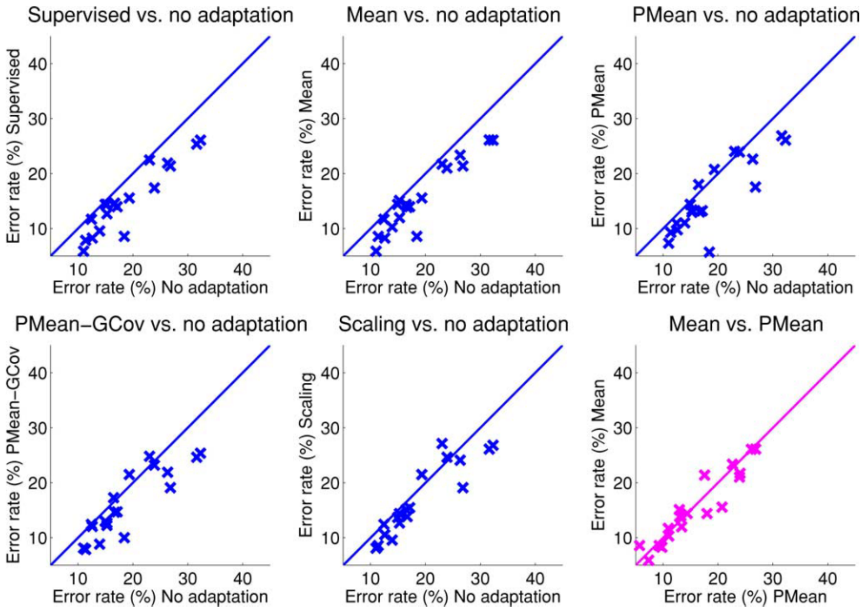


Figure 2.11. BCI Calibration Error. The comparative error rates between the supervised and unsupervised adaptation techniques through changes in the error rate. The pink plot shows the difference between a labeled, mean, and unlabeled, PMean, classification.

state of the art supervised approaches during feedback. The unsupervised technique exhibited robustness as a class was removed from BCI feedback and outperformed the supervised algorithm in Figure 2.12. These results were important because clinicians seldom label BCI datasets and it showed the trade off between supervised and unsupervised may not be that advantageous. This was especially true in this instance as BCI recording sessions were rarely annotated by clinicians, but often as dynamic as seizure or sleep sessions.

2.2.3 Bio-metric Applications

The use of EEG recordings as a means of bio-metric identification was not dominant area of research, but has begun to rapidly advance [105]. Initial efforts focused on being able to discriminate EEG behavior between individuals and between different

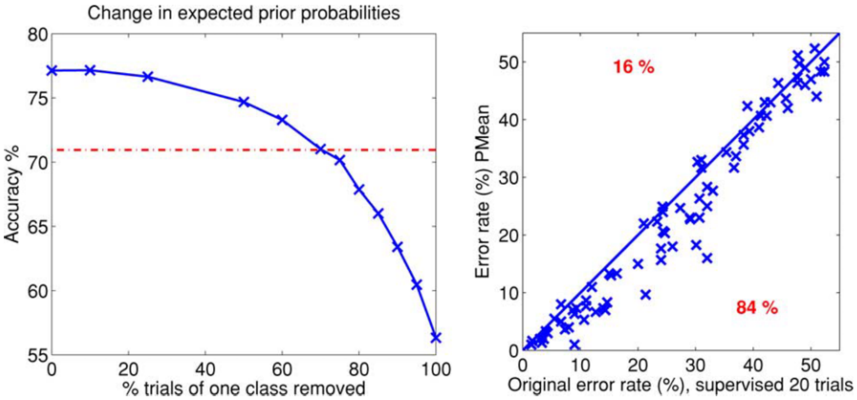


Figure 2.12. BCI Feedback Error. Performance on feedback data after training for supervised adaptation and unsupervised PMean adaptation. The (left) impact of removing one class from the feedback dataset for the supervised algorithm (red line) and unsupervised algorithm (blue line). The (right) error rate between the two algorithms during the online feedback experiments.

brain conditions [160]. This work did not have discrete waveforms to find or frequency ratios to calculate, but instead relied on direct comparison between subjects. Stassen [161] developed computerized methods, borrowed from speech recognition, to recognize normal and schizophrenic individuals based on their EEG spectral pattern. The style of this approach, finding dominant properties in subject epochs, remained in use [59] and was the best corollary to the research proposed in this dissertation.

Advancement of EEGs as a bio-metric tool focused on the statistical properties of each subject [105]. This detached it from the dominant research trends that were reliant on clinical annotations [103]. This made bio-metric applications open-ended as they do not, and in most cases cannot, rely on previously developed feature sets or decision surfaces built from clinical annotations. Researchers were therefore on their own to find features and testing protocols leading to a variety of approaches not seen elsewhere [63, 65, 104, 162, 163].

The initial efforts by VanDis et al.[160] and Stassen [161] focused on subjects at rest with eyes closed and open. Similar experiments contiuned to be carried out [33, 64, 65] but with their aims updated to optimize the accuracy and speed of subject verification as a function of features and channels. Active state recordings, when subjects performed mental tasks, such as imagined hand movements [68, 103], imagined speaking syllables [66], or reading text[39] underwent similar channel and feature optimization testing.

Active and resting based data analysis suggested that the qualities of subject authentication and identification existed regardless of brain state. Other works went as far as suggesting a genetic basis underlies this separability [37, 67]. While interesting, the genetics of brain uniqueness expanded beyond the scope of this work. However, by focusing on the techniques and results of active and resting based data studies comparisons could likely be drawn between the structured waveform based annotations of artifacts, seizures, and sleep.

2.2.3.1 Resting Recordings

The work of La Rocca et al.[33, 64, 65] focused on developing a novel set spatial and temporal patterns as features to improve subject recognition. Brigham et al.[66] used data with imagined activities to test applications of subject identification during mental tasks. These studies represent the adaptation of techniques the worked for other EEG classification tasks, spatial and temporal patterns for BCI and mental tasks for attention/focus/workload performance and ERPs.

In [33] electrode sets of 2, 3, and 5 from the 56 recorded channels were used to find a lower-bound on the number of required channels. The approach used autoregressive stoichastic modeling and polynomial regression to match 3 second epochs broken into features through the 6 standard EEG bands. Classification performance varied as a function of electrode set and the EEG band used. Increasing electrodes trended with an improvement in classification performance. However, regardless of the number of electrodes the alpha band provided the strongest classification accuracy. Performance peaked at 98% classification accuracy when using the alpha, beta, delta, and gamma bands for 5 channel sets. The best single band performance (83%) was seen using only the alpha band across 5 channels.

They followed up this work with ‘bump’ modeling to reduce the amount of data from the 10-20 layout into a parametric model [65]. These bumps were filters that enabling sparse encoding. This generated vectors to control the mapping/weights of the bumps scale the features of the data. These vectors were then classified with LDA based upon features generated from groups of three channels drawn from the six standard EEG bands. The training and testing sets were curated to provide overlapping frames, *jointed*, and without overlapping frames, *disjointed*. This distinction highlighted the impact of frame overlapping with the beta band producing a classification accuracy of 95% for jointed and 74% for disjointed. The alpha band resulted in similar classification accuracy with 96% for jointed and 67% for disjointed. In all bands the jointed feature sets outperformed their disjointed counterparts.

Their final work focused on spatial patterns generated from 1s PSDs epochs from different regions of the brain [64]. This deviated from their earlier attempts at reducing the amount of data through feature and individual channel reduction. Instead it grouped channels together to develop a statistical approach to subject verification using the PhysioNet Database dataset. Classification was carried out by building Gaussian mixtures based upon the distributions of the PSDs. These mixtures were evaluated via a Mahalanobis Distance (MD) classifier to determine likelihood of similarity between subjects. Using the results for each region of the brain, classification accuracy reached 100% for identifying subjects.

2.2.3.2 Active Recordings

In Marcel et al.[103] a nine subject dataset was classified based upon their brain activity during three mental tasks. These tasks required the subjects to imagine carrying out prescribed actions: moving their left hand, moving their right hand, and speaking words with a common leading letter. The feature set was built from 0.5s 50% overlapping epochs of PSDs. These PSDs were spatial filtered over the 10-20 electrode configuration with a surface Laplacian function. The features were given to a GMMs which produced baseline models for subject verification. Evaluation scores were reported as half total error rate (HTER) generated from the false acceptance rate (FAR) and false rejection rate (FRR).

$$HTER = \frac{FAR + FRR}{2} \tag{2.2-1}$$

The results, Table 2.16, of the left and right hand authentication of the subjects suggested performance was dependent on the number of Gaussian mixtures used in the modeling process. This experiment used a large datasets which was collected from the subjects over a three day period. Results using smaller subsets of the dataset showed the imaging word task authentication lagged compared to that of the hand tasks.

In Fraschini et al.[68] phase synchronization was tested as a feature set for identifying subjects. The dataset used the 109 PhysioNet Database subjects' resting eyes closed and resting eyes trials. The features were generated from the standard EEG frequency bands and segmented into 12s non-overlapping epochs. Finding the

Table 2.16. Imagined Activity HTER

Mental Task	Num. Gaussians	FAR	FRR	HTER
Left	4	18.6	32.3	25.4
	8	23.8	25.15	24.5
	16	19.3	19.65	19.5
	32	13.7	24.9	**19.3**
Right	4	18.4	40.5	29.4
	8	20.6	29.5	25.0
	16	15.0	23.6	**19.3**
	32	13.0	30.15	21.6

phase lag index (PLI) relationship between all the channels of an epoch produces distinct mappings between subjects. These topologies were reduced via Eigenvector Centralization to produce a feature vector for each epoch. The Euclidean Distance (ED) between each feature vector was the decision surface used to assert the similarity between the subjects for each frequency band. The results showed the equal error rates (EERs) of Resting Eyes Open and Resting Eyes Closed for the Gamma band were 4.4% and 6.5% and when using the Beta band were 10.2% and 16.9% respectively.

Brigham et al.[66] tested a similar subject identification protocol using two unique datasets. One source of data came from Visually Evoked Potentials (VEPs) in 120 alcoholic and non-alcoholic subjects. The other ws sourced from 6 subjects uttering two syllables, /ba/ and /ku/. Artifacts were removed from each set and processed into PSDs of their respective trial lengths, 1s for the VEP and 10 seconds for the syllables. Using SVMs and KNNs the classification accuracy of each algorithm was averaged from 4-fold cross-validation. After artifact removal the VEP data set contained 9,596 trials for the 120 subjects and 3,787 trials for the 6 syllable subjects.

On the VEP dataset the SVM achieved 98% accuracy and KNN achieved 93% accuracy, both had a 95% confidence interval. The syllable dataset achieved higher accuracy, 99% with SVM and 98% with NN with both, again, at a 95% confidence interval. The strong performance across both datasets indicated the techniques and feature sets worked well on a fundamental level. However, the diminutive number of syllable subjects was not compelling and should have likely be run with more subjects.

In Gui et al.[39] a more contemporary ML technique, Artifical Neural Network (ANN) using feed-forward, back-propagation, and multiplayer perceptron, was used to identify subjects. Their dataset consists of the 6 mid-line channels {Fpz, Cz, Pz, O1, O2, and Oz} of 32 subjects undergoing VEPs. The channels were bandpass filtered, 0Hz to 60Hz, before wavelet packet decomposition (WPD) produces the final three features of mean, variance, and entropy for each 1.1 second epoch. Four experiments are carried out, but only two were of interest in subject classification: (S1) finding a single subject from the set of 32 and (S2) matching all 32 subjects against each other simultaneously. The other experiments consisted of a one versus all classification (S3) and separating small groups of subjects from each other (S4). For S1 the highest accuracy of 10% occurred with 5 neurons and the worst accuracy of 5% occurred with 10 neurons. S2 produced better results with a highest accuracy of 94% with 45 neurons and a worst accuracy of 70% with 30 neurons.

Here the fundamental issues of ML were exposed in terms of dataset source, preprocessing, feature selection, algorithm, and classification task. Across the range of experiments presented nearly every single one used different data or different features or different algorithms. The result was a lack of comparison points from which to drawn definitive conclusions about EEGs data and features or their classification. While many of these experiments produced acceptable results, little was gained and many were often confirming ideas already well documented instead of expanding the knowledge base.

2.3 Identity Vectors

At the center of this dissertation was the introduction of I-Vectors to the EEG classification community. I-Vectors are mathematical models that were designed to reduce the dimensionality of UBMs [117]. UBMs served to reduce a dataset of f-dimensional feature samples into C mixtures of f-dimensional GMMs. Following this, I-Vectors can then be created by enrolling distinct samples into a modeling process involving the UBM and a TVM. The TVM is generated from the enrollment samples and served to constrain the contributions of each mixture within the UBM. Finally, those I-Vectors were evaluated against testing I-Vectors, built from testing samples and the same TVM used to produce the enrollment I-Vectors. This evaluation resolve the distance in the l-dimensional distance between them.

As the technique is entirely data dependent, it could be altered to measure similarities between epochs, channels, individuals, or groups of individuals. I-Vectors were developed originally as an extension of a speech processing method called joint factor analysis (JFA) which split utterances into separate models for speaker, channel, and context [164]. In contrast, I-Vectors collapse those three models into just one.

The principal I-Vector equation is

$$M \approx m + Tw \tag{2.3-2}$$

where M is the feature space of the data, m is the UBM, T is the TVM and w is the I-Vector itself. The specific data used to build the UBM m is referred to as the training data. Once m and T have been defined, they can be used in concert with alternate enrollment targets of size S and testing data sets M to create data-specific I-Vectors, w.

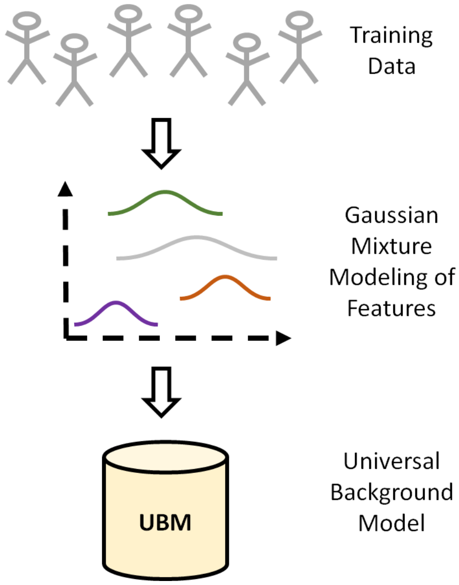

Training Data
Gaussian Mixture Modeling of Features
UBM
Universal Background Model

Figure 2.13. UBM Development. Training data was used to construct C independent Gaussian mixtures over the f dimensional feature space. This transformed the training data into C mixtures each with f means and variances. Taken as a whole these c mixtures were the UBM in addition to a mixture weight parameter. Ultimately this c mixture UBM served as the basis for developing a TVM and the associated I-Vectors.

A typical I-Vector use case might involve determining whether an EEG from a new patient should be diagnosed as epilepsy. First, a large randomized collection of training data drawn from a diverse set of subjects would be used to build a UBM, Figure 2.13. Then, sub-populations of data from known healthy and epileptic patients would be used to construct an enrollment dataset. This enrollment dataset would be used to resolve a TVM and produced enrollment I-Vectors related to the enrollment subjects.

Finally, the new patient's data would be used with the TVM to construct their I-Vectors. Then evaluations between the enrollment and target I-Vectors would inform which population they were more likely to match with, Figure 2.14. Depending on the choice of enrollment and test data, I-Vectors can automatically search for across channels, times, medical conditions, medications, and even entire subjects.

A UBM models the f-dimensional features by representing them with C independent Gaussian mixtures [165]. In general, increasing the number of mixtures captures more nuance, thereby potentially strengthening discrimination. The UBMs provide dimensionality reduction by taking a training dataset of L epochs each with f features each down to C mixtures of f features. As each feature has a mean m, variance σ, and weight ρ, reduction benefits are seen when $L > 3C$. The UBMs can be characterized according to:

$$\Omega_{c=1...C} = \begin{cases} m(c) \\ \sigma(c) \\ \rho(c) \end{cases} \tag{2.3-3}$$

Each parameter is a vector of length f representing a given feature. Each I-Vector is the result of the expectation maximiation (EM) of the available UBM and M.

The I-Vectors are of length $l = Cf$ with many residual elements and are frequently further reduced by the speech community via LDA. This process requires that the final I-Vector length be one less than the number of subjects S or less given the constriants imposed by LDA's algorithm. Thus I-Vectors final length, $l = min(S - 1, Z)$, is frequently controlled by S or Z, where Z is on the order of 100s. The final I-Vectors therefore represent a very dense and robust abstraction to an l dimensional

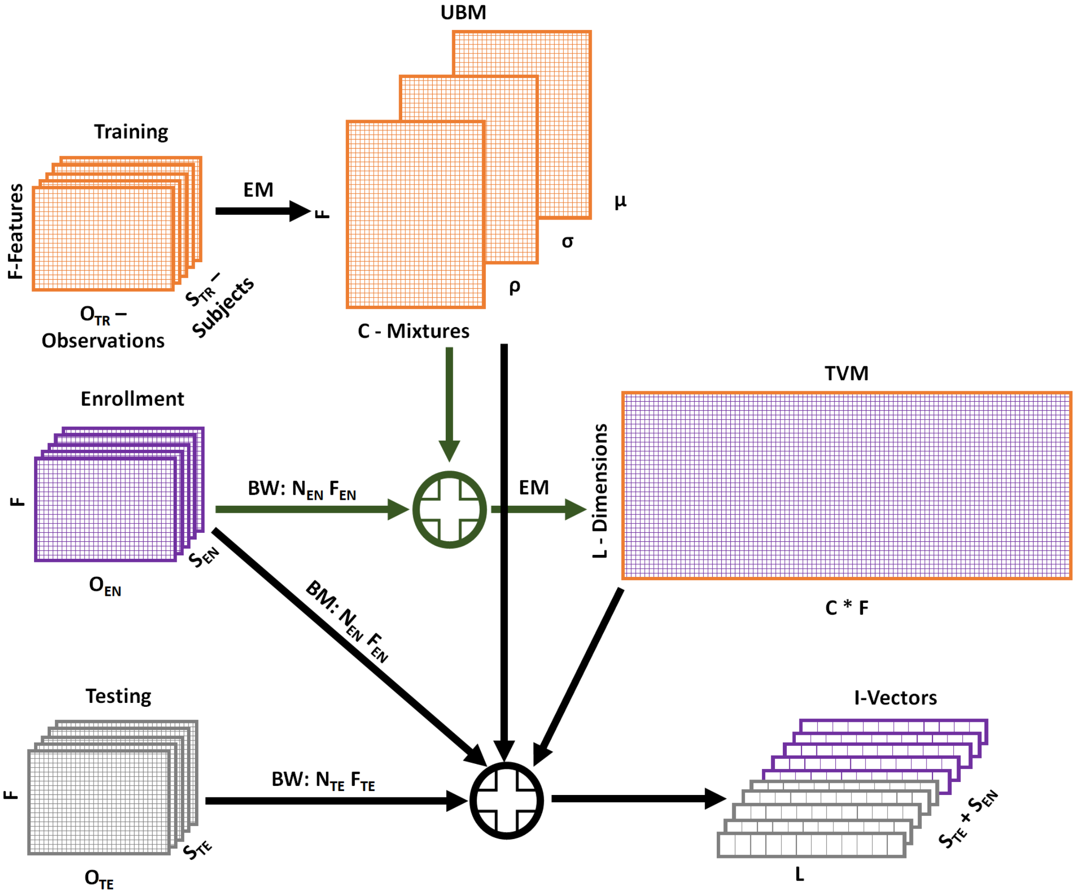


Figure 2.14. I-Vector Development. Using the UBMs as an initialization, the enrollment and training data are transformed into I-Vectors. This process is reliant on the creation of the TVM randomly generated from the variances of the UBMs and refined by adaptation towards the means of the UBMs. The resultant I-Vectors are pairwise evaluated to find the Cosine Distance (CD) between them to rank their similarity.

space. Within this space the similarity between two I-Vectors can be found via any metric evaluation, often CD.

2.3.1 Mathematics

The major components and steps to producing I-Vectors are outlined in reverse starting with the resultant I-Vectors and ending with the original JFA technique. This includes sections on the TVMs, UBMs, maximum a priori (MAP), and GMMs.

2.3.1.1 I-Vectors

The critical component of Equation 2.3-2, is the TVM T. An evolution from the eigenvoice matrix used in JFA, it captures all of the variances present in the UBMs. Generating T from training data requires an iterative EM approach reliant on feedback from the produced I-Vector w.

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_C \end{bmatrix} = \begin{bmatrix} A_1^{-1} * K_1 \\ \vdots \\ A_C^{-1} * K_C \end{bmatrix} \tag{2.3-4}$$

The matrices of A and K represent the updated mean and variance of T. These updates are driven by w and T along with the static values of $N, \hat{F}$, and Σ. The superscript H represents the Hermitian transpose.

$$A_c = \sum_{s=1}^{S} N_s(t) w^{-1}(t) \tag{2.3-5}$$

$$K_c = \sum_{s=1}^{S} \hat{F}_c(s) * \left(w^{-1}(s) * T^H * \Sigma^{-1} * \hat{F}_c(s)\right)^H \tag{2.3-6}$$

The estimation of w uses T a $Cf \times Cf$ matrix. This matrix is formed from the Baum-Welch (BW) statistics $\hat{N}$ and $\hat{F}$, an $l \times l$ identity matrix I, and a model of the UBM variances Σ. As the models are all independent Σ is a diagonal $Cf \times Cf$ matrix of the true variances from the UBMs where as the BW statistics are estimations of the mean N and variance F.

$$w(s) = \left(I + T^t \Sigma^{-1} \hat{N}(s) T\right)^{-1} T^t \Sigma^{-1} \hat{F}(s) \tag{2.3-7}$$

The BW 0th (N) and 1st (F) order statistics are generated from the evaluation of the UBMs against the L epochs in the training data. The higher order statistic must be offset by the preceding orders resulting in a centered 1st order statistic $\hat{F}$. Each statistic models the f features in each of the C mixtures resulting in $C \times f$ matrices. Each epoch, e, from the full epoch set $t = 1 ... L$ is evaluated to generate initial probabilities based on Ω for N and F.

$$\hat{N}(s) = \begin{bmatrix} N_1(s) & & \\ & \ddots & \\ & & N_C(s) \end{bmatrix} \tag{2.3-8}$$

$$\hat{F}(s) = \begin{bmatrix} \widetilde{F}_1(s) \\ \vdots \\ \widetilde{F}_C(s) \end{bmatrix} \tag{2.3-9}$$

$$\widetilde{F}_c(s) = F_c(s) - N_c(s)m_c \tag{2.3-10}$$

$$N_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega) \tag{2.3-11}$$

$$F_c(s) = \sum_{t=1}^{L} P(c \mid e_t, \Omega)e_t \tag{2.3-12}$$

This process resolves a suitable T after approximately twenty iterations of Equation 2.3-4 to Equation 2.3-7. Notice that Equation 2.3-8 to Equation 2.3-12 are needed only once to generate T. Creating I-Vectors from the enrollment and testing data follows Equation 2.3-2 in a modified form. The resultant I-Vector w will be a l row vector where l is a length defined during the creation of the initial estimate of T.

$$w = (M - m)T^{-1} \tag{2.3-13}$$

The number of I-Vectors produced is based upon the enrollment targets h and testing queries q, producing data on the order of $(h + q) \times l$. Therefore dimensionality reduction will not be significant if the data is partitioned such that $h + q \equiv L$.

The I-Vectors are finalized after applying LDA to control for dependencies in the data. This process reduces their length from l to $l = min(S - 1, l)$ elements based upon the transformation matrix produced by the LDA. There are other approaches to normalize the I-Vectors aside from LDA which can be reviewed elsewhere [166]. These final I-Vectors can be compared pairwise using CD to determine similarity between enrollment targets and testing queries.

$$cos(\Theta_{w_1,w_2}) = \frac{w_1^t w_2}{\|w_1\| * \|w_2\|} \tag{2.3-14}$$

2.3.1.2 Total Variability Matrix

After the development of JFA it was discovered that the iterative modeling process was not perfect at separating speaker, channel, and residual effects[166]. In fact the eigenchannel space was collecting information related to the subject when operating on specific utterances. JFA was still considered state of the art, but its performance could be challenged by the total variability space. This space, formally the TVM, was produced by using the first iteration of JFA to generate a low-dimensional speaker- and channel-dependent matrix. As this matrix is the key component in generating I-Vectors a detailed decomposition of its construct and applications is necessary.

The initial form of the T is $f \times C$, GMMs by features, shown in Equation 2.3-15. These parameters were dependent on each other and the training data. The speech community uses a definitive feature set [137], Mel Frequency Cepstral Coefficients (MFCCs), which evolved over time to become the gold standard [167]. This makes determining the number of features straightforward. Settling on an acceptable number of mixtures for the GMM was more difficult given the trade-offs between classification and computational performance[168, 169].

In many studies the number of mixtures is on the order of a base 2 number, often being set to at least 2048 mixtures[170, 171]. The optimization for the number of mixtures was dependent on the best performance, but limited by the dimensions of the training data. Given a number of subjects S each providing u utterances the number of mixtures C would need to be less than $S * u$ to prevent over-fitting.

$$\begin{bmatrix} M_1 \\ \vdots \\ M_f \end{bmatrix} = \begin{bmatrix} m_1 \\ \vdots \\ m_f \end{bmatrix} + \begin{bmatrix} T_{1,1} & \dots & T_{1,C} \\ \vdots & \ddots & \vdots \\ T_{f,1} & \dots & T_{f,C} \end{bmatrix} * \begin{bmatrix} w_1 \\ \vdots \\ w_C \end{bmatrix} \tag{2.3-15}$$

Critically, the TVM was not implemented to mimic utterances, but to map them instead. The technique allowed I-Vectors to be the weights controlling the inclusion of a column of features. In this manner it was possible that one column may contain the dominant features of a low pitched voice and a high pitched voice. If each of the C columns of T represent a unique component of the speakers, then the I-Vector w would be binary. More likely is that the characteristics are spread across mixtures since emergent properties of speech are parameterized via the MFCCs.

Advancing this approach to EEGs may produce a reasonable algorithm for discrimination, but also allow for an understanding of why the discrimination occurs. This is entirely dependent on the chosen features, which are well established for speech, but still open for EEGs. Using a non-linear variation of MFCC maintains the parameterization providing a closed set of features. With features bounded, experiments can then focus on finding an optimal size GMM for the UBM of EEGs.

Working down this chain, further incremental improvements can be made while gaining insight into the discrimination and grouping of EEGs in an unsupervised algorithm. While speech already knows the principal modes of their data [172], how to separate consonants, vowels, words, genders, and ages, such techniques do not meet the needs of the EEG community.

2.3.1.3 Universal Background Models

As mentioned previously UBMs are sets of GMMs created from the features of continuous signals. The GMMs contextualize the varied speech signal segments as independent feature distributions regardless of the spoken text [165]. This technique is suited to the problem of speaker recognition where the goal is to match subjects irrespective of data content. As this process is reliant on the likelihoods of features for a given model or subject sample, it can be used in an unsupervised manner to match and/or separate subjects.

The GMM represents the core component of the UBMs which in turn makes them critical to the performance of I-Vectors. Sets of Gaussian distributions (M) can be represented with a mean (μ) and co-variance (Σ) drawn from each measurement or feature of the D-dimensional raw continuous data [119]. This allows a likelihood calculation equation given a D-dimensional sample x to compare against the model,

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.3-16}$$

where x,μ, and Σ are vectors of length D and w_i corresponds to the weight of each mixture component where $\sum_{i=1}^{M} w_i = 1$. The calculated likelihood provides an unsupervised estimation of the sample relating to the given model(s).

The λ component of $p(\boldsymbol{x}|\lambda)$ represents the GMM and associated parameters: w_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$. While the previous equation does not assign a subscript to λ there would be U GMMs which comprise the fully formed UBM. Just as each GMM attempts to determine the underlying states of the data, the UBM requires depth to account for each class of signal.

As an example suppose one wants to know if the weather on a given day will require a heavy coat, a light coat, a raincoat, or no coat. If the temperature is below 45°F a heavy coat is desired and if the temperature is above 70°F no coat is necessary. In between these two temperatures a light coat may be necessary, but only if the day will be windy. At the same time, at any temperature above 45°F with high humidity levels should warrant wearing a raincoat.

The GMM representing raincoat would have a large variance for wind and temperature, but a small variance for humidity. The temperature means of heavy coat, light coat, and no coat would be unique. However, light coat and no coat would have a similar mean and variance for humidity and overlapping distributions for wind. Meanwhile, the heavy coat model would be insensitive to anything aside from temperature.

The weather conditions (humidity, temperature, and wind) become the three features modeled by the GMMs. Once four, or more, models are created they each categorize the required jacket. This full set becomes the UBM that provides a basis for evaluation of each day's weather. Given a weather report, the UBM would provide the likelihood of each jacket being the correct answer.

To calculate the likelihood for a multivariate normal distribution the follow equation is used, represented as the function $g(\boldsymbol{x}|\mu_i, \boldsymbol{\Sigma}_i)$ from the prior equation,

$$g(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\exp\Big\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \tag{2.3-17}$$

From these equations estimations of underlying modes of the data can be found from which to build a suitable model. Two important assumptions are made in this process, the first is that each Gaussian mixture is independent of the other mixtures and the second is that the underlying modes can me adequately modeled with normal Gaussian distributions. These mixtures are therefore representing a unique hidden set of generators/states that create the resultant signal. Given that the number of hidden states is unknown, GMMs may produce mixtures with marginal weights or mixtures with redundant attributes.

2.3.1.4 Maximum A Posteriori Parameters

With a UBM in place it is possible to tune the model toward specific subjects. The estimation of a subject specific model from a UBM is called MAP estimation[119].

Just as with a UBM, the statistics (weight, mean, and variance) of the subject are found from their data $S = \boldsymbol{s}_t, ..., \boldsymbol{s}_T$. These expectations are derived from the prior model found from the UBM, but operating on the subject specific data.

$$n_i = \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}}) \tag{2.3-18}$$

$$E_i(\boldsymbol{s}) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}})\boldsymbol{s}_t \tag{2.3-19}$$

$$E_i(\boldsymbol{s}^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{s}_t, \lambda_{\text{prior}})\boldsymbol{s}_t^2 \tag{2.3-20}$$

These are then able to adapt each i mixture's weight, mean and variance. The amount of adaptation is based on the expectations and a chosen relevance factor r^ρ.

$$\hat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i\right]\gamma \tag{2.3-21}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\boldsymbol{s} + (1 - \alpha_i^m)\boldsymbol{\mu}_i \tag{2.3-22}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\boldsymbol{x}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 + \mu_i^2) - \hat{\boldsymbol{\mu}}_i^2 \tag{2.3-23}$$

The adaptation coefficient is most often constant for all three statistics, but given unique labeling allowing for decoupling if necessary.

$$\alpha_i^{w,m,v} = \frac{n_i}{n_i + r^\rho} \tag{2.3-24}$$

These new statistics not only provide subject specific models, but present a new set of models for discrimination. An example of this process is shown in Figure 2.15. The models themselves can be compared against each other to determine similarity in addition to evaluating them against new data samples.

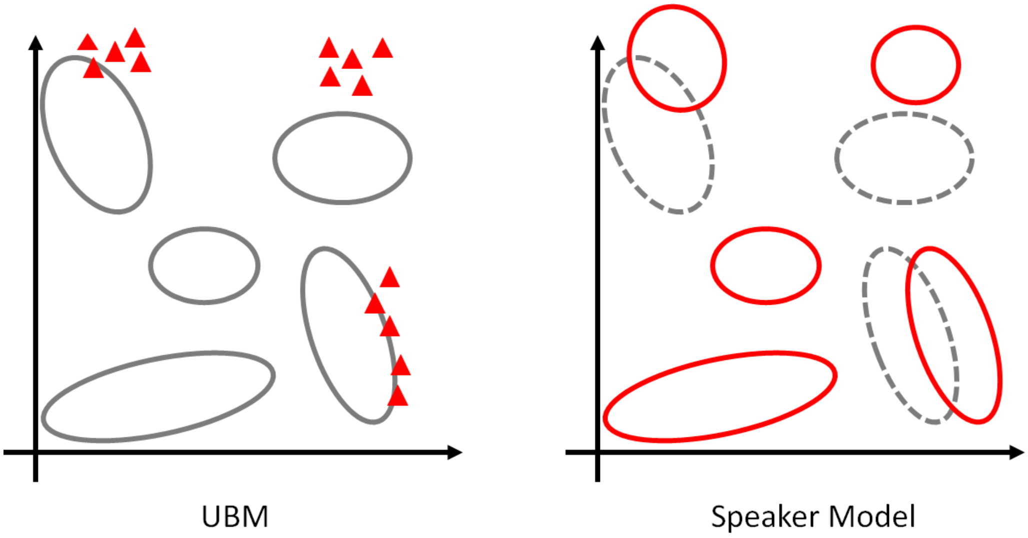


Figure 2.15. Example of MAP of GMM. Results of MAP estimation when speaker data, red triangles, is applied to a UBM, gray mixtures.

This process is also used to carry out EM to each larger mixture iteration of the UBM, Figure 2.16. By using an initial estimate of the means, variances, and weights the next mixture size is constructed by splitting each mixture along the feature axis of its largest variance. The square root of this variance is to used shift one positively and the other negatively producing two new means which share their parent mixture's variance and half their weight.

2.3.1.5 Gaussian Mixture Models

Understanding how GMMs produce likelihoods for a given data sample $\boldsymbol{x}$ informs how each mixture's λ is produced. The more accurate the parameters of λ are for a given GMM, the more insightful the resultant likelihoods. However, unless the parameters are known outright they must be deduced empirically. One of the more prevalent techniques for parameter estimation is maximum likelihood estimation (MLE)[173].

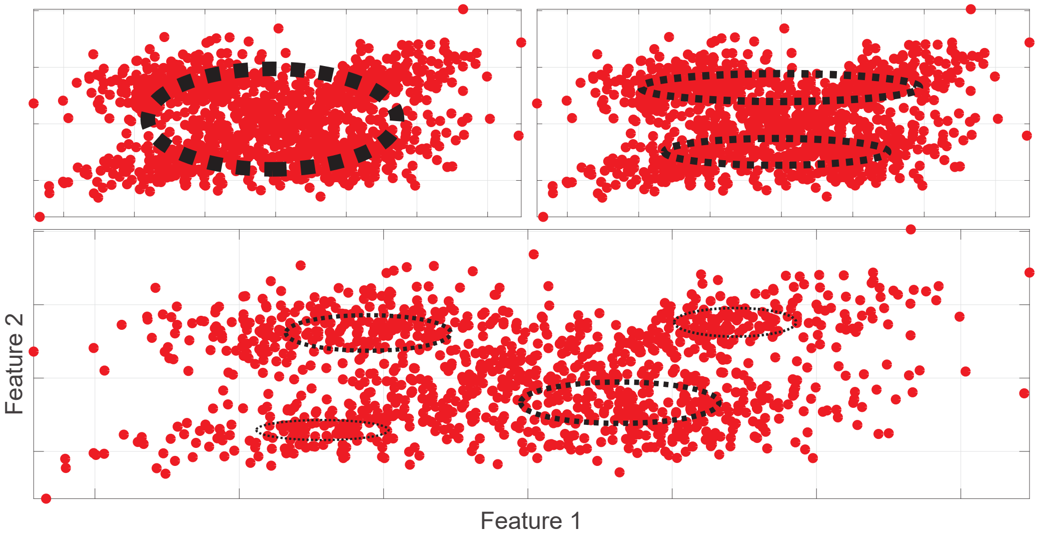


Figure 2.16. UBM EM Example. Using a single mixture UBM built from all of the training observations, EM can be used to produce 2 and then 4-mixture UBMs that accurately model the distribution of the observations. The thickness of each black ring represents the proportional weight associated with each mixture.

The MLE attempts to find a distribution that maximizes each of the T training vectors $X = \{x_1, ..., x_T\}$

$$p(X|\lambda) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\lambda) \tag{2.3-25}$$

this equation assumes that each component of the distribution is independent. This often turns out to be untrue, but is a necessary assumption to provide a functional solution. This function is non-linear as the product of all the training vector evaluations allows for one worsening likelihood to diminish any improvements gained from the remaining vectors. To avoid this problem, a variant of EM can be used to estimate the parameters for each feature independently. This helps isolate the features, in the event that they are not independent, and provides the ability to directly improve the overall likelihood on a feature by feature basis.

With this each parameter of λ can be estimated in an iterative manner with the following equations

$$\bar{w_i} = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda) \tag{2.3-26}$$

$$\bar{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} \tag{2.3-27}$$

$$\bar{\boldsymbol{\sigma}}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)\boldsymbol{x}_t^2}{\sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{2.3-28}$$

these three equations provide updated values for the weights, means, and variances that can feed the next iteration of the EM algorithm. The *a posteriori* probability Pr is found with the following equation

$$\Pr(i|\boldsymbol{w}_t, \lambda) = \frac{w_i g(\boldsymbol{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k g(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{2.3-29}$$

2.3.2 Success in Speech and Adaptation

The deployment of I-Vectors as a tool for speaker recognition/verification[122], language detection[123], accent detection[174], and speaker age[124] showed the growth and trust the speech community put into the algorithm. I-Vectors were developed in 2011 at the Centre de Recherche d'Informatique de Montreal (CRIM) by Dehak, Kenny et al[166]. Prior to this work the group at CRIM developed JFA for use with speech data to address speaker and session variability[175]. I-Vectors were a natural extension of JFA, but proved to be very effective as a feature preprocessing technique and their own classifier when paired with a simple metric like CD [176, 120].

Fundamentally, evaluations of other ML algorithms relied on tracking the sensitivity and specificity of each experiment and I-Vectors were no different. In fact, they performed inline with other approaches achieving over 90% sensitivity and 90% specificity [171]. Given this development, the approach presented here represents the heart of the technique in as simple a manner as possible. The extensive use of I-Vectors has produced a variety of augmentations, but it would have been unwise to start with a more complex system when transporting the technique to a new field of data. It was decided to minimize as many degrees of freedom as possible while developing I-Vectors for EEGs.

Another problem in adapting this technique was that finding valid speech data was relatively easy. If someone was talking, producing sounds, they were likely producing valid data. However, that was not the case with EEGs which have a constant stream of data. It was not clear if EEG recordings since background segments are not devoid of information, essentially all data is data of interest. This naturally leads to an increase in background signals in EEGs compared to speech. A sleep study may last for an entire night only to capture a brief 10 minute seizure. Easy for a clinician to correctly identify, but difficult for a ML technique to recognize.

2.4 Machine Learning Algorithms

The breadth of potential algorithms, supervised and unsupervised, was too much to review in depth. Instead, a review of the most notable algorithms referenced in this section and those critical to the validation of I-Vectors were reviewed in the following section. This was meant to provide necessary context to the present field of EEG classification, but was not comprehensive to the rapidly developing realm of ML. Similarly, a brief discussion of FA was included given the frequent use of LDA in supervised and unsupervised techniques and that I-Vectors were predicated on JFA.

2.4.1 Factor Analysis

At a base level I-Vectors reduced the dimensionality of data by finding the most influential features in the given training dataset. In a general sense this is similar to FA which is used to perform blind source separation (BSS), the decomposition of a signal into a linear representation of statistically independent components [173]. While this was the goal, it is difficult to assure linear independence of all the components. As such the techniques are imperfect given the premise of being blind to the true nature of the data.

Two commonly used techniques to achieve BSS are PCA and ICA. From these algorithms more advanced techniques, LDA and QDA, are capable of separating the components of different known classes. They are not able to operate blind, or unsupervised, as they require knowledge of the classes to define class dependent components. Knowing the dependent components they can then resolve the class independent components in an effort to discern the decisions surfaces between the classes. QDA operates in a more generalized space allowing for separation of two or more classes compared to LDA defining separability of a single class from the dataset.

2.4.1.1 Principal Component Analysis

PCA finds the dominant components in a set of data by maximizing the variance of the given features [177]. For a set of data $\boldsymbol{X}$ composed of p columns of features and n rows of observations there exists a vector $\boldsymbol{w}$ capable of maximizing the variance of a given feature.

$$\boldsymbol{V} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \tag{2.4-30}$$

$$\sigma_w^2 = \boldsymbol{w}^T\boldsymbol{V}\boldsymbol{w} \tag{2.4-31}$$

Here $\boldsymbol{V}$ represents the covariance matrix of the data matrix $\boldsymbol{X}$ which is used to find the eigenvectors that become $\boldsymbol{w}$. As eigenvectors are orthogonal to each other, they are each uncorrelated components and produce the p principle components of the $\boldsymbol{X}$.

There are at most n principle components representing unique weightings of the p features. To find the true number of components, q, the number of zero or near zero eigenvalues, $e_z = p - q$, must be found. This linearly independent q-dimensional space represents the true decision surface of the observations. From these operations it becomes possible to define the critical features and unique observations from the data itself.

2.4.1.2 Independent Component Analysis

ICA separates individual signals from those collected by multiple receivers, commonly known as BSS [173]. The typical example is that of a cocktail party with an equivalent number of microphones and speakers. By using ICA, it is possible to isolate each of the speakers using the data from all of the microphones. This example is referred to as the *Cocktail Party Problem* and exists in many research areas including EEG recordings.

A dataset contains the sequential samples, t, from each recording device and assumes there is a transformation matrix, $\boldsymbol{A}$, that turned the source signals, $\boldsymbol{s}$, into the captured output $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \boldsymbol{A} = \begin{bmatrix} a_1 1 \dots a_1 n \\ \vdots \ddots \vdots \\ a_n 1 \dots a_n n \end{bmatrix} \boldsymbol{s} = \begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} \tag{2.4-32}$$

$$\boldsymbol{X} = \boldsymbol{AS} \tag{2.4-33}$$

From this output, the features of the recorded signals must be *whitened* before the individual signals can be found. Whitening is a process that transforms the data into a matrix. $\boldsymbol{z}$ that is uncorrelated, but not assured to be independent. The approach is similar to PCA in that it requires eigenvalue decomposition to produce the whitening matrix, V. The matrix $\boldsymbol{E}$ is found from the eigenvectors of $\boldsymbol{X}$ and the diagonal matrix D contains the associated eigenvalue for each eigenvector.

$$\boldsymbol{z} = \boldsymbol{Vx} \tag{2.4-34}$$

$$\boldsymbol{V} = \boldsymbol{ED}^{-\frac{1}{2}}\boldsymbol{E}^T \tag{2.4-35}$$

$$\boldsymbol{z} = \boldsymbol{VAs} = \hat{\boldsymbol{A}}\boldsymbol{s} \tag{2.4-36}$$

Now the transformation matrix, $\hat{\boldsymbol{A}}$, contains only orthonormal components instead of the previous correlated components. This process is necessary as it constrains the solution sets when solving for the independent components.

The *kurtosis* of a signal is one of the many ways to solve for the independent components after whitening. As the kurtosis supports the additive property, it provides a natural process for optimization the non-Gaussian portions of the signal. The expectations, E, of the random variable y's second,variance, and fourth moment are used to find the 'tailedness' of the distribution. With a normalized distribution the expectation of the variance would be 1, but for Gaussian distributions kurtosis would always be zero because the fourth moment is always $3(\mathrm{E}\{y^2\})^2$. This is why the independent components must be non-Gaussian otherwise they cannot be separated out.

$$\begin{aligned}\mathrm{kurtosis}(y) &= \mathrm{E}\{y^4\} - 3(\mathrm{E}\{y^2\})^2 \\ \mathrm{kurtosis}(s_1 + s_2) &= \mathrm{kurtosis}(s_1) + \mathrm{kurtosis}(s_2) \\ \mathrm{kurtosis}(\alpha s_1) &= \alpha^4 \mathrm{kurtosis}(s_1) \end{aligned} \tag{2.4-37}$$

When all the random variables are normalized the variance of y is equal to 1 which bounds the solution by the unit circle. This simplifies the solution to finding a vector that produces the largest amplitude of kurtosis for the given distribution. These kurtosis based dimensions indicate projections of non-Gaussian distributions which is where the suspected independent signals reside.

$$|\mathrm{kurtosis}(y)| = |q_2^4 \mathrm{kurtosis}(s_1) + q_2^4 \mathrm{kurtosis}(s_2)| \tag{2.4-38}$$

There are other techniques for discerning the projection space of non-Gaussian distributions, Gram-Schmidt, ML estimation, or negentropy, which focus separating independent non-Gaussian distributions. In all instances the mixing matrix $\boldsymbol{A}$ is chosen to be square to simplify the mathematics. The only constrains on the process, regardless of approach, are on the data being statistically independent and that the underlying signals are non-Gaussian distributions. These both require prior knowledge of the signals in the dataset otherwise the results of ICA will be similar to those of PCA, orthogonal uncorrelated feature vectors.

2.4.1.3 Linear Discriminate Analysis

LDA uses the mean and variance of each class in the data to build decision surfaces between the classes. This is achieved by maximizing the distance between the means $\boldsymbol{S}_B$ and minimizing the variances $\boldsymbol{S}_W$ of the features associated with the classes K. Original developed by Ronald Fisher, often called *Fisher's Linear Discriminant*, it seeks to maximize the discriminant factor $J(\boldsymbol{w})$ by finding the vector w [17].

Given two datasets containing n_i observations of each class, a decision surface $\boldsymbol{w}$ can be found.

$$\boldsymbol{X}_1 = \{\boldsymbol{x}_1^1, ..., \boldsymbol{x}_{n_1}^1\} \ , \ \boldsymbol{X}_2 = \{\boldsymbol{x}_1^2, ..., \boldsymbol{x}_{n_2}^2\}$$

$$\boldsymbol{m}_i = \frac{1}{l_i} \sum_{j=1}^{n_i} \boldsymbol{x}_j^i$$

$$\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T$$

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_j - \boldsymbol{m}_i)(\boldsymbol{x}_j - \boldsymbol{m}_i)^T$$

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{2.4-39}$$

This can be expanded to handle multivarate data by expanding the definitions of $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$. Here $\bar{\boldsymbol{m}}$ represents the mean of the observations n_i across all classes in the training set. Then a sufficient w can be found by maximizing $J(\boldsymbol{w})$ which occurs when $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{S}_W^{-1} \boldsymbol{S}_B$.

$$\boldsymbol{S}_B = \sum_{i=1}^{K} n_i (\boldsymbol{m}_i - \bar{\boldsymbol{m}})(\boldsymbol{m}_i - \bar{\boldsymbol{m}})^T$$

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \boldsymbol{m}_i)(\boldsymbol{x}_{ij} - \boldsymbol{m}_i)^T$$

Classification based off LDA requires an additional step to set thresholds for each class with respect to the resultant eigenvalues produced by $\boldsymbol{w} \cdot \boldsymbol{x}$. Through this metric many approaches can be used to distinguish between the K classes in the multivariate data such as individual or one-versus-all classification.

The multivariate approach often assumes a common global covariance matrix S_X to ensure that $S + W^{-1}S_B$ is diagonalizable. This assures that the eigevenvectors will be caused by the features within the data. To approximate a global covariance matrix the pooled within-class covariance matrix is scaled by the degrees of freedom between the observations and classes.

$$S_X = (n - K)^{-1}\boldsymbol{S}_W \tag{2.4-40}$$

This results in $K - 1$ eigenvectors as diagonalizablity of a matrix does not ensure unique eigenvectors. In general, LDA is frequently used to perform dimensonality reduction similar to PCA based upon the eigenvalues associated with each eigenvector. Even without reviewing the eigenvalues, LDA always produces one less feature dimension than classes to force discrimination upon the next eigenvector axis.

2.4.2 Algorithms

Numerous algorithms were introduced while reviewing the applications of EEG recordings. The following section highlights the more common algorithms used in ML and those to be compared against I-Vectors. From training datasets the algorithms are able to classify unknown samples by providing a likelihood of a match or a discrete label if given labeled data. These introductions serve only to address the nature of the algorithm, unsupervised or supervised, the process of discrimination, and show the input parameters and type of classification produced.

2.4.2.1 Gaussian Classifiers

Once created, GMMs can be used as the basis for discrimination. As discussed in Section 2.3.1.5, the data is broken down into a series of estimated Gaussian distributions. These distributions strive to model classes defined by the data. To identify new data, a likelihood score is generated based upon the distance between each model and the new data sample. Calculating the distance, and thus likelihood, can be done in a number of ways. Assuming the distributions are Gaussian in nature, the following equation provides the likelihood the point belongs with the model.

Here x is the location in d dimensional space with a known mixture modeled by its mean μ and co-variance Σ.

$$likelihood(x, \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}}{\sqrt{|\Sigma|(2\pi)^d}} \tag{2.4-41}$$

This general form produces the likelihood a sample x could come from a given mixture. The end result becomes a set of likelihoods of the known classes from which to draw a classification label. However, there is no assurance of a data sample exceeding 50% likelihood of any of the classes.

This classifier functions based on the modeled distributions. If the GMMs are created via EM or another clustering method the entire process is unsupervised. However, it is possible make the process supervised by knowing the class means and variances in advance or using labeled data to manual cluster the data. The evaluation of a likelihood based upon a distribution is a fundamental technique used by many

ML algorithms. It serves as a natural comparison point for I-Vectors as a preliminary step in their development is to produce GMMs.

2.4.2.2 Naive Bayes Classifier

Naive Bayes Classifiers (NBCs) make use of probabilities to classify based on discrete conditions. The classifier is built out from Bayes' Theorem which describes the probability of an event occurring given the current conditions. This approach requires knowledge about the events that inform the probabilities making it a supervised algorithm. The two class form of a NBC is

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} \tag{2.4-42}$$

which provides the likelihood of A given B. In this equation $\mathrm{P}(A)$ and $\mathrm{P}(B)$ represent the independent probabilities of events A and B and the probability of B given A is given as $\mathrm{P}(B|A)$. This expands to multiple conditions T by taking into account the likelihoods of each possible condition with

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_i^T \mathrm{P}(B|A_i)\mathrm{P}(A_i)} \tag{2.4-43}$$

The expansion of the unitary case shows that as the number of conditions increases probabilities for each condition with respect to each class are needed. In a sense the conditions could be features representative of classes or the classes themselves.

The approach is a natural tool for evaluating any modeling technique that produces discrete probabilities assuming they are all independent. Since this cannot always be assumed the technique's performance is dependent on adequate feature selection and class separation. The outcome is a probability of the test event or class occurring that is bounded between $(0\% - 100\%)$.

2.4.2.3 K-Nearest Neighbor Classifier

A KNN classifier uses labeled datasets to assume the class of an unknown sample. This approach is similar to using GMMs, but KNN can only operate with labeled data. Given the k closets neighbors class, the unknown sample is labeled as the highest counted class. The algorithm relies on mapping distances between the data points in their f dimensional feature space [178].

Determining the distance between unique samples provides flexibility in handling non-Gaussian distributions. Unlike GMMs classifiers and similar to NBCs, this algorithm operates directly on the data and not through a model when fed training data. The trade-off becomes having enough data and selecting a sufficient value of k to produce acceptable classifications. The previous two algorithms relied on the statistics drawn from the training data, but KNN is directly dependent on samples in the training data.

The simplistic nature of and ease of conceptulizing lead KNN to be used in a variety of experiments as a comparative benchmark [14, 152].

2.4.2.4 Support Vector Machines

Another kernel based classifier, SVMs, creates a hyperplane between a target class and all other data. The use of a kernel allows linear and non-linear decision surfaces to be transformed onto a hyperplane for discrimination. This hyperplane maximizes the distance between a target cluster and a non-target cluster [179]. Development of the technique stemmed from considering two normal distributions $\boldsymbol{N}_1 : m_1, \Sigma_1$ &

$\boldsymbol{N}_2 : m_2, \Sigma_2$ and an target location x.

$$F_{sq}(x) = \text{sign}\Big[\frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1) - \frac{1}{2}(x-m_2)^T\Sigma_2^{-1}(x-m_2) + \ln\frac{|\Sigma 2|}{|\Sigma_1|}\Big] \quad (2.4\text{-}44)$$

In this case $F_{sq}(x)$ resolves to a positive sign indicative point x is in $\boldsymbol{N}_1$ and a negative sign for $\boldsymbol{N}_2$. From this initial equation may variations developed to address non-normal distributions and how to simplify the equation by approximating $\Sigma_1 \approx \Sigma_2$.

Results of SVMs are a binary one-versus-all classification. This provides no way to produce clusters of data nor known the strength of the classifications. As with the other classifiers it builds the hyperplane used for separation from a labeled training set, making it a supervised classifier. As it seeks to maximize the space between clusters additional data is most beneficial when it represents boundary conditions of each class. It has been used on I-Vectors in the speech community [180] and numerous EEG classification tasks [14, 102, 66].

2.4.2.5 Dirichlet Process

A Dircihlet Processes (DP) allows for distributions of distributions to be built in an unsupervised manner. The process produces random variables G_K as sub-distributions from the full dataset's distribution G_0 given a concentration parameter α. In this manner an unlimited number of distributions can be produced from a closed dataset containing $T_1 ... T_K$ partitions[17] of the data Θ[181].

$$G \approx \text{DP}(\alpha, G_0) \quad (2.4\text{-}45)$$

$$\Big[G(T_1), ..., G(T_K)\Big] \approx \text{Dir}(\alpha G_0(T_1), ..., \alpha G_0(T_K)) \quad (2.4\text{-}46)$$

[17] A partition of Θ defines a collection of subsets whose union is Θ. A partition is measurable if it is closed under complementation and countable union.

Generating new distributions in this manner assures that the average distribution properties are maintained. Those distributions with large α will contribute more heavily, but have a greater likelihood of exemplifying the full dataset's true distribution. Through iterative measures it is possible to produce distributions that separate into naturally defined classes based on the dataset alone.

The clustering of the data occurs via the atoms at each level. An atom is a model of the statistical patterns of some phenomena in the data. At the lowest clustering level only atoms relevant to that level are present, but the next highest level contains these atoms plus their own atoms. Building up towards the highest clustering level means collecting all the atoms along the way. By sharing the atoms across the dataset, it becomes possible to then map similarities based upon the mixture of these atoms at each level [182].

The version used in Wulsin et al.[51], Heirarchical Dirichlet Process (HDP), allows distributions to be drawn across multiple levels of the data at once. This exemplifies the use case of a DP for clustering data on multiple levels with minimal prior knowledge. Wulsin built clusters at each level of the data (subject, seizure, and channel) so the knowledge was about the structure of the data and not the contents of the data. This is similar to I-Vectors as features are clustered in the GMMs and then the resultant samples are clustered based on the feature models.

2.4.2.6 Artificial Neural Networks

By applying the functional structure of brain neurons, an algorithm that behaves as a NN can be trained to perform non-linear classification. Each node in the network takes in information from the preceding layer, evaluates an equation to determine its state, and then contributes this activation to the ensuring layer. The connections between nodes have their own weights and the number and depth of layers is based upon the needs of the network. The algorithms referenced thus far included DBNs, RBFNNs, multilayer perceptron neural networks (MLPNNs), and MLPNNs represent a small sample of breadth of NNs.

Depending on the type of data and intended classification goal one NN may perform better than another. The trade-offs between the algorithms stem from the characteristics of the data related to the number of classes and any temporal relationships. At the crux of these algorithms is the need for a large diverse amount of labeled data. Like other algorithms, they learn directly through each sample of data which enables them to be non-linear classifiers. The training methodology is driven by reducing the error in the training dataset through adjusting the weights connecting the nodes and the biases of activation in each node. The complexity of the problem to be solved is often matched by the complexity of the NN.

Of interest to the development of I-Vectors is a Long Short-Term Memory Neural Network (LSTMNN) adaptation capable of quantifying the similarity between two inputs [183]. By training on ranked input vectors, in the case of Mueller et al. [183] sentences, the algorithm can learn to produce a discrete similarity score. This approach is highly dependent on the initialization parameters and the quality and quantity of training data available given the need to operate on variable length input vectors that represent the same classification.

2.4.2.7 X-Vectors

The I-Vector methodology was improved upon while this work was ongoing by research in the speech community that augmented it with a deep neural network (DNN) [184]. This combined system used I-Vectors and embeddings from a feed-forward deep neural network (the x-vectors) to surpass both of their individual speaker verification performances. The premise of the x-vectors was to utilize the post-statistics pooling layers of the DNN to generate feature vectors. The first embedding was taken from the first affine layer after the statistics pooling and second embedding was taken from the affine layer that received the output of a ReLU (rectified linear units) layer driven by the previous embedding. This made the first embedding a linear representation of the speaker's statistics and the second embedding a non-linear representation of the same statistics. The embeddings and I-Vectors are evaluated using the same process of LDA followed by length normalization and probabilistic linear discriminant analysis (PLDA) to produce the classification scores.

The original research group further refined the technique to the point that it surpassed its I-Vector counterparts [185]. These results were promising for advancing speaker recognition on text-independent datasets, where I-Vectors had been the standard classification technique. However, the x-vector approach is a supervised ML algorithm which relied on speaker labels to build the embeddings from the training dataset before generating the embeddings from the test dataset. They are clearly superior to I-Vectors for speaker/subject recognition, but this approach would be reliant on clinical annotations to expand subject recognition. Additionally, the computational requirements of x-vectors is orders of magnitude beyond that of the presented I-Vector technique, as the original work by Snyder et al [184] utilized 4.4 million parameters over all layers of the DNN.

Chapter 3

METHODS

Those who fail to plan, plan to fail.

Attributed to Benjamin Franklin

The application of I-Vectors on EEGs is a novel concept given that I-Vectors were designed for speech processing. Therefore, there is minimal guidance on how to use I-Vectors on EEG data. As indicated in the background, the foundations of the experiments proposed here came from following the development of I-Vectors within the speech community. The two fields are related in terms of their signal analysis goals, and subject and condition discrimination (Research Aim 1), but their optimization processes may be different (Research Aim 2). In both Aims, the desired goals afford insight into the classification process, which in turn is leveraged into insight about the features, datasets, and EEGs themselves.

3.1 Experimental Outline

The ultimate goal of this research is to provide subject and condition discrimination of EEGs. Prior to this work, this goal was not possible using I-Vectors given the lack of a software tools specifically for EEGs. The first experiments provided classification performance showing that I-Vectors met or exceeded performance of equivalent techniques. Providing competitive classification required an understanding of the technique's trade-offs in terms of features, datasets, and parameters. Running experiments to sweep through the features, datasets, and parameters provided operational thresholds for the datasets, UBMs, and UBMs for using I-Vectors based classification on EEGs.

In this work the experiments are classified as *Algorithm Benchmarks*, *Parameter Sweeps*, and *UBM-TVM Relationship*. The Algorithm Benchmarks addressed Research Aim 1 (RA1) by testing the performance of I-Vectors against benchmark classifiers, specifically Mahalanobis distance and GMM-UBM. The initial comparisons were carried out using parameters borrowed from speech recognition, which then required optimization through the Parameter Sweeps that addressed Research Aim 2 (RA2). Using the optimal classification parameters, the mechansims by which I-Vectors carried out their classification was resolved through analysis of the relationships between the UBMs, TVMs, and feature sets. These UBM-TVM Relationship experiments addressed Research Aim 3 (RA3) and represented the major contribution to understanding EEGs and multi-modal signal analysis.

Each experiment operated on the same fundamental features, datasets, and evaluations as they built upon each other. This chapter details all the components used to build out the experiments. The ensuing three chapters organize present each of the experiments: Chapter 5 - Parameter Sweeps, Chapter 6 - Algorithm Benchmarks, and Chapter 7 - UBM-TVM Relationship.

3.2 Data

Using heterogeneous data is necessary for validating any statistically rigorous method such as I-Vectors, but EEG data is difficult to obtain. Typically, new data is generated as part of research experiments and/or acquired from hospitals, but rarely if ever enters the public domain. This limits innovation to specific combinations of data and techniques. To mitigate this, only the publicly available datasets from PhysioNet Database[100] and TUH-EEG[18] were used in this work. While not comprehensive in terms of the variety of subjects and conditions used in other studies this collection provided the necessary breadth to validate the goals of this work. These data include EEG from imagined and actual hand, arm, and foot motion, and normal, abnormal, and seizure clinical EEGs from over 600 subjects.

3.2.1 PhysioNet Database

This EEG data comes from the New York State Department of Health's Wadsworth Center [94] and is a component of the PhysioBank archive maintained by MIT's Lab for Computational Physiology[1]. Within the data bank are EEG recordings pertaining to resting states, imagined motion, and motion tasks. The data consist of 64 channel EEGs from 109 subjects performing 14 trials: 12 motion and 2 resting calibration outlined in Figure 2.3. Information about the subjects (age, gender, handedness, etc) is not provided, making subjects and trials the most applicable decision surfaces.

Each 2-minute imagined-motion/motion trial consists of a series of 30 4.1 second tasks. These alternate between rest states and the computer prompted tasks (T1-T4). The tasks consist of opening/closing left or right fist (T1), imagine opening/closing left or right fist (T2), opening/closing both fists or feet (T3), and imagine opening/closing both fists or feet (T4). The two resting state trials, TR1 Eyes Opened (EO) and TR2 Eyes Closed (EC), are one minute recordings of unprompted subject recordings. From this, three dataset

1. **Physio Full** - All fourteen trials (TR01-TR14)

2. **Physio Single** - One trial of each type (TR01-TR06)

[1] https://www.physionet.org/pn4/eegmmidb/

3. **Physio Motion** - One of each motion trial (TR03-TR06)

These datasets allowed classification experiments on distinct levels of the data. The highest level was subject classification across trials. Beneath that was subject-trial classification, dependent on matching the correct subject and trial. Finally, within-subject trial classification was possible given the grouping of the repeated trials.

The recordings consist of 64 electrodes sampled at 160Hz following a standard 10-20 layout. A 65th channel provides labels for each task during the trials. Since its introduction in 2009, the PhysioNet Database has been used in biometric classifications [105] with respect to task sensitivity [86], subject independence [186], various subject classification schemes [68, 104], and attempts at content based retrieval [187].

3.2.2 TUH Corpus

The Temple University EEG Corpus (TUH-EEG) contains over 25,000 EEGs with their associated medical evaluations. All data comes from patients seen by Temple University Hospital in Philadelphia, Pennsylvania [18]. These recordings represent considerable breadth and depth in terms of patients, medical conditions, and recording conditions. Seizures were the most common diagnosis for patient's with medical records, but stroke and concussion patients are represented as well, while the majority of all recordings are simply indeterminate. In addition to these patients, there are subsets consisting of normal patients and those with indeterminate conditions considered abnormal. These latter classifications (abnormal/normal) along with seizure patients were used to organize 3 distinct datasets:

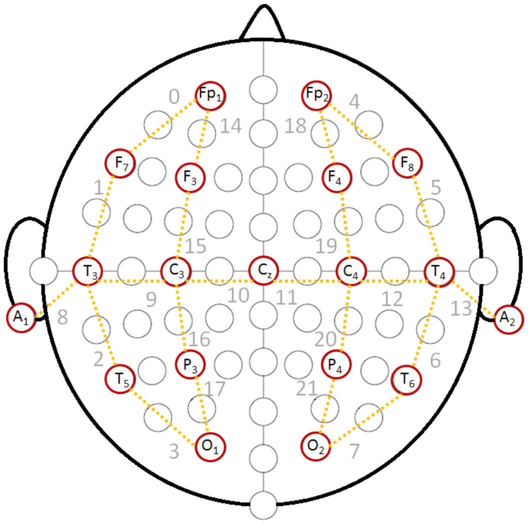


Figure 3.1. Layout of TCP CEP Montage. The Trans-Cranial Parasagittal (TCP) montage uses a rostral to caudal differential between electrodes to produce channel data. This differential is applied from the ears inward as well to produce 22 distinct channels. Common electrode names are provide with intermediate electrodes left blank. The gray numbers represent the channel index found in the Temple University EEG Corpus (TUH-EEG).

1. **TUH Normal** - 50 normal patient sessions

2. **TUH Abnormal** - 50 abnormal patient sessions

3. **TUH Seizure** - 411 seizure patient sessions

These datasets allowed for two types of classification experiments. The first was on the subject level, as each was built from unique subjects. The second was developed by combining the datasets to classify them based upon their condition, abnormal/normal/seizure. Further analysis was possible given the associated medical reports, but beyond the time and scope of this research.

Unlike the PhysioNet Database, the TUH-EEG is *in vivo*, leading to a wide array of recording variation. The electrode configurations, sampling rates, and session counts are at the discretion of medical professionals and not a structured research protocol. As addressed in its public release [18], the most common recording configuration consists of 31 electrodes at a 250Hz sample rate. This is substantially fewer electrodes than the PhysioNet Database, but is enough to produce clinically common EEG montages[2].

3.2.3 Synthetic Dataset

Developing and testing on experimental data alone would make it impossible to provide validation of the software's efficacy; therefore, a synthetic dataset was built. This controlled dataset allowed for two 'ideal' configurations: (1) a dataset with a common feature across all subjects and (2) a dataset with an unique feature for each subject. These datasets were labeled as *simulated*, *static* (simulated with an additional common feature across subjects) , and *unique* (simulated with a unique feature for each subject). Each one contained 10 minutes of data for the simulated 12 subjects and their 22 channels, matching the number of channels in the AutoEEG dataset.

Production of the synthetic datasets relied on a Gaussian Mixture Model based Hidden Markov Model (GMMHMM) consisting of 3, 4, or 5 Gaussian models drawn from UBMs. The baseline UBM came from 12 TUH-EEG AutoEEG V1.1.0 subjects using a 16-mixture UBM. The common and unique features came from a single random subject in the PhysioNet Database, also using a 16-mixture UBM. Simulated data contained either 3 or 4 mixtures, allowing the static and unique to add an additional feature containing 4 or 5 mixtures depicted by Figure 3.2.

This produced six unique synthetic data sets: Sim3, Sim4, Sta3, Sta4, Uni3, Uni4, outlined in Figure 3.2. Data was generated for each one-second epoch of each channel as CEP features directly. The distribution of the simulated data followed the

[2] ACNS - Guideline 3: `http://www.acns.org/UserFiles/file/EEGGuideline3Montage.pdf`

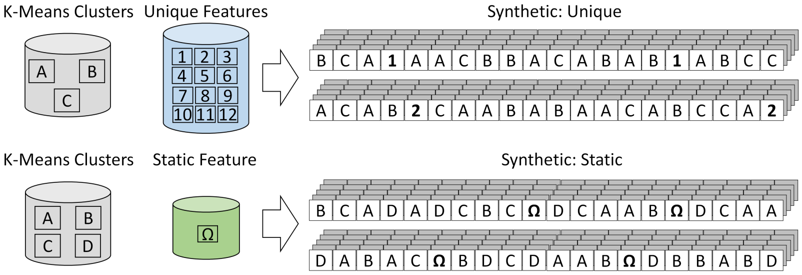

K-Means Clusters
Unique Features
Synthetic: Unique
A B C
1 2 3 4 5 6 7 8 9 10 11 12
B C A 1 A A C B B A C A B A B 1 A B C C
A C A B 2 C A A B A B A A C A B C C A 2
K-Means Clusters
Static Feature
Synthetic: Static
A B C D
Ω
B C A D A D C B C Ω D C A A B Ω D C A A
D A B A C Ω B D C D A A B Ω D B B A B D

Figure 3.2. Generation of synthetic data from the TUH-EEG. The GMMHMM modeled data (gray) and the unique (blue) or static (green) features enable the creation of unique and static synthetic data sets. Only 10% of the simulated data is replaced by the external PhysioNet Database feature. The modeling produced features for each epoch's 22 channels simultaneously to keep the channel-epochs temporal synchronized for each of the 12 simulated TUH-EEG subjects.

weighting of the initial 16 mixture UBM. When the static and unique features were added they overwrote 10% of the simulated data with the new PhysioNet Database-based feature. Authenticity of the raw data was preserved by keeping the synthetic data as similar to the TUH-EEG AutoEEG V1.1.0 dataset as possible, highlighted in Table 3.1.

Table 3.1. Composition of Synthetic Data Sets

Name	Type	Features	Channels	Sampling Rate (Hz)	Duration (s)
AutoEEG	Real	∞	22	100	1200
PhysioNet	Real	∞	64	160	120
Sim3	Simulated	3	22	100	600
Sta3	Static	4	22	100	600
Uni3	Unique	4	22	100	600
Sim4	Simulated	4	22	100	600
Sta4	Static	5	22	100	600
Uni4	Unique	5	22	100	600

3.2.4 Feature Sets

In addition to using multiple datasets, three feature sets were applied to the PhysioNet Database and TUH-EEG: Cepstral Coefficient (CEP), spectral coherence (COH), and Power Spectral Density (PSD). Using multiple feature sets was important because there is no consensus on an optimal feature set for EEGs. PSD features have a long history of use with EEGs [91, 188, 189], as do COH features [64, 162]. CEP are well-established features in the speech processing domain [190, 123]; their application to EEG research was introduced by the Neural Engineering Data Consortium (NEDC) [41].

The COH and PSD features were computed according to the work of LaRocca [64]. The CEP features were built following the standards developed by the speech community [41] and their channels modified to conform with a TCP montage used by neurologists [7]. Thus the feature sets are distinct not only in their mathematical construction, but also their topographical configurations, Table 3.2.

Table 3.2. Feature Set Configurations

Name	Type	Features	Channels
CEP	Original	26	22
	Slim	26	22
PSD	Original	40	56
	Slim	40	19
COH	Original	40	1540
	Slim	40	22

As discussed in the background, EEG recordings can use a variety of electrode configurations. For example, the PhysioNet Database contains 64 electrodes of data, while the TUH-EEG contained a myriad of electrode configurations. Therefore the TUH-EEG set was aligned with the most common standard, the TCP montage, resulting in 19 electrodes organized as 22 differential channels. La Rocca's features consisted of 56 PSD channels and 1540 COH channels making for a larger disparity in channels for each feature set. To address this channel imbalance, the TUH-EEG configuration layout was replicated for the PSD and COH feature sets producing two groups of features. The first was the 55 electrode layouts used by La Rocca [64] and the second time was a mirror of the 19 electrodes from the TUH-EEG TCP montage.

This resulted in a slim feature set consistent of the 22 channel CEP, 19 channel PSD, and 22 channel COH. The CEP and COH confirmed to the TCP layout, but the PSD were not converted to keep them as distinct from the COH features as possible. The benchmark testing against La Rocca's worked used the full feature sets, while all Algorithm Benchmarks and UBM-TVM Relationship experiments used the slim feature sets.

3.2.4.1 Cepstral Features

The CEP-based features were predicated on the success of similar MFCC used in speech recognition. Their adoption for EEG required shifting from a log frequency scale to linear frequency and adjusting the time windows for the Δ and $\Delta\Delta$ differentials. Generation of these features was introduced and detailed by Harati et al in [41], but is outlined here.

The base feature vector consisted of of nine coefficients (seven cepstral coefficients, the frequency domain energy, and the differential energy). The filter banks actually produce eight spectral coefficients covering the following frequency ranges: {0, 1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 71-80 Hz}. However, the zeroth coefficient is discarded and replaced with the frequency domain energy; the differential frequency energy becomes the ninth term. These filters provided a single energy value after bandpass filtering (Hamming) the FFT for each of the listed frequency ranges.

The two energy terms: frequency domain (E_f) and differential frequency energy (E_d) are given as:

$$E_f = \log\Big(\sum_{k=0}^{N-1} |X(k)|^2\Big) \tag{3.2-1}$$

$$E_d = \max\Big(E_f(m)\Big) - \min\Big(E_f(m)\Big) \tag{3.2-2}$$

E_f was derived from the outputs of the filter banks where N are the number of filters and X is the filtered cepstrum frequency output. Using these values within the prescribed 0.9s window of samples, the E_d is found by comparing the maximum and minimum E_f values over the range of m elements in the signal window. These built the first nine features with the remaining 17 coming from the first derivative (Δ) and second derivative ($\Delta\Delta$).

The Δ and $\Delta\Delta$ features used the same equations, but with different window sizes:

$$d_t = \frac{\sum_{n=1}^{N} \left[c_{t+n} - -c_{t-n}\right]}{2\sum_{n=1}^{N} n^2} \tag{3.2-3}$$

Here each sample n in the window N was used to produce a derivative for a given coefficient c centered around time t. Zero padding was used to pad the vector near the beginning and ending of the data. The first derivative Δ used $N = 0.9$. Once resolved, the second derivative $\Delta\Delta$ used the Δ values with a new window of $N = 0.3$. In Harati's work [41], the optimal configuration was found to be a 26 feature vector where the $\Delta\Delta$ for E_d was excluded. This configuration was adopted by the research group and became the consistent feature for the experiments in this work.

3.2.4.2 Power Spectral Density Features

PSD features are derived from the sum of energy over a frequency range for a given time sample. Variation in their creation can be found in their frequency range, number of FFT samples, and filtering of the time signal. The variation of PSD based features used in this work are identical to those of La Rocca et al. [64] which used a frequency range of 0-100Hz, a 100-point FFT, and Hanning windows for filters. The final features were 10-second epochs with 40 PSD values evenly spanning 1-40Hz.

The time series data was filtered with 1 second Hanning window using a 0.5 second overlap. This produced 20 filtered samples for each 10 second epoch centered around each 1 second interval from 0 to 9.5 seconds. These filtered samples were evaluated using Welch's averaged modified periodgram (built into Matlab) with a 100 point FFT to produce 1Hz resolution over the range of 0 to 100Hz. La Rocca's work used the PhysioNet Database data which first had to be resampled from 160Hz to 100Hz prior to the filtering.

The resultant 100 energy levels were reduced down to only those spanning 1-40Hz. This reduction in frequencies is necessary given (a) the resampling and (b) that the EEG oscillations of interest Delta (0.5-4Hz), Theta (4-7Hz), Alpha (8-14Hz), Beta (15-29Hz), and Gamma (30-40Hz) fall within that range. This resulted in 40 features per EEG channel. The channel count was reduced to 56 from PhysioNet Database's original 64. The discarded channels, highlighted in Figure 3.3, were AF_7, AF_8, FT_7, FT_8, T_9, T_{10}, O_Z, and I_Z.

While originally designed with the PhysioNet Database in mind, these features were readily adapted to the TUH-EEG. Recordings were resampled to 100Hz and pared down to the match the abbreviated 56 channel layout.

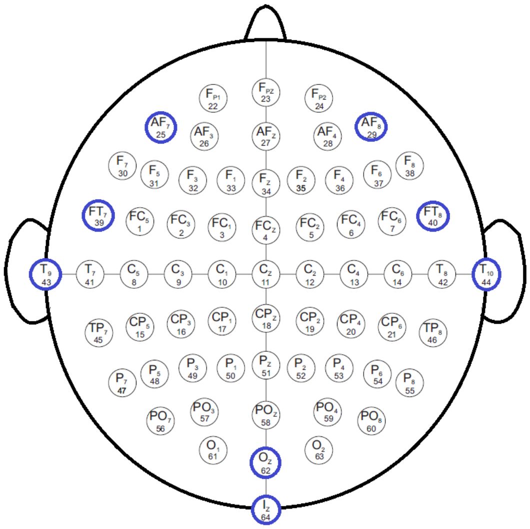


Figure 3.3. Layout of La Rocca's PSD and COH Channels. The channel layout La Rocca et al used removed 8, highlighted in blue, channels from the overal 64 channel configuration of the PhysioNet Database.

3.2.4.3 Spectral Coherence Features

The COH features were proposed by La Rocca as an improvement over PSD features for subject classification. Measuring coherence between electrodes had been used prior for distinguishing ADHD [139], a general connectivity measure of the brain [34] and auditory oddball paradigms for BCI/P300 responses [82]. Thus they were not novel features, but applied to a broad range of applications beyond subject classification.

These features were generated by quantifying the amount of synchronous energy at each frequency band of each electrode. This was achieved by first building the PSD features and then using them to generate a COH value for each frequency f between two different electrodes i and j, outlined as follows:

$$\mathrm{COH}_{i,j}(f) = \frac{|S_{i,j}(f)|^2}{S_{i,i}(f) \cdot S_{j,j}(f)} \tag{3.2-4}$$

The resultant values were scaled by arctan to normalize their distribution making them bounded on the range $(0, \frac{\pi}{2})$. This configured the final feature set as 1540 'channel' which La Rocca called elements. Each with the 40 distinct frequency bins found through the PSD feature process.

3.2.4.4 Aggregated Datasets

The UBM-TVM Relationship experiments needed subject and condition variation to test classification performance. To achieve, this aggregated datasets were built by combining the PhysioNet Database and TUH-EEG datasets. The combinations of PhysioNet Database's motion data and the TUH-EEG's normal, TUH-EEG's abnormal and normal, or TUH-EEG abnormal, normal, and seizure datasets allowed classification of subjects and known characteristics within a single experiment. This was important to address algorithm robustness and to mitigate any benefits conferred based upon a given dataset-feature-algorithm combination. Each combination was given a designation, Table 3.3 to streamline documentation and discussion.

3.3 Evaluation Metrics

All experiments were run as subject verification tests. This was inline with La Rocca's experiment which used Correct Recognition Rate (CRR) as their sole evaluation metric. However, given the depth of the datasets and parameter testing to be conducted it was necessary to also include the EER as well. The performance of I-Vectors has typically been reported in terms of EER, while the EEG research community is typically more broadly focused more on CRR. Exceptions in the literature [86, 103, 163] show results in terms of EER, FAR, FRR, HTER, or

Table 3.3. Combine Dataset Designations

Designation	Dataset 1	Dataset 2	Dataset 3
AbnNrm	TUH Abnormal	TUH Normal	-
AbnSzr	TUH Abnormal	TUH Seizure	-
NrmSzr	TUH Normal	TUH Seizure	-
AbnMot	TUH Abnormal	Physio Motion	-
NrmMot	TUH Normal	Physio Motion	-
SzrMot	TUH Seizure	Physio Motion	-
AbnNrmSzr	TUH Abnormal	TUH Normal	TUH Seizure
AbnNrmMot	TUH Abnormal	TUH Normal	Physio Motion
NrmSzrMot	TUH Normal	TUH Seizure	Physio Motion
AbnSzrMot	TUH Abnormal	TUH Seizure	Physio Motion

Detection Error Tradeoff (DET) curves. For the purposes of this research results were reported in terms of CRR and EER to facilitate readers from both the I-Vector and EEG communities being able to contextualize the experiment performances.

In this work CRR was calculated based on the testing data correctly matching into the enrollment data. The EER was calculated over the entire distance matrix ensuring it evaluated the strength of all matches. This meant if subject 100s second best score was stronger than subject 4's score the EER would be none zero. This is why it was critical to include it for the parameter sweeps, as the CRR masked the majority of the nuance of the full system.

Even with the importance of both metrics, the intended parameter sweeps and comparison points made always displaying both CRR and EER cumbersome and ineffective to the end goal of comparative performance. AS such, the C Metric was defined which combined the CRR and EER by subtracting the EER from the CRR. Thus the threshold for an acceptable C Metric score was set at 0.75 which could represent a CRR of 85% and an EER of 10%. This was primarily used for the expansive Algorithm Benchmarks to showcase performance differences between GMM-UBMs, MD, and I-Vectors.

3.3.1 Mixture Size

For UBMs, TVMs, and I-Vectors the dimension of the underlying mixture model is a critical parameter than can affect performance. Effectively, the n dimensional feature space is modeled by m gaussians; these gaussians are used to train the I-Vectors. As has been the case in the speech community [124, 168, 191], it was necessary to determine the size of the mixture model that would optimize I-Vector performance under different circumstances. While some experiments applied GMM-UBMs previously, their protocols and datasets were not a sufficient starting point[42, 163].

These experiments were used to inform the initial mixture sweep range {2, 4, 8, 16. 32, 64, 128, 256, 512, 1024} used as part of the Parameter Sweeps. After which it was expanded to {2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048} for the Algorithm Benchmarks and UBM-TVM Relationship experiments. The smallest datasets contained 50 subjects, but each dataset had at least 19 channels per subject amounting to a lower bound of 950 distinct subject-channels each with 40 features. From this lower subject-channel bound there the number of epochs in the training and enrollment datastets would change based upon the epoch duration. With the largest epoch duration of 10 seconds, there would be at minimum 9 epochs for each of the 950 subject-channels producing 8,550 unique subject-channel-epochs to model. This value exceeded the upper limits of the two mixture sweeps ensuring overfitting was not a major influence on performance.

3.3.2 TVM Dimensions

The size of the TVM was bounded by the number of mixtures and a dimension factor called l from section Section 2.3.1. As stated, this l value had to be less than or equal to the number of subjects, otherwise models would be built specifically for each subject. Overfitting concerns were addressed with respect to the mixture sizes, but limiting the TVM depth to the number of subjects assured overfitting was impossible in the production of I-Vectors.

This was not strictly required, as the examples used to inform this work would build TVMs with a depth beyond that of the number of subjects [166]. As the TVM is an intermediate step before finalizing the I-Vectors with LDA, the dimension of the TVM could be 1200 for processing data from 75 subjects [192]. However, such options were based on datasets with an order of magnitude more epochs and contained feature vectors double in size than was proposed in this work

Bounding the upper limit was necessary given the dynamic between mixture size, TVM depth, and LDA depth. An upper bound of 200 was chosen because the majority of datasets and aggregated datasets would not exceed 200 subjects. Additionally, producing the TVM was the most computational intense components of the algorithm requiring a tradeoff of the sweep range and execution time. The lower bound was set at 25, half the smallest subject count. Three incremental values were used to step between the lower and upper bound which resulted in the following sweep range: {25, 50, 75, 100, 200}.

3.3.3 LDA Dimension

The use of LDA to finalize the I-Vectors was well documented by the founders of I-Vectors [121, 166] highlighting their own sweep for optimization with speech data.

Thus LDA depth represented a third parameter to consider when building and evaluating I-Vector performance. The upper bound of LDA is determined by the size of its paired TVM. As the range of TVM dimensions was being aligned with the various aggregated dataset subject counts, the LDA dimensions were aligned to operate on a similar scaling.

The lower bound for the LDA dimensions was set to 15, slightly less than the TVM lower bound, and the upper bound was set to 100, half the TVM upper bound. Five intermediate values were chosen between the bounds which resulted in the following sweep range: {15, 30, 45, 60, 75, 100}. By focusing on smaller increments this parameter was designed to be less influential than the mixture size and the TVM depth. This sweep range would later be adjusted following the results of the Parameter Sweeps to: {5, 15, 20, 25, 25, 50, 75, 95, 100, 150, 195}.

3.3.4 Epoch Configuration

The final controllable parameters were the number and duration of epochs. Drawing the experiments from the work of La Rocca et al, the initial epoch duration was 10 seconds with 6 epochs per subject, based around the resting trials of the PhysioNet Database. The epoch durations were expanded to include 5, 2, and 1 second epochs. This naturally altered the number of epochs as the PhysioNet Database contained 1 minute and 2 minute trials which split into a various numbers of epochs for each epoch duration recording combination, show in Table 3.4.

Based upon reviewer feedback to a prior publication [193] epoch generation was altered to enabled the number of epochs to be independent of epoch duration. This provided another parameter to sweep, number of epochs, which was previously conflated with the epoch duration and trial duration.

Table 3.4. Epoch Duration Configuration

Trial Duration (s)	Epoch Duration (s)			
	10	5	2	1
60	6	12	30	60
120	12	24	60	120

3.3.5 Datasets and Features

Each experiment used all three feature sets, but not every combination of datasets was explored. This was because finding an optimal feature set was beyond the scope of the proposed work. There were not enough available resources in terms of datasets, features, and time to satisfy a robust feature search. However, it was understood that the proposed experiments could offer insight into feature selection which is why every experiment used all three feature sets.

Despite this limitation, using all three feature sets for each experiment provided a comparison point for understanding algorithm-dataset performance. It was hypothesized that one feature set would generally outperform the others, independent of data. Variations in relative performance triggered by mixture size, TVM depth, LDA depth, or epoch settings were used to define areas of interest with respect to these controllable parameters. Additionally, using multiple features mitigated any potential bias generated for stumbling upon an ideal dataset-feature-algorithm combination and being able to identify it as such given the number of dataset-feature-algorithm pairings.

3.4 Implementation

In keeping with the theme of publicly available datasets, the software and hardware solutions were developed to be open sourced. As the research intersects multiple communities it was important that access be given to all regardless of expertise in software development or hardware support. Many of the latest data science solutions required every updating tool kits running on large computing clusters which can limit the use of novel tool kits.

3.4.1 Software

The initial search for I-Vector toolboxes yielded bob.spear[194], Kaldi[195], and Microsoft Research (MSR) Identity Toolbox[196]. The bob.spear toolbox did not work on Windows based machines and Kaldi had proven difficult to implement on the NEDC computing cluster. However, the MSR Identity Toolbox was developed with MATLAB and was easily setup locally and on the computing cluster.

The majority of software was developed specifically for this research with minor components drawn from public sources. A MATLAB toolbox called VOICEBOX[3] was used to support handling of the CEP features generated as Hidden Markov Toolkit (HTK) files. All EDF EEG files were manipulated using edfREAD available through Mathworks MATLAB File Exchange[4].

The decision was made to build using Matlab because it provided a known functional model in the MSR toolbox, would be accessible to both the speech processing and EEG communities, and be robust to hardware/software configurations, and scalable for use on computing clusters. In hindsight there were

[3] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

[4] `https://www.mathworks.com/matlabcentral/fileexchange/31900-edfread`

tradeoffs in terms of performance and flexibility that may have been mitigated by developing the software tools in Python, but the development of this software package was a tertiary goal. Over the duration of the research the Matlab versions started with R2015A and finished on R2017B.

A review of the major facets of the software's workflow is provided in this section. All experiments started with feature creation as the data already existed within the NEDC file system. Once features were produced, a parameter file was written to control the experiment. This file outlined how each process would operate. The experiments were run sequentially to assure each algorithm used the same randomly generated epoch splits for training, enrollment, and testing data. Ultimately, all experiments were run on the NEDC clustering requiring Bash scripts to interface with our Slurm Workload Manager. Those interested in the individual classes and functions should refer to public Git repository's ReadMe[5].

3.4.1.1 Feature Creation

The conversion of EEG recordings, stored as EDF files, into CEP, COH, and PSD features was independent of the experiments. This was done to ensure static feature sets and simplified the structure of processing the features during the experiments. Given the number of 'channels' produced from COH features, all feature data was indexed and saved in relation to their epochs.

Thus the number of files produced for each feature set was dependent on subject and number of epochs with channel data organized inside each epoch file. These file lists were the inputs to the experiments where they were aggregated. This tool was written to run with multiple Matlab workers and was supported via a Slurm base script.

[5] `https://github.com/izlandman`

3.4.1.2 UBM Class

The use of UBMs was handled through the development of a Universial Background Model (UBM) class in Matlab called *UniBacMod.m*. This simplified the generation, evaluation, and loading/saving of existing models. The generation of the UBMs leveraged Matlab's Single Program Multiple Data (SPMD) parallel computing feature to carry out the EM process on the training data. The enrollment models were built using a parallel MAP adaptation from the generated models and enrollment data. These enrollment models were compared against the testing data to produce log-likelihood ratios which were scored for CRR and EER.

This class controlled the number of UBMs mixtures, the number of EM iterations, and the downsample factor. In addition it held the number of epochs and the resultant UBMs. All of these variables were saved after converting the class to a structure enabling subsequent I-Vector experiments to use the same UBMs.

3.4.1.3 TVM Class

The use of TVMs was handled through the development of a total variability matrix (TVM) class in Matlab called *TotVarMat.m*. This simplified the generation, evaluation, and loading/saving of existing models. Again the EM process used to build the TVM was run using the same parallel processes for the UBM class. The generation of I-Vectors was done in parallel as well, with the option to produce a set of LDA constrained I-Vector in addition to the native TVM I-Vectors. Final evaluations between the I-Vectors were carried out through a parallel cosine distance function to produce the CRR and EER metrics.

The class retained the enrollment and testing I-Vectors and performance metrics binary files, with all other parameters saved as a Matlab structure. Control over the depth of the TVM, depths of the LDA variants, and training steps for the TVM EM were carried out in this class. The constraints previously laid out by the imported UBM are inherited by the TVM class. This assured the number of epochs and UBM parameters were consistent between algorithms. Critically, this allowed the production of I-Vectors from a static UBM produced in a prior experiment.

3.4.1.4 Mahalanobis Evaluation

The use of Mahalanobis Distance as a classifier was borrowed from the work of La Rocca et al. [64]. They developed their experiments using Matlab using the built-in `Mahal` function from the Statistics and Machine Learning Toolbox. Each training/enrollment subject's epochs were used to produce a subject mean. The variances for each feature were drawn from a pooled covariance matrix built from all subject's epoch data.

Evaluation of the the distance matrix between all subjects was used to produce CRRs and EERs aligned to the same epoch, mixture, LDA depth process as the I-Vectors. The resultant distance matrices were saved for each step of the cross-validation process as a binary file. No class was built for this process as it was not the main focus of the proposed research.

3.4.2 Hardware

All of the experiments were run on the NEDC computing cluster, Neuronix. While the cluster supported CPU and GPU parallel processes, the toolkit was written to only support CPU parallelization. Neuronix contained four main identical CPU compute nodes and two minor identical CPU compute nodes. The main nodes consisted of two AMD Opteron 6378s with 16 cores supported by 128GB of DDR3 Ram. The minor nodes consisted of two Intel Xeon E5-2603s with 8 cores supported by 128GB of Ram. The data server consisted of over 2TB of disk space shared by all the users of NeuroNix.

Chapter 4

I-VECTOR DEVELOPMENT

I-Vectors were developed for use by the speech recognition community as in improvement to JFA[197]. At that time, JFA was “a model used to treat the problem of speaker and session variability in [speech] GMMs” [198]. These GMMs were built from UBMs that showed strong performance in speaker verification tasks [120]. Similar work was attempted on EEGs [103] indicating the potential of this technique on EEG data [199] which has properties similar to speech data [137].

While I-Vectors were being actively developed for speech data, there was no such published research at the onset of this work in 2016 [200]. Thus the preliminary research goals were (1) to build and verify a software package capable of producing and evaluating I-Vectors from EEG recordings, (2) explore the impact of UBM mixture size on performance, (3) compare performance on a subject and session levels, and (4) contrast with existing modeling techniques, GMM-UBM.

With minimal documentation on the feasibility of I-Vectors as an EEG discriminator, preliminary research was necessary to show efficacy. This was achieved by carrying out a series of experiments using both the synthetic data and the PhysioNet Database. First, the synthetic dataset was used to validate the the performance of the developed software in a controlled environment. Secondly, subject verification across the PhysioNet Database was carried out. Finally, intra-subject trial and channel verification testing was performed to test the classification and modeling processes on different facets of the data.

Each of these experiments used a range of UBM mixtures for the I-Vector and GMM based verification tests. In this verification testing scheme, all data used (training, enrollment, and testing) was identical to focus on performance independent of data variations. This helped to satisfy the four preliminary research goals: build and verify software, explore impact of mixtures, compare classification on distinct facets of the dataset (channel, subject, and trial), and evaluate against a known technique. Using GMM-UBM as the alternative classifier provided verification for the creation of UBMs as well as the I-Vector performance.

These results were used to inform the research aims and gave rise to the Core, Principal, and Comparison Experiments. Therefore this preliminary research used a different experimental protocol as La Rocca's experimental protocols were adopted later with the Parameter Sweeps. All data was sourced from the PhysioNet Database and only CEP features were used. Each subject verification experiment was carried out once as the datasets were identical making it impossible to perform any sort of leave one out cross validation (LOOCV).

The initial experiment utilized synthetic features and a small dataset to test and validate the construction of the algorithm in software. Following on this were two experiments, one testing all of the PhysioNet Database on the subject level and another testing on the trial level. This lead to trial specific experiments where the two calibration trials were withheld and included and also subject specific trial verification. Lastly, the data was partitioned on the channel level and evaluated against matching to its native subject-trial and subject.

The range of UBMs was 2 to 1024 mixutres and the TVM dimension was set for the number of subjects. LDA was used to finalize all I-Vectors to a maximum size of one less than the number of subjects. This resulted in an I-Vector dimension of 11 for the synthetic data and 108 for the PhysioNet Database. Only EER was used, as that was the defining performance metric in the majority of I-Vector papers.

4.1 Synthetic Experiments

The initial version of the software was validated by testing on the controlled synthetic data set outlined in section 3.2.3. There was concern over proper epoch duration, mixture size, channel organization, and data quality issues that needed to be addressed. Using a synthetic set based upon real world data mitigated issues of epoch duration, channel organization and data quality because they became dependent variables in the experiments (see Table 3.1). This enabled the synthetic experiments to focus on mixture size versus dataset composition and the performance of GMM-UBMs versus I-Vectors.

4.1.1 GMM-UBM

The GMM-UBM classification provided a performance target for the I-Vectors technique to compare against. Figure 4.1 provides the EERs of subject verification for each of the six synthetic datasets and the original PhysioNet Database subject. The original and synthetic datasets classifications using GMM-UBM improved with larger mixture sizes. The rate of EER reduction in the two unique datasets (Unique 3GMM and Unique 4GMM) was the strongest, reaching zero at 32 mixtures. The remaining datasets reduced their EER by roughly 10% over the 10 mixture sizes.

4.1.2 I-Vector

The I-Vectors classification results, Figure 4.2, showed the majority of datsets achieved near zero EER with 4 mixtures in the UBM. The two exceptions, Static

3GMM and Static 4GMM, required 8 mixtures to achieve equivalent performance. Unlike their UBM based counterpart none of the datasets ever settle to an EER of zero. However, the overall performance indicated that the mixture size was more influential than dataset composition.

4.1.3 Discussion

The classification results of the GMM-UBM and I-Vector experiments showed the initial software development was successful. While not presented in a figure, all of the subject verification matches produced CRRs of 100%. The reported EERs came from inconsistency in the strength of the verification matches. As the best CDs produced for a given subject's verification match many not have exceeded those values produced

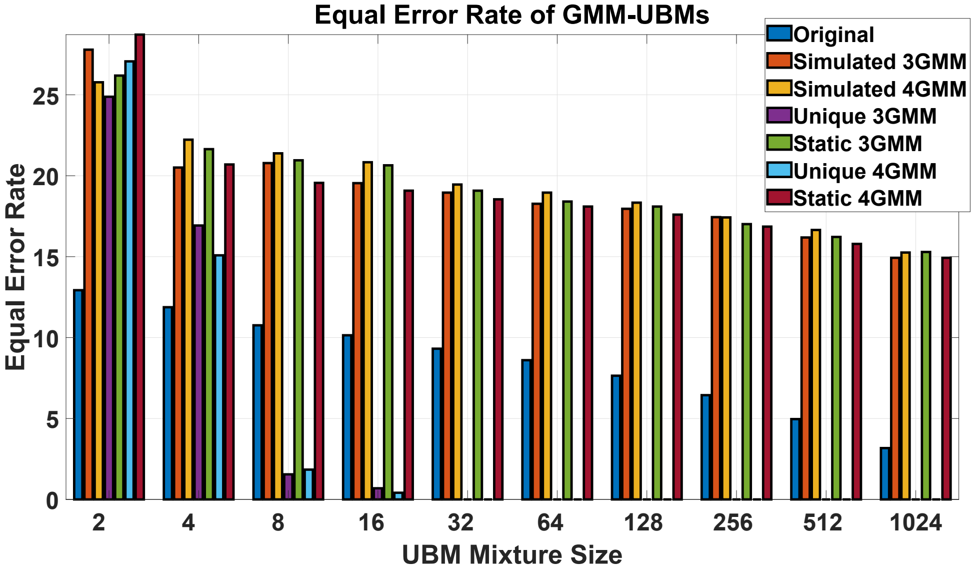


Figure 4.1. Synthetic TUH Corpus GMM-UBM Results. EER of UBMs on the seven data sets (l to r) Original, Simulated 3GMM, Simulated 4GMM, Unique 3GMM, Static 3GMM, Unique 4GMM, Static 4GMM. The EER for two unique data sets reaches 0% when the models exceed 16 mixtures.

from secondary matches of other subjects. Thus the best 109 CDs were not linked with the primary matches for each subject.

The I-Vectors produced more robust CDs which lead to a strong EER performance, figure 4.2, it was clearly difficult for the GMM-UBMs to produce consistently strong evaluations for all 109 subjects when the dataset lacked distinct features, figure 4.1, regardless of UBM mixtuer size. The I-Vectors performed well across all mixture sizes and datasets. The ability of the I-Vectors to reduce EER below 5% for all datasets showed the impact of the TVM in further differentiating the subjects. This stark contrast in overall performance supported the hypothesis that I-Vector would exceed GMM-UBM performance. However, this was not true for all combinations of datasets and mixtures as the GMM-UBM achieved an EER of zero on the unique datasets when using a 32 mixture UBM or larger.

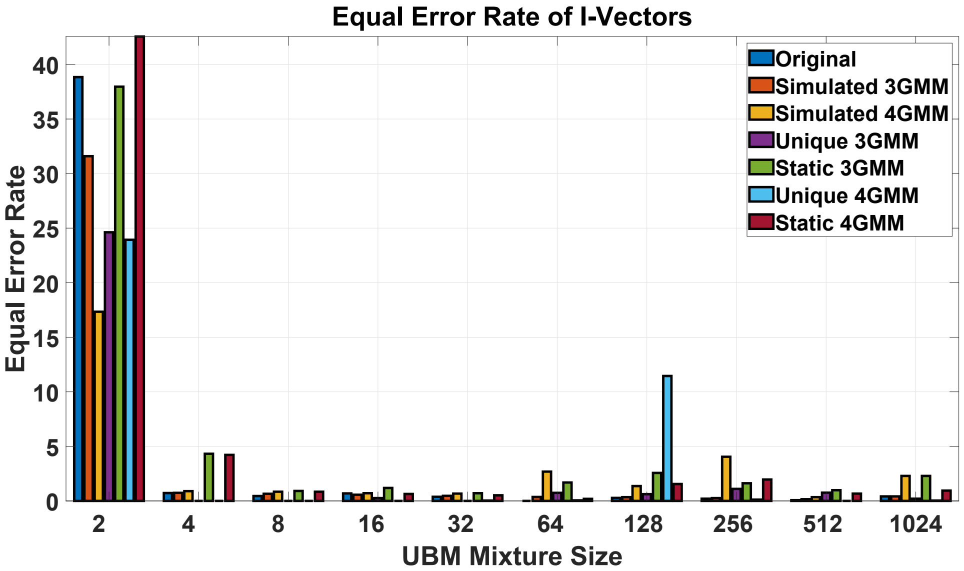


Figure 4.2. Synthetic TUH Corpus I-Vector Results. EER of I-Vectors on the seven datasets (l to r) Original, Simulated 3GMM, Simulated 4GMM, Unique 3GMM, Static 3GMM, Unique 4GMM, Static 4GMM. A strong reduction in EER is seen when transitioning from 2 to 4 mixtures for modeling. Beyond this transition changes to the EERs at higher mixture sizes are minimal.

In these experiments the datasets were randomly generated, leading to each algorithm-mixture combination using distinct training, enrollment, and test datasets. This mean a distinct UBMs was built for each of the algorithms, the 8 mixture UBMs used for the I-Vector and GMM-UBM classifications were only as similar as the UBM training process made them. Once a UBM was developed it was determined that the TVM generation process was deterministic. This was tested as a response to a paper revision question by confirming the similarity between five repetitions of building a TVM from a static UBM and enrollment dataset pairing.

This indicated any variance in the UBM or enrollment data would be passed on to its TVM. This likely attributed to the outliers in the results, such as the EER of the unique 4GMM dataset in Figure 4.2. The GMM-UBM EER trends were shared, in that each dataset experienced equivalent performance improvements as a function of mixture size. However, the strength of that performance improvement was dependent on the content of the dataset with the unique datasets improving more rapidly with increases in UBM mixture size.

From this initial experiment it was evident that a variety of datasets would be necessary to understand the impact of mixture size on performance. While the performance of the GMM-UBM classification appeared more dependent on dataset composition, it still served as an acceptable performance goal of I-Vector classification.

4.2 Verification Experiments

After confirming the development of the software and its performance, verification experiments were run on real data drawn from the PhysioNet Database. The purpose of these experiments was to see if the synthetic performances could be replicated on a real EEG dataset. This increased the complexity of the task given each of the 109 subjects had 14 trials to process, far more than the 12 subjects in the synthetic dataset. In total the PhysioNet Database contained 33,572 channels drawn from 1,526 trials belonging to 109 subjects making it possible to test the upper limits of the I-Vectors which needed only a 4 mixture UBM on the synthetic dataset. The layers of the data (channels, trials, and subjects) also allowed for classification beyond features built at the subject level.

The first experiments, subject and trial verification, simply replicated the synthetic experiments on the subject and trial levels using the full PhysioNet Database. This was meant to assess the influence of the dataset hierarchy on classification performance, since speech-based applications had found I-Vectors capable of performing more than just subject verification, such as language recognition [121] and word recognition [201]. Additionally, the synthetic experiments showed that dataset composition impacted performance so the verification experiments used a full dataset and one devoid of resting state trials.

As each subject's recordings were split into 14 trials, intra-subject trial verification was performed to determine the full dataset's trial verification performance. This required outlining the potential groupings of the trials which could be classified as as resting Eyes Closed (EC) or Eyes Opened (EO), and one of four repeated motion tasks depicted in **??**.

This created a natural split between the repeated motion tasks and the resting states shown in table 4.1 as distinct groups (G) based upon their content. Given this protocol it was possible to generated an expected likelihood of the outcomes for matching each grouping, table 4.2. This experiment used only the data within a given subject providing far less training, enrollment, and testing data than those experiments using all 109 subjects.

Table 4.1. [PhysioNet Trial Cohort Groups

Label	Data Group	Cohort Groups	Search Interval

Motion 3/3	G1, G2,
	G3, G4
		{G1}{G2}
		{G3}{G4}
			[1-3] of 12

Full 3/3	G0, G1,
	G2, G3, G4
		{G1}{G2}
		{G3}{G4}
			[1-3] of 14

Full 3/5	G0, G1
	G2, G3, G4
		{G1}{G2}
		{G3}{G4}
			[1-5] of 14

Full 5/5	G0, G1
	G2, G3, G4
		{G0 G1}{G0 G2}
		{G0 G3}{G0 G4}
			[1-5] of 14

The Data Group specifies the trials given for the search space. The Cohort Groups show which trials are considered a distinct group. The Search Interval defines the acceptable positions [a-b] out of the available trials presented in the Data Group.

Table 4.2. Expected PhysioNet Database Cohort Likelihoods

Combinations	1 Match	2 Matches	3 Matches	4 Matches	5 Matches
Motion 3 of 3	65.45	32.73	1.82	-	-
Full 3 of 3	70.51	28.21	1.28	-	-
Full 3 of 5	49.45	24.73	2.75	-	-
Full 5 of 5	17.62	47.00	30.18	5.04	0.14

The probabilities are generated from *p choose n* using the parameters set forth in Table 4.3.

Lastly, subject verification at the channel was tested by evaluating how well channel I-Vectors matched into their native subject and trial I-Vectors sets. This used the full PhysioNet Database split into channels whose cosine distances were averaged across channels to find a resultant subject and trial match. Operating on the channel level allowed for matches on the trial level to match into any of the subject's trials. This provided a comparison to the trial matching of the intra-subject trial verification. In addition, these classifications were reversed to also test subject and trial groupings against their native channels.

4.2.1 Subject and Trial Verification

The subject verification experiment of the synthetic data was replicated using all of the subjects in the PhysioNet Database. This was expanded to include verification of subject-trials into subjects as well. For the subject experiment 109 I-Vectors were built from each of the enrollment and testing data. For the subject-trial experiment 1526 acpIV were built, one for each trial of each subject, from each of the enrollment and testing data. The subject I-Vectors were evaluated directly by their CD, while the subject-trial I-Vectors were evaluated based upon the consensus of the enrollment I-Vectors voting in favour of match to the testing I-Vector. As this organization was carried out by altering what the algorithm defined as a 'subject', the GMM-UBM operated on an identical datasets performing equivalent evaluations.

The performance of the GMM-UBM classifier improved on both datasets as the UBM mixture size increased, consistent with the behavior seen on the synthetic data. However, the I-Vectors classification EERs was inconsistent across the two datasets. When evaluating the subject verification the general trend of increased mixtures leading to reduced EER was observed. The subject-trial experiments yielded equivalent or worsen performance beyond an eight mixture UBM. This resulted in the GMM-UBMs being better suited for the subject-trial data, and the I-Vectors being better suited for the subject data. Although, the GMM-UBMs required a 512 mixture UBM to match the I-Vectors performance before exceeding it with 1024 mixtures.

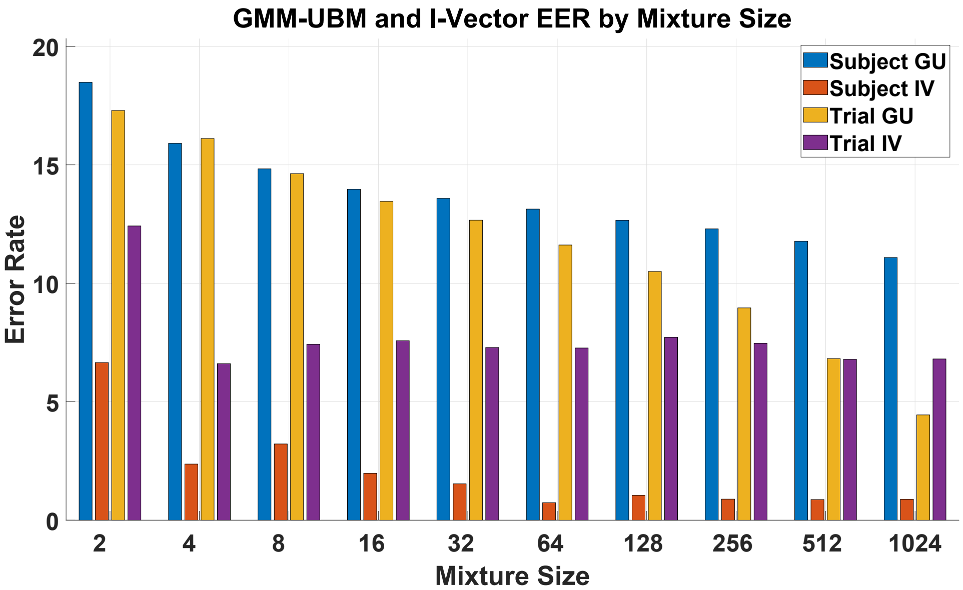


Figure 4.3. PhysioNet Verification Testing Results. PhysioNet Database UBM and I-Vector verification test results as a function of mixture size. The leftmost bars represent Subject UBMs and I-Vectors and the rightmost bars represent Trial UBMs and I-Vectors.

This behavior was further examined by evaluating on the subject-trials directly against each other, removing the consensus evaluation. Now a subject-trial in the testing dataset would be required to match to its twin in the enrollment dataset. Given the discrepancy in duration of the trials (1 minute for resting trials and 2 minutes for motion trials) two different datasets were evaluated, one with the resting trials and one without them, figure 4.4. These results exhibited a gradual GMM-UBM improvement over the range of UBMs. Meanwhile, the I-Vectors displayed a near step-wise improvement over the first three UBM mixture sizes as performance leveled off with the sixteen mixture UBM. The performance of the GMM-UBMs reached that of the I-Vectors when given a 256 mixture UBM and then exceeded it by 1024 mixtures.

The removal of the resting trials reduced the amount of data for the training, enrollment, and testing datasets, which manifested as a worse EER for the GMM-UBMs and an improved EER for the I-Vectors (after reaching 8 mixtures). However,

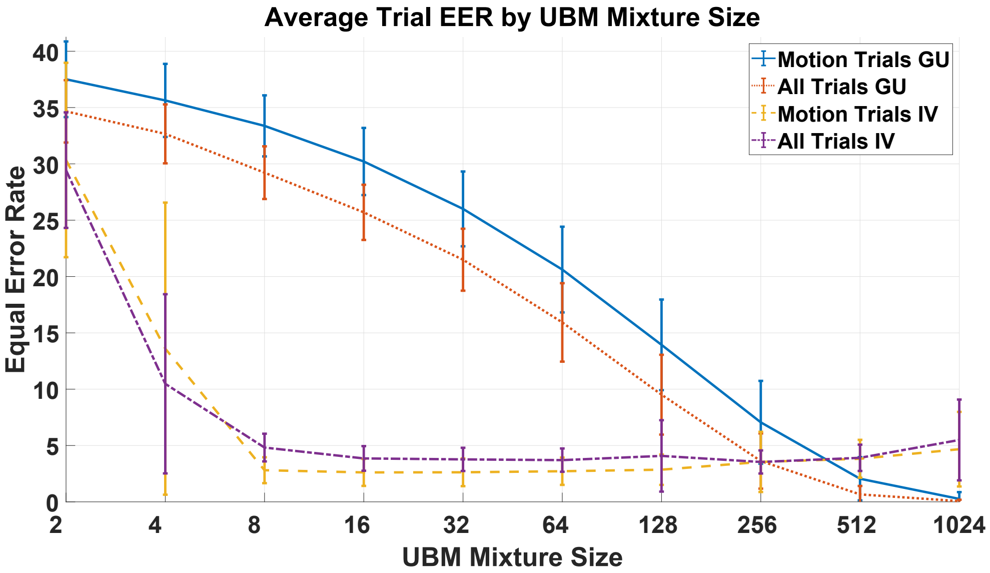


Figure 4.4. PhysioNet Trials' EER]. The averaged EERs of each PhysioNet Database subjects' trials as a function of mixture size. Error bars represent +/- one standard deviation across the entire subject set. The I-Vectors' EER plateaus after 8 mixtures, while the UBMs' EER decreases as the mixture size grows.

the one sigma error bars of the I-Vectors results show an overlap between the means of the two data sets, while the GMM-UBM error bars do not reach the other dataset's mean. This suggested the I-Vectors produced stronger classifications given the two datasets were well within each others variances. The same was not possible for the GMM-UBMs, as the addition of the resting trials pushed performance beyond the standard deviation of the motion data significantly improving performance.

4.2.2 Intra-Subject Trial Verification

The strong I-Vector subject-trial performance suggested the possibility of clustering a the trials unique to a given subject based upon their groupings; Recall that the four motion trials were repeated 3 times in the original experimental protocol, table 4.1. Evaluating the within-trial clustering for the 109 subjects provided enough samples to generate statistical significance, table 4.3, with respect to the expected distributions, table 4.2. The previous subject-trial results indicated an eight mixture UBM would be ideal, there was concern that the number of mixtures was greater than half the number of trials. This lead to the experiment being conducted with the four mixture UBM data which likely increased the difficulty, but ensured the number of models would be less than the number of unique trials (EC,EO, and four motions).

Table 4.3. Experimental PhysioNet Database Cohort Likelihoods

	1	2	3	4	5
Motion 3 of 3	55.66ˆ	41.13ˆ	3.21*	-	-
Full 3 of 3	66.44ˆ	33.72ˆ	2.14*	-	-
Full 3 of 5	38.00ˆ	51.07ˆ	10.93ˆ	-	-
Full 5 of 5	15.29ˆ	43.88ˆ	33.64ˆ	6.88ˆ	0.31*

Values indexed with * are significant at $p<0.005$ and with ˆ are significant at $p<0.001$.

Recall that value values represent the likelihood of these events given the expected probabilities of 4.2.

All results were found to be statistically significant in terms of improving the ability to cluster the subject's trials. For each cohort, the top ranking match was the native trial with the remaining cohort trials filling out the additional matches. The most important result was that the inclusion of the two resting trials ("Motion 3 of 3" versus "Full 3 of 3") reduced the performance of 2+ from 44.34% to 35.86%. However, Full 3 of 5 resolved 62.00% of the subjects with 2+ matches and Full 5 of 5 matched 84.71% of subjects with 2+ matches and 40.83% with 3+ matches.

The inclusion of the resting trials, Full 3 of 5, complicated the classification because the motion trials are comprised of rest states between the motion tasks, Figure **??**. Despite their addition, the I-Vectors were able to find an additional motion trial at a rate of 60% which was in excess of the expected 27.48%. These results further validated the ability of I-Vector as a classification technique while suggesting motion trials were distinct from each other.

4.2.3 Channel Verification

Channel-based subject and trial verification testing extended the previous experiments by operating on the lowest level of the data hierarchy (subject → trial → channel). In these experiments each of the 33,572 total channels was turned into an I-Vector. By comparison, for the subject and subject-trial verification in 4.2.1, each of the 109 subjects' data was converted into its own I-Vector; for the subject-trial verification, each of the 1526 subject-trials was converted into an I-Vector. The results, given in Table 4.5, were built by ordering the CDs for a given channel from strongest to weakest. This allowed the remaining 21 channels to be evaluated for their similarity to their native subject's trial cohort (primary) and any subject's matching trial cohort matches (secondary).

Figure 4.5. PhysioNet Subject and Trial Matching

Data	Verification	Primary	Secondary
Sub to Chan	76.15	-	-
Chan to Sub	82.57	-	-
Trial to Chan	81.06	52.43	9.48
Chan to Trial	96.66	63.23	7.61

Despite the large number of channels, subjects, and trials to model, the verification was performed with an 8-mixture UBM, since Figure 4.4 showed no improvement beyond 8 mixtures. The results are reported as true positive percentage, as the experiment focused on the ability of I-Vectors to match/cluster into subsets of relevant data. Channel I-Vectors performed well when verified against their native subject (82.57%) and subject-trial (96.66%). Reversing the verification process showed it was more difficult to verify subjects (76.15%) and trials (81.06%) to individual channels. This was likely caused by averaging the channel cosine distances to build the models, making channel outliers more distant than their median counterparts.

Of the 21 remaining channels, the majority (70.84%) matched with a channel from the same subject's trial cohort or the same trial cohort from a different subject. This was not as strong (61.91%) when classifying a trial toward it's channels or similar trials. This suggested the I-Vectors had captured data relevant to their subject and trial, which was consistent with the capabilities reported in speech recognition.

4.2.4 Discussion

Performing verification experiments on each of layer PhysioNet Database data hierarchy showed that I-Vectors met or exceeded the performance of their GMM-UBM precursor. Only when evaluating trial data for subject verification did the I-Vectors perform poorly in terms of EER, 4.3. The follow-on experiments focused on intra-subject verification suggested this could have been caused by the strength of a subject-trial to subject match overpowering the ideal subject-trial to subject-trial match. Thus the subject was likely correct, but the incorrect subject-trial was chosen. After all, simple intra-trial verification did not perform perfectly, 4.3. Still, when the subject-trial verification was evaluated it showed robust I-Vector performance, suggesting the organization of the datasets directly impacts performance across the hierarchy.

The influence of the hierarchy was why the channel verification experiments were critical. They took this one step farther by indicating that channel I-Vectors could be used to find their original subject and trial. In addition, they were capable of finding the majority of their associated trial channels too. These results showcased the power of I-Vector when applied with minimal domain knowledge to EEGs and to the technique itself. However, the limited experimental scope provided no insight into why or how this was possible.

These results confirmed the already well documented application of EEG as a biometric application and hinted at being simultaneously capable of BCI applications given its intra-channel performance. Further experiments should have been conducted using epochs built from the channel data and from diverse datasets. However, there are few substantial EEG datasets for BCI, seizure, sleep, and artifacts. Without those, and without a standardized EEG feature set, direct algorithm comparisons were impractical with current published research.

4.3 Conclusion

The preliminary research experiments showed that I-Vectors could be applied to EEG for biometric applications and potentially clustering/cohort retrieval tasks. The verification performance on the subject, trial, and channel levels indicated that complexity of EEG signals can be discerned by this factor analysis technique. However, the limited dataset and feature set could have been the driving factors in performance and not the algorithms themselves. It was therefore necessary to build on these experimental results to validate the efficacy of I-Vectors as an EEG classification technique.

A main concern was the variety of EEG data. Many studies operate on a specific dataset tailored to their experimental protocol, but are seldom made public. This makes it difficult to run experiments with multiple datasts and to offer comparative results across diverse types of EEG data. However, the NEDC is amassing the largest publicly available EEG database, making it possible to develop diverse datasets between the PhysioNet Database and the TUH-EEG. This allowed for the creation of four new distinct datasets: TUH Abnormal, TUH Normal, TUH Seizure, and Physio Motion.

These datasets are converted to three feature sets which allowed the more commonly used COH and PSD features to serve as a benchmark against the novel CEP features. In addition, the datasets were partitioned into epochs and structured on the subject and trial level. With three feature sets and four proposed epoch durations, each experiment had to be run twelve times for a given dataset. This was done to track the impact of the feature sets and epoch duration across each dataset.

The range of UBM mixtures was enough to capture the major trends of the GMM-UBM and I-Vector performance. While some experiments were minimally impacted by larger mixtures, others showed strong gains with larger mixtures (see Figures 4.3 and 4.4). If the number of subjects were to increase, in the instance of building the aggregated datasets, it would be suggested to increase the range of mixtures, but otherwise the range was acceptable.

Despite all of the positive results, nothing was done to refine the I-Vectors in any meaningful way. Their lengths were set by the number of subjects in the datasets and optimized by LDA after the generation of the TVM. While the dimension of the TVM and the dimension of the resultant I-Vectors were frequently the focus advances in the speech community, none of those adjustments were used in these experiments. It is entirely possible a set of UBM, TVM and LDA parameters could improve the I-Vector performance. This concern was the seed for the majority of test conditions in the Parameter Sweeps.

These preliminary experiments were far from robust in terms of thoroughness or diversity, but provided basic assurances about the applicaption of I-Vectors on

EEGs. This foundation would need to be extended not only in terms of parameters and datasets, but also in terms of features and protocol. Interest in EEGs as a bio-metric tool focused on subject verification problems and was a natural starting point for furthering the testing of I-Vectors. One such group's work, La Rocca et al [64], provided a streamlined experimental protocol that aligned with the intent of the Parameter Sweeps.

Thus their research served as a benchmark for I-Vectors and the CEP features as their experiments carried out subject verification using PSD and COH with a MD classifier. Integrating their work to the development of I-Vectors formed the basic feature set and algorithm comparisons that underpinned the Parameter Sweeps, Algorithm Benchmarks, and UBM-TVM Relationship experiments in the ensuing chapters.

Chapter 5

PROTOCOL REPLICATION AND PARAMETER SWEEPS

The purpose of the Experiment 2: Protocol Replication was to touch upon Research Aim 1, competitive I-Vector performance, before expanding upon Experiment 3: Parameter Sweeps. Ideally Research Aim 2, determining the optimal operating conditions for I-Vectors by isolating the effects of the various I-Vector parameters on performance, would have come first but the results of Experiment 1: I-Vectors Development were not sufficient to prove the efficacy of I-Vectors in Chapter 4. Those initial results showed that I-Vector based classification could match the performance of a comparative classification algorithm, namely GMM-UBM, but evaluation against published research was missing. This meant it was necessary to compare I-Vectors and CEP features performance within a protocol similar to that of the intended cohort retrieval process, which lead to the inclusion of La Rocca et al's work [64].

Thus Experiment 2: Protocol Replication was based on reproducing La Rocca's results with the inclusion of I-Vectors, GMM-UBMs and CEP features. La Rocca's systematic search for an optimal channel set for PSD and COH features was readily adapted to include additional algorithms and features. This helped address fundamental components of Research Aims 1 and 2, while drawing from well-understood EEG techniques such as PSD features (commonly used in EEG signal processing [105]) and MD based classifiers [110].

Despite its adaptation to the current work, the original intent of the La Rocca methodology was to find an optimal reduced set of channels for EEG analysis. Channel reduction was (and remains) an active research area for the EEG bio-metric community, as it decreases the computational and data needs of the classification system [105, 42]. In this work, the La Rocca methodology was rigorously adhered to, going so far as to reach out to the original La Rocca experimental team in order to ensure that the Protocol Replication could be directly compared to La Rocca's published results. The Protocol Replication was a recreation of La Rocca's search for the minimum set of channels for each feature set that produced the strongest CRR. The only modification in the current work was to include EER as a performance metric of each algorithm and feature set combination.

Following completion of the Protocol Replication, all of the Experiment Sweeps could be run. The purpose of these experiments was first to sweep epoch durations (to determine the optimal size for working with I-Vectors) and then to sweep the actual I-Vector parameters themselves. Unlike La Rocca's experiments, which only used the *resting* trials from the PhysioNet Database, these experiments used *all* of the PhysioNet Database trials. In addition to sweeping the epoch duration, the UBM mixture size was also varied. These changes in protocol required that instead of independently testing each channel, the channel data was combined under the banner of each test subject.

The I-Vector parameter sweeps evaluated the the size of the UBM, TVM dimension, and LDA dimension using datasets from both the PhysioNet Database and TUH-EEG. This provided broader context through the increased diversity of the data (normal, abnormal, and seizure) and subject counts (over 400 in the seizure dataset). These experiments used the same protocol as the previous sweeps, but their results are separated based upon the parameter (UBM, TVM, LDA and I-Vector dimension).

5.1 La Rocca Based Protocol Experiments

The results of Replication and Sweep Experiments were published in 2019 [193]. They represented the initial link between existing bio-metric verification experiments and the formal introduction of I-Vectors for EEGs.[1] The figures followed a format similar to those in La Rocca's work, focusing on the number of channels used to achieve a resultant level of verification.

5.1.1 Results: Protocol Replication

The results from the Protocol Replication are shown in Figures 5.1–5.4. Specifically, the CRR results for the MD, GMM-UBM, and I-Vector classifiers are shown in Figures 5.1–5.3, respectively, while the EER results are shown in Figure 5.4. The original experiment iteratively tested channels for inclusion in a super set where each additional channel provided the strongest increase in classification performance. This mean the strongest performance improvements were with the first few channels, and the evaluation used mean Correct Recognition Rate (mCRR) which was a product of each channel's individual CRR.

Each of the CRR figures contains two plots. The left plots recreate La Rocca's presentation of the results of match score fusion by using mCRRs. The right plots shows the distribution of the underlying CRRs used to build the mCRRs. These distribution plots highlight the mean (central white dot), standard deviation (colored bar), and range (markers) of CRRs for each algorithm and dataset. For clarity,

[1] Previous research was published in 2016 showing the nascent development of the technology [202].

the GMM-UBM and I-Vector plots all use a UBM of 32 mixtures, since minimal performance gains occurred with larger mixtures.

Figure 5.1 shows the results of the MD classifier. Classification of the COH and PSD features achieved a mCRR of 95% or better by the 6$^{\text{th}}$ channel, whereas the CEP features reached mCRRs of 28.13% for EC and 32.57% for EO.

For the GMM-UBM classifier (Figure 5.2), the COH and PSD features achieved a mCRR of 90% or better by the 6$^{\text{th}}$ channel while the CEP based features only reached mCRRs of 8.10% for EC and 9.33% for EO.

Finally, for the I-Vector classifier (Figure 5.3), COH and PSD based features were correctly classified with mCRR of 90% or better by the 6$^{\text{th}}$ channel, and the CEP based features reached mCRRs of 4.13% for EC and 4.13% for EO. Figure 5.4 displays the mean Equal Error Rate (mEER) results on the EO data set. The three best mEERs all corresponded to PSD features (6.64% PSD-GMM, 6.86% PSD-IVEC, and 8.60% PSD-MHAL). Note that the CEP features all yielded mEERs values over 40%.

5.1.2 Results: Sweep Experiments

Figures 5.5–5.7 summarize the results from Parameter Sweeps. In these plots, the CRR and EER means and standard deviations are presented as a function of epoch size (bottom axis) and mixture size (top axis). Note that Mahalanobis classifiers do not use mixtures, and therefore performance was independent of mixture size.

Figure 5.5 charts the performance of the classifiers when using the first four PhysioNet Database motion trials. This made the dataset 8 times larger than the EO and EC dataset used in the Protocol Replication. The strongest CRR of 91.4% was achieved using 2s epochs with a GMM-UBM of 16 mixtures. For 1s epochs a

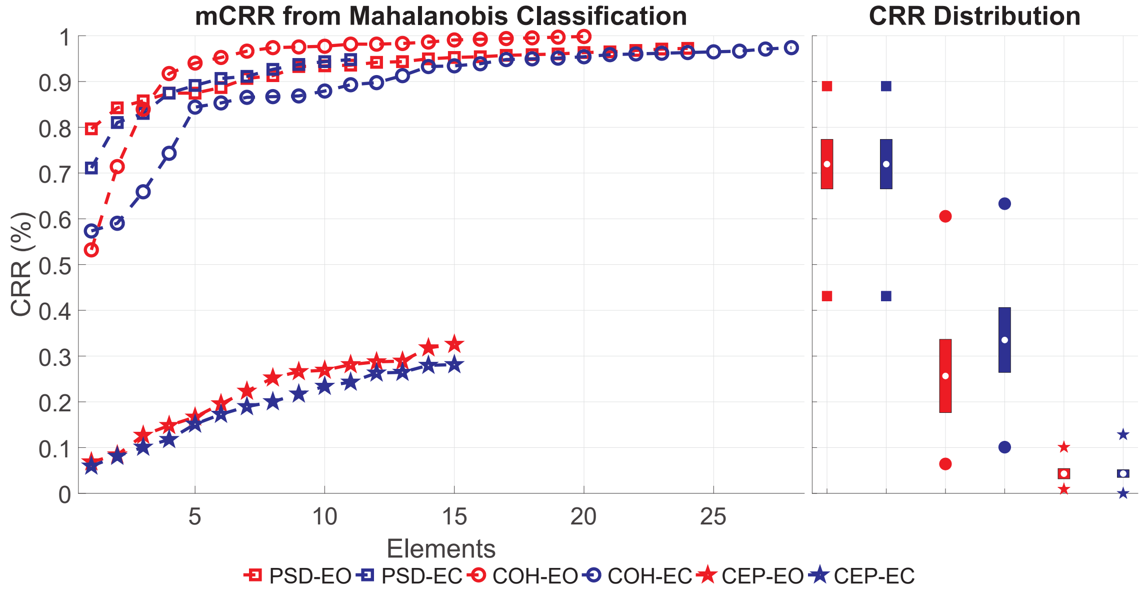


Figure 5.1. Replication MD Performance. The results of the Protocol Replication provide the match-score fusion mean Correct Recognition Rate (mCRR) (y-axis) in the left plot as a function of the number included elements (x-axis) for Mahalanobis classification. Elements in these plots refer to channels for Power Spectral Density (PSD) and Cepstral Coefficient (CEP) and channel pairings for Cepstral Coefficient (CEP) features. The right plot shows the distribution of the 6-fold cross validation Correct Recognition Rates (CRRs) (y-axis) with mean (dot), standard deviation (bar), and full range (marker) for each classifier-feature pairing. Its x-axis represents the Eyes Opened (EO) and Eyes Closed (EC) pairings of each feature (PSD, COH, and CEP).

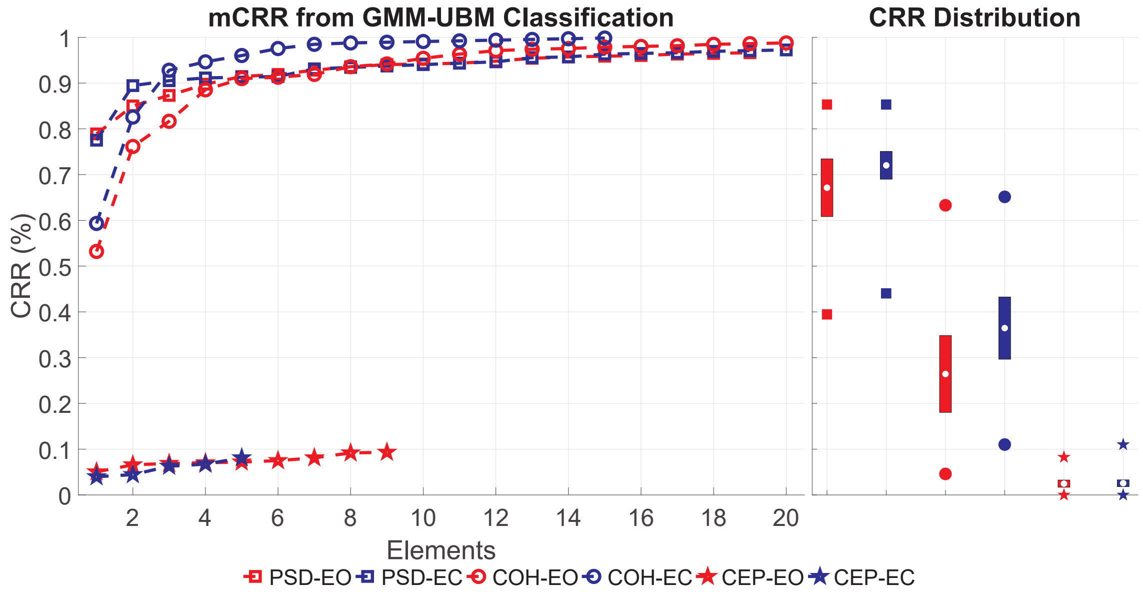


Figure 5.2. Replication GMM-UBM Performance. Gaussian Mixture Model-Universal Background Model (GMM-UBM) classification Correct Recognition Rate (CRR) (y-axis) versus element count and feature set (x-axis).

CRR of 87.46% was achieved using I-Vectors based on a 256 mixture UBM. For 5s epochs a CRR of 89.79% was achieved using GMM-UBM with 64 mixtures.

Figure 5.6 charts the performance of the classifiers when using the first four PhysioNet Database motion trials *and* the EO and EC trials. The resulting dataset was therefore 10 times larger than the individual EO and EC datasets used in the baseline experiments. The strongest CRR of 87.92% was achieved using 1s epochs with I-Vectors based on a 512-mixture UBM. For 2s epochs a CRR of 84.2% was achieved using I-Vectors based on a 512-mixture UBM. For 5s epochs a CRR of 87.05% was achieved using GMM-UBM with 64 mixtures.

Figure 5.7 shows classifier performance for *all* data in the PhysioNet Database (26 times more data than the baseline experiments). The strongest CRR of 90.52% was achieved using 1s epochs with I-Vectors based on a 512-mixture UBM. For 2s

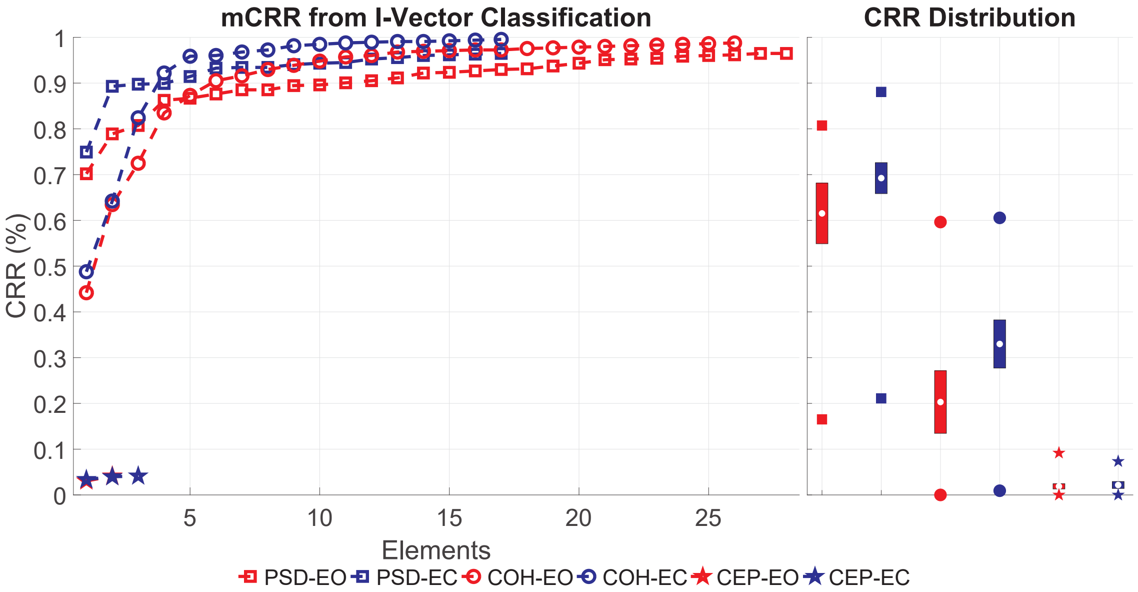


Figure 5.3. Replication I-Vector Performance. Identity Vector (I-Vector) classification Correct Recognition Rate (CRR) (y-axis) versus element count and feature set (x-axis).

epochs a CRR of 80.42% was achieved using I-Vectors with a 512-mixture UBM. For 5s epochs a CRR of 84.05% was achieved using GMM-UBM with 64 mixtures.

5.1.3 Discussion

The goal of these experiments was to demonstrate the ability of I-Vectors in increasing the robustness and flexibility of EEG based bio-metrics/subject verification. Overall the performance of the I-Vector based classifications met or exceed those of GMM-UBM and MD when limited by channels, UBM size, or epoch duration. However, the I-Vector process was limited by the cap of 100 elements for the TVM and LDA dimensions, which was chosen to fit under the 109 subject limit.

Despite using a small and homogenous dataset, the impact of feature and algorithm choice were starkly evident. The epoch and UBM sweep was carried out only on the PSD features because the CEP features performed poorly and the COH

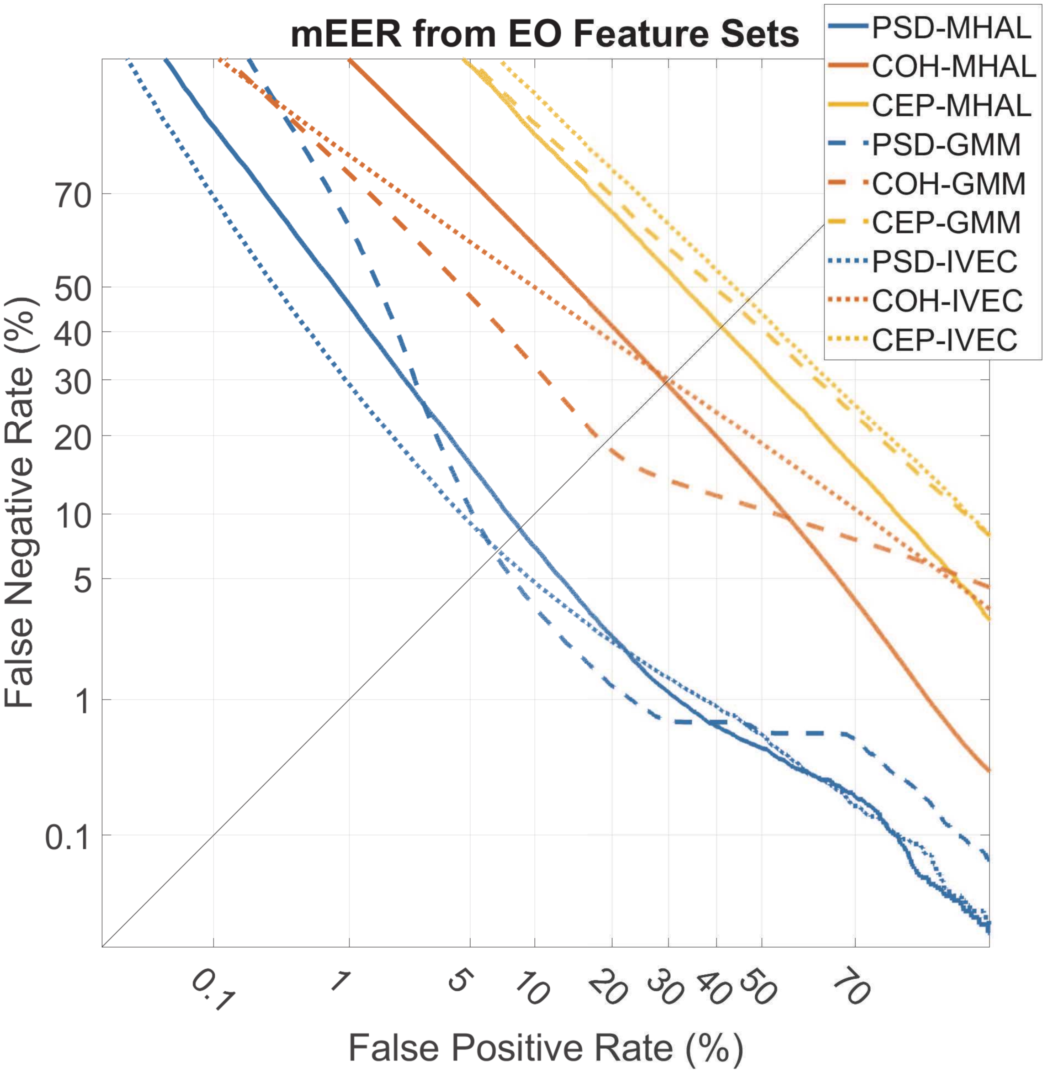


Figure 5.4. ROC for Algorithms. ROC curves based on the EO trial false negatives (y-axis) versus false positives (x-axis). The reported best in class mEERs range from 6.64%, PSD-IVEC, to 46.97%, CEP-IVEC, indicated by the diagonal.

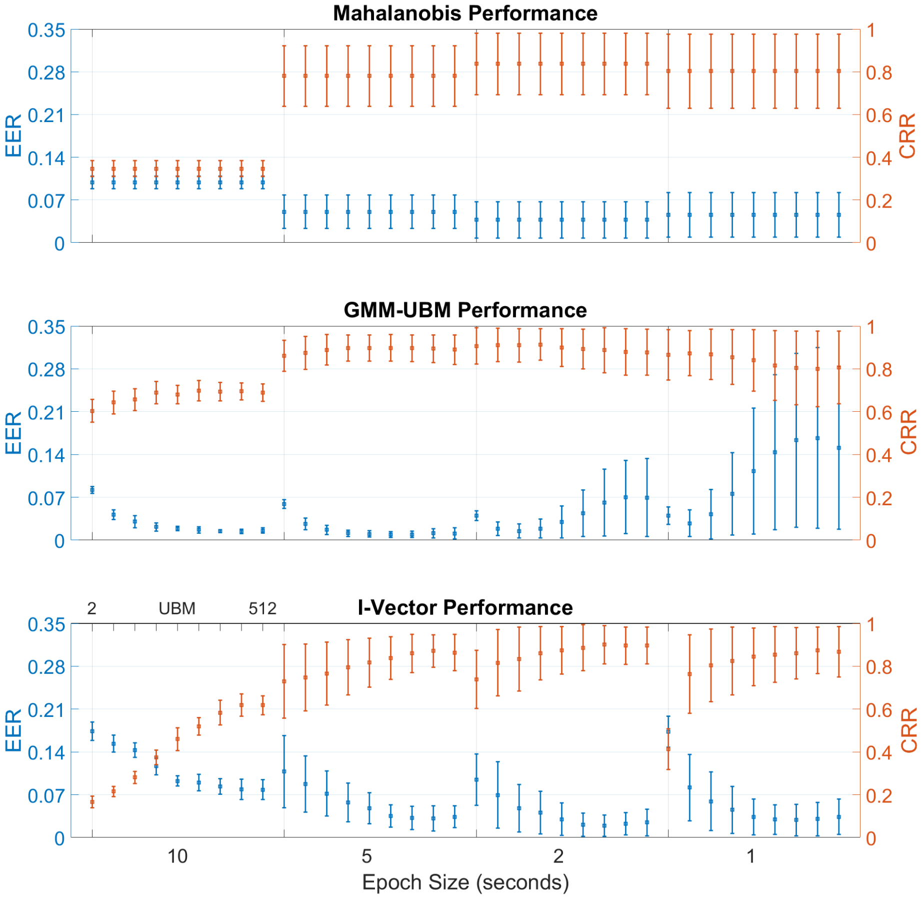


Figure 5.5. Four Trial PhysioNet Database Epoch Sweep. Classification performance on the first four PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) motion trials. The equal error rate (EER) (y-axis left) and Correct Recognition Rate (CRR) (y-axis right) are evaluated as a function of epoch duration (x-axis) for each classifier. The Universial Background Model (UBM) mixtures are shown as individual distributions within each epoch duration.

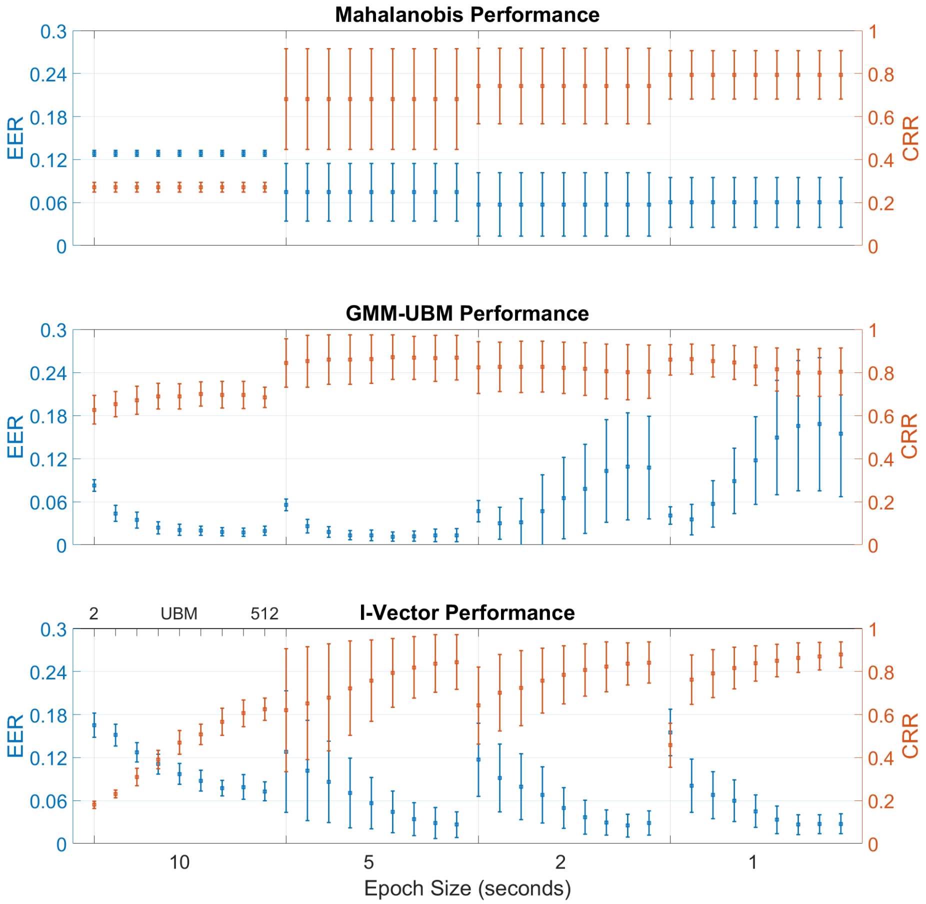


Figure 5.6. Six Trial PhysioNet Database Epoch Sweep. Classification performance on the first four PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) motion trials and the resting trials Eyes Opened (EO) and Eyes Closed (EC). The equal error rate (EER) (y-axis left) and Correct Recognition Rate (CRR) (y-axis right) are evaluated as a function of epoch duration (x-axis) for each classifier.

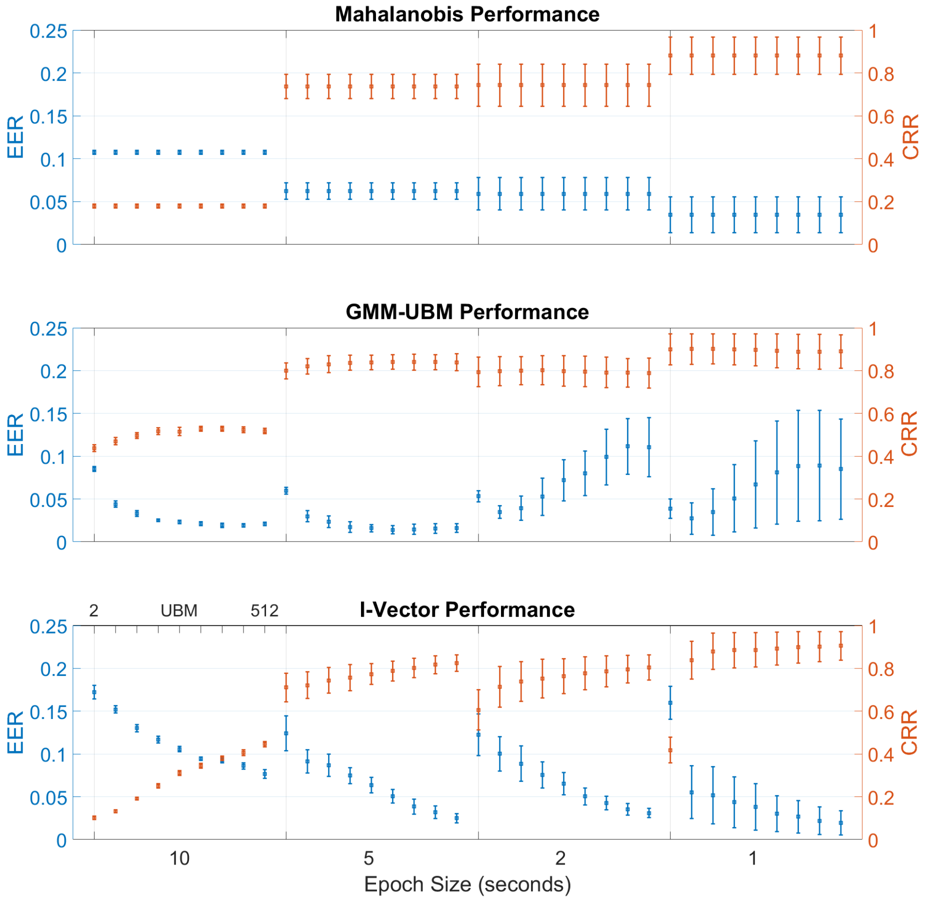


Figure 5.7. Full Trial PhysioNet Database Epoch Sweep. Classification performance on all 14 PhysioNet EEG Motor Movement/Imagery Database (PhysioNet Database) trials, motion and resting. The equal error rate (EER) (y-axis left) and Correct Recognition Rate (CRR) (y-axis right) are evaluated as a function of epoch duration (x-axis) for each classifier.

features were computational exhaustive at 1540 ‘elements’ per epoch. The number of samples available for training the models was the number of channels times the number of trials. Thus the inclusion of all the trial data would have produced 20,560 samples, introducing computational complexity beyond the scope of introducing I-Vectors for use on EEGs.

Conversely, the CEP features may have benefited from the increase in samples; with only 22 channels per trial, the 14 PhysioNet trials per subject would provide 308 samples. The PSD features drawn from La Rocca’s protocol were produced from 56 of the original 64 electrodes in the PhysioNet Database. Therefore the PSD feature set contained 784 samples when using all 14 trials which exceeded all UBMs of 512 mixtures or less. By starting with 4 trials and then expanding to 6 and 14 performance should have provided too few samples for many of the UBMs. This suggested that the results of Figure 5.6 represent over-fitting of the 6 trial PSD dataset because it only provided 336 samples, making results beyond 512 mixtures effectively over-fitted. However, there was no guarantee that even with more samples all of the mixtures would be used.

Overall, the results suggested that I-Vectors can perform on an equivalent level as the other algorithms. Despite unintended setbacks in adequate sampling volumes, feature set nuances, and data source variability, I-Vectors appeared the strongest and produced the most consistent EERs. This performance was not strong enough to suggest I-Vectors should replace other techniques, but that investing time in their development would not be wasted. To that end, development of TVM, LDA, sample pools, and more data variation could be pursued knowing the technique showed viability in EEG applications.

5.1.3.1 Protocol Replication

Replicating the original La Rocca experiment with an additional feature set (CEP) and evaluation criteria (EER and mEER) helps contextualize the use cases of the features and classification techniques. The successful implementation of the La Rocca methodology [64] was established by achieving equivalent match-score fusion based mCRRs for the MD trials (Figure 5.1) . The difficulty in perfectly replicating La Rocca's work could be attributed to using all 109 subjects whereas they used 108, the process of averaging to the mean ear electrodes, or differences when using a 'proper anti-aliasing low-pass filter' to restrict signals to 50Hz.

However, each classifier-feature pair's mEER (Figure 5.4) suggests La Rocca's COH features require a considerable performance trade-off to achieve a 5% improvement in mCRR. The PSD features (blue lines) produced the best EERs for each algorithm. The next strongest was when the algorithms were paired with the COH features (orange lines), but incurred at least a 10% increase in EER. In fact, for the MD algorithm the increase in EEG was over 20%.

The range of CRRs (the secondary plots of Figures 5.1–5.3) confirms La Rocca's finding that aggregating channels is a simple but effective method for improving performance relative to the mCRRs. Figures 5.1–5.3 indicate that, regardless of trial, it takes at least MD 13 elements, GMM-UBM 6 elements, and I-Vectors 11 elements to reach 90% mCRR for COH and PSD feature sets. Despite these successes, the best performance of the CEP features barely reaches 30% mCRR with the MD classifier and performs even worse for the other classifiers. However, within these benchmarks there are trade-offs between mCRR and mEER for the COH and PSD features based on the classifier.

The GMM-UBM classifier provides equivalent mCRR performance to the MD classifier while improving the mEER of PSD by 1.96% and of COH by 10.70%. The I-Vector classifier provides similar performance to the other two classifiers, but fails to significantly improve either mEERs, with PSD decreasing by 1.74% and COH increasing by 0.57%. This may explain why MD is not frequently used for EEG classification despite its acceptable classification performance [91, 98]. It also suggested, in this instance, that I-Vectors could an acceptable alternative to both classifiers given their mCRR on the dominant feature set, PSD.

Feature set performance was found to be clustered according to its mEER. Notably, the mEERs of PSD are 15% stronger than those of the other features (Figure 5.4), independent of classifier. This suggests that feature selection is the most critical component in deciding EER. Conversely the mCRRs present with no such trend, other than the overall poor performance of the CEP features and that additional channels improves mCRR. Despite this, classifier choice does appear to marginally improve the mEERs on these small datasets, a finding which was not addressed in La Rocca's original work.

5.1.3.2 Parameter Sweeps

Expanding the original experiments by incorporating smaller epochs and larger datasets tested the modeling process of each algorithm. The inclusion of multiple distinct motion trials as well as the resting trials increased the diversity of the testing datasets, while producing smaller epochs tested the modeling of each classifier. Typically EEG biometric experiments were more tightly regulated to simplify the characteristics of the data, such as La Rocca using only the resting trials or other groups using only the tasks and not the resting phases between them [45, 66]. The speech community has addressed similar issues through tools like

I-Vectors that address variations in the recording channel (landline, mobile phone, microphone, and other speech modifiers) and residual noise to identify the speaker [166, 203].

The impact increasing the characteristics of the dataset was immediately apparent in the performance decrease seen when adding the resting trials, Figure 5.6, to the first four motion trials, Figure 5.5. The addition of the resting trials caused a decrease in CRR compared to the pure motion trials for all algorithms. Additionally, the EERs of the MD and GMM-UBM classifiers were worse for the larger dataset. Both of these occurrences were likely attributed to the presence of rest periods in the motion trials being classified as resting trials. Furthermore, a false positive being more likely than a true positive drove up the EER. This was more likely with the expanded datasets because classification occurred across multiple recordings from the same target subject, but the within trial subject remained the only correct answer.

Across the three datasets, Figures 5.5–5.7 show that the 10s epoch performance decreases for all classifiers as the amount of data increases. However, smaller epochs provide varying levels of improvement to the MD and I-Vector classifiers. As the epoch duration shrinks, more epochs become available in the enrollment data to build the subject-trial models which was the likely driver of the improving performance. This was an interesting result because it does not fully align with the view that longer sessions produce more stable subject verification results, as presented by Maiorana et al [98].

Their work, however, limited feature generation to 5s epochs with 40% overlap, which were evaluated after epoch-based fusion scoring algorithms, similar to the match-score fusion, for a given recording duration (10s to 90s). Thus they did not produce a 90s epoch, but rather evaluated a series of 5s epochs drawn from fewer subjects (50) using only within-trial data (EO or EC). Their work clearly validates the longitudinal stability of the approach, with days and weeks between recording sessions, but that is beyond the scope of introducing I-Vectors to the EEG community. In contrast, the scope of this work was to introduce and establish I-Vectors as a tool that could remove the requirements of specifically matched datasets, arbitrary epoch lengths between experiments, and channel/epoch based fusion schemes from EEG classification.

Toward those goals, the I-Vector CRR and EER improved with each subsequent mixture across all epoch sizes. This was most evident in Figures 5.7 and 5.8, where I-Vector CRR and EER (80.42%, 3.11% and 90.52%, 1.96%) exceed those of the GMM-UBM (80.25%, 5.26% and 90.10%, 3.48%) for 2s and 1s epochs, respectively. The performance margin is smaller in Figures 5.6 and 5.9, but I-Vectors (84.20%, 2.854% and 87.92%, 2.75%) are again superior to GMM-UBMs (82.7%, 4.67% and 86.24%, 3.52%) for the two smallest epochs. It is only when the resting EO and EC datasets are removed that GMM-UBM outperforms the I-Vectors for the 2s epochs.

The seemingly varied GMM-UBM EERs for 2s and 1s epochs probably occurs because at most there are 512 mixtures of 40 features working to account for numerous subject-trial variations. Adjusting the MAP relevance factor, r, could help correct this behavior by relying more on the enrollment data. The smaller epochs make for deeper training and enrollment datasets, so it is likely the UBMs are producing articulated mixtures that the MAP adaptation is unable to generate sufficiently diverse models because the relevance factor α_c is too small. It is difficult to optimize the relevance factor because it applies to the weight of each mixture, in the UBM, produced during BW estimation, where $\boldsymbol{N}_c$ comes from Equation 2.3-8.

$$\alpha_c = \frac{\boldsymbol{N}_c}{\boldsymbol{N}_c + r} \tag{5.1-1}$$

Furthermore the performance discrepancy of the 2s and 1s epochs could have been attributed to the epoch duration being less than the duration of the tasks in the motion trials (4s). This meant that epochs could straddle rest and motion tasks. As all three classifiers used data pooled across the channels, they could encounter difficulty if an epoch split a tasks. The GMM-UBM and I-Vector classifiers were built on the sensitivities of the individual variances by adapting the underlying UBMs, which meant they may under-perform if the UBMs fail to capture all modes of the data. The MD combined the variances making it sensitive to the combined variance of the model.

It could be that the combined weighting of the MD prioritizes different facets of the features than those of the UBMs, giving rise to different sets of outliers between the classifiers. Given that outliers are likely to appear when epochs overlap tasks in the recordings, 5s and 2s results could be a more unbiased measure of the classifier's strength. This may explain why there was minimal improvement, or often a decrease, in classification performance between between 5s and 2s epochs, as the means are not changing, only the variances. This would be problematic if the testing epochs were drawn all drawn either rest or motion making them devoid of half the information the models had learned. Even so, shorter epochs cannot be disregarded as the root cause because, for event classification tasks, short epochs have proven successful [102, 204].

There are no such trends with the GMM-UBM classifier, as its EER grows for each mixture size in the 2s and 1s epoch sets, which is the opposite of the 10s and 5s epochs. Its CRRs shows minimal improvement, and often even a regression, at these larger mixture sizes. This suggests the subject-trial models are generating scores from non-matching subjects that exceed those found when matching subjects. The models may be unable to overcome these edge cases despite increasing the mixture sizes, suggesting the problem is inherent to the data itself and not to the modeling

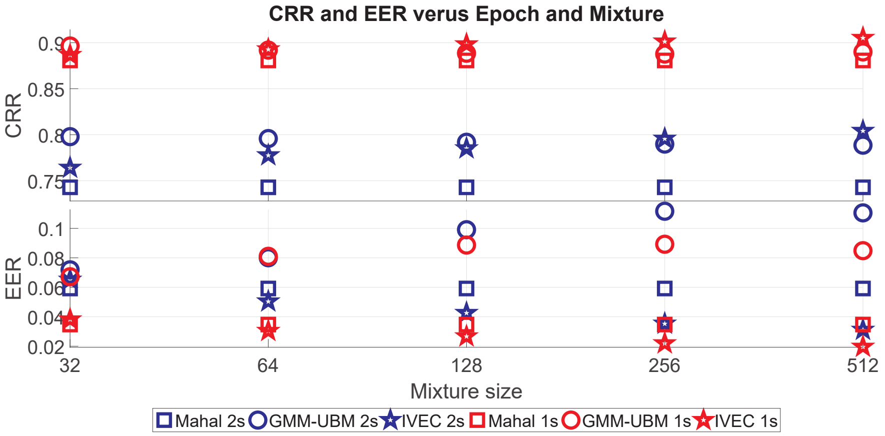


Figure 5.8. Full Trial PhysioNet Database Epoch Sweep, Expanded Mixtures. Mean classifier performance, Correct Recognition Rate (CRR) and equal error rate (EER), as a function of epoch duration and mixture size limited to the five largest mixtures using all 14 trials. The color of lines represent 2s epochs (blue) and 1s epochs (red). The shape of markers represents the algorithms with Mahalanobis (square), Gaussian Mixture Model-Universal Background Model (GMM-UBM) (circle), Identity Vector (I-Vector) (star).

process. While none of the algorithms avoid performance problems over all epochs and datasets, only the I-Vectors show a continued progression towards improved CRR and EER for all datasets, epochs, and mixture sizes. I-Vectors would therefore be best suited for generalizing across additional data, similar to its adaption from speech to EEGs.

5.1.4 Constraints

La Rocca's findings showed COH features could improve subject verification via individual channels and through their match-score fusion when compared to PSD features. Reducing the required amount of data to perform subject verification is an important goal for developing biometric-based EEG applications [68, 104, 105], which is often achieved by incorporating novel features, new algorithms, or unique dataset configurations. La Rocca focused on novel features by replacing PSD

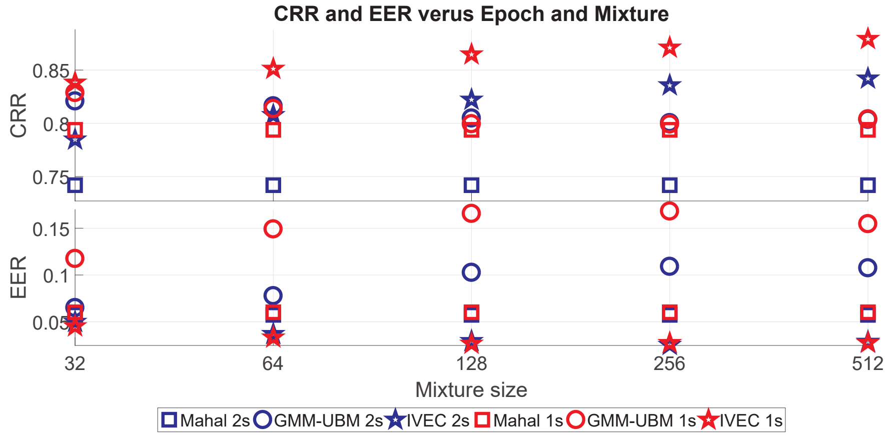


Figure 5.9. Six Trial PhysioNet Database Epoch Sweep, Expanded Mixtures. Mean classifier performance, Correct Recognition Rate (CRR) and equal error rate (EER), as a function of epoch duration and mixture size limited to the five largest mixtures using the 4 motion trials and 2 resting trials. The color of lines represent 2s epochs (blue) and 1s epochs (red). The shape of markers represents the algorithms with Mahalanobis (square), Gaussian Mixture Model-Universal Background Model (GMM-UBM) (circle), Identity Vector (I-Vector) (star).

features for COH features given the influence of each channel's "dynamic relationship with other regions" of the brain [64].

Using the substantially larger channel set improved match-score fusion classification performance to 100%, but only because channel reduction was exchanged for an increase channel search space. Both feature sets were constructed from the same raw data and classified using the same Mahalanobis distance classifier. This suggests that the features were the driving factor in classification improvement. However, by replicating the original work in the baseline experiments, GMM-UBMs and I-Vectors provided equal if not better performance compared to Mahalanobis classifiers where COH features were outperformed by PSD features. As the epoch sizes became smaller and the datasets increased in the Baseline Experiments, only I-Vectors maintained their performance. This suggests the epoch duration and dataset size were the strongest performance factors for PSD features.

The sweep experiments highlighted the impact of epoch duration and dataset size/composition in classification performance. Needing to control for dataset composition, dataset size, and epoch duration increases the difficulty of finding an optimal feature set and classification algorithm pairing [52, 103]. La Rocca's work mitigated this with a static epoch duration, brain regions, and controlled datasets because they understood their dataset and classification goal. Other research agreed that performance could be improved and maintained over time through external knowledge and adequate data preparation [66, 97, 98]. Despite success with such measures, these steps were often laborious and some researchers chose not to address the issues much like the goal of this work in developing a technique agnostic to such concerns citeYang2016,Marcel2007a.

Conditions mitigated by data pre-processing are not always present, which makes understanding the impact of epoch duration and dataset composition critical for experimental success. The presence of *a priori* knowledge is only useful if it is properly understood, otherwise it creates an unequal testing ground that impedes the development of robust features and algorithms. When given an environment lacking in this knowledge, I-Vector classification is able to mitigate these shortfalls. This reduces the impact of epoch duration, dataset composition, and algorithm performance so that the focus can be on feature selection and the level of discrimination (subject, trial, channel, etc) leading to fundamental knowledge gains about EEG behavior [200].

The results of Baseline Experiments showed that I-Vector EERs decreased and CRRs increased in terms of both mean and variance with each subsequent mixture size. Although the experiment stopped with a UBM of 512 mixtures, the speech community has tested mixtures of size 2048 [171] and 4096 [174]. It is possible that larger mixture sizes could produce stronger classification results for Baseline and

Sweep Experiments. This would bringing their performance into line with the best case results of La Rocca's original experiment. Therefore future experiments should increase the range of UBM mixture sizes and ensure enough samples are present in the enrollment and testing datasets to support such modeling. This would increase the computational needs of the experiments, but could also improve performance.

5.2 Parameter Sweeps

The results of the La Rocca-inspired Protocol Replication and Parameter Sweeps indicated a new protocol was necessary for characterizing the performance of I-Vectors. Testing each parameter in an isolated fashion was not tenable given their connected nature. The number of UBM mixtures could vary, but exceeding the available samples for enrollment and testing sets did not produce reliable results. The number of epochs available for a given dataset was dependent on the channels, trials, and epoch duration. The number of available epochs and trials was dependent on the dataset. Lastly, the performance of each feature set was dependent on all of these conditions as well as the algorithm it was paired with for the given experiment.

To gain any understanding of one parameter with respect to the others, static configurations were organized around epoch duration and number of samples for all experiments. This enabled sweeps of UBM mixture, TVM dimension, and LDA dimension. There was no way to control for algorithm or feature set influence, as no benchmarks existed for the proposed experiments. This required each variation of dataset and algorithm to be tested for each parameter configuration. A few considerations were made to make the experiments tractable.

First, a “slim” version of each feature set was developed based upon the most common electrode configuration found in the TUH-EEG corpus, recall Figure Figure 2.2. This linked all feature sets to the same number of recording electrodes, equalizing the disparity in samples encountered previously. This made the CEP and COH feature sets contain 22 matching channels, and the PSD features were limited to 19 original electrodes. This sped up the PSD and COH based experimental computations and better aligned the dimensions of each feature set which helped mitigate performance variation based upon the feature sets.

Secondly, bounds had to be set on the ranges of the TVM and LDA dimensions. While the step size and range of the UBMs were organized in base 2 increments, the TVM and LDA dimensions only had a ceiling defined by the number of subjects. The software was built to ensure the LDA dimension would never exceed $subjects - 1$, but the TVM was allowed to be larger than the number of subjects as LDA would reduce the length of the I-Vector.

Table 5.1. Identity Vector Parameter Sweep

UBM	TVM	LDA
2 4 8 16 32 64 128 256 512 1024 2048	25	15
	50	45 30 15
	75	60 45 30 15
	100	75 60 45 30 15
	200	100 75 60 45 30 15

Thirdly, the number of samples was indirectly tested during the epoch sweep experiments, but was directly tested during these experiments. Sample depth for the test datasets was controlled for with experiments focusing on 1, 4, 6, and 14 epochs with a duration of 10 and 2 seconds. This was done because 10s represented the original case tested in La Rocca's work, while 2s appeared the strongest for both I-Vectors and GMM-UBMs. In addition, 5s duration epochs only provided 20 epochs for PhysioNet Database's 1 minute recordings making the testing dataset the majority of the data (12 of 20 epochs) which was not ideal.

The results of these experiments are presented using a novel metric formulated by subtracting the EER from the CRR. Recall that the CRR represented the subject verification performance of a correct matching meaning there can be no false positive. This prevented the use of an F1 score as a comprehensive metric. Thus the so-called "C metric", representing the combination of CRR and EER, is reported; a C score of 0.75 represents with a minimum acceptable threshold. A system with a CRR of 80% and an EER of 5% or one with a CRR of 90% and an EER of 15% would be considered acceptable given the early stages of this research.

5.2.1 Results: I-Vector Parameters

The first sweep experiment, see Figures 5.10–5.12, was done to show the performance when using a single epoch in the testing dataset. This was sourced from the TUH-EEG datasets of abnormal, normal, and seizure over the three feature sets. Performance was poor across all datasets and feature sets aside from the TVM built from CEP based normal data (see Figure 5.10). The use of PSD and COH features provided similar results which is as expected given that the COH features are built off of the PSD features.

Those initial results served as a benchmark allowing comparisons as the number of samples in the training dataset was increased. Performing the same analysis, but focusing only on the normal dataset, the impact of increasing the samples is shown in Figures 5.13–5.15. Here the impact of altering epoch duration and number was

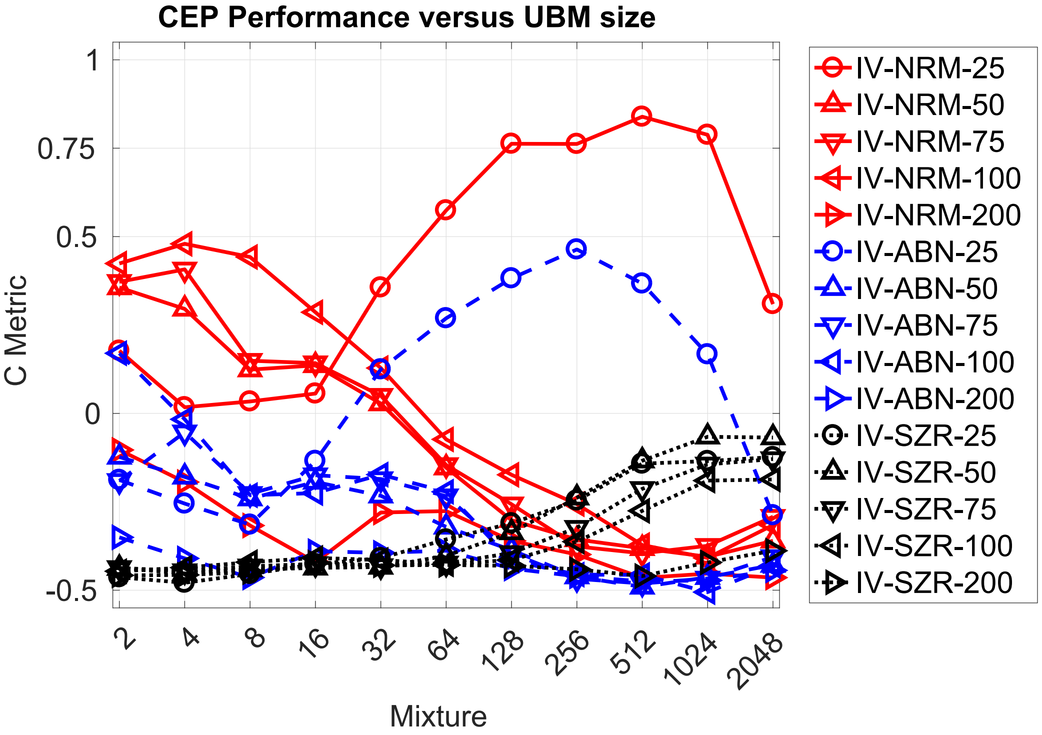


Figure 5.10. I-Vector with CEP TUH-EEG, Single Test Epoch. This C Metric plot shows the CEP based TUH-EEG abnormal, normal, and seizure data performance as a function of UBM mixture and TVM dimension when using one epoch in the testing data. The peak performances achieved were 0.84 with the normal (NRM) dataset using a 512 mixture UBM with a TVM dimension of 25, 0.46 with the abnormal (ABN) dataset using a 256 mixture UBM with a TVM dimension of 25, and -0.07 with the seizure (SZR) dataset using a 1024 mixture UBM with a TVM dimension of 50.

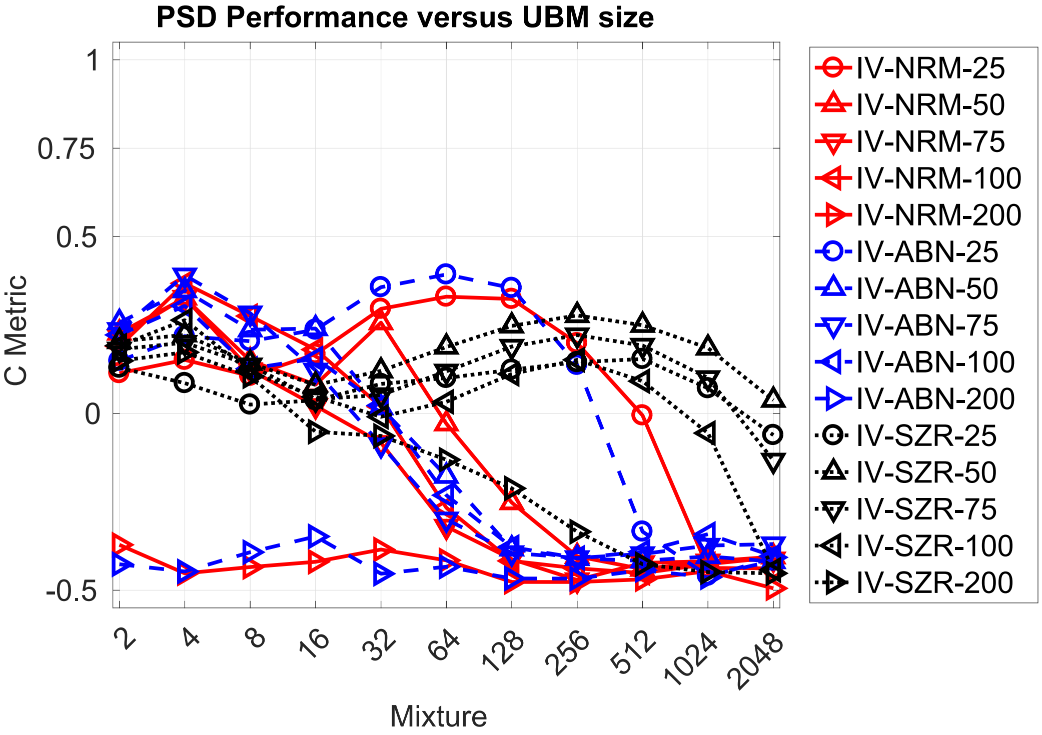


Figure 5.11. I-Vector with PSD TUH-EEG, Single Test Epoch. This C Metric plot shows the PSD based TUH-EEG abnormal, normal, and seizure data performance as a function of UBM mixture and TVM dimension when using one epoch in the testing data. The peak performances achieved were 0.368 with the normal (NRM) dataset using a 4 mixture UBM with a TVM dimension of 100, 0.39 with the abnormal (ABN) dataset using a 64 mixture UBM with a TVM dimension of 25, and 0.28 with the seizure (SZR) dataset using a 256 mixture UBM with a TVM dimension of 50.

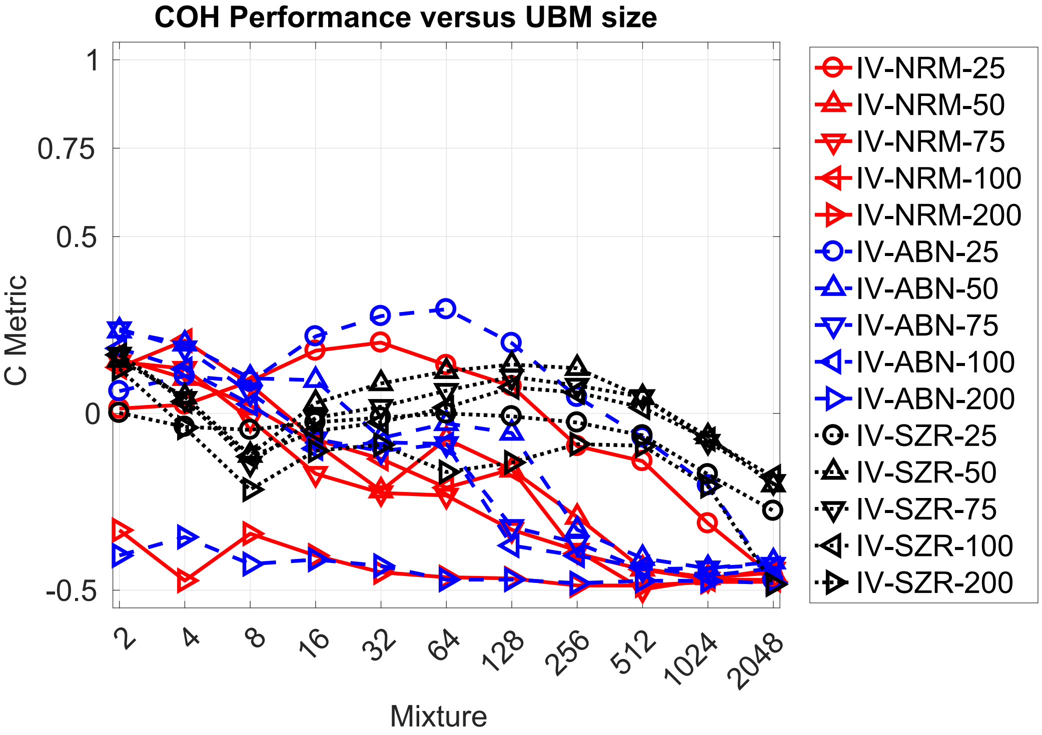


Figure 5.12. I-Vector with COH TUH-EEG, Single Test Epoch. This C Metric plot shows the COH based TUH-EEG abnormal, normal, and seizure data performance as a function of UBM mixture and TVM dimension when using one epoch in the testing data. The peak performances achieved were 0.2059 with the normal (NRM) dataset using a 4 mixture UBM with a TVM dimension of 100, 0.30 with the abnormal (ABN) dataset using a 64 mixture UBM with a TVM dimension of 25, and 0.17 with the seizure (SZR) dataset using a 2 mixture UBM with a TVM dimension of 75.

detailed as sets of 2 epochs 10s in duration, solid red lines, and 4 epochs 10s in duration, blue dashed lines, and 4 epochs 2s in duration, black dotted lines, epochs.

Doubling the number of epochs from one to two had a marginal impact across all feature sets. However, increasing to four boosted the performance of CEP and PSD features for TVMs using 32 mixtures or more. The performance of the CEP features at the lowest TVM size provided near perfect performance for 4 epochs of 2s duration. This was followed closely by 4 epochs of 10s duration.

When the dataset shifted to using 2s epochs, a spike in performance was found for the COH features at 128 mixtures for the TVM of size 50. A similar event was seen in the PSD features for the TVM of size 50 between 32 and 256 mixtures. This did not make the overall performance acceptable, but was significant given the preceding mixtures showed declining performance.

The COH features failed to produce a metric score over 0.5, despite their underlying PSD features reaching this threshold across the sample and epoch permutations. Instead the overall trend was declining performance as the UBM mixture size increased.

Given the trend of normal TUH-EEG performance increasing as sample size increased, sample sets of 6 and 14 were added. This was done to mirror the nature of the PhysioNet Database based experiments where 4 motion trials, 4 motion trials and the two resting trials, and all 14 trials were used in Section 5.1.3.2. Here all the sample sets use the 2s duration epochs to produce Figures 5.16–5.18. Using 14 samples matched the natural number of trials in the PhysioNet Database data, providing a normalized comparison point.

With the inclusion of more samples in the training dataset, the CEP results continued to show strong performance for TVMs of size 25 for a range of mixture sizes, but also a performance spike at 128 mixtures when using a TVM of size 50.

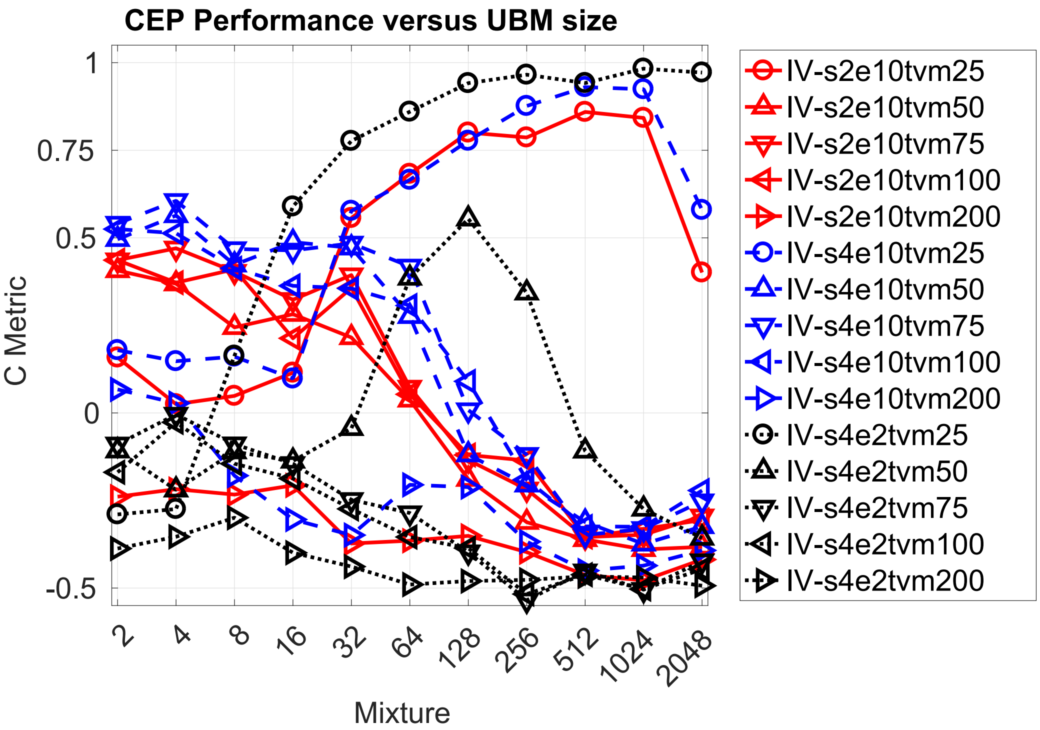


Figure 5.13. I-Vector with CEP `Nrm`, Multiple Test Epochs. This C Metric plot shows the CEP based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 2 testing epochs of 10s duration, 4 testing epochs of 10s duration, and 4 testing epochs of 2s duration. The peak performances achieved were 0.86 with the 2 epochs 10s in duration configuration using a 512 mixture UBM with a TVM dimension of 25, 0.93 with the 4 epochs 10s in duration configuration using a 512 mixture UBM with a TVM dimension of 25, and 0.98 with the 4 epoch 2s in duration configuration using a 1024 mixture UBM with a TVM dimension of 25.

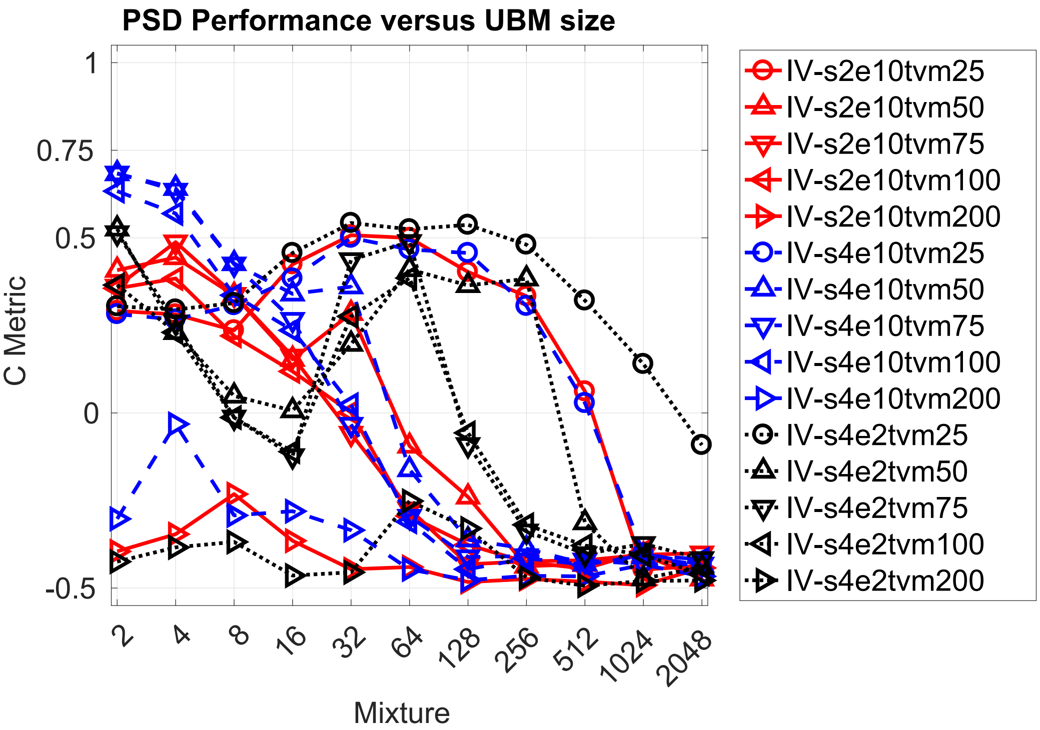


Figure 5.14. I-Vector with PSD `Nrm`, Multiple Test Epochs. This C Metric plot shows the PSD based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 2 testing epochs of 10s duration, 4 testing epochs of 10s duration, and 4 testing epochs of 2s duration. The peak performances achieved were 0.51 with the 2 epochs 10s in duration configuration using a 32 mixture UBM with a TVM dimension of 25, 0.68 with the 4 epochs 10s in duration configuration using a 2 mixture UBM with a TVM dimension of 75, and 0.54 with the 4 epochs of 2s duration configuration using a 32 mixture UBM with a TVM dimension of 25.

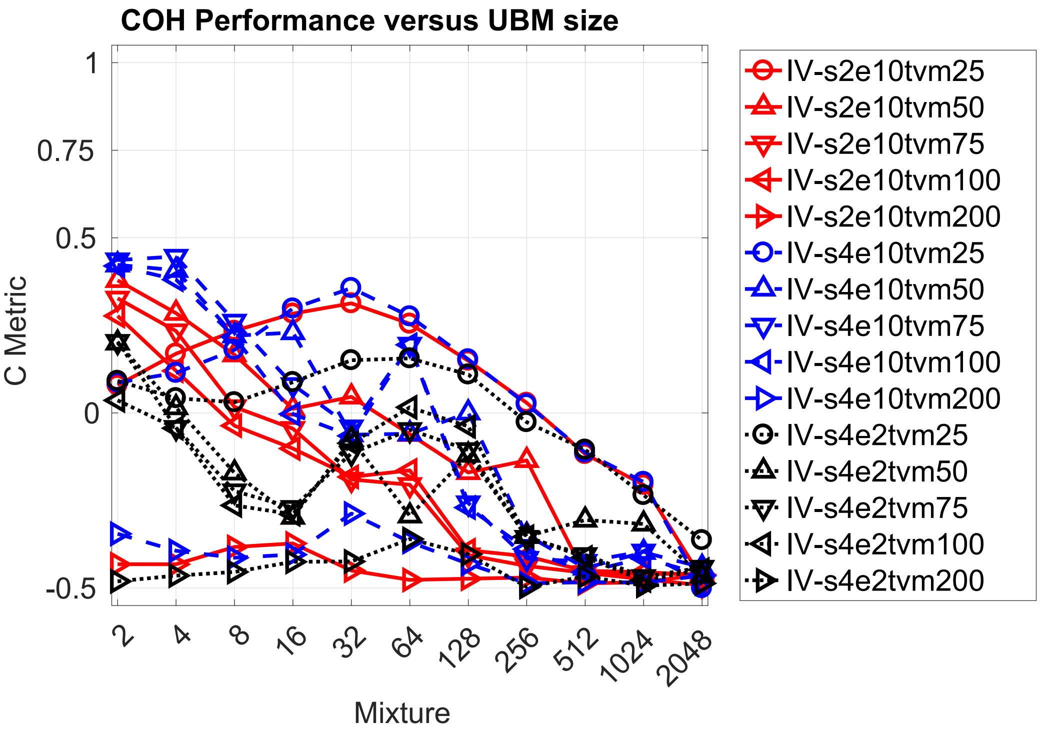


Figure 5.15. I-Vector with COH `Nrm`, Multiple Test Epochs. This C Metric plot shows the COH based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 2 testing epochs of 10s duration, 4 testing epochs of 10s duration, and 4 testing epochs of 2s duration. The peak performances achieved were 0.38 with the 2 epochs 10s in duration configuration using a 2 mixture UBM with a TVM dimension of 50, 0.45 with the 4 epochs 10s in duration configuration using a 4 mixture UBM with a TVM dimension of 75, and 0.20 with the 4 epochs of 2s duration configuration using a 2 mixture UBM with a TVM dimension of 75.

However, the increased samples are lowering the performance of smaller sized UBMs.

The trend of performance declining for smaller mixture UBMs was seen for the PSD and COH feature sets as well. In the same manner, the performance spike around 128 mixtures was found across all mixtures. Regardless of feature set, the strongest performances came from the 14-sample sets and the weakest from the 4-sample sets.

The results of the PhysioNet Database experiments, Figures 5.19–5.21, drew additional samples from different trials. Thus each increase in samples represented an increase in trials with the 4-sample sets being only motion trial data, the 6-sample sets being motion trial data and calibration trial data, and the 14-sample sets containing all trial data. By using 10s duration epochs, these results provided a baseline for subsequent epoch duration variations and a comparison point for the initial TUH-EEG results of Figures 5.13–5.15. Within these results, distinct trends were seen for each feature set.

When using CEP features (see Figure 5.19) performance improved as UBM mixture size increased for all but the largest sized TVM. For each combination of UBM and TVM size the best reported result came from the dataset with the larger number of samples, with the 14-sample set outperforming the 6-sample set which in turn surpassed the 4-sample set.

The PSD results, Figure 5.20, showed a slight performance increase with UBM mixture size for the smaller sized TVMs. However, there was a strong performance roll-off of the larger TVMs at larger UBM mixture sizes. This was more pronounced for the smaller sample sets of 4 and 6. The majority of this behavior occurred at UBM mixture sizes of 16 and 32.

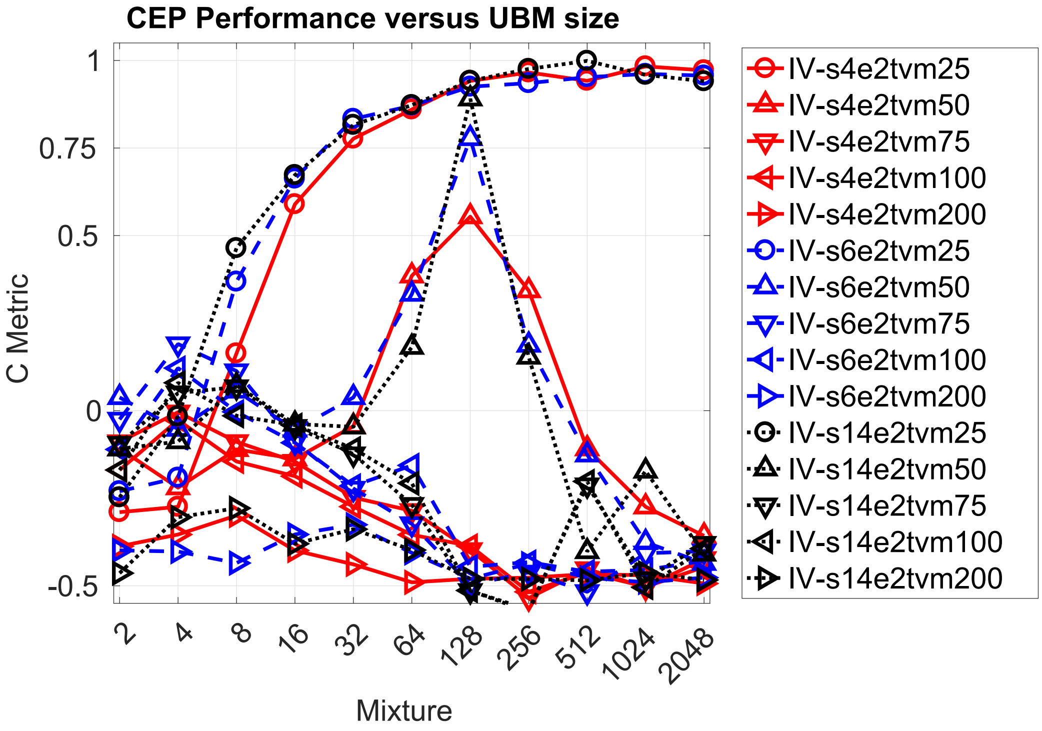


Figure 5.16. I-Vector with CEP `Nrm`, Multiple Test Epochs. This C Metric plot shows the CEP based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.98 with the 4 epochs of 2s duration configuration using a 1024 mixture UBM with a TVM dimension of 25, 0.9615 with the 6 epochs of 2s duration configuration using a 1024 mixture UBM with a TVM dimension of 25, and 0.99 with the 14 epochs of 2s duration configuration using a 512 mixture UBM with a TVM dimension of 25.

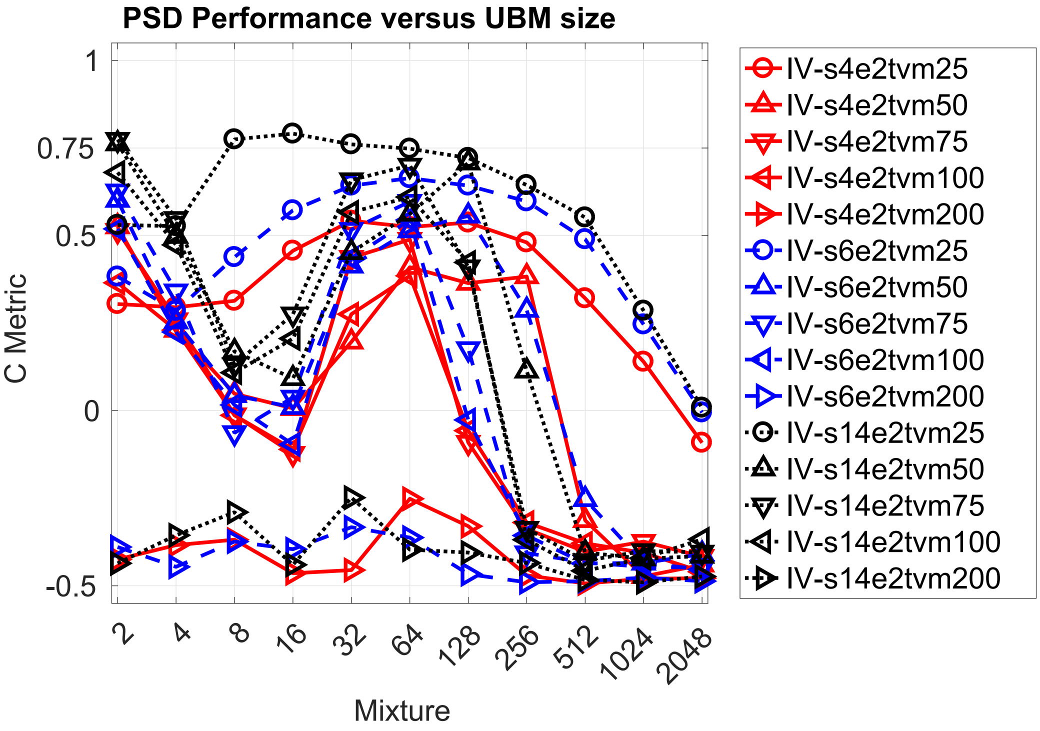


Figure 5.17. I-Vector with PSD `Nrm`, Multiple Test Epochs. This C Metric plot shows the PSD based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.54 with the 4 epochs of 2s duration configuration using a 32 mixture UBM with a TVM dimension of 25, 0.66 with the 6 epochs of 2s duration configuration using a 64 mixture UBM with a TVM dimension of 25, and 0.79 with the 14 epochs of 2s duration configuration using a 16 mixture UBM with a TVM dimension of 25.

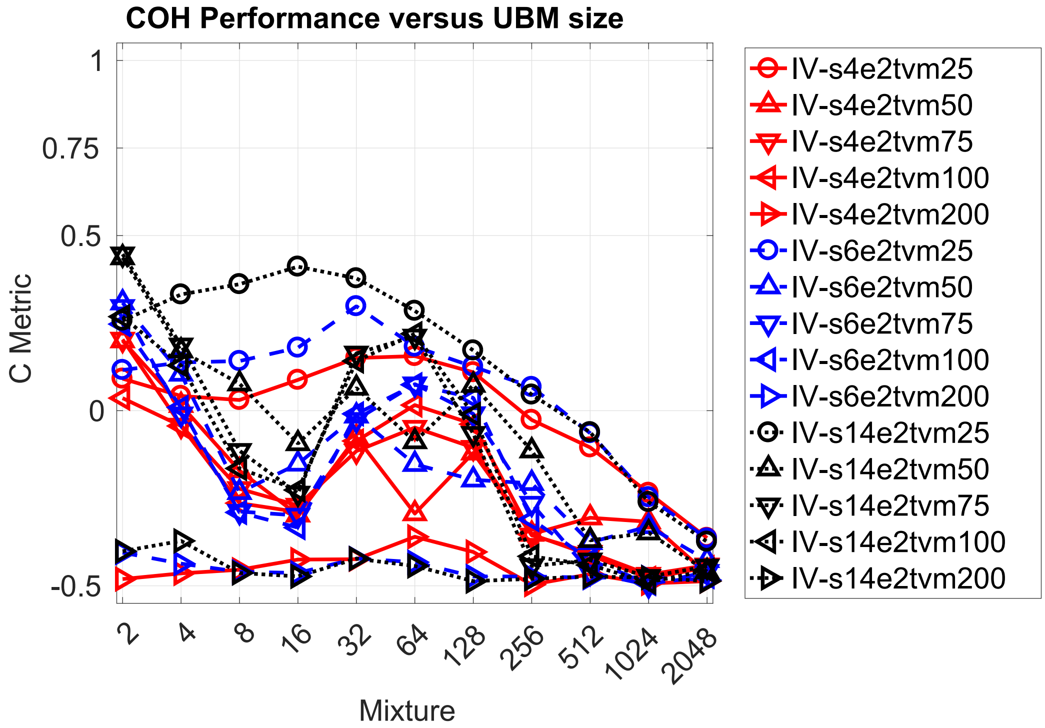


Figure 5.18. I-Vector with COH `Nrm`, Multiple Test Epochs. This C Metric plot shows the COH based TUH-EEG normal dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.20 with the 4 epochs of 2s duration configuration using a 2 mixture UBM with a TVM dimension of 75, 0.31 with the 6 epochs of 2s duration configuration using a 2 mixture UBM with a TVM dimension of 50, and 0.45 with the 14 epochs of 2s duration configuration using a 2 mixture UBM with a TVM dimension of 75.

The COH results showed a similar overall strong performance at lower UBM mixture sizes as the PSD features. However, the only performance roll-off was attributed to TVMs of size 200 until the UBM reach 2048 mixtures.

The results of TUH-EEG seizure data experiments, Figures 5.22–5.24, tracked the growth of the testing data from 4 epochs to 6 epochs where the epoch duration was reduced from 10s to 2s. This overlapped with previously tested configurations for the PhysioNet Database data and the TUH-EEG normal dataset. While the TUH-EEG normal dataset contained 50 subjects, and the PhysioNet Database dataset contained 109 subjects, the TUH-EEG seizure dataset contained 411 subjects. This mitigated dimensional constraints previously imposed on the TVM of dimension 200.

The use of CEP features, Figure 5.22, for any combination of TVM, sample size or epoch duration failed to reach a C score of 0.5. Increasing the size of the UBM improved performance until 1024 mixtures after which it appeared to roll-off or plateau at 2048 mixtures. The smaller sized TVMs performed best with the 6-sample sized experiments as the peak performer. Conversely, the PSD and COH feature sets performed far better even at smaller UBM sizes.

The PSD features, Figure 5.23, were able to produce scores between 0 and 0.5 for the majority of the UBM mixture sizes. Again, the best performances were seen from the 6 epoch 2s duration epoch dataset. However, moving between epoch durations for the 4 epoch sets showed the 10s epochs with better performance at smaller UBM mixture sizes and 2s epochs with better performance at higher UBM mixture sizes.

The COH features, Figure 5.24, produced scores below their PSD counterparts, but higher than those of the CEP features. Using 10s epochs and 4 epoch sets outperformed the other approaches for the majority of the UBM mixtures. The notable exceptions was at a mixture size of 8, where only the TVM of size 25 did

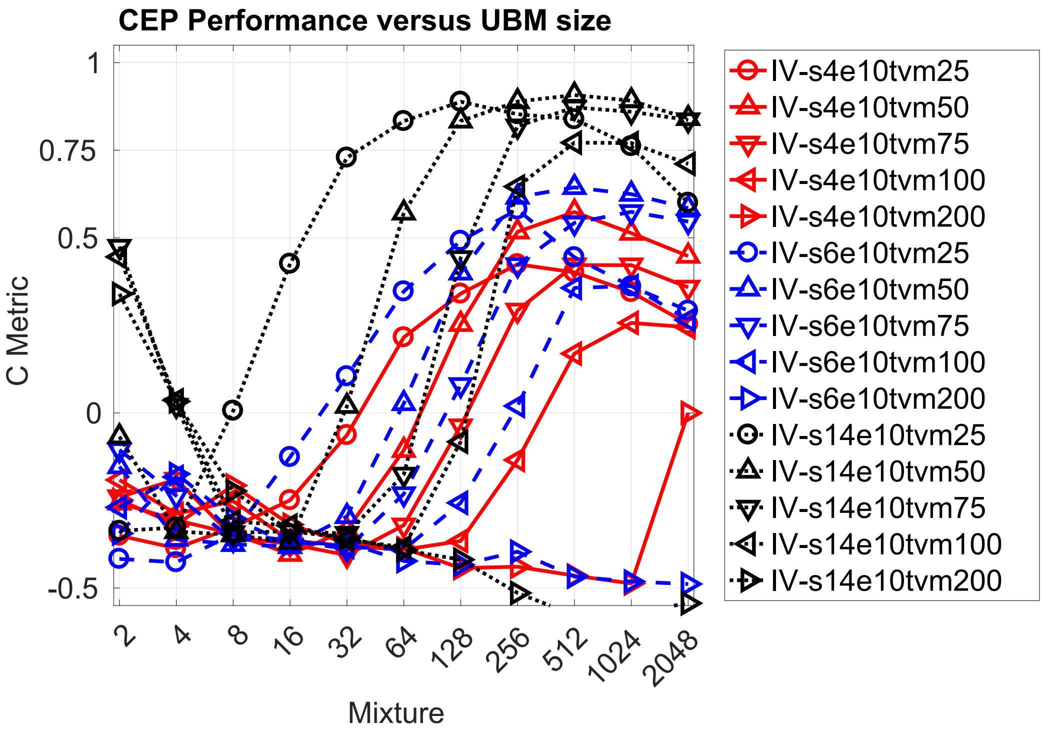


Figure 5.19. I-Vector with CEP Mot, Multiple Test Epochs. This C Metric plot shows the CEP based PhysioNet Database dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 6 testing epochs of 10s duration, and 14 testing epochs of 10s duration. The peak performances achieved were 0.57 with the 4 epochs of 2s duration configuration using a 512 mixture UBM with a TVM dimension of 50, 0.64 with the 6 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50, and 0.91 with the 14 epochs of 2s duration configuration using a 512 mixture UBM with a TVM dimension of 50.

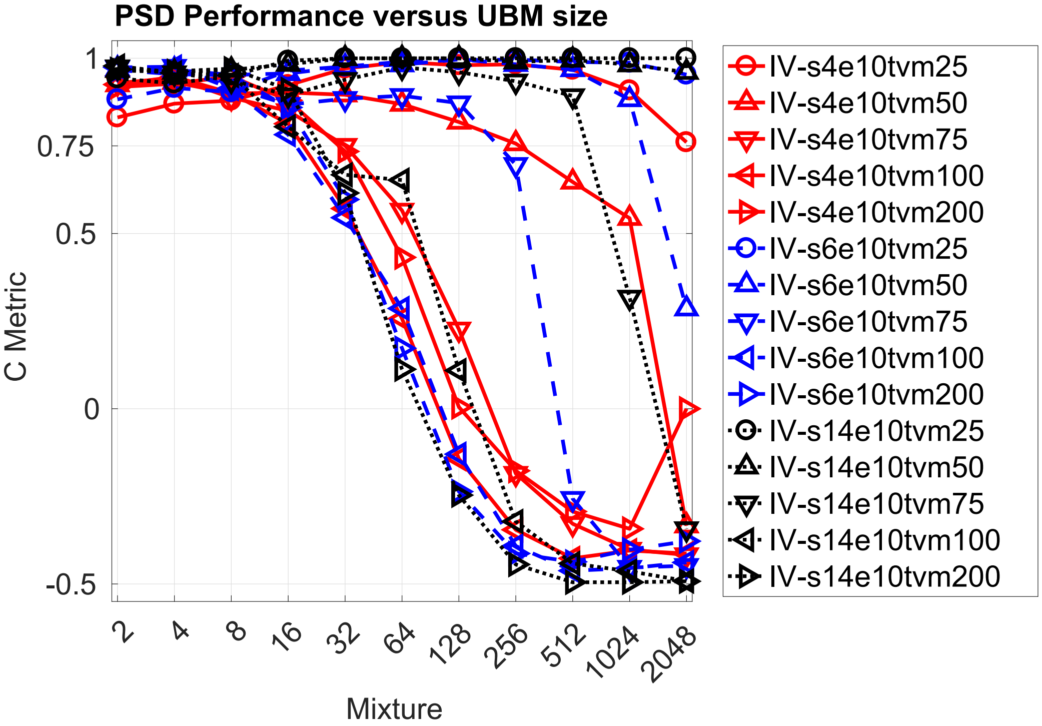


Figure 5.20. I-Vector with PSD Mot, Multiple Test Epochs. This C Metric plot shows the PSD based PhysioNet Database dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 6 testing epochs of 10s duration, and 14 testing epochs of 10s duration. The peak performances achieved were 0.99 with the 4 epochs of 2s duration configuration using a 64 mixture UBM with a TVM dimension of 25, 1 with the 6 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 75, and 1 with the 14 epochs of 2s duration configuration using a 64 mixture UBM with a TVM dimension of 100.

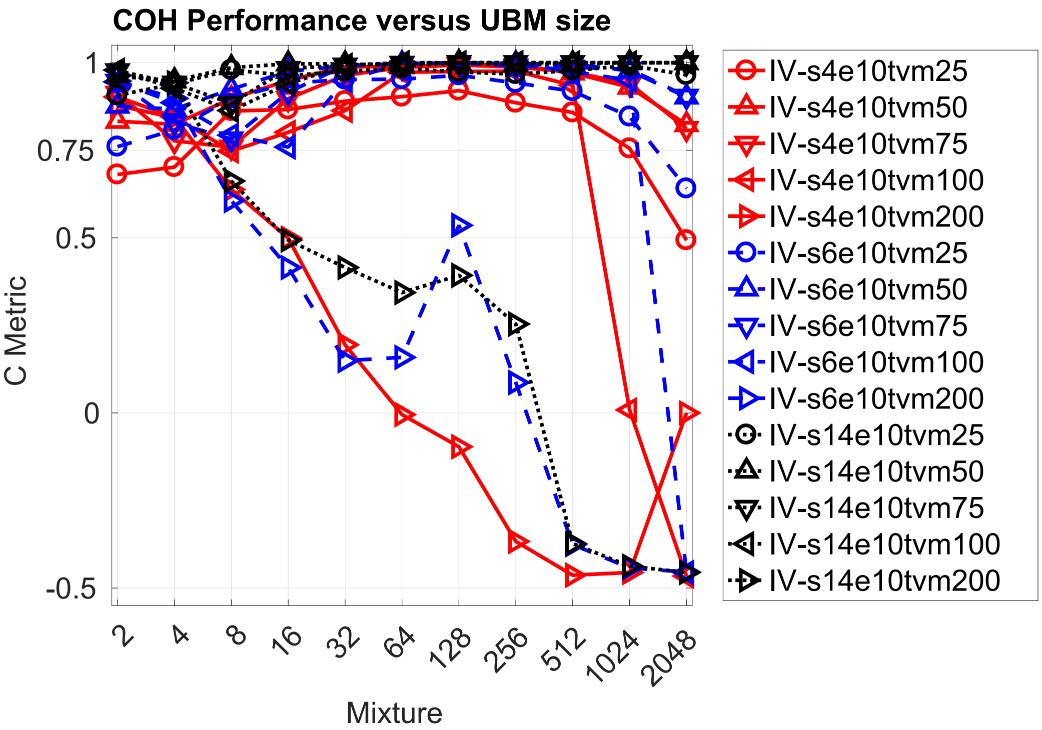


Figure 5.21. I-Vector with COH `Mot`, Multiple Test Epochs. This C Metric plot shows the COH based PhysioNet Database dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 6 testing epochs of 10s duration, and 14 testing epochs of 10s duration. The peak performances achieved were 0.99 with the 4 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50, 1 with the 6 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 75, and 1 with the 14 epochs of 2s duration configuration using a 64 mixture UBM with a TVM dimension of 100.

not decrease, before a performance roll-off for all tested configurations started at a mixture size of 128.

The TUH-EEG seizure dataset was tested again, Figures 5.25–5.27, with sample sets on the order used for the PhysioNet Database datasets. Using the 2s duration epochs, an additional 14 epoch set was included. Overall the scoring trend was that the larger sample sets performed better, but the extent varied by feature set.

For the CEP features, Figure 5.25, the increased samples allowed TVMs of size 25, 50, and 75 to exceed those previously reported for the 6-sample sets before rolling off when the UBM contained 2048 mixtures. The PSD features, Figure 5.26, met and exceed the 0.75 score threshold when the UBM was between 32 and 1024 mixtures. This occurred for TVMs of size 50, 75, and 100.

Previously, the COH features were best served by the 10s duration 4 epoch set, but the 2s duration 14 epoch set included here matched their performance. However, this was only for a TVM of size 50 despite the larger sample set out performing its counterparts in Figure 5.27.

5.2.2 Results: LDA Parameters

The performance impact of the LDA was run on a subset of the optimized parameter space of epoch duration, sample set parameters, and UBM size. The UBMs built from 32 to 128 mixtures appeared the strongest across datasets and feature sets. Epoch duration was chosen as 2 seconds with a sample set size of 14 for the TUH-EEG normal and seizure datasets. For the PhysioNet Database dataset the epoch duration was 10s with an epoch set size of 14.

The scores of the TUH-EEG normal dataset, Figures 5.28–5.30, are limited in scope by the number of subjects (50), making the largest LDA dimension 49. Overall performance appeared linked to TVM size more than UBM mixture size.

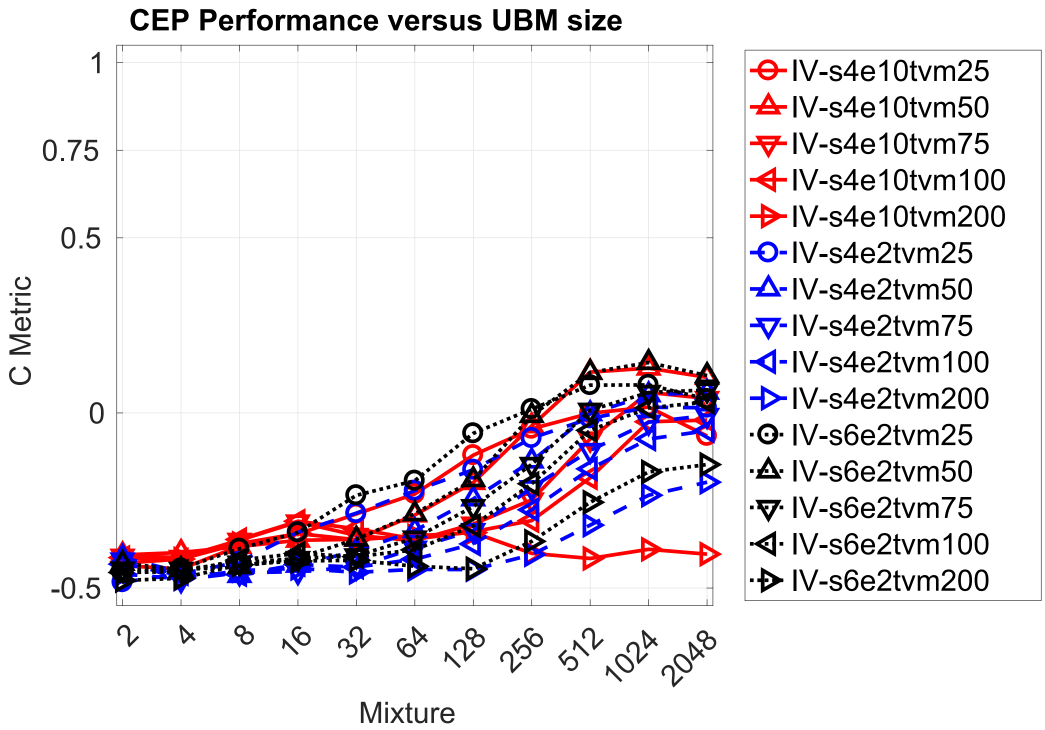


Figure 5.22. I-Vector with CEP `Szr`, Multiple Test Epochs. This C Metric plot shows the CEP based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 4 testing epochs of 2s duration, and 6 testing epochs of 2s duration. The peak performances achieved were 0.13 with the 4 epochs 10s in duration configuration using a 1024 mixture UBM with a TVM dimension of 50, 0.06 with the 4 epochs of 2s duration configuration using a 2048 mixture UBM with a TVM dimension of 50, and 0.14 with the 6 epochs of 2s duration configuration using a 1024 mixture UBM with a TVM dimension of 50.

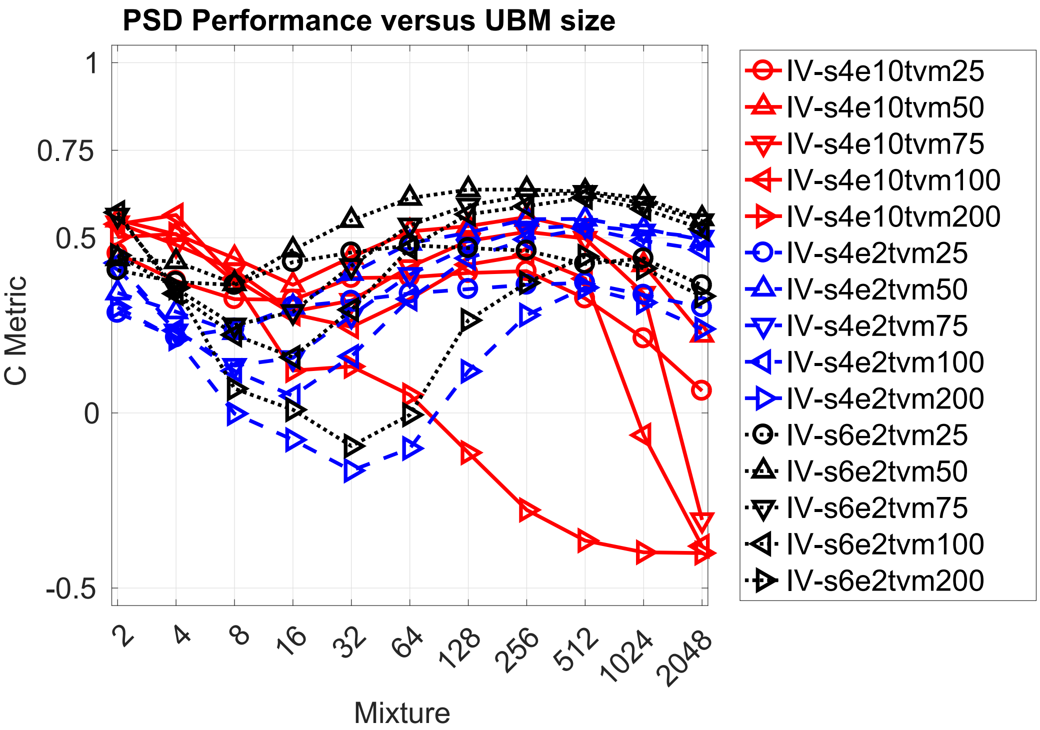


Figure 5.23. I-Vector with PSD `Szr`, Multiple Test Epochs. This C Metric plot shows the PSD based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 4 testing epochs of 2s duration, and 6 testing epochs of 2s duration. The peak performances achieved were 0.57 with the 4 epochs 10s in duration configuration using a 4 mixture UBM with a TVM dimension of 100, 0.55 with the 4 epochs of 2s duration configuration using a 512 mixture UBM with a TVM dimension of 50, and 0.64 with the 6 epochs of 2s duration configuration using a 256 mixture UBM with a TVM dimension of 50.

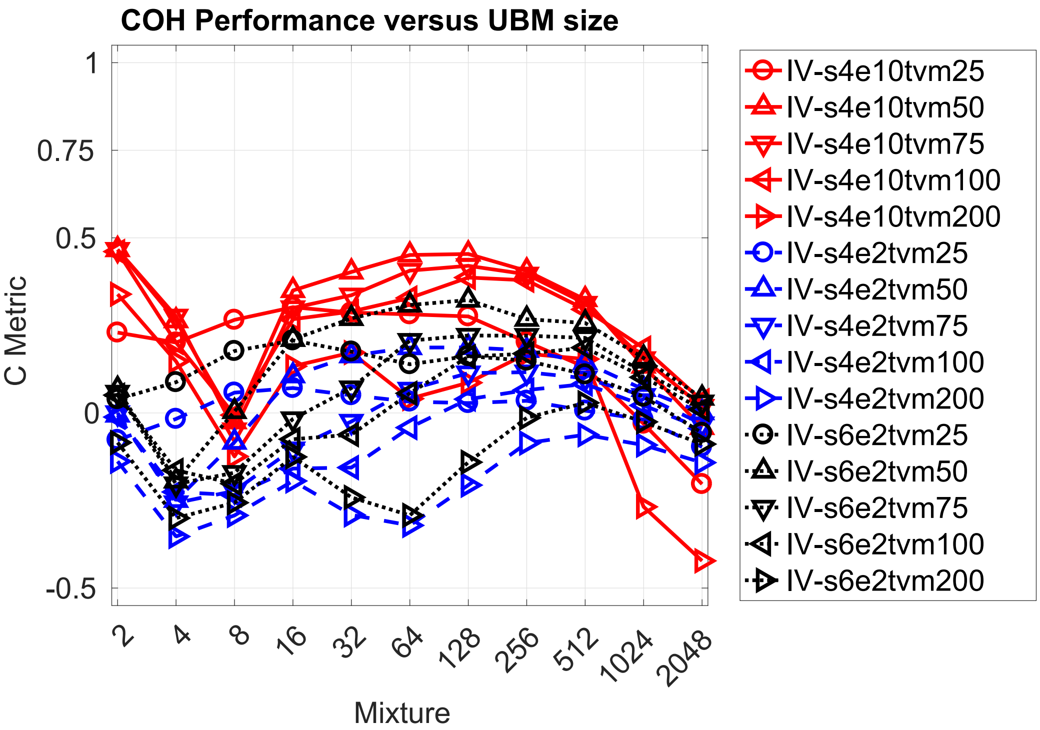


Figure 5.24. I-Vector with COH `Szr`, Multiple Test Epochs. This C Metric plot shows the COH based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 10s duration, 4 testing epochs of 2s duration, and 6 testing epochs of 2s duration. The peak performances achieved were 0.19 with the 4 epochs 10s in duration configuration using a 2 mixture UBM with a TVM dimension of 50, 1 with the 4 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 75, and 0.32 with the 6 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50.

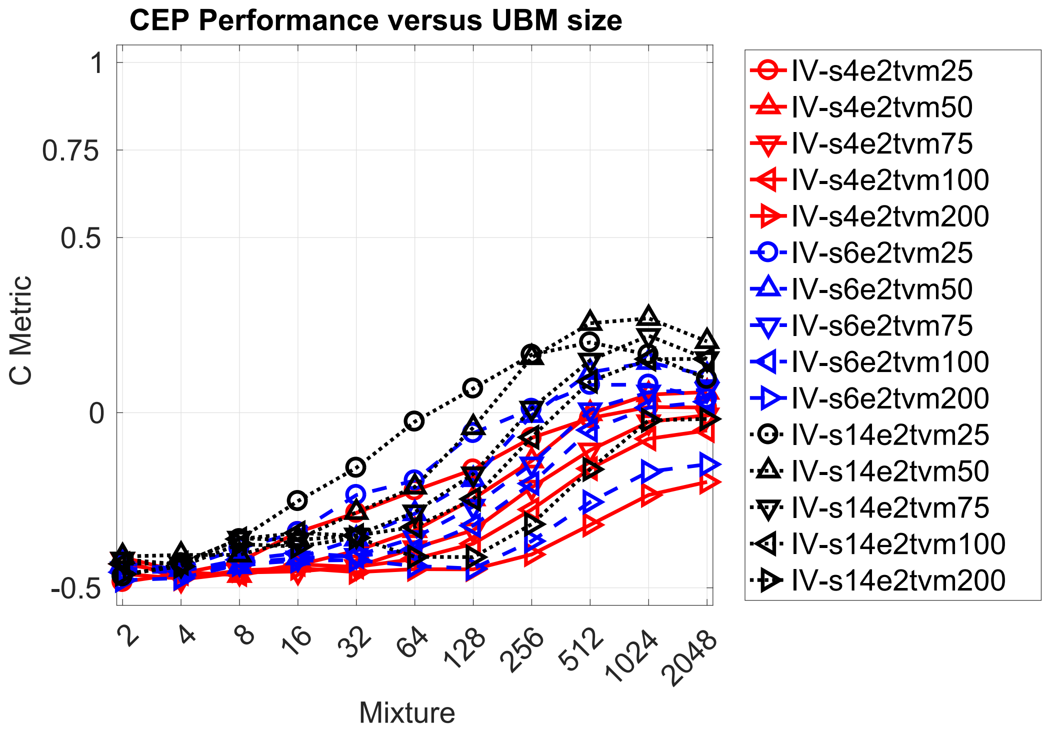


Figure 5.25. I-Vector with CEP `Szr`, Multiple Test Epochs. This C Metric plot shows the CEP based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.0589 with the 4 epochs of 2s duration configuration using a 2048 mixture UBM with a TVM dimension of 50, 0.14 with the 6 epochs of 2s duration configuration using a 1024 mixture UBM with a TVM dimension of 50, and 0.27 with the 14 epochs of 2s duration configuration using a 1024 mixture UBM with a TVM dimension of 50.

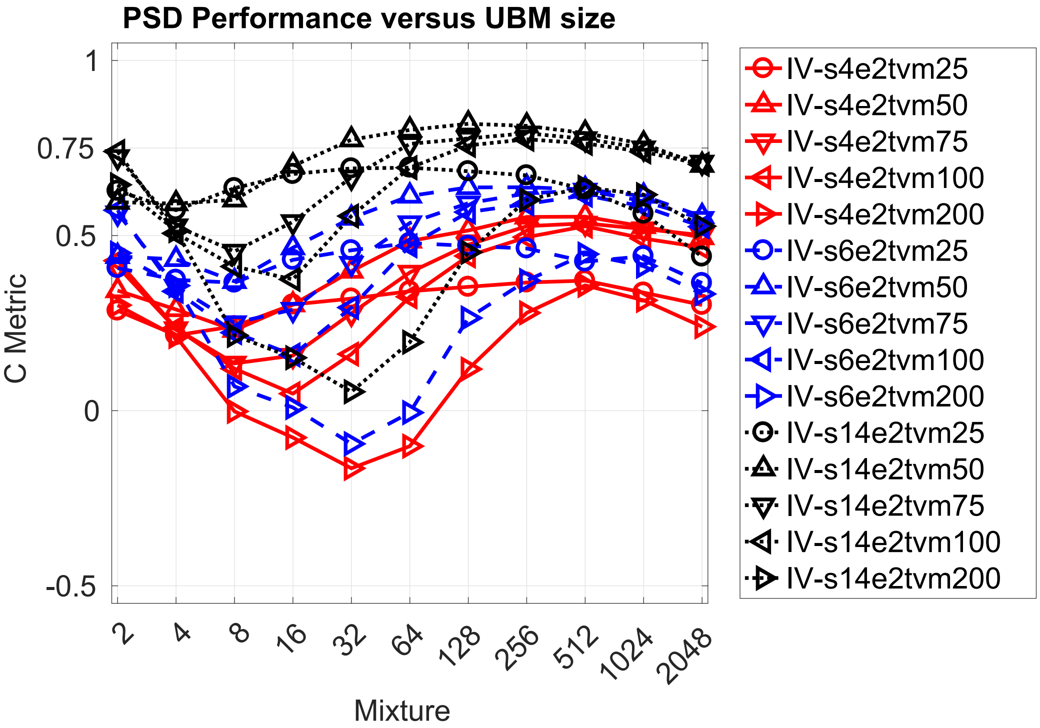


Figure 5.26. I-Vector with PSD Szr, Multiple Test Epochs. This C Metric plot shows the PSD based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.55 with the 4 epochs of 2s duration configuration using a 256 mixture UBM with a TVM dimension of 50, 0.64 with the 6 epochs of 2s duration configuration using a 256 mixture UBM with a TVM dimension of 50, and 0.82 with the 14 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50.

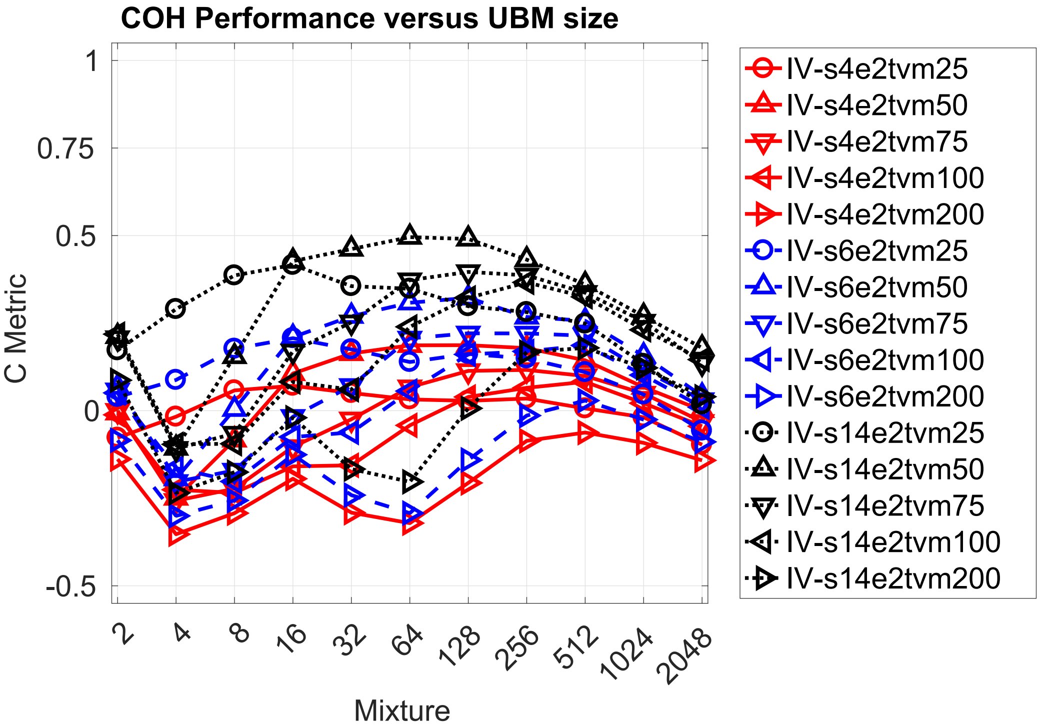


Figure 5.27. I-Vector with COH `Szr`, Multiple Test Epochs. This C Metric plot shows the COH based TUH-EEG seizure dataset performance as a function of UBM mixture and TVM dimension. This data was split into three groups: 4 testing epochs of 2s duration, 6 testing epochs of 2s duration, and 14 testing epochs of 2s duration. The peak performances achieved were 0.19 with the 4 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50, 0.32 with the 6 epochs of 2s duration configuration using a 128 mixture UBM with a TVM dimension of 50, and 0.50 with the 14 epochs of 2s duration configuration using a 64 mixture UBM with a TVM dimension of 50.

The reduction in LDA dimensions had minimal impact on performance over the range of 49, 45, and 30. The major loss occurred at stepping down to 15 dimensions for any TVM that was larger than 25.

For the CEP features, the best scores came from TVMs of size 25 for any UBM and a TVM of size 50 when using a 128 mixture UBM. The PSD features showed similar scores, with the TVMs of size 25 setting the benchmark near a score of 0.75 that the others attempted to reach. The COH features also had the top scores for the TVMs of size 25. However, when using a UBM with 64 mixtures the scores of TVMs of size 75 and 100 were on par with that of the 128 mixture UBM's size 25 TVM.

The same experiment with the PhysioNet Database dataset, Figures 5.31–5.33, contained 109 subjects enabling the full spread of LDA dimensions listed in Table 5.1. Again the smallest TVM appeared to provide consistently strong scores, however the impact of the remaining TVMs varied based upon feature set.

For the CEP features, decreasing the LDA dimensions weakened performance for TVMs of size 50. However, the same dimensional reduction applied to UBMs of size 32 and 64 resulted in improved performance for larger TVMs at smaller LDA sizes. This behavior partially occurred for the 128 mixture UBM as the performance of the 75 TVM decreased as the LDA dimensions were reduced, while the 100 TVM performance eventually increased.

The unique performance of the 128 mixture UBM continued for the PSD features with the 100 TVM score increasing and decreasing as the LDA dimension was decreased. However, a strong performance loss was seen for all 200 TVMs and minimal loss was seen for TVMs below 100.

This behavior was most stark in the COH features as only the 200 TVMs showed significant changes with a reduction in the LDA dimensions. For all TVMs below

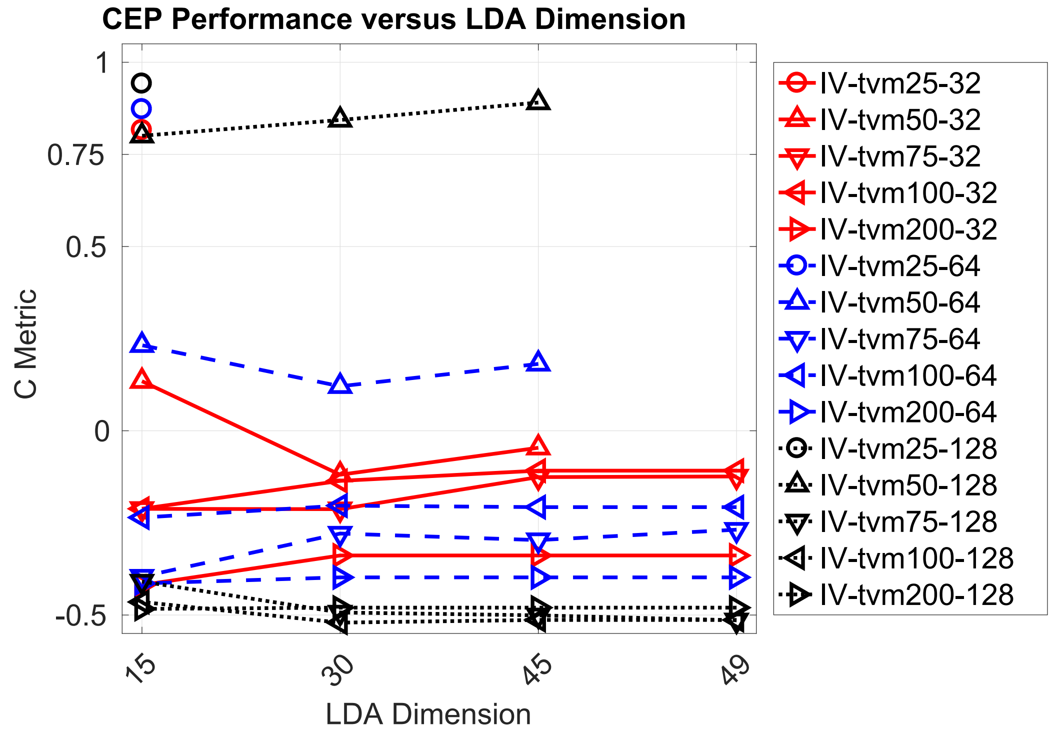


Figure 5.28. I-Vector and LDA with CEP `Szr`. This C Metric plot shows the CEP based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.16. The best 32 mixture UBM score was 0.82 using a 25-15 TVM LDA pairing, 0.87 using a 25-15 TVM LDA pairing with 64 mixtures, and 0.9419 using a 25-15 TVM LDA pairing with 128 mixtures.

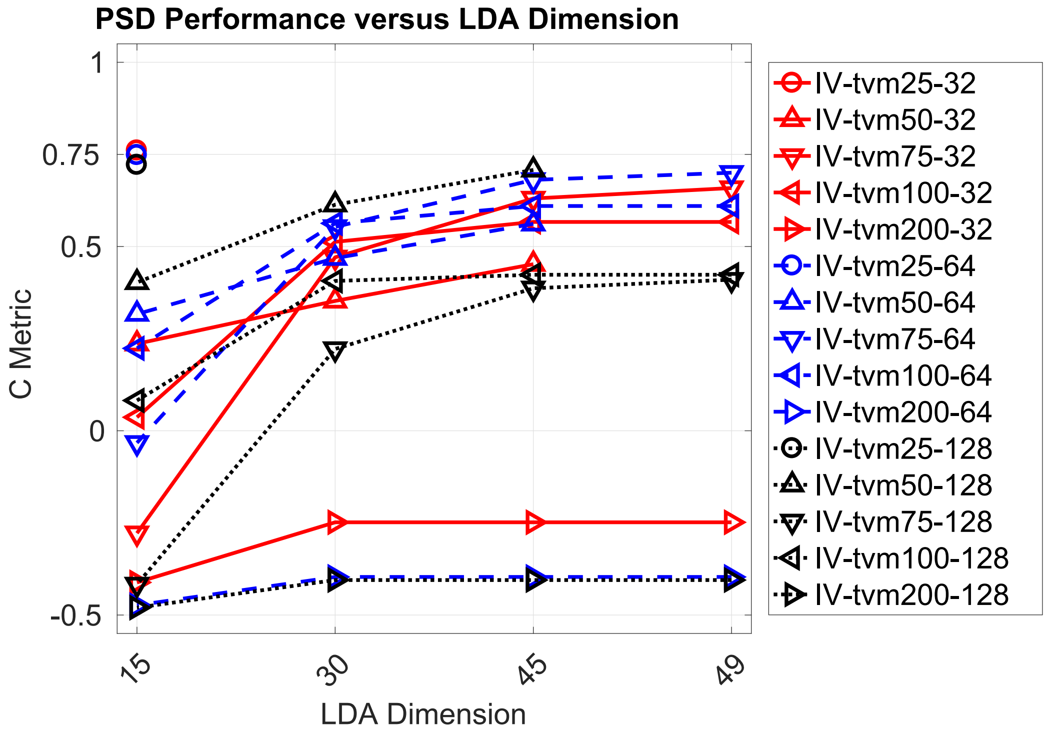


Figure 5.29. I-Vector and LDA with PSD `Szr`. This C Metric plot shows the PSD based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.17. The best 32 mixture UBM score was 0.76 using a 25-15 TVM LDA pairing, 0.7481 using a 25-15 TVM LDA pairing with 64 mixtures, and 0.72 using a 25-15 TVM LDA pairing with 128 mixtures.

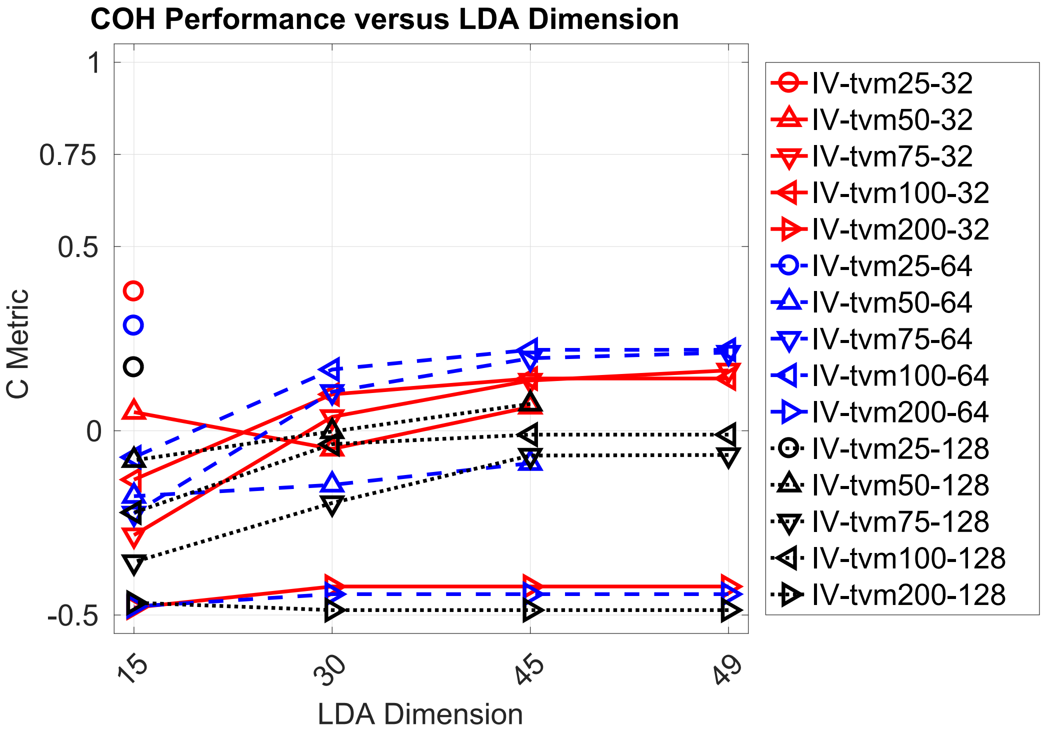


Figure 5.30. I-Vector and LDA with COH `Szr`. This C Metric plot shows the COH based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.18. The best 32 mixture UBM score was 0.378 using a 25-15 TVM LDA pairing, 0.2851 using a 25-15 TVM LDA pairing with 64 mixtures, and 0.172 using a 25-15 TVM LDA pairing with 128 mixtures.

200, the only significant performance changes occurred at 15 and 30 dimensions while at 200 the shift occurred between 45 and 60 dimensions.

The final LDA experiment involved the TUH-EEG seizure data, Figures 5.34–5.36. This dataset contained 411 subjects ensuring all variations of TVM and LDA could be tested. Breaking the trend of the previous experiments, the best performance was not found when using a TVM of size 25 for all feature sets.

When using the CEP features, Figure 5.34, performance was poor for all parameter configurations. Only the TVM of size 25 built from the 128 mixture UBM were able to exceed a score of zero. Even with such poor performance the larger UBMs scored better and decreasing the LDA dimensions had a noticeable impact of further reducing performance.

The PSD features, Figure 5.35, produced results that met or exceeded the size 25 TVMs across all UBM mixture sizes. These included the 128 mixture UBM using TVM LDA pairings of 100-75, 75-60, 75-45, 50-45, and 50-30. As well as the 64 mixture UBM using TVM LDA pairings of 75-60, 75-45, 50-45, and 50-30. Only a pairing of 50-45 for the 32 mixture UBM met or exceeded the 25 TVMs performance. All TVMs showed the largest reduction in performance when operating with with an LDA of 15.

The impact of LDA on the COH features, Figure 5.36, produced similar results to the PSD features with the larger UBMs mixtures outperforming the smaller mixtures. The TVM LDA pairings of 50-45 and 50-30 for all UBMs met or exceed the performance of the 25-15 TVM LDA pairing. While only the 75-60 pairing of the 64 and 128 mixture UBMs met or exceeded that benchmark. The 200 TVM results produced scores at or below zero for all UBM mixtures.

To validate that the selection of 32, 64, and 128 were the optimal UBM mixture sizes, two alternative ranges were tested using the TUH-EEG normal dataset. The

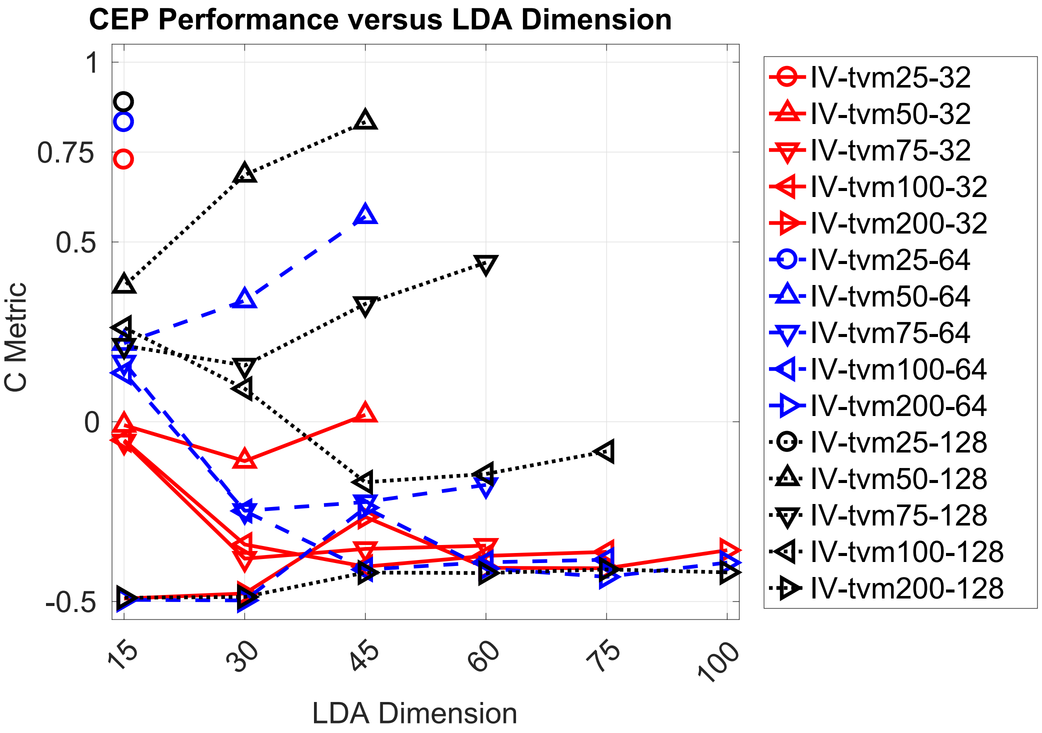


Figure 5.31. I-Vector and LDA with CEP `Mot`. This C Metric plot shows the CEP based PhysioNet Database dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 10s duration presented previous in Figure 5.19. The best 32 mixture UBM score was 0.7289 using a 25-15 TVM LDA pairing, 0.8333 using a 25-15 TVM LDA pairing with 64 mixtures, and 0.8885 using a 25-15 TVM LDA pairing with 128 mixtures.

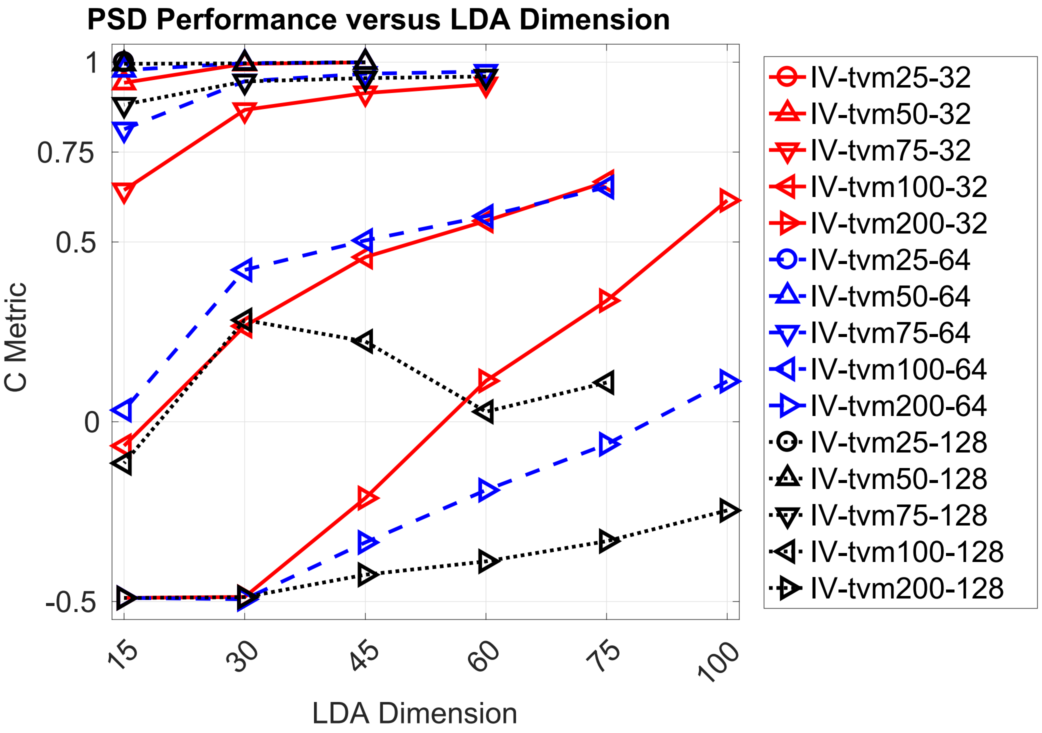


Figure 5.32. I-Vector and LDA with PSD Mot. This C Metric plot shows the PSD based PhysioNet Database dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 10s duration presented previous in Figure 5.20. The best 32 mixture UBM score was 0.9998 using a 25-15 TVM LDA pairing, 0.999 using a 25-15 TVM LDA pairing with 64 mixtures, and 1 using a 25-15 TVM LDA pairing with 128 mixtures.

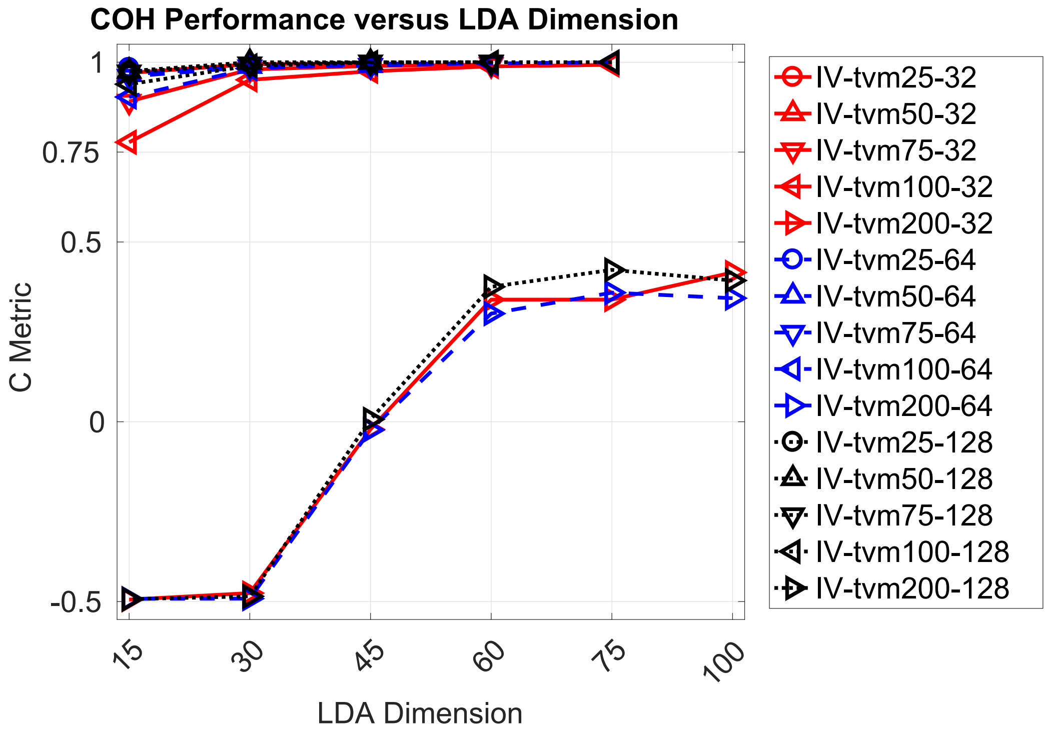


Figure 5.33. I-Vector and LDA with COH `Mot`. This C Metric plot shows the COH based PhysioNet Database dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 10s duration presented previous in Figure 5.21. The best 32 mixture UBM score was 0.99 using a 50-45 TVM LDA pairing, 1 using a 100-75 TVM LDA pairing with 64 mixtures, and 1 using a 75-45 TVM LDA pairing with 128 mixtures.

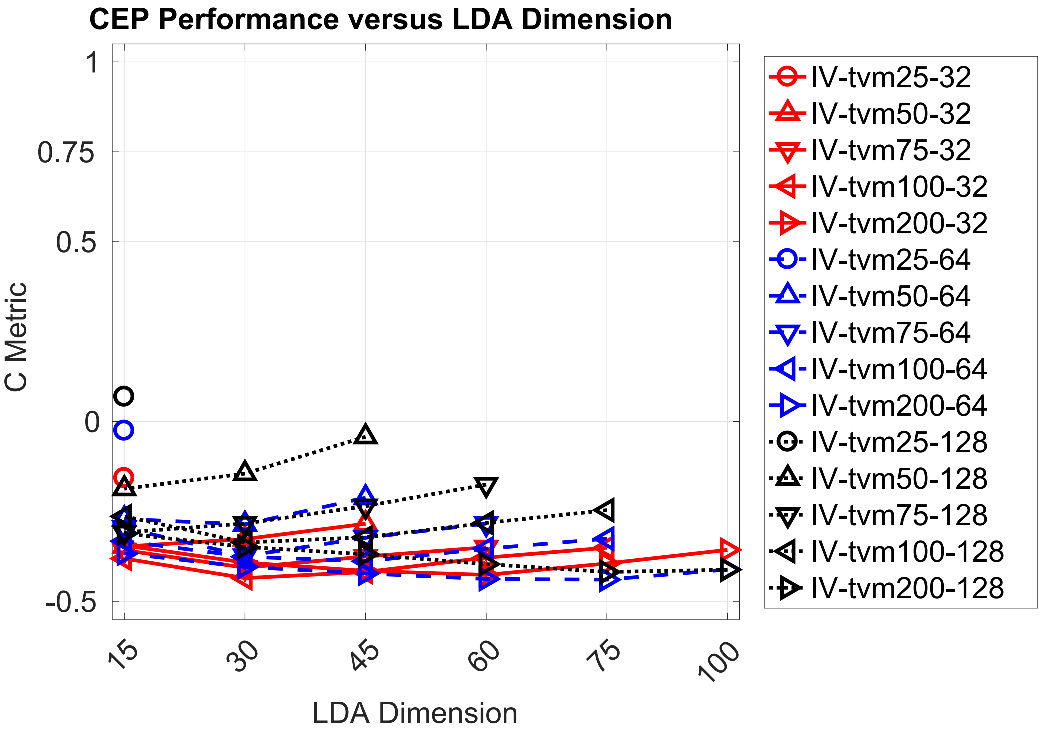


Figure 5.34. I-Vector and LDA with CEP `Szr`. This C Metric plot shows the CEP based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.25. The best 32 mixture UBM score was 0.07 using a 25-15 TVM LDA pairing, -0.03 using a 25-15 TVM LDA pairing with 64 mixtures, and 0.09 using a 25-15 TVM LDA pairing with 128 mixtures.

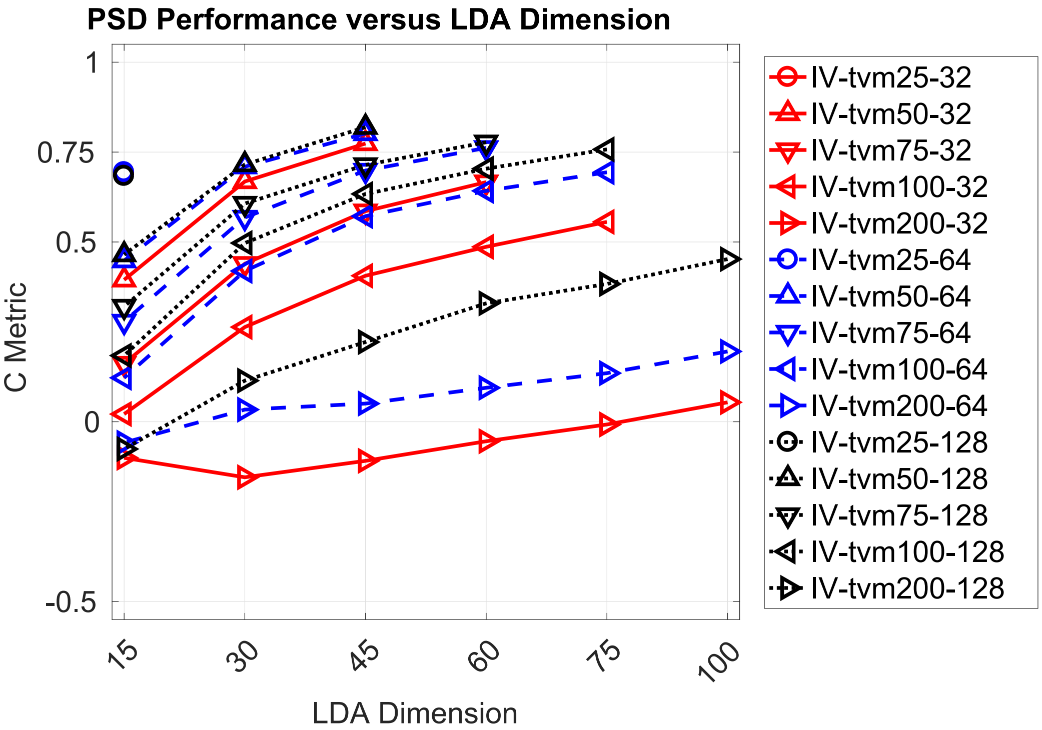


Figure 5.35. I-Vector and LDA with PSD Szr. This C Metric plot shows the PSD based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.26. The best 32 mixture UBM score was 0.77 using a 50-45 TVM LDA pairing, 0.8016 using a 50-45 TVM LDA pairing with 64 mixtures, and 0.82 using a 50-45 TVM LDA pairing with 128 mixtures.

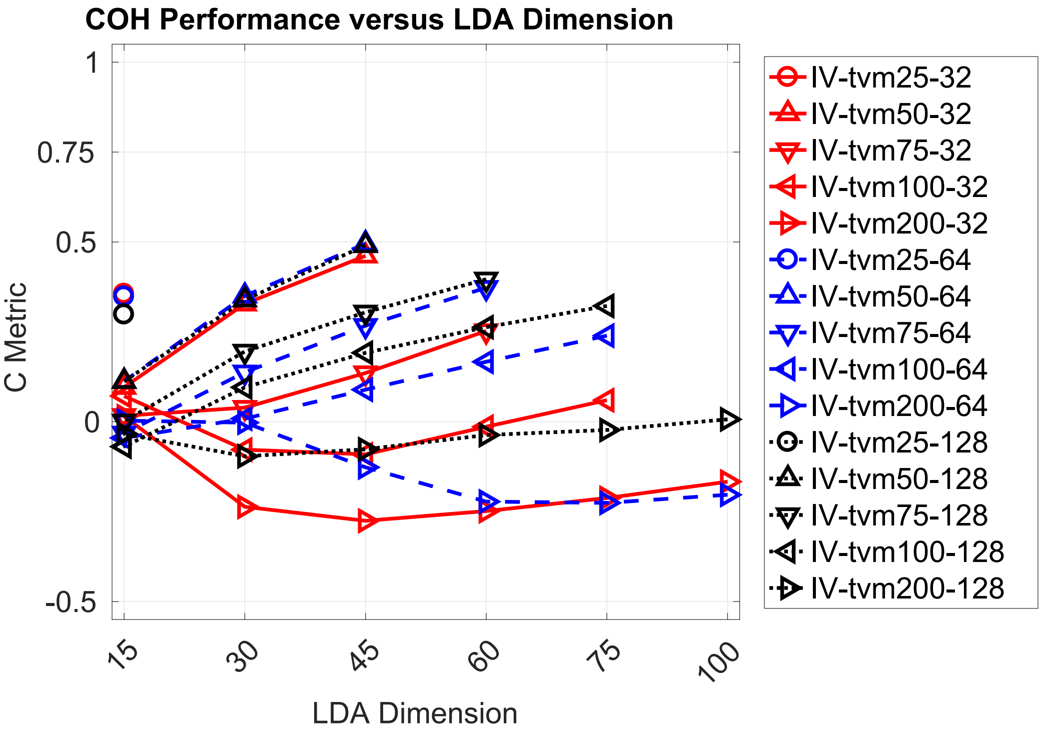


Figure 5.36. I-Vector and LDA with COH `Szr`. This C Metric plot shows the COH based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.27. The best 32 mixture UBM score was 0.4903 using a 50-45 TVM LDA pairing, 0.4312 using a 50-45 TVM LDA pairing with 64 mixtures, and 0.3604 using a 50-45 TVM LDA pairing with 128 mixtures.

first alternative used 8, 16, and 32 mixture UBMs and the second alternative used 128, 256, and 512 mixture UBMs. The first alternative repeated the smallest mixture size and the second alternative repeated the largest mixture size providing context for the results.

When using the smaller mixtures, Figures 5.37–5.39, the strongest scores came from the 32 mixture UBM based TVMs. Only in the case of the CEP features was the performance of the 32 mixture UBM exceed by the two smaller mixtures. However, the only acceptable scores came from the 25 TVM which clearly favored the 32 mixture UBM. For the PSD and COH features the 32 mixture UBM significantly outperformed its counterparts for all TVM LDA combinations, aside from at the 25-15 pairing.

Expanding the UBM to larger mixtures, Figures 5.40–5.42, produced distinct behavior for each feature set. The CEP feature performance improved with each increase in the UBM mixture size. However, these improvements were not enough to match the performances of the PSD or COH feature sets.

Conversely, the PSD feature set showed minimal if any difference in performance for the TVM LDA pairings, aside from the 200 TVM which improved with the larger UBM mixtures. When the TVM LDA pairings of 100-75, 75-50, and 50-45 were used, their performance exceeded a score of 0.75 for all UBM mixture sizes.

The COH feature set showed that performance was better or equal when using the 128 mixture UBM except when using the 200 TVM. However, none of the pairings exceeded a score of 0.5.

5.2.3 Discussion

The initial sweeps of dataset, UBM mixture, TVM dimension, and sample set size were natural follow up experiments to the La Rocca based experiments. While the

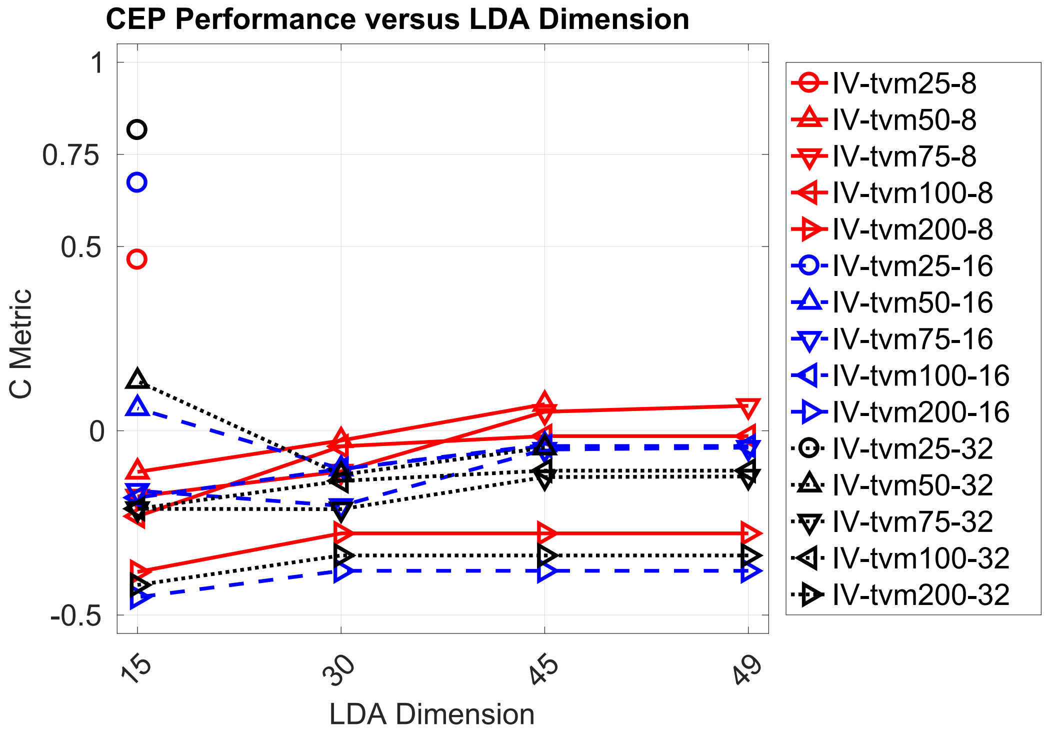


Figure 5.37. I-Vector and LDA with CEP Nrm. This C Metric plot shows the CEP based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.16. The best 8 mixture UBM score was 0.4641 using a 25-15 TVM LDA pairing, 0.6728 using a 25-15 TVM LDA pairing with 16 mixtures, and 0.8159 using a 25-15 TVM LDA pairing with 32 mixtures.

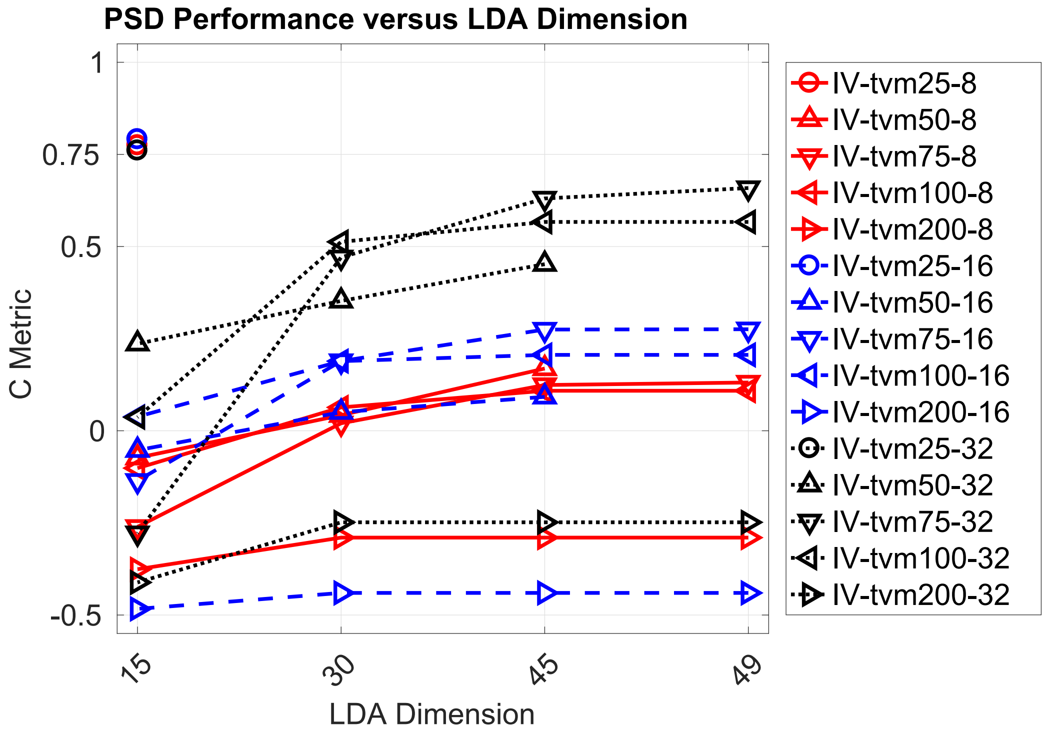


Figure 5.38. I-Vector and LDA with PSD `Nrm`. This C Metric plot shows the PSD based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.17. The best 8 mixture UBM score was 0.7746 using a 25-15 TVM LDA pairing, 0.791 using a 25-15 TVM LDA pairing with 16 mixtures, and 0.7604 using a 25-15 TVM LDA pairing with 32 mixtures.

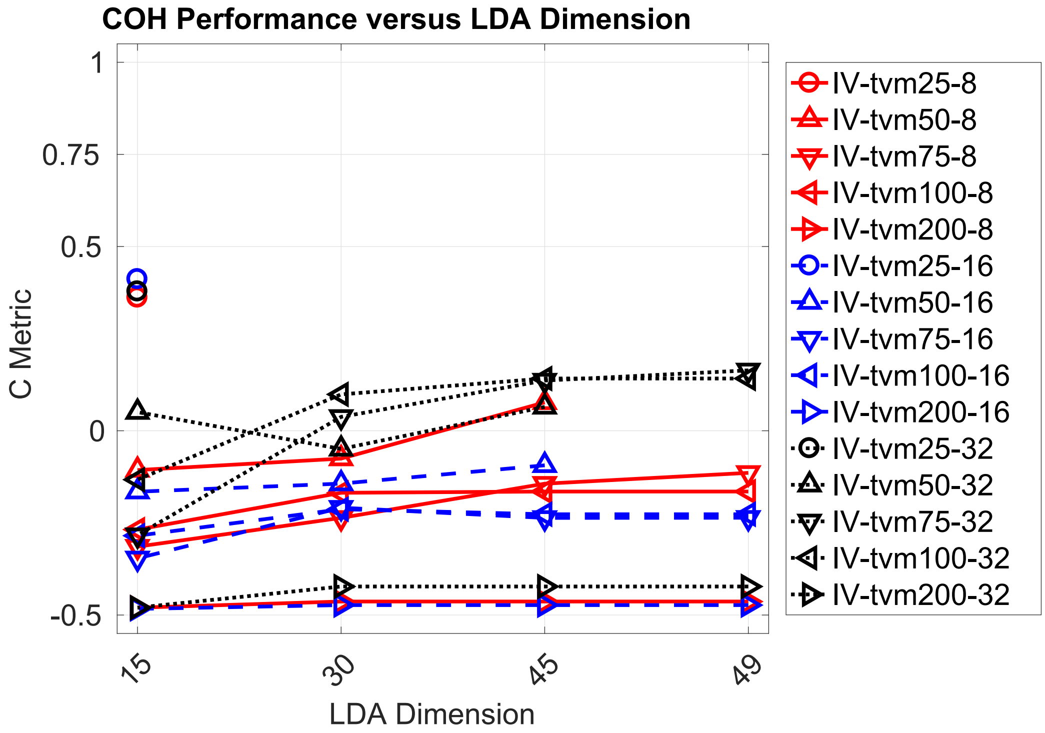


Figure 5.39. I-Vector and LDA with COH `Nrm`. This C Metric plot shows the COH based TUH-EEG normal dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.18. The best 8 mixture UBM score was 0.3612 using a 25-15 TVM LDA pairing, 0.4112 using a 25-15 TVM LDA pairing with 16 mixtures, and 0.378 using a 25-15 TVM LDA pairing with 32 mixtures.

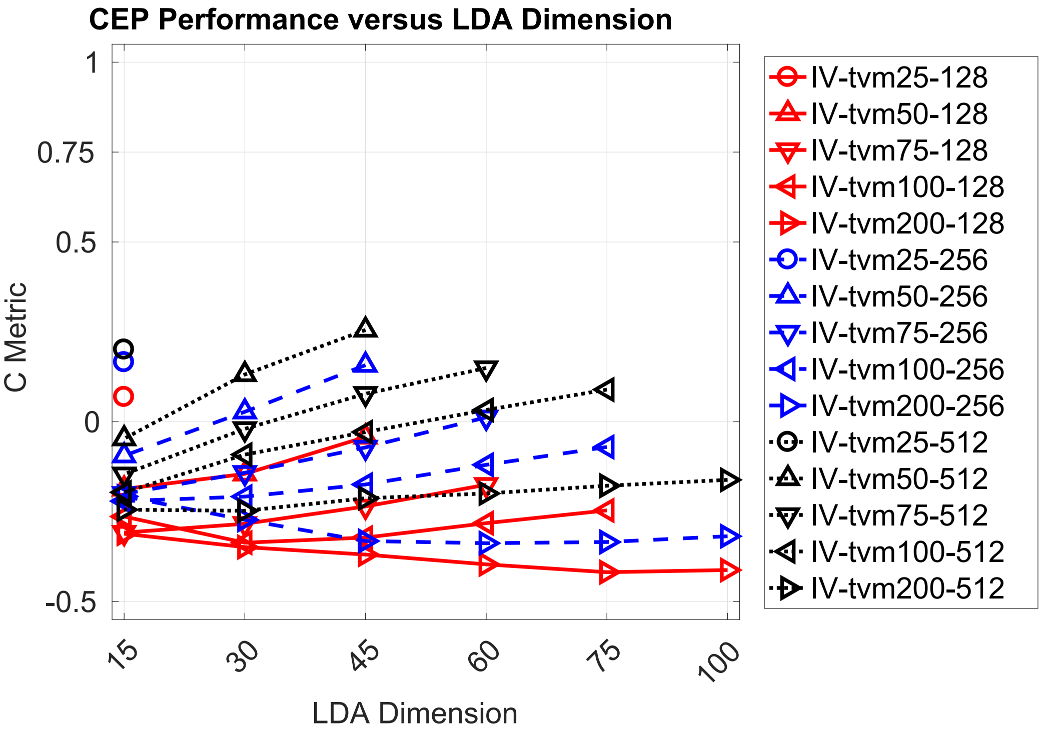


Figure 5.40. I-Vector and LDA with CEP `Szr`. This C Metric plot shows the CEP based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.25. The best 128 mixture UBM score was 0.0688 using a 25-15 TVM LDA pairing, 0.1655 using a 25-15 TVM LDA pairing with 256 mixtures, and 0.2007 using a 25-15 TVM LDA pairing with 512 mixtures.

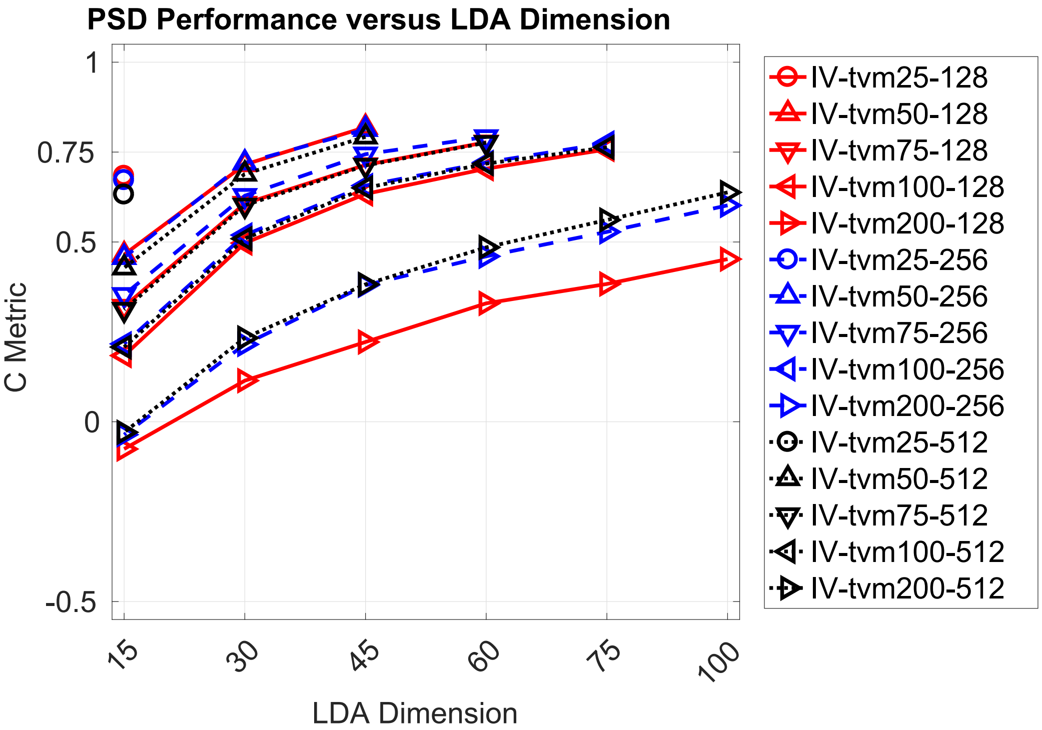


Figure 5.41. I-Vector and LDA with PSD `Szr`. This C Metric plot shows the PSD based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.26. The best 128 mixture UBM score was 0.8192 using a 50-45 TVM LDA pairing, 0.8133 using a 50-45 TVM LDA pairing with 256 mixtures, and 0.7924 using a 50-45 TVM LDA pairing with 512 mixtures.

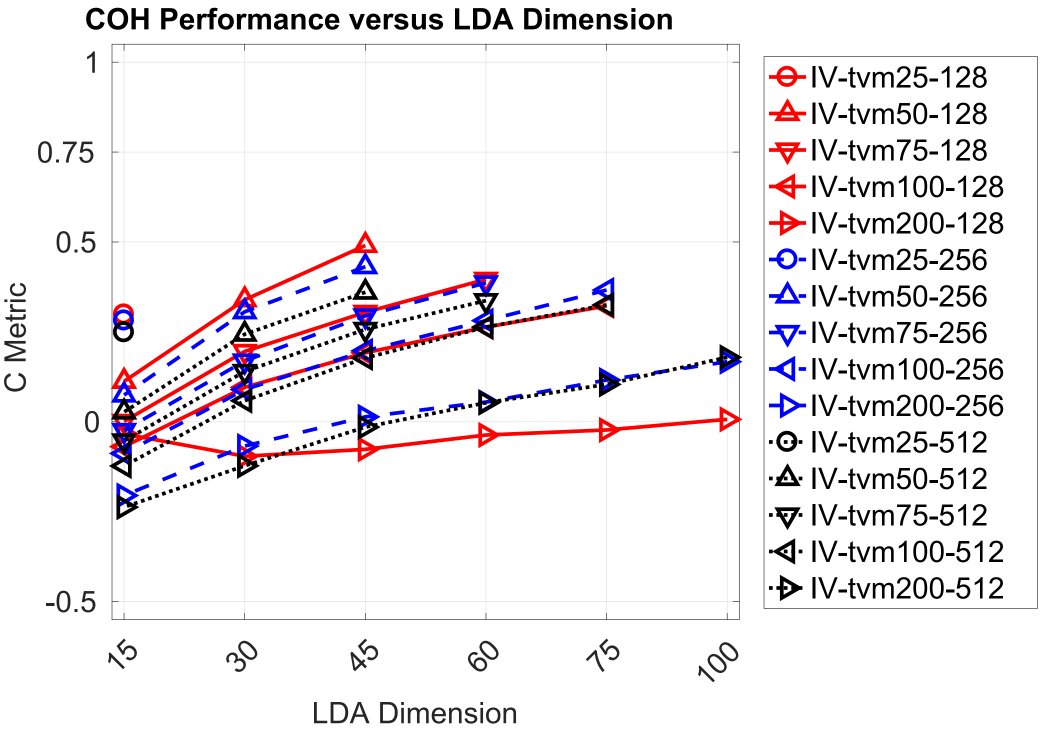


Figure 5.42. I-Vector and LDA with COH `Szr`. This C Metric plot shows the COH based TUH-EEG seizure dataset performance as a function of LDA dimension for specific UBM mixture sizes over a range of TVM dimensions. This data was drawn from the 14 epoch set of 2s epochs presented previous in Figure 5.27. The best 128 mixture UBM score was 0.4903 using a 50-45 TVM LDA pairing, 0.4312 using a 50-45 TVM LDA pairing with 256 mixtures, and 0.3604 using a 50-45 TVM LDA pairing with 512 mixtures.

three feature sets were tested along with each iteration, there was no intent to select a best feature. As such the three feature sets were reported throughout these experimental parameter sweeps. Following the major parameter sweeps, attention was paid to the impact of LDA on the resultant TVMs. Those LDA sweeps highlight the performance of I-Vectors when using the optimal configuration of UBM mixture, epoch duration, and sample set size.

5.2.3.1 I-Vector Parameters

The goal of these experiments was to find an optimal parameter range for the UBM mixture size, TVM dimension, LDA dimension, samples in a testing set, and epoch duration. Two distinct datasets, the PhysioNet Database dataset of Figures 5.5–5.7 and TUH-EEG dataset of Figures 5.10–5.12, were used to ensure the robustness of these solutions. Given the scope of the experiments it was necessary to determine if all three TUH-EEG datasets needed to be tested in addition to the PhysioNet Database dataset.

From the results of Figures 5.10–5.12 it was determined that the TUH-EEG abnormal and normal datasets produced similar scores for each feature set. Conversely the TUH-EEG seizure dataset produced scores distinct from both the abnormal and normal datasets. The CEP features appeared best suited for the abnormal and normal datasets and the PSD features for the seizure dataset. However, the overall performances were far from acceptable aside from the scores produced by the normal dataset using a TVM dimension of 25 with UBMs between 128 and 1024 mixtures.

During these experiments the epoch duration was decreased from 10 seconds to 2 seconds, see Figures 5.13–5.15. This provided insight into the influence of altering the balance of enrollment and testing data, as decreasing the epoch duration increased the total number of samples in the recordings. Using 2 minute recordings with 10s duration epochs produced 12 total epochs, which meant 4 epochs represented 25% of the data. With 2s duration epochs, withholding 4 epochs meant 5.7% of the data was reserved, 6 epochs meant 10%, 12 epochs meant 20%, and 14 epochs meant 23% of the data. Across all three feature sets, the 4 sample testing set met or outperformed the 2 sample testing set when using 10s duration epochs. Comparatively, the 2s duration 4 epoch testing set was the optimal choice only for CEP as it was also bested by the 10s duration 4 epochs on the PSD and COH feature sets.

The choice of testing sets was designed around the PhysioNet Database trial configuration (2 resting trials, 4 motion trials repeated 3 times) allowing for a variety of intermediate epoch sets with maximum of 14 epochs in the testing data. The reported scores using epoch sets of 4, 6, and 14 with 2s durations in Figures 5.16–5.18 indicated similar performance trends even as the size of the training data grew. In fact, the 14 sample sets for all features matched or exceeded scores of smaller sets across all features for each TVM dimension. The majority of these performance gains happened when progressing from a 32 mixture UBM up to a 128 mixture UBM.

However, comparing these performances was difficult because each epoch set also corresponded to a distinct dataset. The 6 and 14 epoch datasets contained resting data which was absent from the 4 epoch dataset. Likewise, the 15 epoch dataset contained an overwhelming amount of motion data, 12 trials to the 2 resting trials. While performance clearly improved as more data was added, it was difficult to determine how much impact the characteristics (resting versus motion) of the data had on performance.

Similar performance trends were seen in the TUH-EEG datasets. While 2 epoch testing datasets were tested on the normal data, seizure data started with 4 epoch testing datasets using 10s duration epochs, just like the PhysioNet Database experiments, as well as 4 epoch 2s duration and 6 epoch 2s duration testing datasets. However, these tests also tracked the impact of the feature sets. Contrasting with the TUH-EEG normal results, the CEP features produced the worst scores while the PSD features showed the most promise with the 6 epochs of 2s duration, Figures 5.22–5.24. However, the strongest scores for the COH features, Figure 5.24, were the 4 epoch of 10s duration. It was difficult to infer anything from these results as only the PSD features scored above 0.5, which lead to repeating this experiment with a 14 epoch testing dataset, Figures 5.25–5.27.

The trend of larger epoch sets improving performance was verified through these seizure based experiments, but it also suggested feature sets might be suited for specific datasets. The PhysioNet Database scored better using the PSD and COH features than the CEP features. The COH tracking with the PSD was expected as they are built on the PSD values, however the seizure results showed COH were better than PSD, Figures 5.22–5.24. For the 14 epoch 2s duration testing seizure dataset, the PSD feature set exceeded a score of 0.75 with multiple TVMs dimensions (50, 75, and 100) while the COH features could only reach 0.5 with a TVM of dimension 50. Meanwhile, the CEP features failed to match the performance of the other features directly opposing the performance seen for the equivalent normal data experiments.

Despite the presence of 411 subjects in the seizure dataset, performance peaked when using between 32 and 128 mixture UBMs for the PSD and COH features. This trend was consistent across datasets, while the CEP features often continued to improve as the number of mixtures in the UBM increased. This was the case in Figure 5.25, as well as the prior CEP Figures 5.13 and 5.19. However, the majority of the classification gains were made in the 32 to 128 mixture range with additional mixtures providing less significant improvement.

5.2.3.2 LDA Parameters

The LDA parameter sweep used 14 samples in the testing data with 2 epochs for the TUH-EEG datasets and 14 samples with 10s epochs for the PhysioNet Database dataset. The PhysioNet Database experiments used a larger epoch duration because the smaller duration epochs performed exceptionally well leaving minimal remove for improvement, Figures 5.19–5.21. By reducing the number of UBM mixtures, the importance of TVM dimension became more apparent given the subsequent reduction via LDA. Of course, the smaller sized TVMs had fewer reduced configurations, but the intent was to see how performance shifted as the number of classification elements was reduced. This directly addressed if the natural discriminators of the the data could be reduced to a closed reproducible set of elements regardless of initial TVM dimension.

The initial focus was on the TUH-EEG normal dataset, Figures 5.28–5.30, which had poor performance for large TVM dimensions. This behavior was mostly consistent across feature sets, but for the CEP features. They were capable of producing a C score over 0.75 for all LDA dimensions when using a 128 mixture UBM and a TVM of 49 dimensions. This suggested many of the original elements were not critical to classification, which was supported by the strength of the other 15 dimension LDA scores across the 3 UBM mixture sizes. However, the PSD and COH feature results had a more pronounced decrease in performance at the lowest LDA dimension. Despite the disparate results at the lowest LDA dimension, reducing the dimension from 49 to 45 showed little adverse impact to the reported score for any configuration across the feature sets.

The TUH-EEG normal data had 50 subjects which limited the number of LDA dimensions. This was not the case for the 109 subject PhysioNet Database dataset, Figures 5.31–5.33, which allowed the full range of the LDA dimensions, Table 5.1. With the inclusion of LDA dimensions, the performance of CEP features was the worst of all three feature sets. As the TVMs were reduced via LDA, the performance for each configuration decreased until rebounding at 15 dimensions. The 32 mixture UBM performed the worst with the 64 mixture UBM only surpassing it for TVMs of dimension 25, 50, and 75. The 128 mixture UBM produced the top score overall of 0.89 for a TVM of dimension 25 reduced to 15 and matched the best performance of the 64 mixture UBM with a score of 0.83 for a TVM of 50 reduced to 45 dimensions.

The PSD features produced acceptable performance when using a TVMs of dimension 75, 50, and 25. However, the larger TVMs were inconsistent with the 64 mixture UBM being optimal with a TVM of dimension 100 and the 32 mixture being best with a TVM of dimension 200. The 128 mixture UBM was notable because a TVM of dimension 100 improved when reduced to LDA dimensions of 45 and 30. All TVM configurations larger than 50 showed significant performance loss when reduced to 15 dimensions.

Conversely, the COH features performed well for TVMs dimensions of 100 or smaller. Even the TVMs of dimension 200 showed promise in the higher LDA dimensions. Aside from the largest TVMs, all configurations produced near perfect scores until reduced to an LDA dimension of 15. These features were the clear favorite of the PhysioNet Database dataset in terms of performance and stability over the LDA dimensions.

The TUH-EEG seizure dataset performance as a function of LDA dimensions, Figures 5.34–5.36, highlighted the difficulty of using CEP features. Performance appeared to improve with the larger UBM mixture sizes and larger LDA dimensions. Only the 128 mixture UBM was able to exceed a score of 0 when using a TVM of dimension 25 with 15 LDA dimensions with a score of 0.0688. This was anticipated as the original UBM mixture sweep indicated the CEP features performance peaked at 1024 mixtures, Figure 5.25, with a score of 0.27.

The PSD features, however, peaked within the 32 to 128 mixture range producing a number of acceptable scores for LDA reduced I-Vectors. Despite this, every tested configuration showed gradual performance loss as the LDA tested small dimensions. The results, Figure 5.35, showed an increasing performance gap among the UBMs as the TVM dimension was increased. The performance of each UBM when using a TVM of dimension 25 was identical. As the TVM dimension increased, performance of the UBMs' scores widened. This also occurred during the LDA dimension reduction, with larger differences at the lower dimensions.

The trend of improved performance with larger sized UBMs, TVMs, and LDA dimensions continued with the COH features. While not nearly as pronounced as those in the PSD features, the 32 mixture UBM was competitive for only the TVM of dimension 25 and outpaced by the 64 mixture UBM for second place against the 128 mixture UBM results. Overall this feature set was unable to break to the 0.75 threshold, as the PSD features did, but they did perform better than the CEP features.

Given the TUH-EEG normal dataset experiments showed an early decrease in performance before rising across the 8 to 32 mixture UBMs, an alternative set of LDA sweeps were performed, Figures 5.37–5.39. Their results showed that regardless of feature set, a TVM of dimension 25 outperformed all other configurations at these mixture sizes. For the larger TVM dimensions, the 32 mixture UBM provided superior performance when paired with PSD and COH features. The case for CEP features was the opposite with the 8 mixture UBM producing superior scores with larger TVMs and LDA dimensions. However, changes to any of the UBMs via LDA, regardless of feature set, showed minimal impact. This suggested the UBM modeling had generally been a failure, with only the 32 mixture UBM showing major shifts in performance as the LDA dimensions reached 15 for the PSD and COH features.

In contrast, the TUH-EEG seizure data needed larger UBMs to maximize performance, Figures 5.25–5.27. This was explored in a set of expanded UBM mixtures, Figures 5.40–5.42. The CEP features showed improvement from the larger UBM mixtures, but it was not enough to exceed a score of 0.5. This suggested the larger mixtures had captured additional information making them distinct from each other.

The PSD features met and exceeded the 0.75 threshold, but the improvement of the 256 and 512 mixture UBMs was marginal compared to the scores of the 128 mixture UBM. Only when a TVM of dimension 200 was used did the 128 mixture UBM fail to be competitive with its counterparts, but none of these exceeded a score of 0.75. The COH features showed the 128 mixture UBM outperformed the other mixtures when the TVM dimension was at or below 75. It was competitive at a TVM dimension of 100, and struggled to break a score of 0 at a dimension of 200. Thus these two features were not readily gaining new insight into the datasets as their performances tracked across TVM dimension.

5.2.4 Constraints

Performance over each feature set was improved by increasing the number of samples in the training data. However, each sample assigned to the testing data was removed from the training and enrollment dataset. This trade-off is common, but had unique considerations given the modeling of the datasets during UBM generation and BW statistics estimation. The UBM needed to be given enough samples to at least exceed the desired number of mixtures. However, there was no assurance that the samples would be unique to the point of requiring the number of desired mixtures. This was why using the 2s duration epochs was critical. It provided a sufficient number of epochs for the training, enrollment and testing datasets.

By splitting the 2 minute recordings into 2s epochs, 60 epochs were available from each channel. At 19 channels per subject, this provided, at minimum, 1140 epochs per subject in total or a split of 266 epochs for the testing data and 874 epochs for the training/enrollment data. These values were outside the range of UBM mixtures tested, but this was acceptable because the UBMs were trained using the aggregated subject training/enrollment dataset. Therefore, a given UBM constructed its models using at least 43,700 epochs if built on the small 50 subject TUH-EEG datasets. With a 2048 mixture UBM this provided more than 20 times the minimum number of epochs required. Thus the suggested over-fitting int the epoch duration sweeps of Section 5.1.3.2 was not well founded.

Once constructed the UBMs were used to guide the BW estimation process. The extracted statistics generated log-likelihoods of each UBM mixture's presence in the given dataset. If presented with data lacking in epochs or data lacking in diversity of epochs, very low probabilities were returned. Therefore the number of testing epochs was important, not because a statistical threshold needed to be exceeded, but more epochs ensured an accurate representation of the testing subject.

The same functions were used to model the UBMs and TVMs as were used to extract data for I-Vector generation. Given the occurrence of these near zero probabilities, the software would generate poor matrices in terms of `NaN` or `Inf` values, or poorly conditioned ones that could not be inverted. This behavior required the BW statistical estimation to have an artificial floor, a minimum likelihood to avoid computational under and overflows. This impacted the construction of the UBM and TVM because they relied on the same log-likelihood generator so changes to this floor would likely impact performance making it an untested parameter.

Another constraint on the I-Vector generation process was that the original TVM I-Vectors were not produced; only I-Vectors corresponding to the LDA dimensions were generated. This was done to ensure the I-Vector dimension was was always less than the number of tested subjects. However, the TVM dimensions was similarly constrained to be one less than the number of subjects to protect against a GMM being built on a per subject basis. However, in instances when the number of subjects exceed the rows of the TVM, the TUH-EEG seizure dataset and UBMs with less than 8 mixtures, the solution was likely overdetermined. Thus the extreme cases of few subjects large UBMs and many subjects small UBMs had the potential to produce poorly conditioned TVMs.

This was likely why smaller TVMs dimensions performed poorly when paired with large mixture sized UBMs. Using only 25 or 50 elements to control the means of over 1000 mixtures became difficult when the dataset was large and diverse, Figure 5.25, compared to a smaller procedural dataset, Figure 5.19. This happened across all feature sets, but appeared to be influenced more by the dataset as the TUH-EEG seizure dataset was the only one capable of leveraging the 200 dimension TVM across all UBMs.

5.3 Conclusion

The results of this work indicated that, although the choice of feature set affected subject verification, epoch duration was a stronger predictor of performance. While reducing the epoch duration generated more epochs for the training and enrollment datasets, it did not change the number of epochs in the testing datasets for the sweep experiments, Figures 5.5–5.7. Expanding on this with the TUH-EEG normal dataset, Figures 5.13–5.15, showed that adding two epochs to the testing dataset (and removing two from the training/enrolllment dataset) improved performance, but shifting from 4 10s duration epochs to 4 2s duration epochs was not an improvement across all feature sets.

Once all epochs were 2s in duration, increasing the number of testing data epochs showed improved performance across all feature sets, Figures 5.16–5.18. When the TUH-EEG seizure dataset explored 4 epochs of 10s and 2s duration, the 10s epochs met or exceeded the performance of the 2s epochs, Figures 5.22–5.24. And again, when using only 2s duration epochs an increase in the number of testing data epochs resulted in improved performance across all feature sets, Figures 5.25–5.27.

It was thus re-affirmed that for PSD features an epoch duration of 2s was better than 10 seconds [193]. However, it appeared that by decreasing the epoch duration the overall number of epochs increased which allowed better UBMs to be constructed. Minor shifts in the number of testing data epochs, 2 to 4 or 4 to 6, had minimal impact on overall performance when the epoch duration was held constant. Conversely a constant number of testing epochs could be improved by lowering the epoch duration. Despite the intent to resolve the impact of epoch duration and testing epochs, it was clear far more expansive experiments were necessary.

Understand the role of epochs would have also required a deeper understanding of the datasets and feature sets given the impact they have on each other, Figures 5.43–5.45. In these figure, where the UBM mixtures were swept using a TVM dimension of 25 with an LDA dimension of 15, performance was shown to be dependent on dataset. The aggregated datasets were split up into sets of two (blue) and three (red), with the extra AbnNrm dataset being assigned red as well. The CEP features showed improved performance at larger UBM mixtures, while the

PSD and COH features performance waned with larger UBM mixtures. Critically, the best performing aggregated datasets, AbnMot and NrmMot datasets, contained the PhysioNet Database dataset and the worst contained the TUH-EEG seizure dataset for the PSD and COH features. Whereas the CEP features performance subverts this with the AbnSzr dataset performing as well as the AbnNrm dataset. The top performances of the PSD and COH featrue sets had the fourth and fifth highest scores when using CEP features. For all feature sets, the AbnNrmMot dataset had the third highest score.

Even by combining the datasets, it remained difficult to determine the main driver (feature, epoch, dataset) of performance with respect to each algorithm. Evaluating each dataset independently was thought to be a way to isolate these components, but in hindsight it seemed the experiments were not exhaustive enough to form a consensus on the ideal dataset parameters. Ideally, the number of experiments would have quadrupled by performing sweeps over each epoch duration (1s, 2s, 5s, and 10s) in addition to an incremental number of epochs per testing dataset (1, 2, 4, 6, 8, 10, 12, 14). The entirety of this was impracticable, but Research Aim 1 could be partially address through Experiment 4: Algorithm Benchmarks based on the refinements to the I-Vector algorithm gained within these Parameter Sweeps.

A poorly chosen epoch duration appeared to negatively impact an otherwise acceptable feature set, Figures 5.5–5.7. This made it difficult to select an appropriate classifier-feature pairing in the La Rocca Based Protocol Experiments which was why the follow-on Identity Vector Parameter Experiments used all three feature sets. Through the course of these experiments it was noted that each feature had a natural affinity for a specific type of EEG dataset: the TUH-EEG normal and CEP features, Figure 5.16; the PhysioNet Database and the COH features, Figure 5.21; the TUH-EEG seizure data and the PSD features, Figure 5.26. This was not

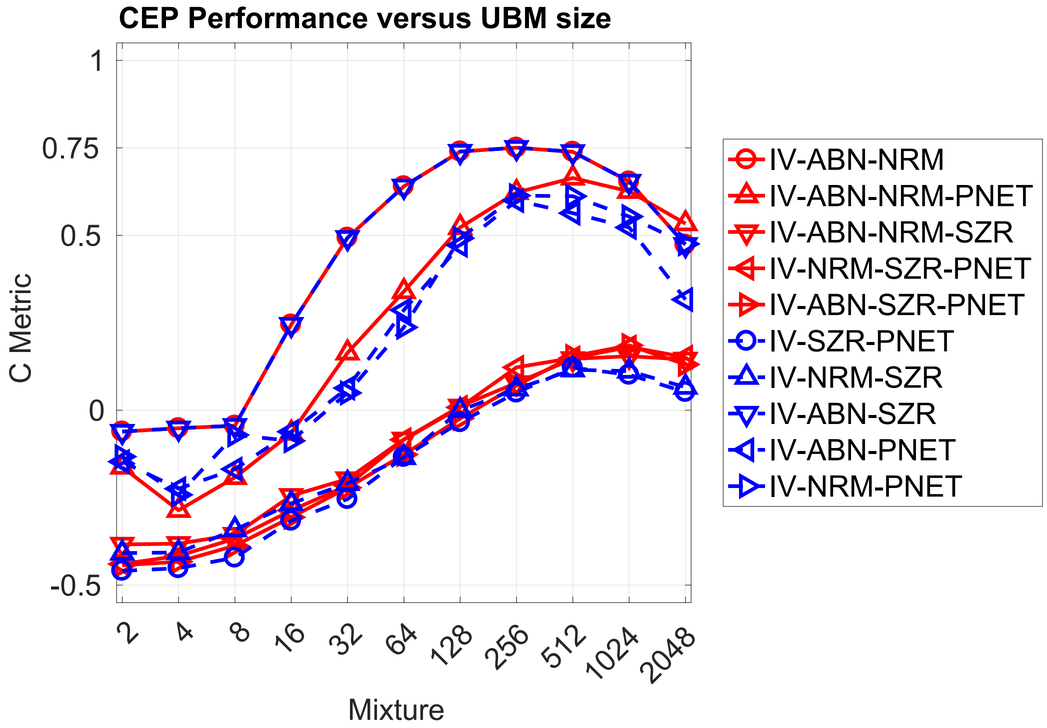


Figure 5.43. I-Vector on aggregated datasets using CEP. This C Metric plot shows the CEP based aggregated datasets performance as a function of UBM mixture size. This data was drawn from the 4 epoch set of 10s duration.

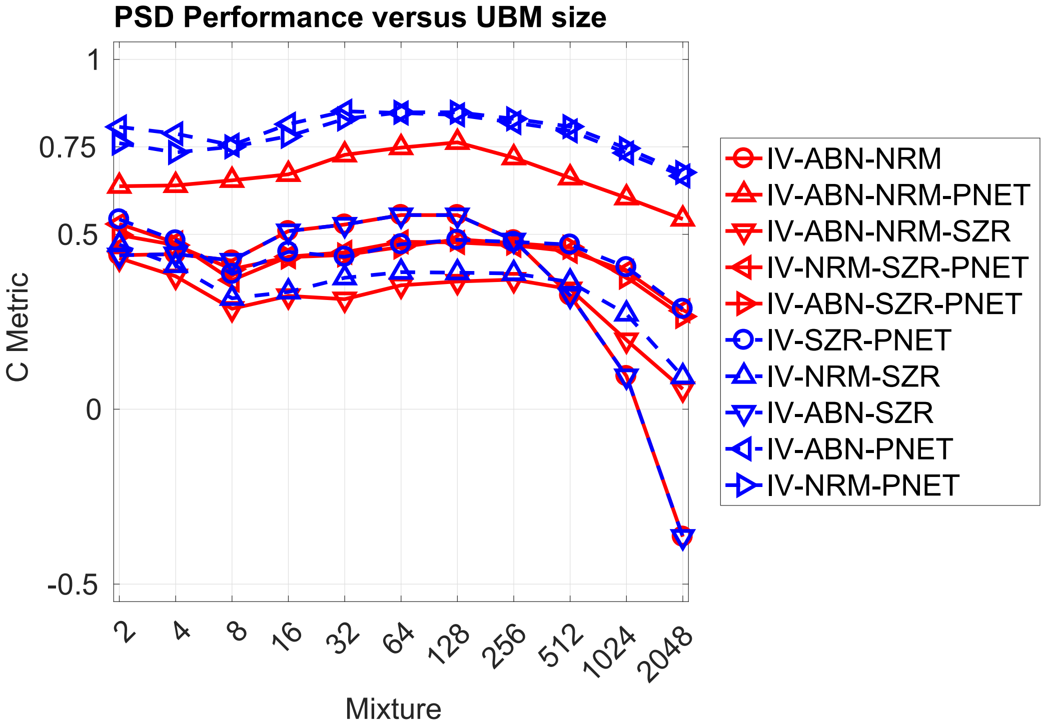


Figure 5.44. I-Vector on aggregated datasets using PSD. This C Metric plot shows the PSD based aggregated datasets performance as a function of UBM mixture size. This data was drawn from the 4 epoch set of 10s duration.

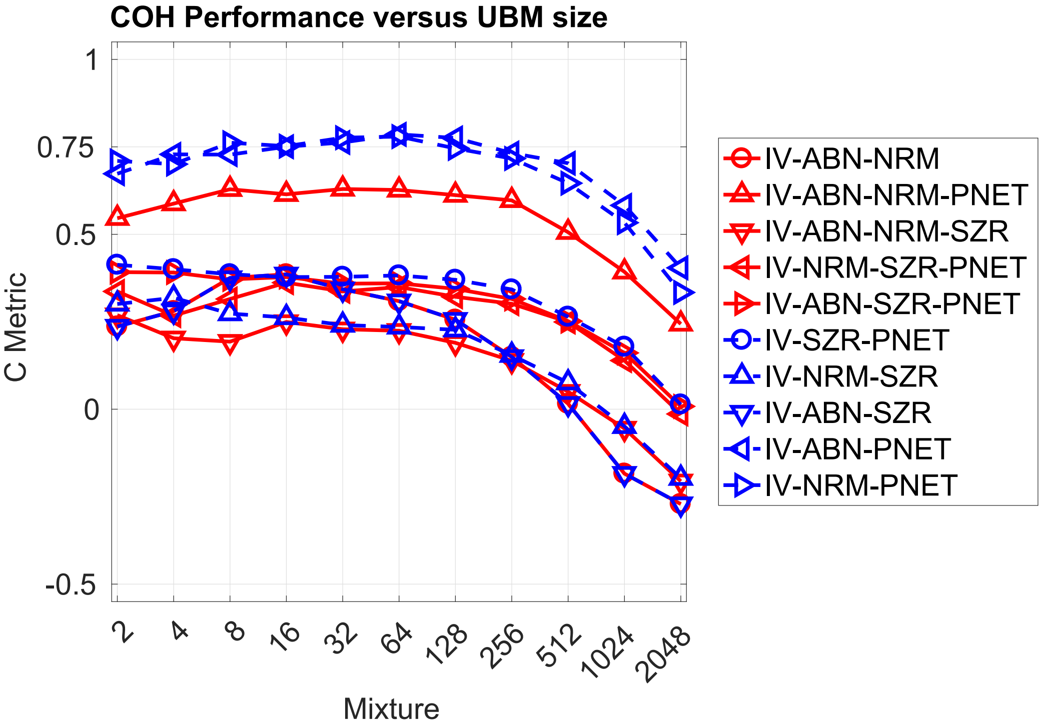


Figure 5.45. I-Vector on aggregated datasets using COH. This C Metric plot shows the COH based aggregated datasets performance as a function of UBM mixture size. This data was drawn from the 4 epoch set of 10s duration.

unexpected as many studies have attempted to improve performance though novel feature sets [64, 81, 98, 110].

Epoch duration was known to influence performance [13, 42, 48, 91] and did in these experiments, but feature selection and UBM mixture size appeared to have a more dominant role in classification. Given a single testing epoch, I-Vectors were able exceed a score of 0.75 on the TUH-EEG normal dataset using CEP features, Figure 5.10. I-Vectors were able to exceed a score of 0.75 on the PhysioNet Database dataset with a variety of COH feature epochs because performance appeared dependent on TVM and UBM mixture size, Figure 5.19. Acceptable performance on the TUH-EEG seizure dataset was dependent on increasing the UBM mixture size and using as many epochs in the testing dataset as possible. Therefore the 2s duration epochs were set as the basis for the UBM-TVM Relationship experiments with 12 epochs per testing dataset. The number of epochs was reduced from 14 because it was felt the two shorter non-motion resting trials weakened the divergent nature of the four datasets (abnormal, normal, seizure, and motion).

To say this applied to all EEG circumstances would be premature as there was no sleep data [141], no P300 data [138], nor emotion/workload data [107] included in the Parameter Sweeps. Additionally, accounting for longitudinal data was not addressed despite its presence in the TUH-EEG seizure data. As a whole the tested data represented new ground within EEG subject verification, but left much to be desired for classifying beyond subjects. The PhysioNet Database and some of the TUH-EEG data lacked external documentation to characterize the subjects (age, gender, handedness, etc) or their conditions with complete clinical annotations. Ideally all of these conditions would be met which would enable such a dataset to enable a comprehensive analysis of features, algorithms and parameters.

Therefor the best path forward was to aggregate the datasets into groups of two or three in an effort to provided an unbiased experimental protocol. This increased the complexity of the classification task via larger subject counts, 50 to 570, and their associated statistical patterns, such as motion and seizure phenomena. The goal was to provide a more competitive set of benchmarks for evaluating the three classification algorithms and feature sets. Critically, EEG data was rarely processed from such disparate sources making these result a nexus of comparison across a variety of commonly tested data types because of them being publicly accessible. Successful classification by I-Vectors could then be inferred as the missing benchmark tool capable of linking EEG subject identification/verification tests together.

A technique capable of handling such variety needed to be developed because of the multimodal nature of EEG data itself and the amount of variability from subject to subject in larger datasets [205]. The majority of signal processing techniques for EEGs have typically been focused on specific use cases such as seizure detection, BCIs, and biometrics. Specifically, investigators have sought techniques for both data optimization and modeling that were unique to their particular EEG signals. By tuning data parameters such as channel count, feature type, and recording duration, investigators have been able to successfully classify data according to subject [105] and waveform [162].

Thus the work presented here was consistent with the goals of the EEG signal processing community in that it supported techniques for subject classification [103, 64, 105, 68], BCI applications [206, 151], session variability [66, 86, 42], and noise classification [11]. The most logical application of I-Vectors would within EEG biometrics [64, 68, 86] because those align with I-Vectors' original intent of speaker verification, defined here as one-to-one matching of an unknown recording to a known recording. However, the aim was to develop I-Vectors as a broad data agnostic subject verification tool. The EEGs community had no such tool and it was clearly necessary given the diversity of solutions for classification (BCI [205], seizure detection and prediction [6], and cohort retrieval [200]) tasks found throughout the EEG processing community.

These Parameter Sweeps outlined how I-Vector parameters should be tuned for use given the tested feature sets and datasets. Ultimately, a set of UBM mixtures (32, 64, 128) was chosen instead of a single mixture given how each feature and dataset performance varied, Table 5.2. It was clear only the TUH-EEG seizure dataset took advantage of the range of TVM dimensions so they remained unchanged going forward. The LDA were also rarely influential and so they were retained but reworked to align better to the TVM dimensions. The major parameter reduction took place by resolving to use 2s duration epochs and use 12 epochs in the testing datasets. This mean the training and enrollment data would be given 48 epochs for a reasonable 80% training and 20% testing split of the data.

Table 5.2. Identity Vector Parameter Sweep, Optimized

UBM	TVM	LDA
32 64 128	25	20 15 5
	50	45 25 15
	75	70 50 25
	100	95 75 50 25
	200	195 100 75 50 25

Chapter 6

ALGORITHM BENCHMARKS

The Parameter Sweeps in Chapter 5 provided an understanding of how to configure the datasets, UBMs, TVM dimensions and LDA dimensions to apply I-Vectors for use with EEGs. A portion of Chapter 5 presented a limited comparison of the GMM-UBM, I-Vector, and MD algorithms within the context of La Rocca's experimental protocol and the Parameter Sweep protocol. From these experiments, it was found that I-Vectors performed competitively relative to MD and GMM-UBM classification using PSD features of the PhysioNet Database dataset. Results were also comparable to the performance of MD-based [7, 64, 98, 110] and GMM-UBM classifiers [42, 163] as published by other groups. Thus even without optimizing epoch duration and TVM configurations, I-Vectors provided competitive performance, suggesting strong results would be possible for the Algorithm Benchmarks reported in this chapter.

The comparative results of the Protocol Replication in Section 5.1.2 were limited to the PSD feature set and PhysioNet Database dataset. This meant fully inclusive experiments, using the CEP and COH features in addition to the PSD features and the TUH-EEG datasets in addition to the PhysioNet Database dataset, would be used for the Algorithm Benchmarks. Unlike the single-dataset experiments carried out in the Parameter Sweeps Section 5.2.1, the Algorithm Benchmarks in this chapter were applied to larger and more heterogeneous aggregated datasets, as outlined in Chapter 3 and summarized here again in Table 6.1. As established in the conclusions of Chapter 5, each of these datasets fixed at epochs of 2s duration (with 12 epochs used for the testing dataset) and UBMs consisting of 32, 64, and 128 mixtures.

Table 6.1. Combined Dataset Designations and Subject Count

Designation	Dataset 1	Dataset 2	Dataset 3	# Subjects
AbnNrm	TUH Abnormal	TUH Normal	-	100
AbnSzr	TUH Abnormal	TUH Seizure	-	461
NrmSzr	TUH Normal	TUH Seizure	-	461
AbnMot	TUH Abnormal	Physio Motion	-	159
NrmMot	TUH Normal	Physio Motion	-	159
SzrMot	TUH Seizure	Physio Motion	-	520
AbnNrmSzr	TUH Abnormal	TUH Normal	TUH Seizure	511
AbnNrmMot	TUH Abnormal	TUH Normal	Physio Motion	209
NrmSzrMot	TUH Normal	TUH Seizure	Physio Motion	570
AbnSzrMot	TUH Abnormal	TUH Seizure	Physio Motion	570

Consequently, the free parameters of the Algorithm Benchmarks in this chapter were the feature sets, TVM dimensions, and LDA dimensions. All three feature sets (PSD, COH, and CEP) were included because the CEP features performed well in Chapter 4's I-Vector development, as did the PSD features in Chapter 5's Parameter Sweeps. The COH features had worked well for La Rocca and others in bio-metric and BCI classification tasks [59, 64]. Although the larger TVM dimensions (100 and 200) had performed poorly for the majority of the tested datasets, they showed promise on the 411-subject TUH-EEG seizure data. This suggested that the largest subject pools of the aggregated datasets would benefit from the larger TVM dimensions. Likewise, the increased complexity (via the mixing of abnormal, normal, motion, and seizure data) was thought to provide an opportunity for LDA to refine the native TVMs and thus improve the I-Vector classification.

The purpose of evaluating these algorithms and parameters over the 10 unique datasets listed in Table 6.1 was to produce a strong understanding how the relationships between datasets, features, and algorithms affects I-Vector performance. This was in line with Research Aim 1: Can I-Vectors perform as well as or better than similar techniques. Although this chapter focuses on the pre-selected UBM sizes (34, 64, and 128), summary results are also presented without limits on the UBM mixture sizes. This assured the transparency of the results with the hope that these datasets would be used by others in an effort to promote a common evaluation platform for algorithm and feature development.

The experiments followed the same protocol of Parameter Sweeps (see Chapter 5) by presenting the C-Metric scores of the algorithms as a function of their UBM sizes and TVM dimensions when applicable. Even though the the MD classifier produced a single score for each dataset, this score was repeated for each mixture size, even though mixtures are not applicable to its operation. Similarly, the GMM-UBM classifier, while dependent on the UBMs, was not impacted by the TVM dimension. This was also true when analyzing the impact of the LDA dimensions, since neither MD and GMM-UBM techniques were impacted.

The only difference in methods introduced in this chapter was that a set of I-Vectors was generated for each TVM prior to any LDA. This enabled a 'raw' evaluation of the TVMs prior to refinement by LDA. This assured as direct as possible of a comparison between the three algorithms labeled as the "Native TVM Performance". In contrast, the impact of LDA on these native TVMs was labeled LDA Enhanced Performance. These LDA experiments were the final component of Research Aim 2's parameter testing while Research Aim 1 was directly addressed by the Native TVMs experiments.

6.1 Native TVM Performance

The Native TVM experiments tracked the performance of I-Vectors based on the TVM dimension. In the existing literature, I-Vectors were not typically derived directly from a TVM as post processing, like LDA, channel normalization, or length normalization, was shown to further improve performance [170, 192, 207]. While this was established practice within the speech community, such steps would have added too much complexity (degrees of freedom and algorithm choices) to the benchmarking process. The restriction on not allowing the TVM dimension to exceed the number of subjects in the dataset meant that the `AbnNrm`, `AbnMot`, and `NrmMot` had maximum TVM dimensions of 99, 158, and 158 respectively.

6.1.1 Results

Just as in Chapter 5, each dataset was evaluated using each feature set, producing a total of 30 figures. The first experiment, Figures 6.1–6.3, tested the algorithms' performances across the three feature sets when using the TUH-EEG Abnormal and Normal datasets. This combined dataset served as a basline comparison point as it consisted of smallest number of subjects, 100. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three feature. I-Vectors were also able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the CEP and PSD features. All three algorithms reported their best scores when using the CEP features.

The second experiment, Figures 6.4–6.6, tested the algorithms' performances when using the TUH-EEG Abnormal and Seizure datasets, consisting of 461 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features.

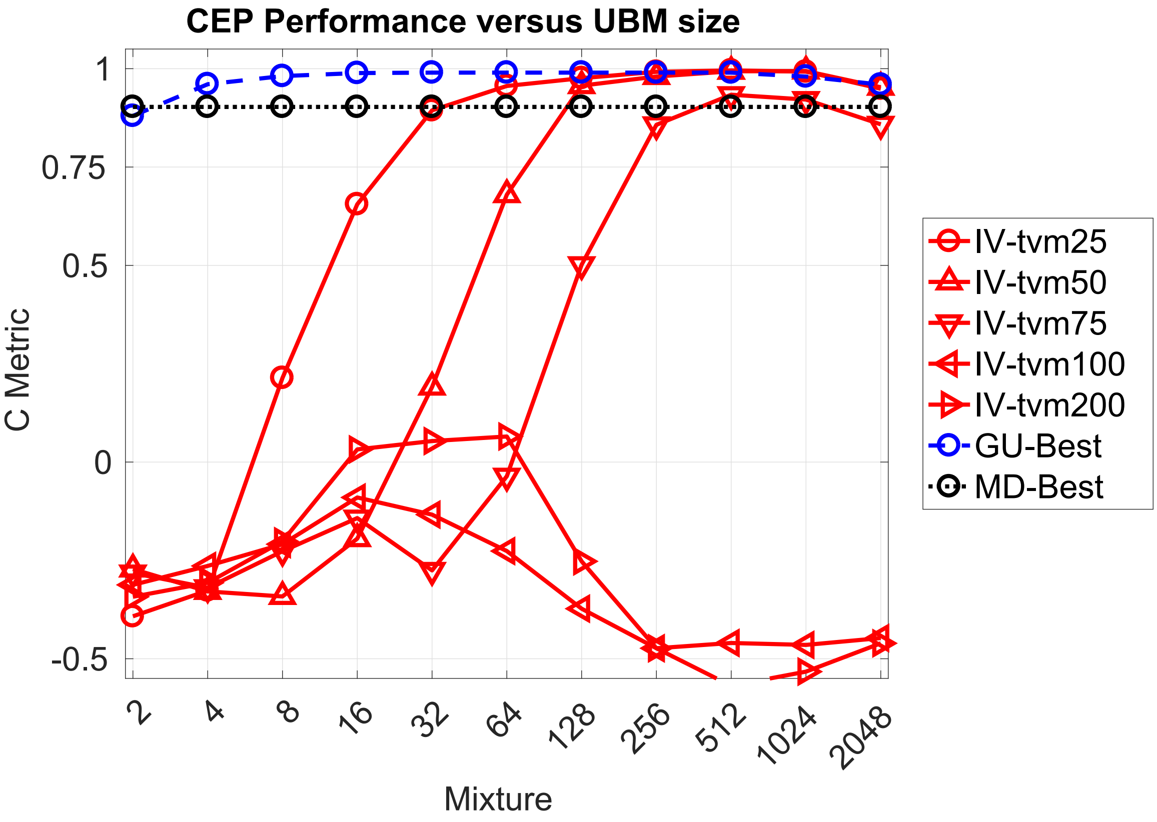


Figure 6.1. C Metric Plot of CEP `AbnNrm`. This C Metric plot shows the CEP based TUH-EEG Abnormal and Normal datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector, red solid line, and GMM-UBM, blue dashed line, results. The MD, black dotted line, results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the TVM dimensions to a maximum of 99.

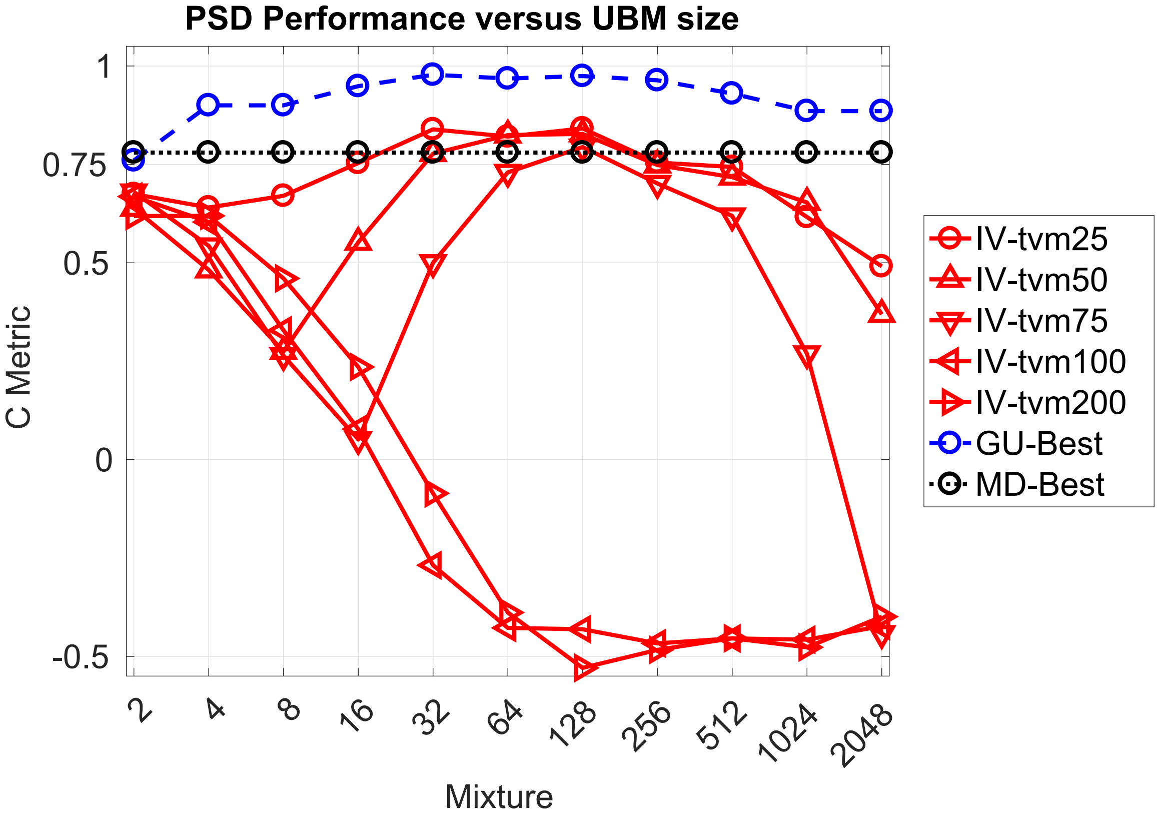


Figure 6.2. C Metric Plot of PSD `AbnNrm`. This C Metric plot shows the PSD based TUH-EEG Abnormal and Normal datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the TVM dimensions to a maximum of 99.

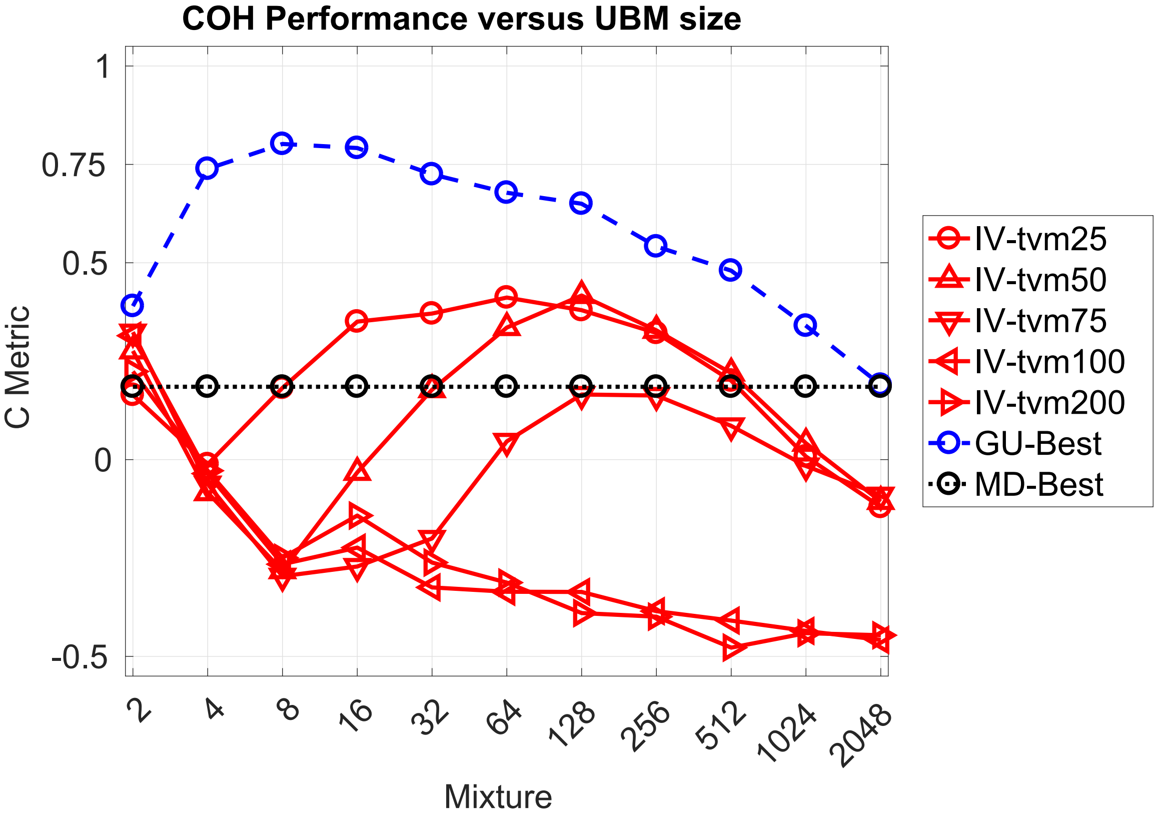


Figure 6.3. C Metric Plot of COH `AbnNrm`. This C Metric plot shows the COH based TUH-EEG Abnormal and Normal datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the TVM dimensions to a maximum of 99.

The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

The third experiment, Figures 6.7–6.9, tested the algorithms' performances when using the TUH-EEG Normal and Seizure datasets consisting of 461 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

The fourth experiment, Figures 6.10–6.12, tested the algorithms' performances when using the TUH-EEG Abnormal and the PhysioNet Database Motion datasets consisting, of 159 subjects. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD and COH features, but only at a TVM dimension of 25 for the CEP features. The MD algorithm reported its best score with the CEP features, while the GMM-UBM and I-Vectors reported their best scores with the PSD features.

The fifth experiment, Figures 6.13–6.15, tested the algorithms' performances when using the TUH-EEG Normal and the PhysioNet Database Motion datasets, consisting of 159 subjects. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the

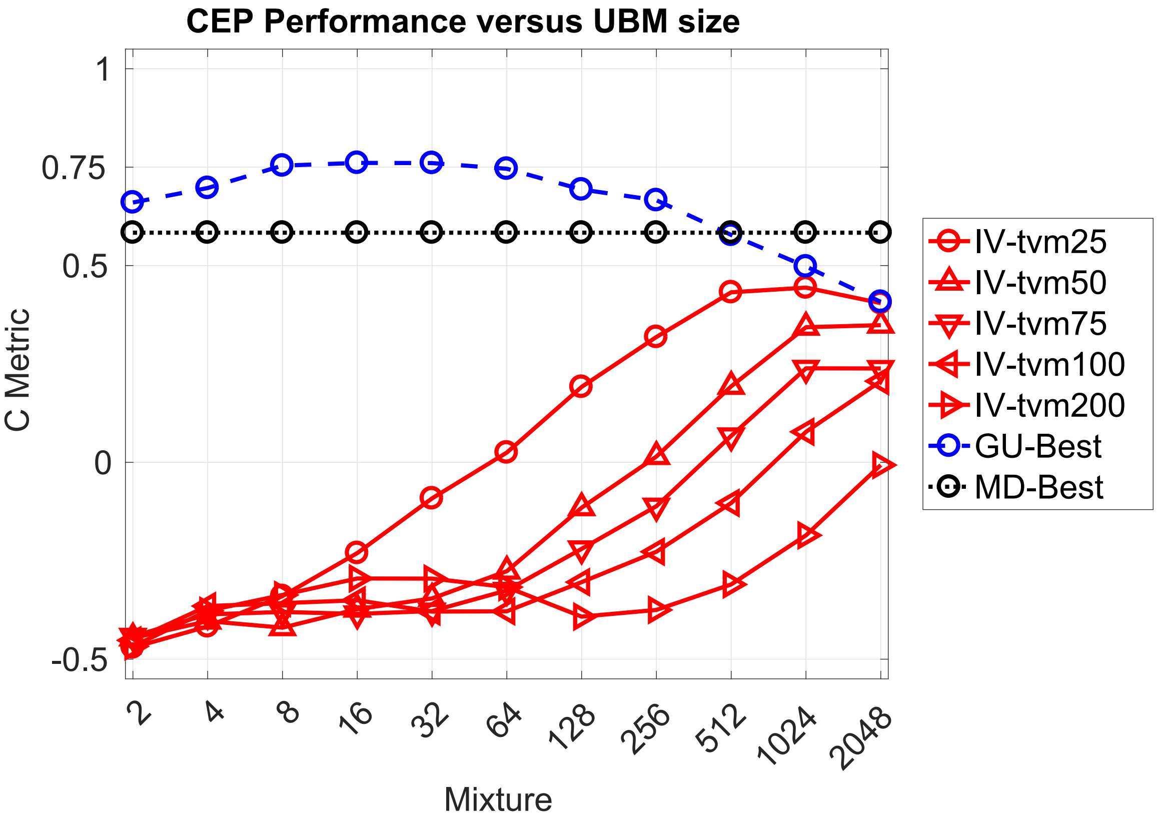


Figure 6.4. C Metric Plot of CEP `AbnSzr`. This C Metric plot shows the CEP based TUH-EEG Abnormal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

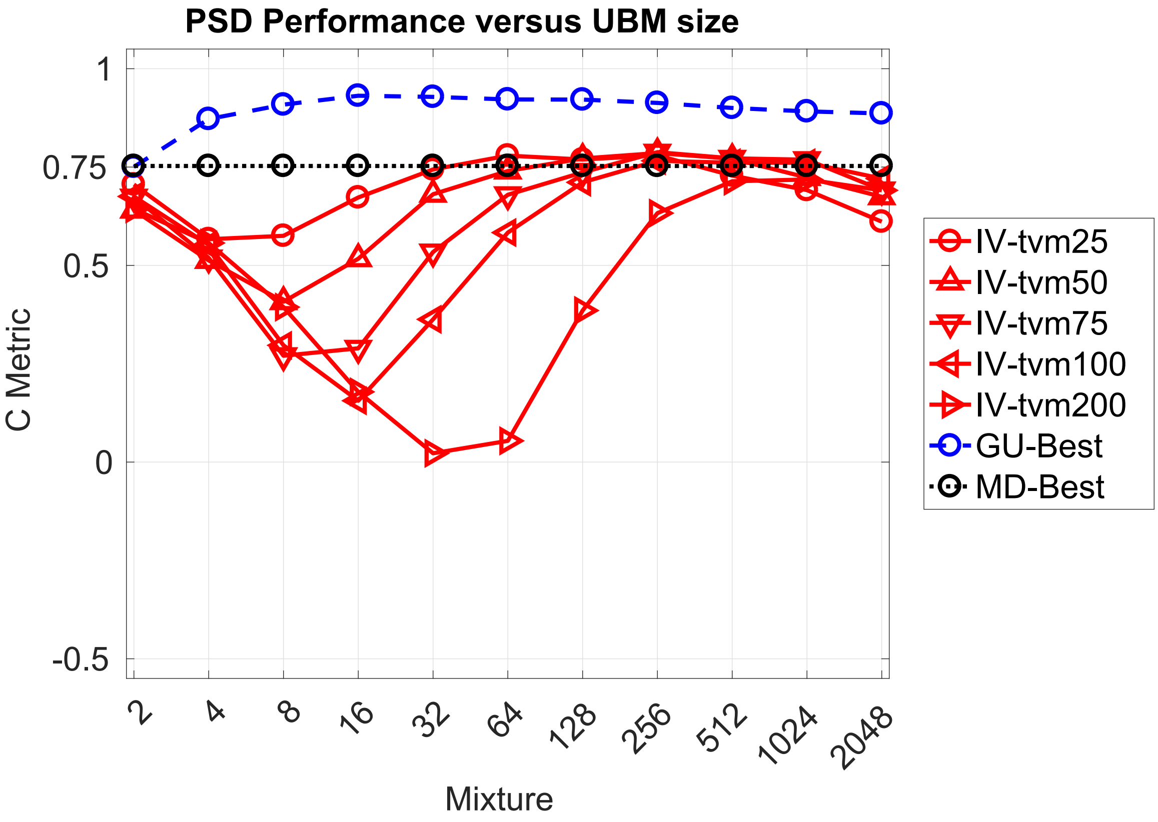


Figure 6.5. C Metric Plot of PSD `AbnSzr`. This C Metric plot shows the PSD based TUH-EEG Abnormal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

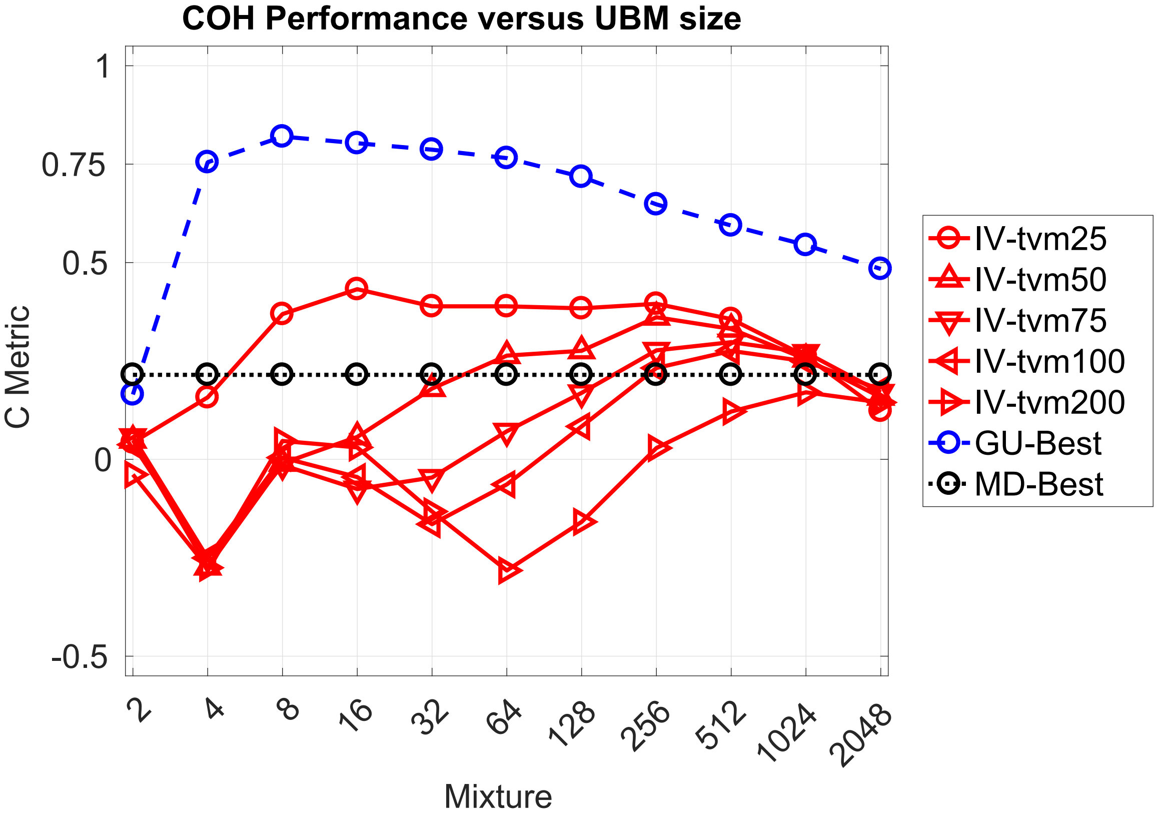


Figure 6.6. C Metric Plot of COH `AbnSzr`. This C Metric plot shows the COH based TUH-EEG Abnormal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

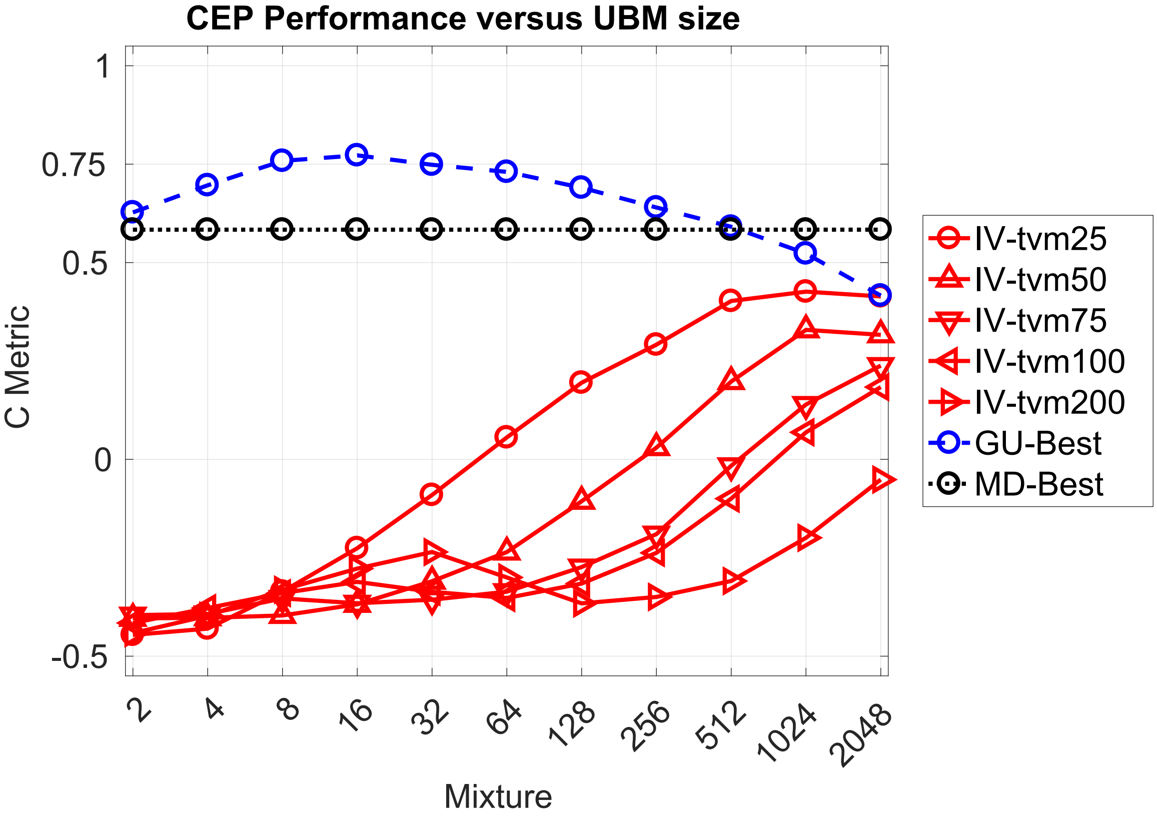


Figure 6.7. C Metric Plot of CEP `NrmSzr`. This C Metric plot shows the CEP based TUH-EEG Normal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

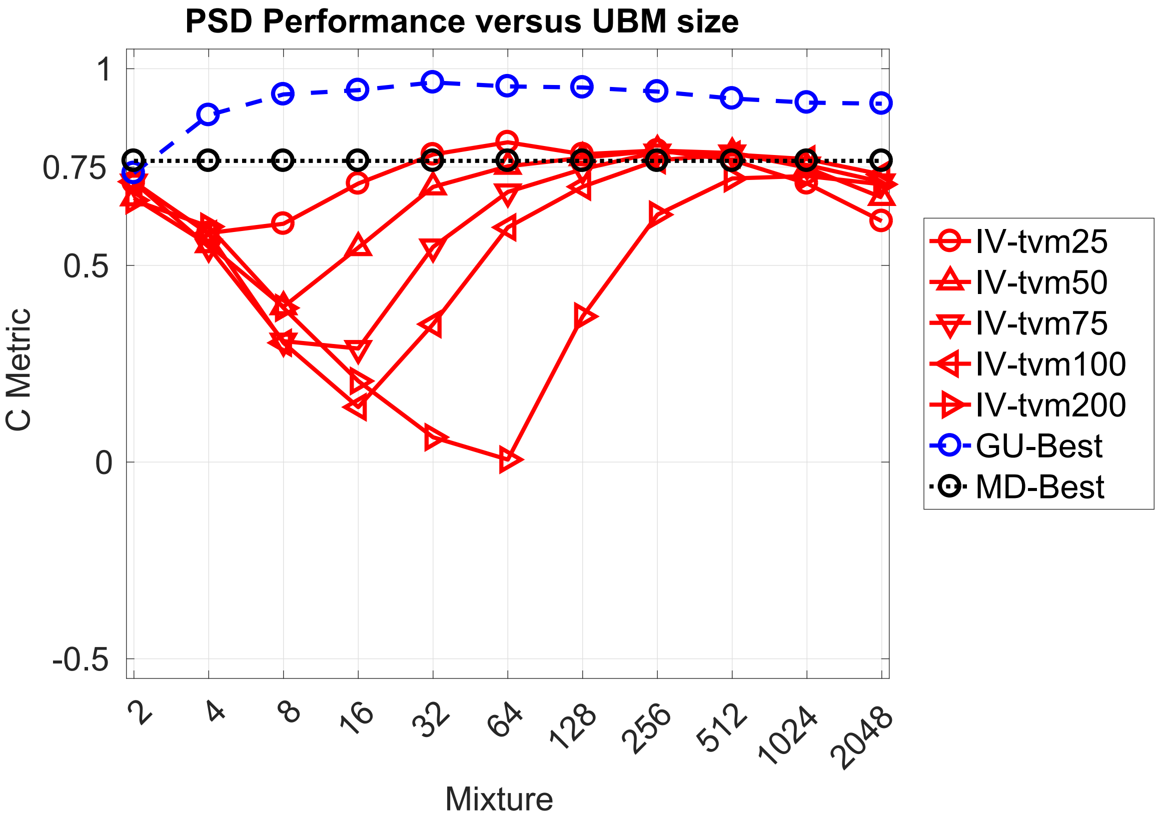


Figure 6.8. C Metric Plot of PSD NrmSzr. This C Metric plot shows the PSD based TUH-EEG Normal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

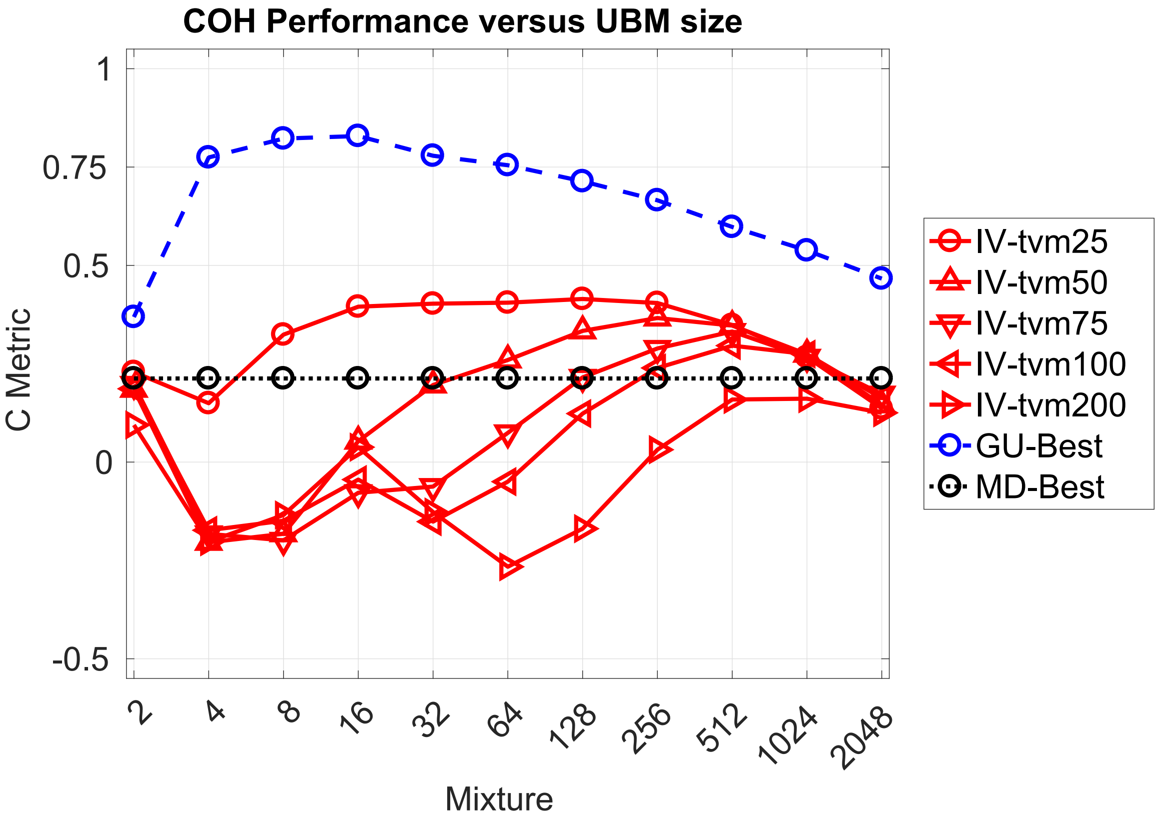


Figure 6.9. C Metric Plot of COH `NrmSzr`. This C Metric plot shows the COH based TUH-EEG Normal and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

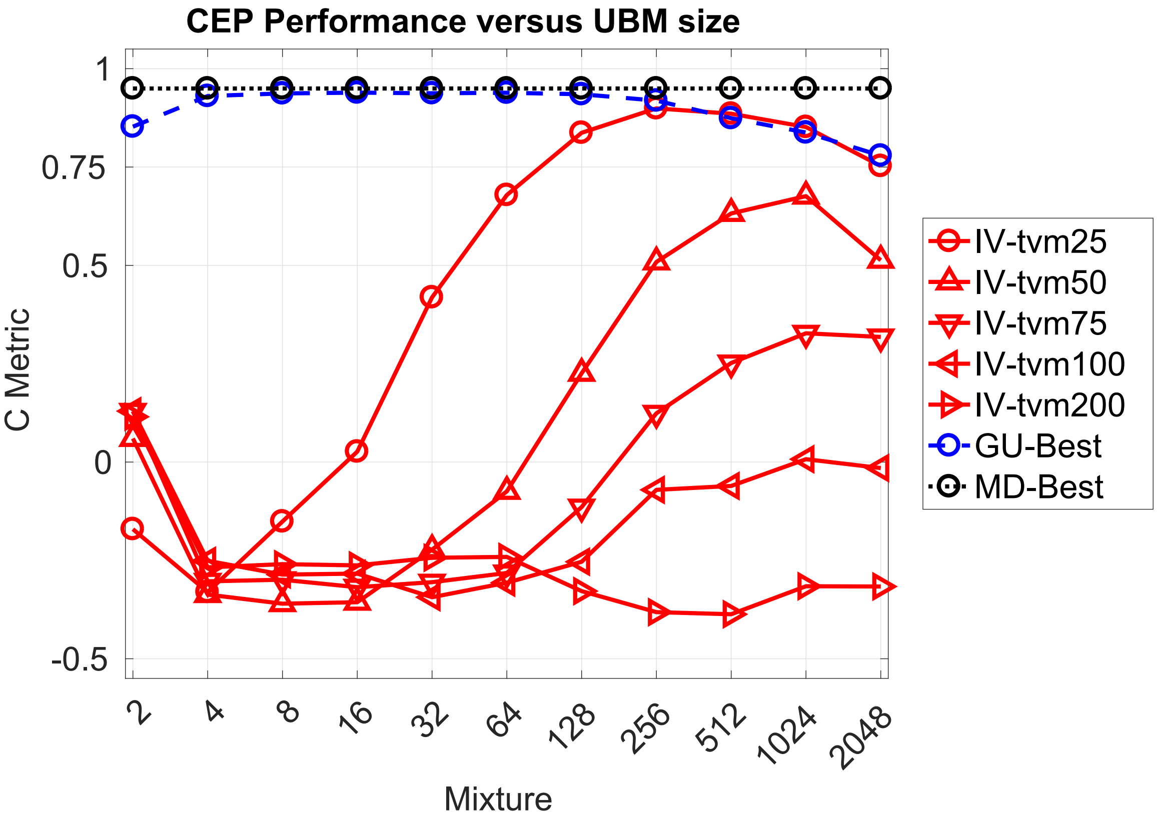


Figure 6.10. C Metric Plot of CEP `AbnMot`. This C Metric plot shows the CEP based TUH-EEG Abnormal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

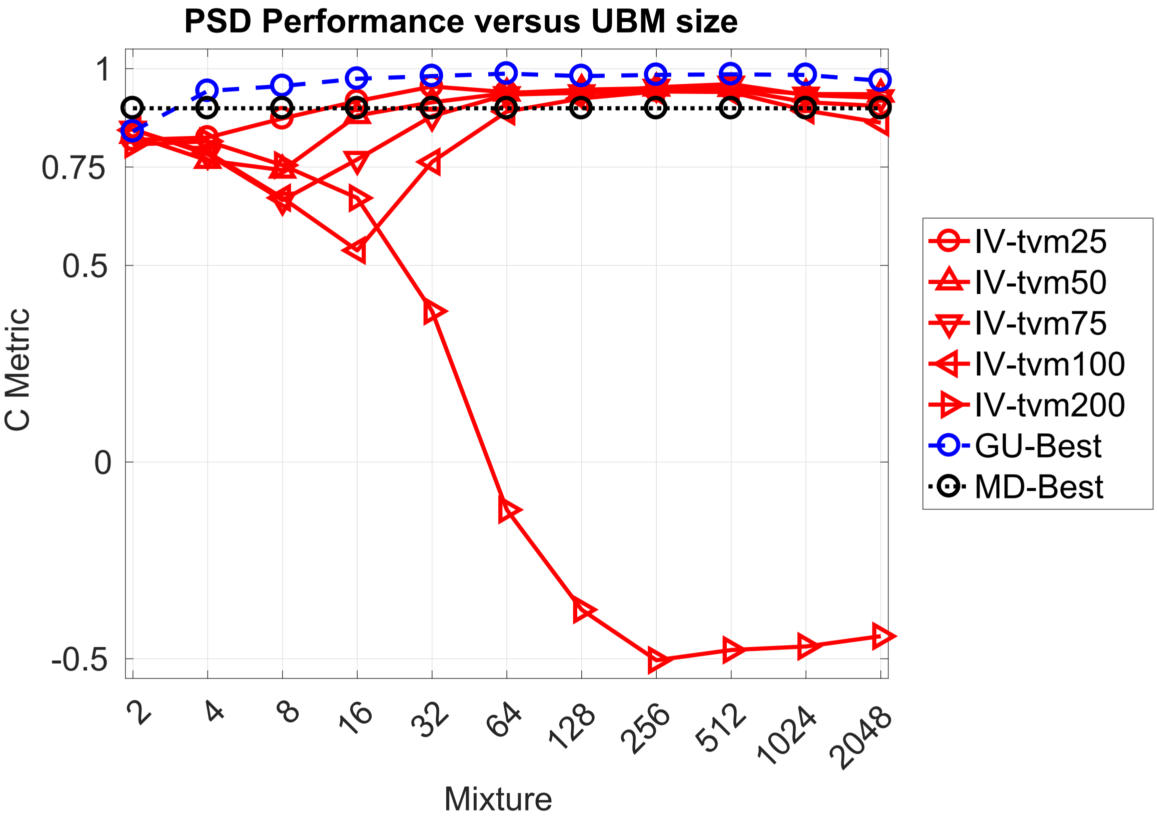


Figure 6.11. C Metric Plot of PSD `AbnMot`. This C Metric plot shows the PSD based TUH-EEG Abnormal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

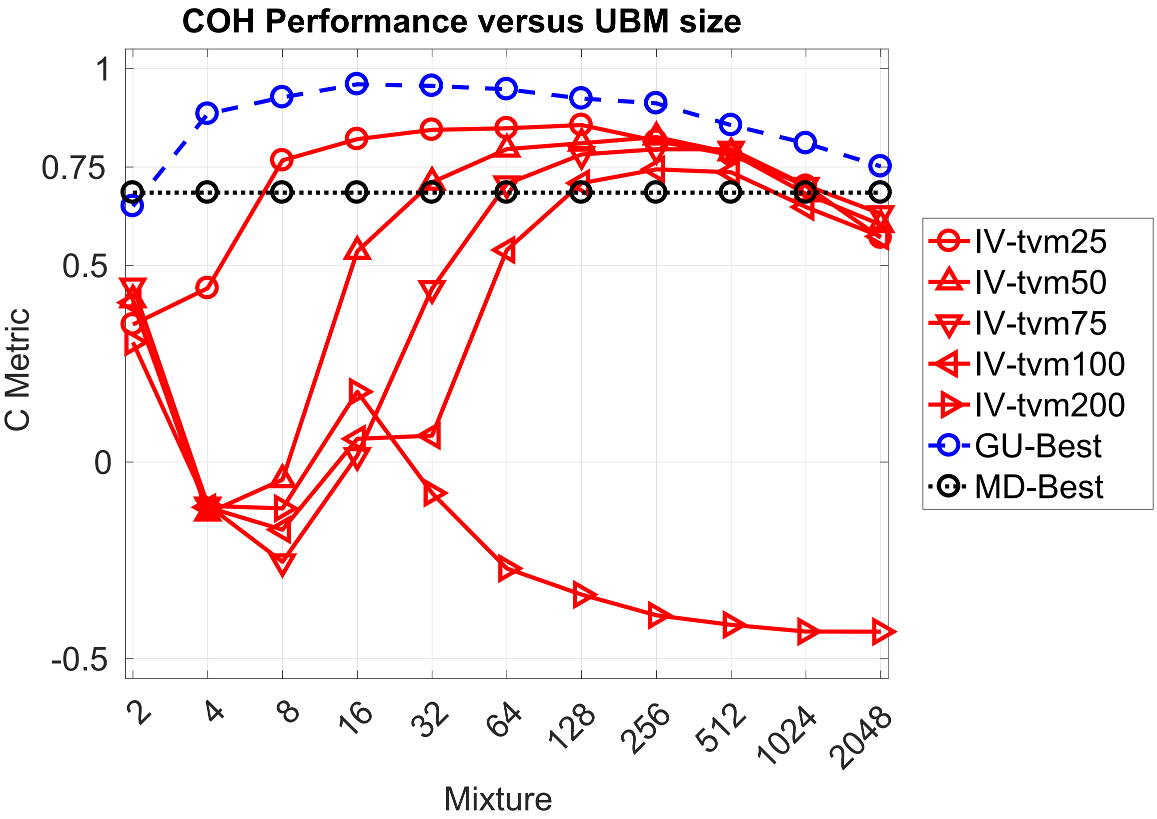


Figure 6.12. C Metric Plot of COH `AbnMot`. This C Metric plot shows the COH based TUH-EEG Abnormal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

0.75 score threshold with a variety of UBM mixtures and TVM dimensions for all features. The MD algorithm reported its best score when using the CEP features, while the GMM-UBM and I-Vectors algorithms reported their best scores when using the PSD features.

The sixth experiment, Figures 6.16–6.18, tested the algorithms' performances when using the TUH-EEG Seizure and the PhysioNet Database Motion datasets, consisting of 520 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

The seventh experiment, Figures 6.19–6.21, tested the algorithms' performances when using the TUH-EEG Abnormal, Normal, and Seizure datasets, consisting of 511 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

The eighth experiment, Figures 6.22–6.24, tested the algorithms' performances when using the TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets consisting of 209 subjects. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the CEP and PSD features. The MD algorithm

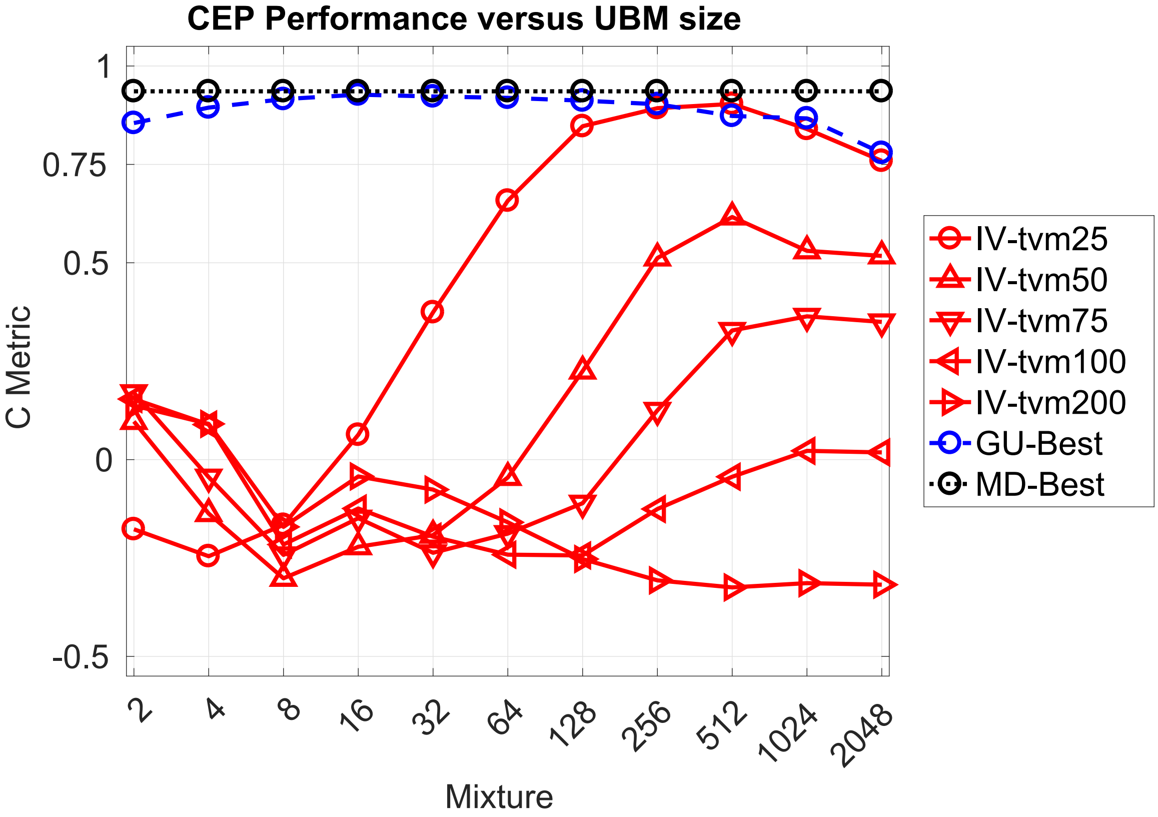


Figure 6.13. C Metric Plot of CEP `NrmMot`. This C Metric plot shows the CEP based TUH-EEG Normal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

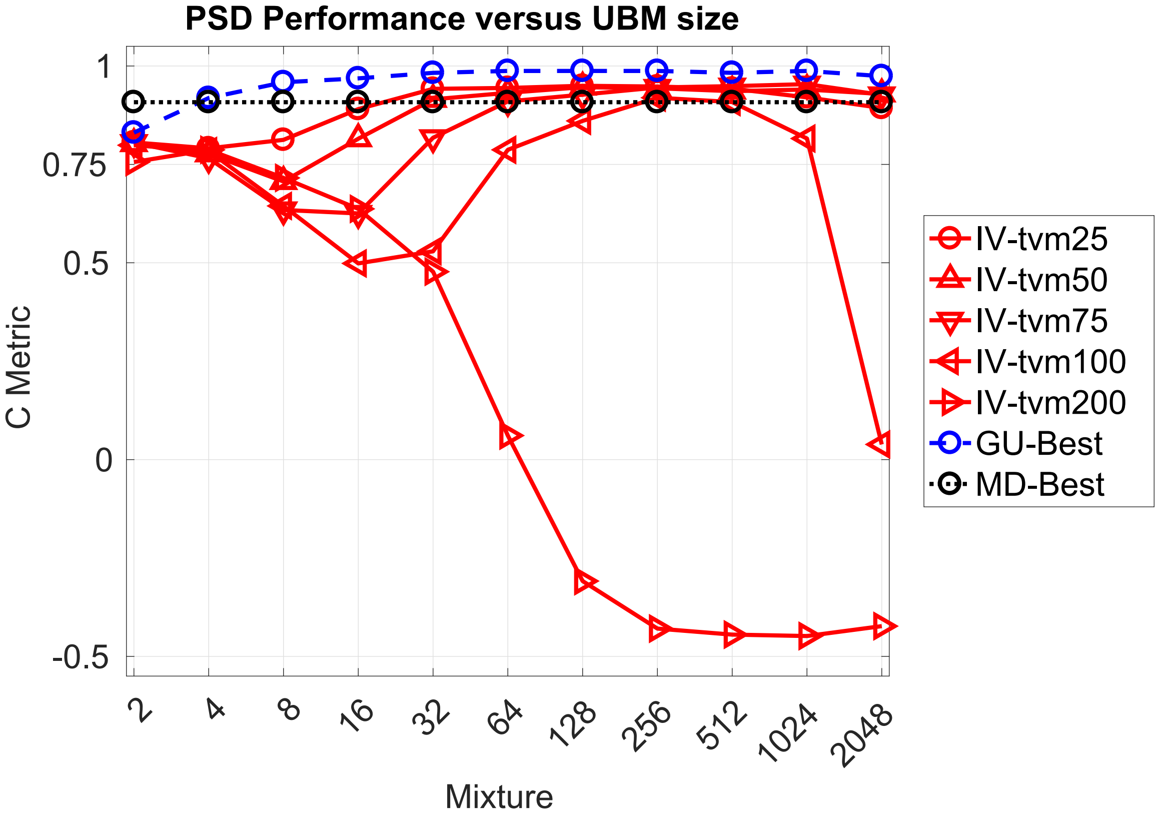


Figure 6.14. C Metric Plot of PSD `NrmMot`. This C Metric plot shows the PSD based TUH-EEG Normal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

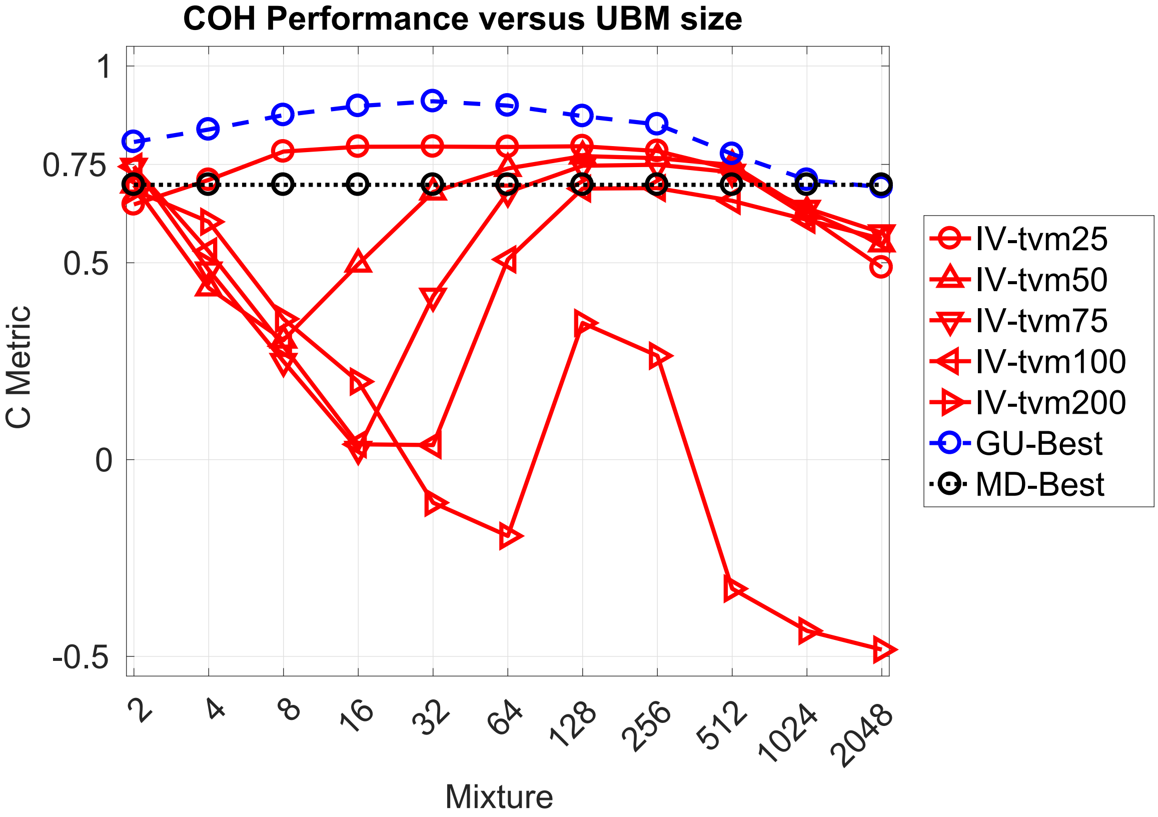


Figure 6.15. C Metric Plot of COH `NrmMot`. This C Metric plot shows the COH based TUH-EEG Normal and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the TVM dimensions a maximum of 158.

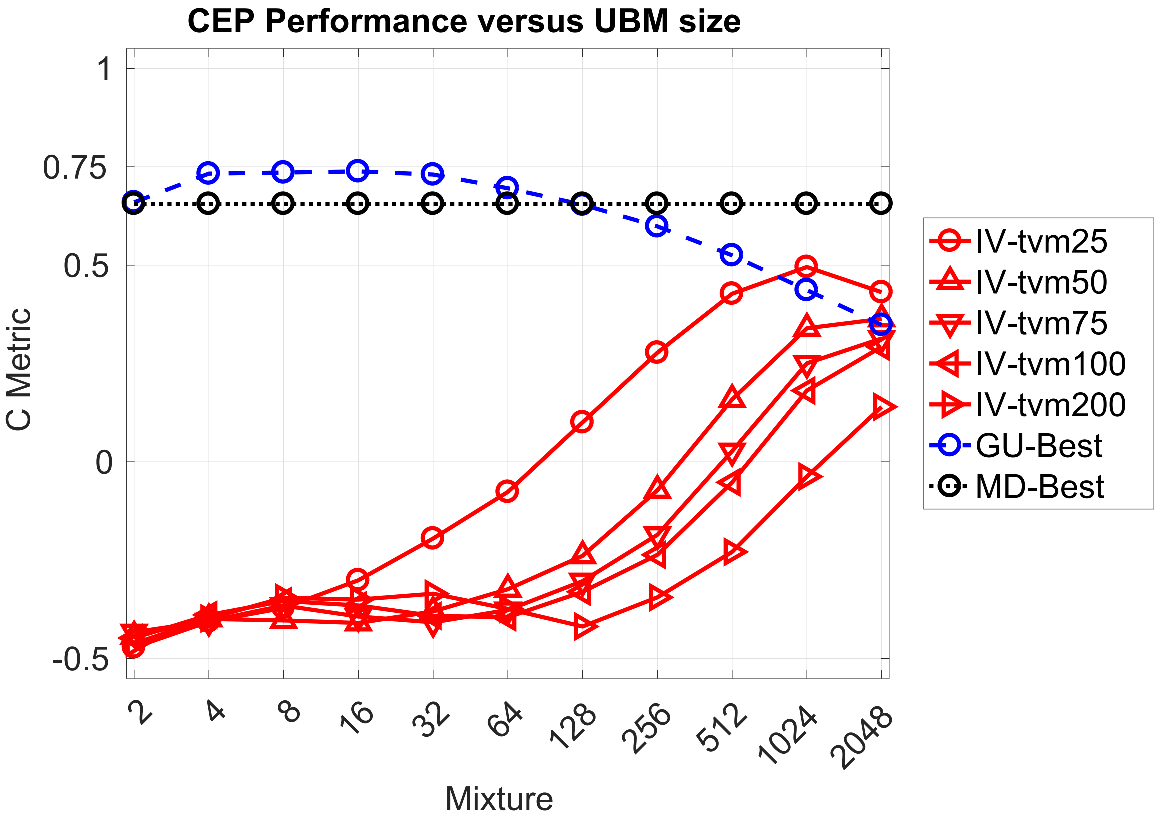


Figure 6.16. C Metric Plot of CEP `SzrMot`. This C Metric plot shows the CEP based TUH-EEG Seizure and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

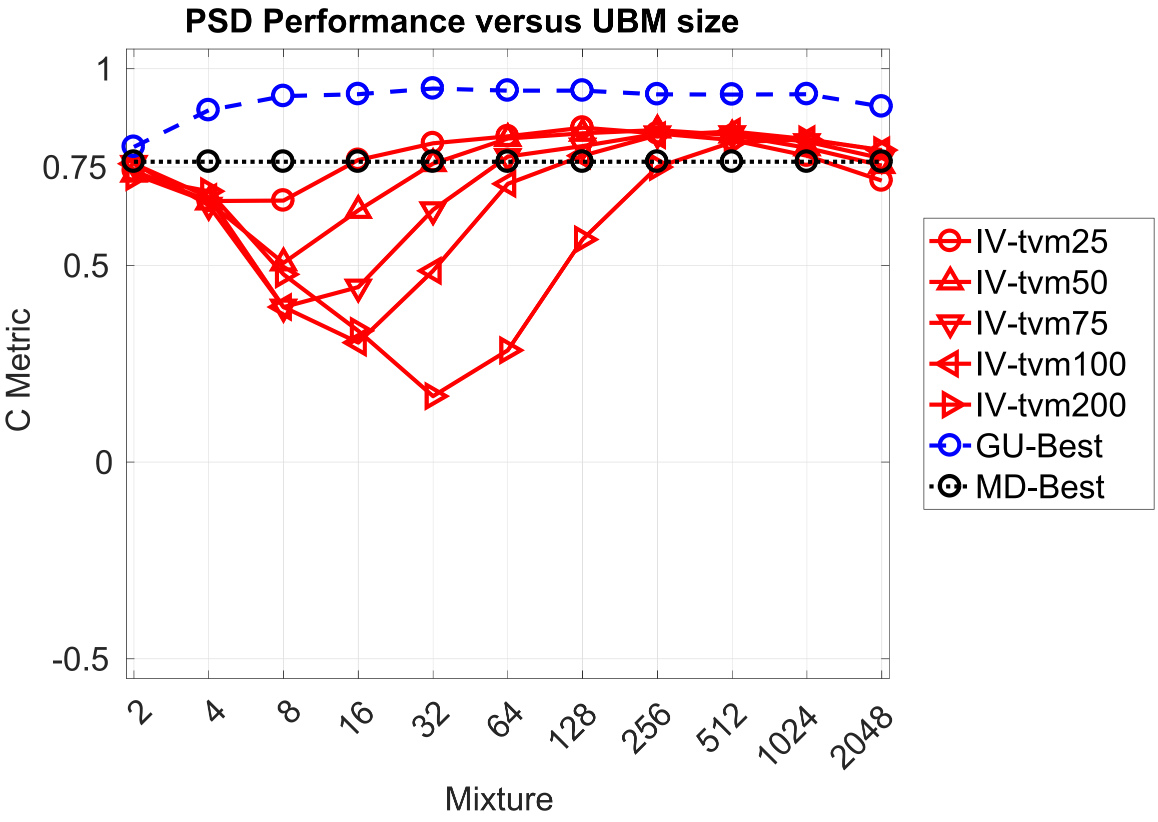


Figure 6.17. C Metric Plot of PSD `SzrMot`. This C Metric plot shows the PSD based TUH-EEG Seizure and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

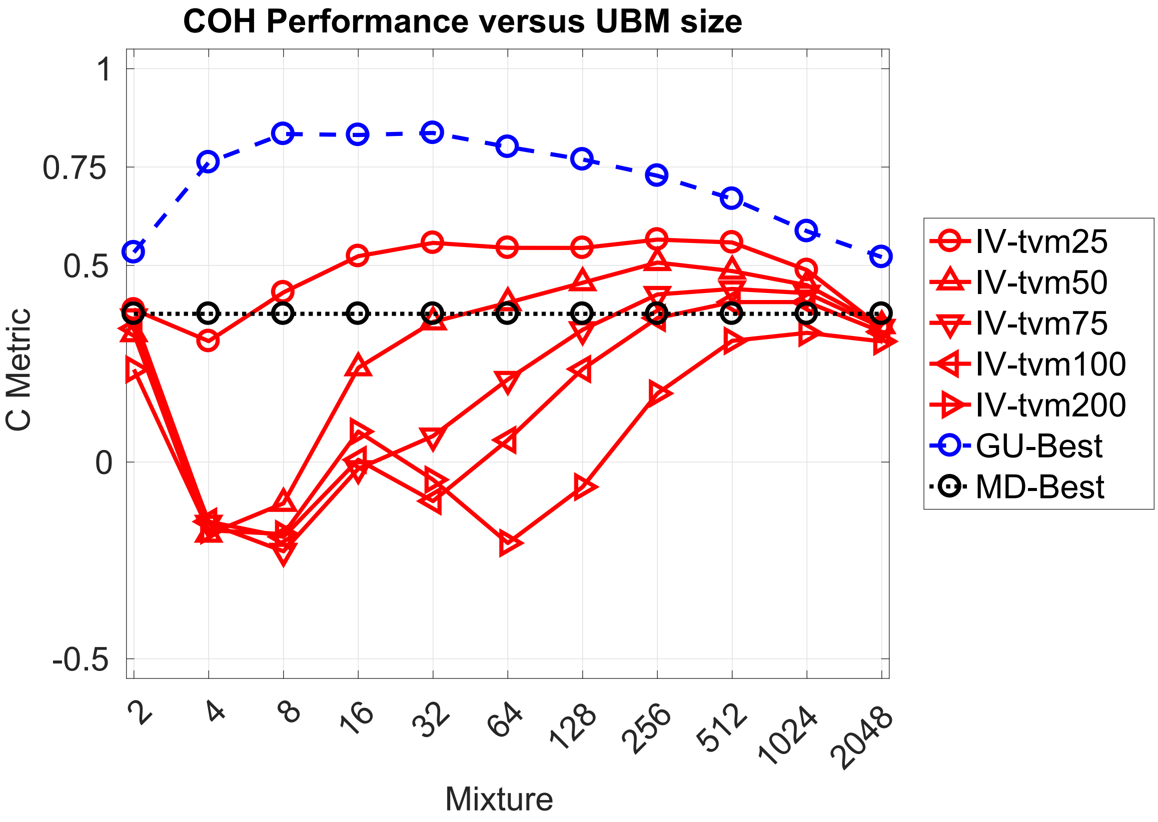


Figure 6.18. <u>C Metric Plot of COH `SzrMot`.</u> This C Metric plot shows the COH basedTUH-EEG Seizure and the PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

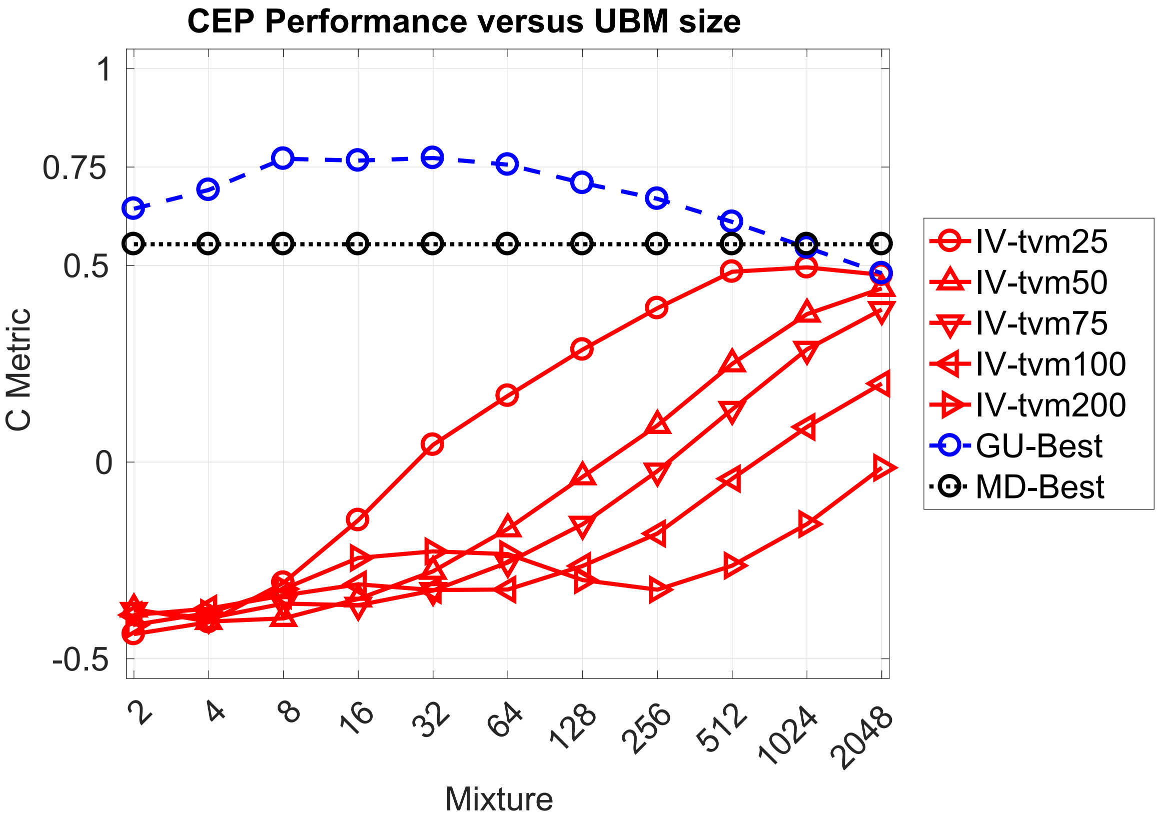


Figure 6.19. C Metric Plot of CEP `AbnNrmSzr`. This C Metric plot shows the CEP based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

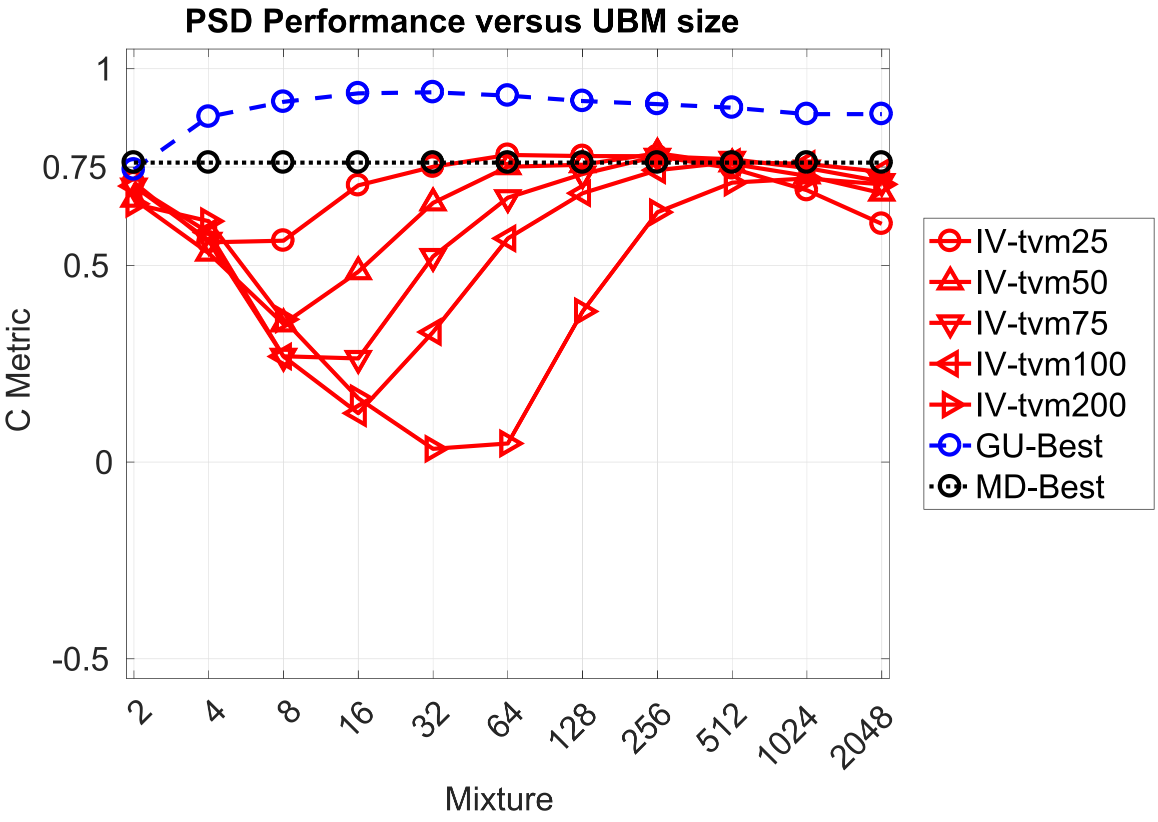


Figure 6.20. C Metric Plot of PSD `AbnNrmSzr`. This C Metric plot shows the PSD based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

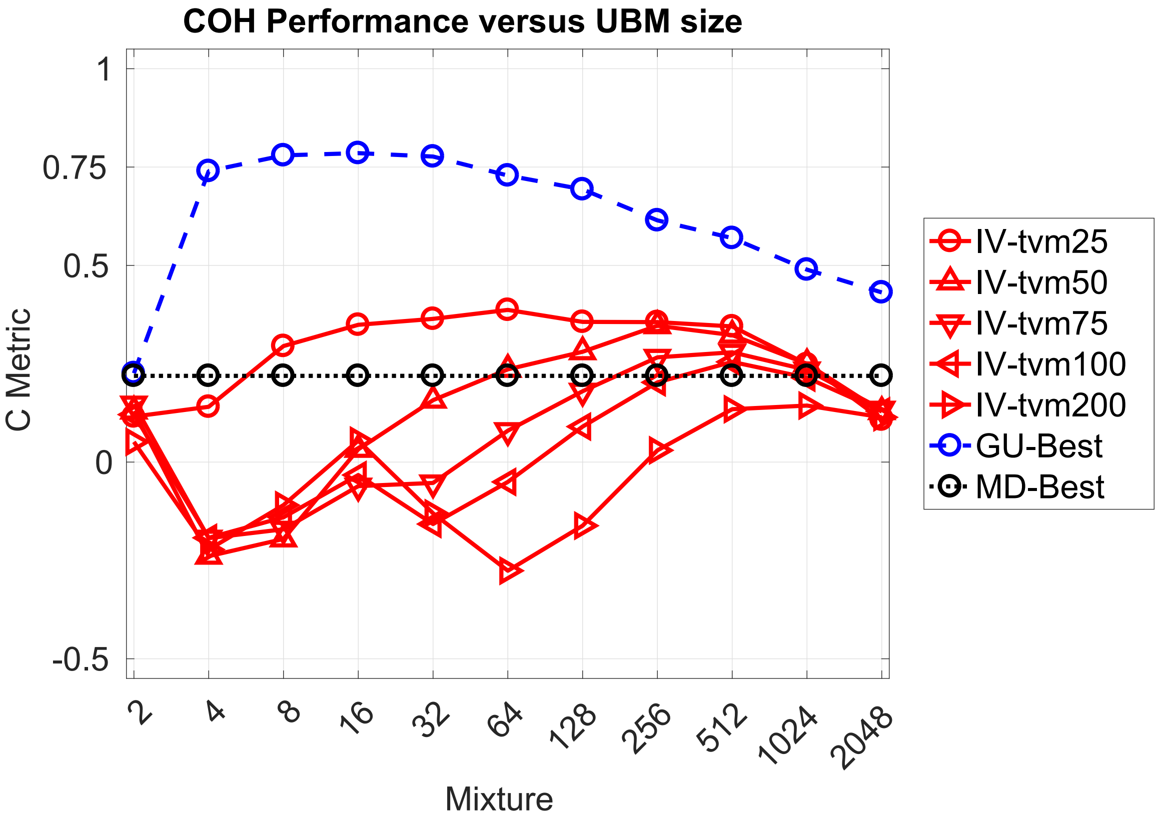


Figure 6.21. C Metric Plot of COH `AbnNrmSzr`. This C Metric plot shows the COH based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

reported its best score when using the CEP features, while the GMM-UBM and I-Vectors algorithms reported their best scores when using the PSD features.

The ninth experiment, Figures 6.25–6.27, tested the algorithms' performances when using the TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets, consisting of 570 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

The tenth experiment, Figures 6.28–6.30, tested the algorithms' performances when using the TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets consisting of 570 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold with multiple UBM mixtures for all three features. The I-Vectors were able to exceed the 0.75 score threshold with a variety of UBM mixtures and TVM dimensions for the PSD features. All three algorithms reported their best scores when using the PSD features.

6.1.2 Discussion

The top scores for each algorithm, dataset, and feature set pairing are given in Table 6.2. These represent the peak performance of each system within the tested range of datasets, feature sets, and, when applicable, UBM mixture size and TVM dimensions. The minimum acceptable score of 0.75 was not indicated in the table; instead, the top two scores were highlighted for each dataset. The GMM-UBM algorithm had the most high scores and the I-Vector had the most second highest

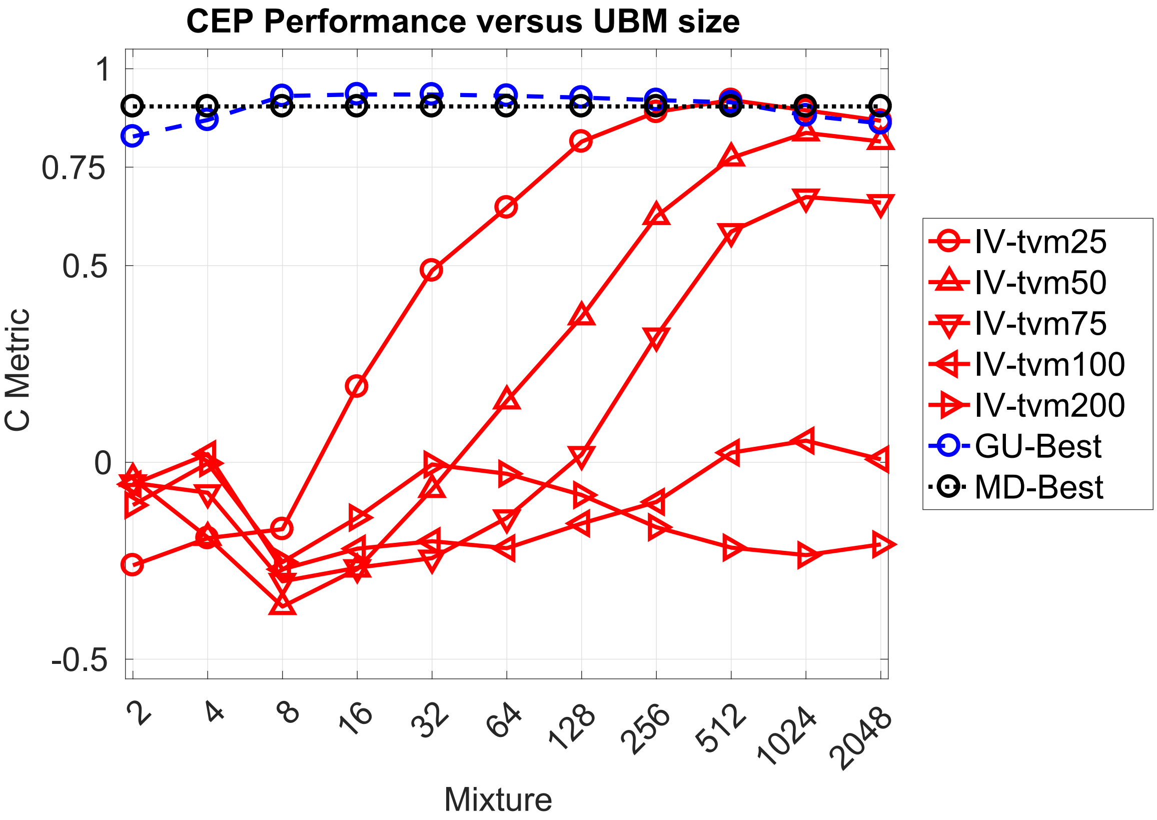


Figure 6.22. C Metric Plot of CEP AbnNrmMot. This C Metric plot shows the TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

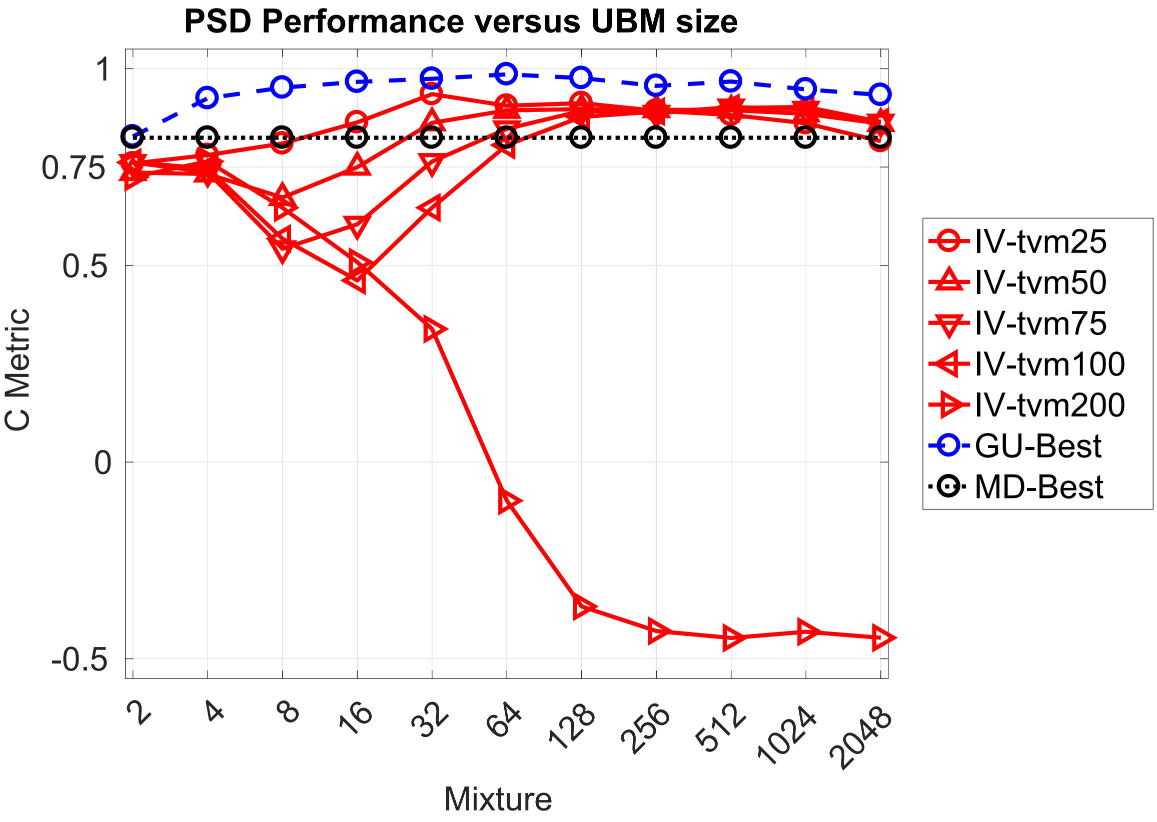


Figure 6.23. C Metric Plot of PSD `AbnNrmMot`. This C Metric plot shows the PSD based TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

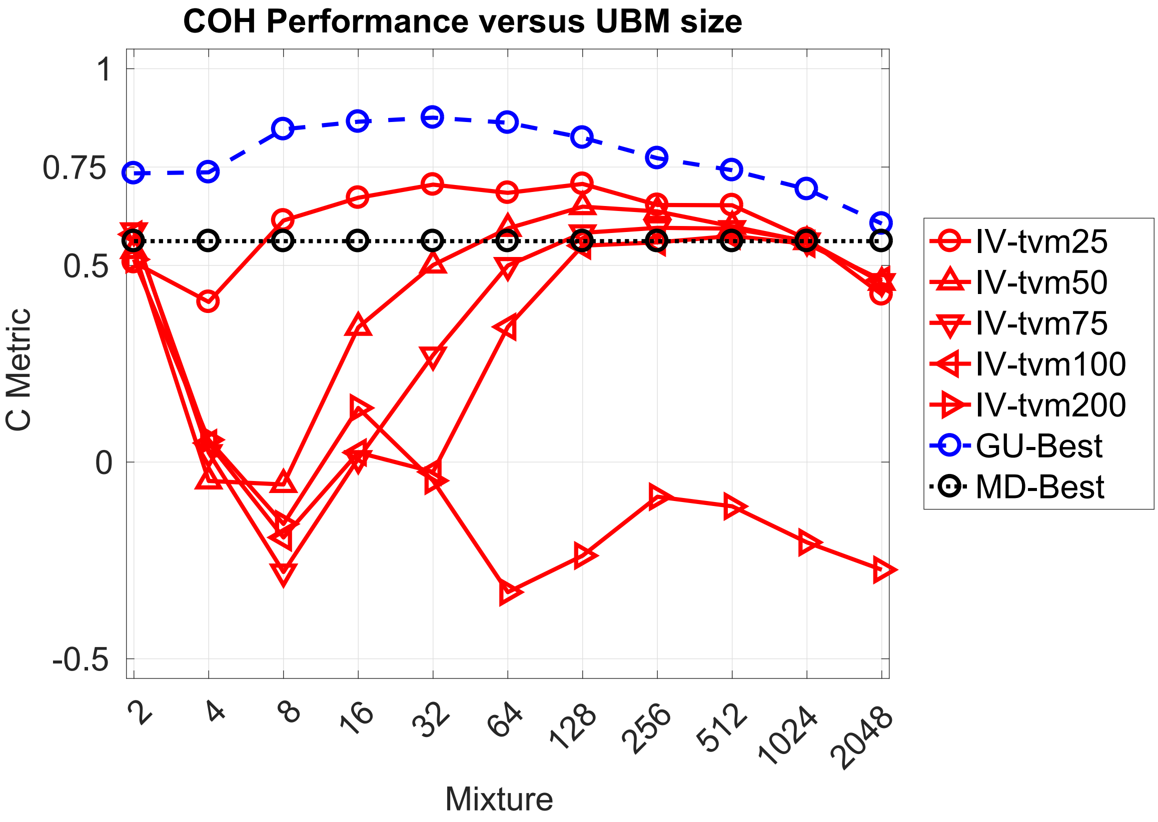


Figure 6.24. C Metric Plot of COH `AbnNrmMot`. This C Metric plot shows the TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

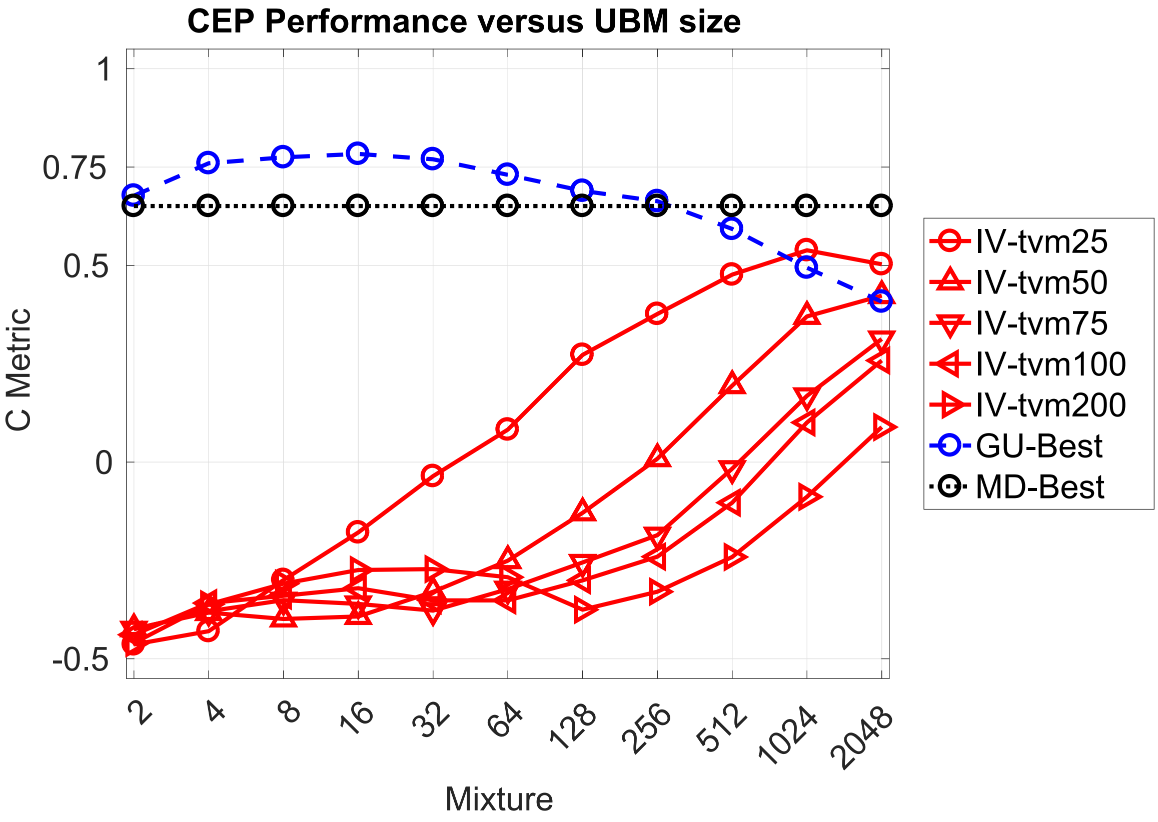


Figure 6.25. C Metric Plot of CEP `NrmSzrMot`. This C Metric plot shows the CEP based TUH-EEG Normal, Seizure, and PhysioNet Database Motion dataset performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

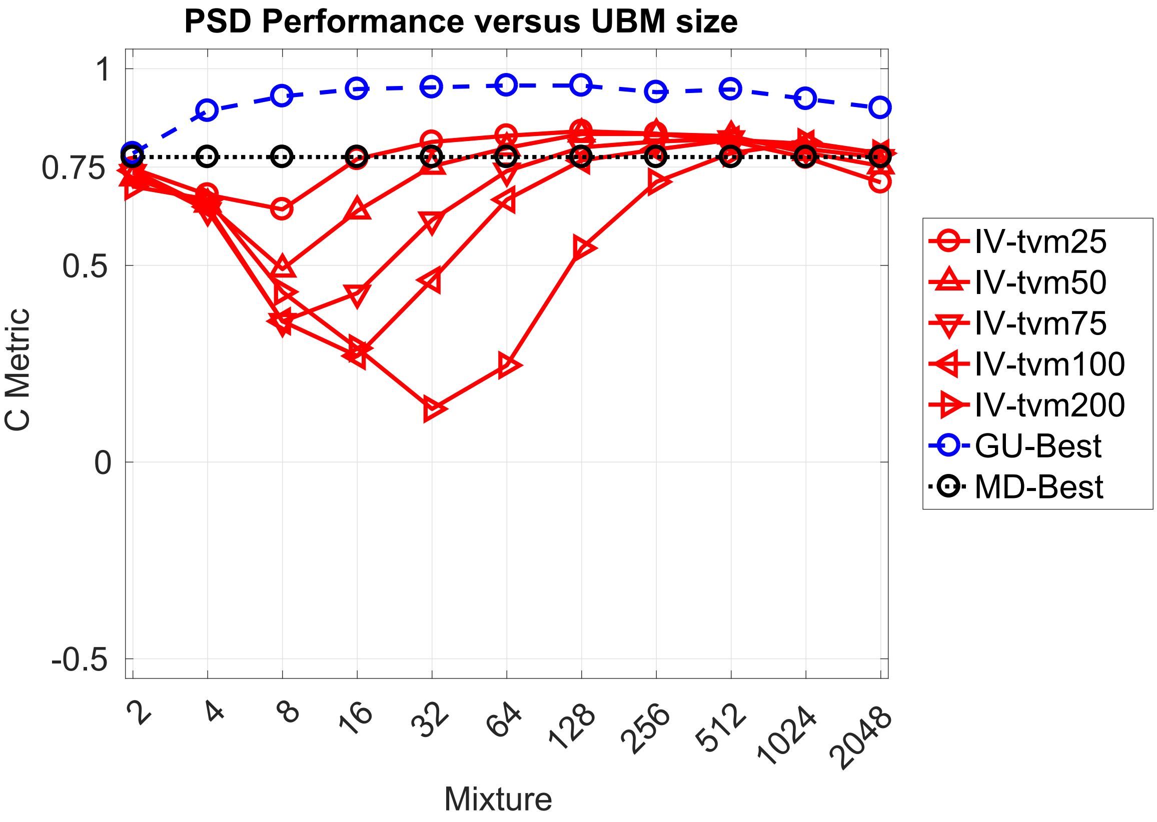


Figure 6.26. C Metric Plot of PSD NrmSzrMot. This C Metric plot shows the PSD based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

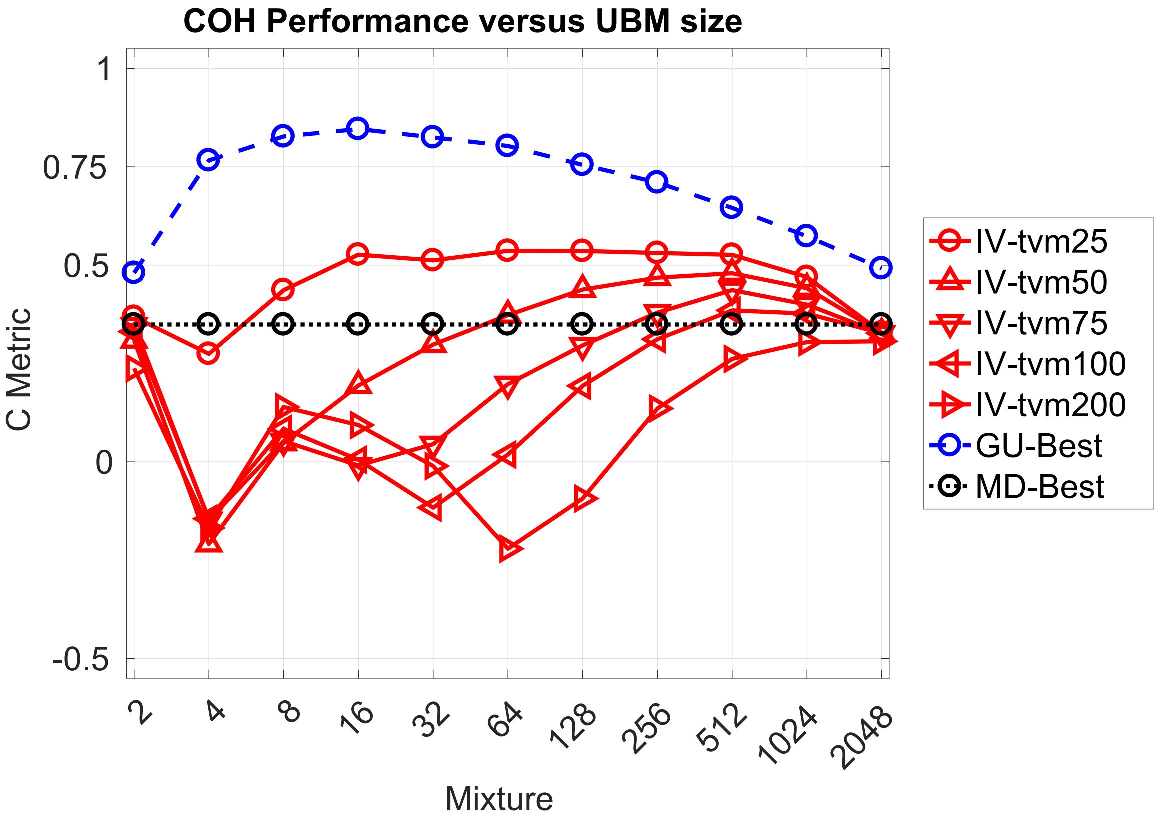


Figure 6.27. C Metric Plot of COH `NrmSzrMot`. This C Metric plot shows the COH based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

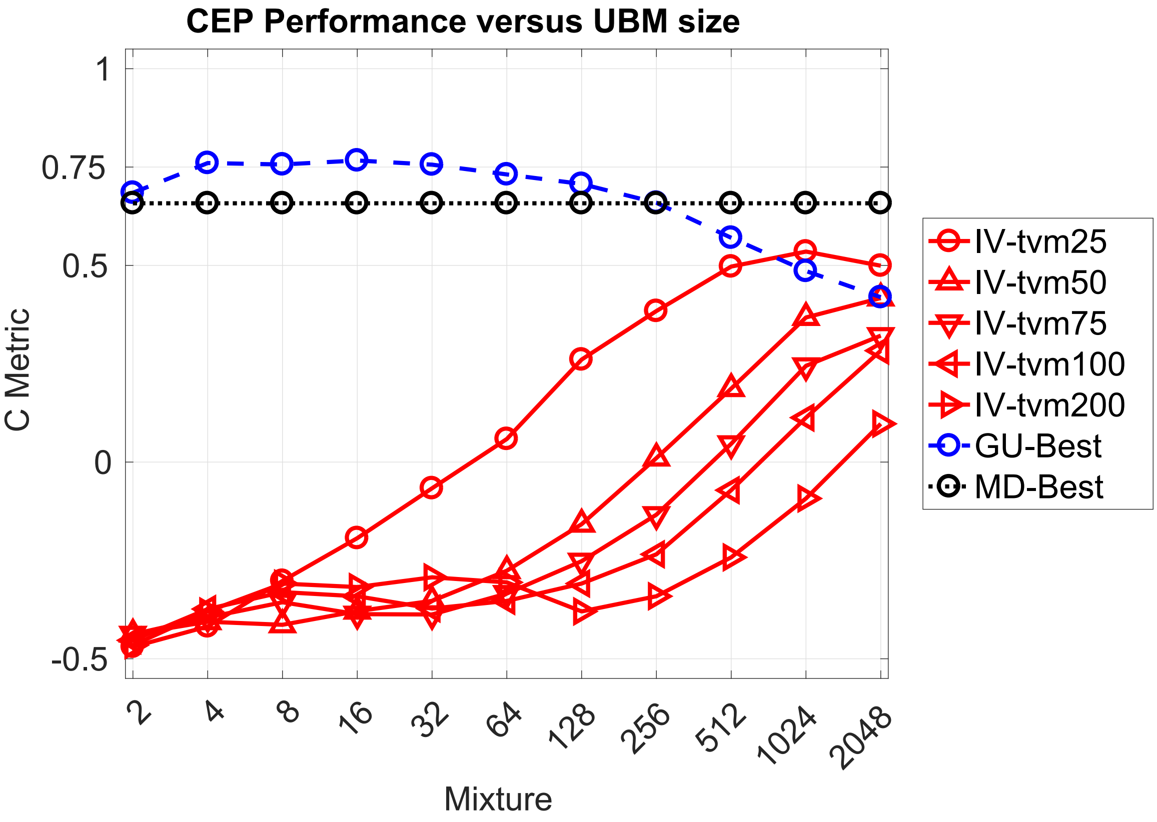


Figure 6.28. C Metric Plot of CEP `AbnSzrMot`. This C Metric plot shows the CEP based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

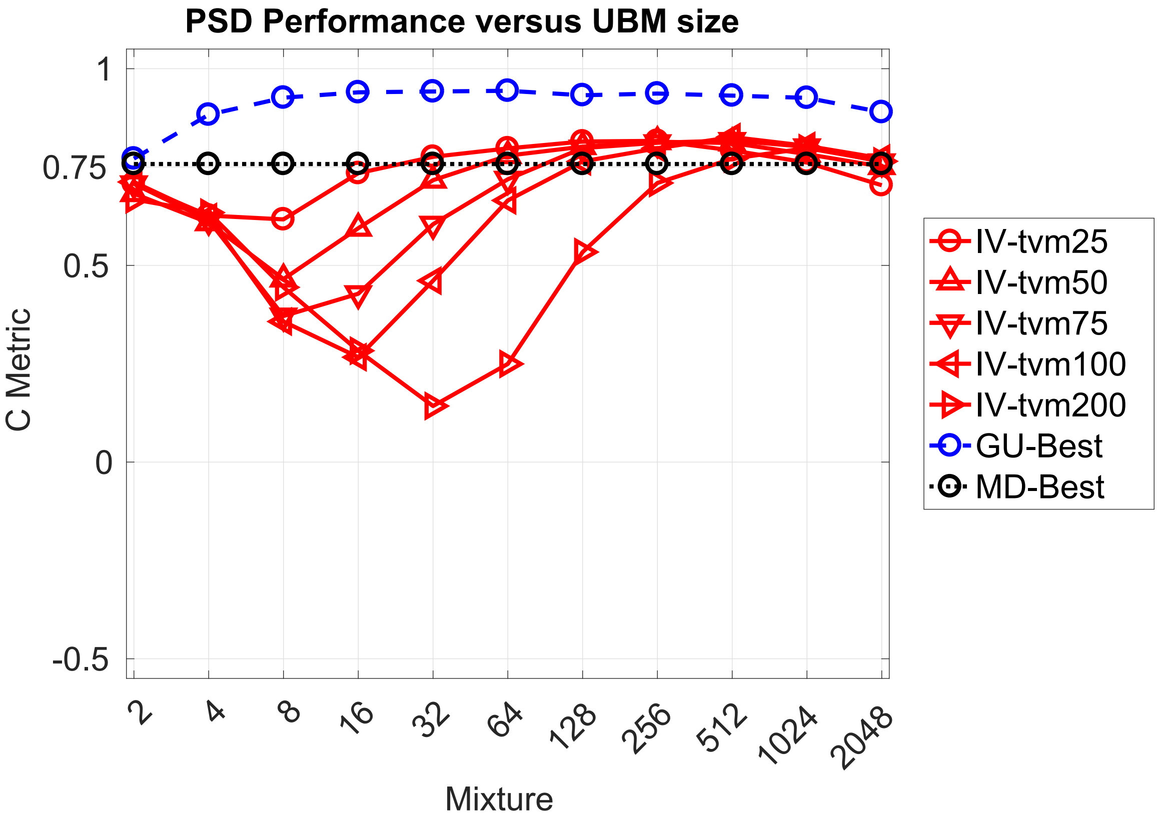


Figure 6.29. C Metric Plot of PSD `AbnSzrMot`. This C Metric plot shows the PSD based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

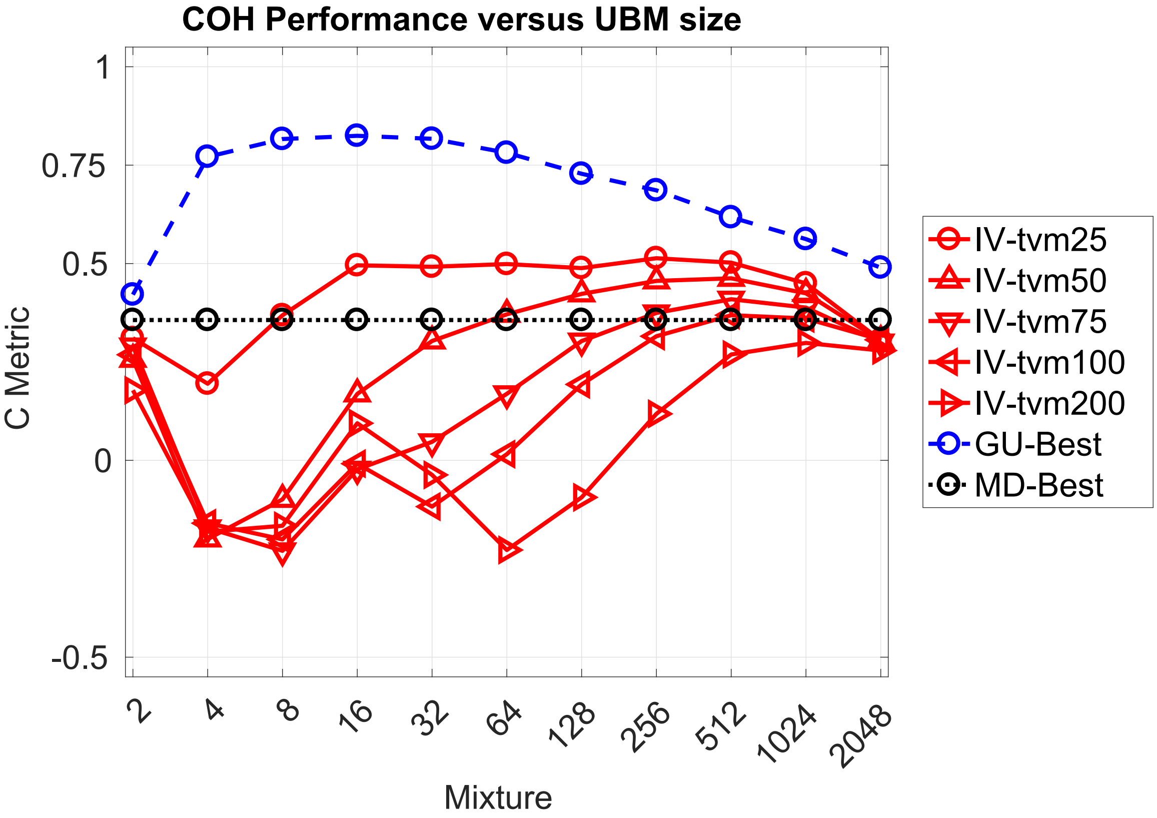


Figure 6.30. C Metric Plot of COH `AbnSzrMot`. This C Metric plot shows the COH basedTUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of algorithm selection. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

scores, while the MD algorithm failed to produce any such scores. Regardless of classifier, these scores were predominantly tied to the PSD features. Unlike the algorithms' themselves, each feature set found itself among the best reported at least once.

It was not necessary to break out scores by TVM dimension because the same dimension (25) consistently produced the best results. The exceptions were when a TVM of dimension 50 outperformed 25 on the PSD AbnSzr and PSD AbnNrmSzr datasets. The strength of the 25-dimension TVM was particularly noticeable when paired with the CEP features, as exemplified in Figures 6.13, 6.16 and 6.22. However the PSD features scores frequently peaked at the same score regardless of the TVM dimension, as exemplified in Figures 6.5, 6.8 and 6.14. This occurred with the COH features as well, but only once the UBM was built from 512 or more mixtures.

Generally all three algorithms performed worse when operating on the CEP or COH features compared to their PSD counterparts. This trend was most pronounced when the TUH-EEG seizure dataset was included. The scores for the AbnNrm AbnMot, and NrmMot datasets were consistently strong across feature sets, but the AbnSzr and NrmSzr scores were substantially worse for the CEP and COH features. In worst case comparisons the GMM-UBM saw a 0.21 drop in score from the PSD to the next strongest feature on the SzrMot dataset, a 0.72 drop for the MD on the AbnNrm dataset, and 0.58 drop for I-Vectors on the AbnNrm dataset.

The relationship between dataset and feature was apparent from the scores of AbnNrm dataset which had poor COH scores, but the AbnSzr and NrmSzr datasets showed improved COH performance while decreasing the CEP and PSD performance. COH feature performance improved further when the PhysioNet Database motion dataset was added for two or three dataset aggregations. The best two dataset performances of the COH features were AbnMot

Table 6.2. Top C Metric Performances

Dataset	Feature	GU	MD	IV
AbnNrm	CEP	*0.9900*	0.9028	**0.9956**
	PSD	0.9777	0.7800	0.8413
	COH	0.8019	0.1851	0.4117
AbnSzr	CEP	0.7608	0.5835	0.4440
	PSD	**0.9313**	0.7527	0.7796
	COH	*0.8200*	0.2148	0.4323
NrmSzr	CEP	0.7583	0.5835	0.4261
	PSD	**0.9651**	0.7657	0.8130
	COH	*0.8289*	0.2126	0.4146
AbnMot	CEP	0.9391	0.9496	0.8991
	PSD	**0.9874**	0.8994	0.9547
	COH	*0.9604*	0.6849	0.8567
NrmMot	CEP	0.9272	0.9357	0.9036
	PSD	**0.9874**	0.9080	*0.9497*
	COH	0.9104	0.6981	0.7960
SzrMot	CEP	0.7383	0.6558	0.4953
	PSD	**0.9496**	0.7635	*0.8501*
	COH	0.8367	0.3769	0.5656
AbnNrmSzr	CEP	0.7730	0.5583	0.4950
	PSD	**0.9401**	0.7613	0.7808
	COH	*0.7854*	0.2192	0.3868
AbnNrmMot	CEP	0.9349	0.9043	0.9205
	PSD	**0.9856**	0.8245	*0.9354*
	COH	0.8756	0.5616	0.7072
NrmSzrMot	CEP	0.7834	0.6509	0.5384
	PSD	**0.9574**	0.7754	0.8412
	COH	*0.8461*	0.3491	0.5364
AbnSzrMot	CEP	0.7667	0.6579	0.5352
	PSD	**0.9439**	0.7577	0.8166
	COH	*0.8246*	0.3565	0.5135

The best score is in **bold** with the runner up in *italic*.

followed closely by `NrmMot`. This was the case as well with the three dataset performances of `AbnNrmMot`. Unsurprisingly, this was also the dataset with the best CEP feature performance as well.

Overall, the PSD features were consistently strong regardless of algorithm and dataset. Coupled with the strength of the GMM-UBM, this pairing represented the optimal subject verification technique among the tested configurations. In instances where this was not true (most notably the `AbnNrm` dataset where CEP features and I-Vectors were the optimal combination) the second best options were the COH features paired with GMM-UBMs or PSD features paired with I-Vectors. The hierarchy of features for GMM-UBMs and I-Vectors was always PSD, followed by COH, and then CEP. Contrary to this, the MD performance produced three instances outside the `AbnNrm` dataset where the CEP feature produced better performance than the PSD features: `AbnMot`, `NrmMot`, and `AbnNrmMot`.

Despite I-Vectors not outperforming the GMM-UBMs, they were consistently the reported the second best score for PSD features. At worst they trailed by 0.16 for the `AbnSzr`, `NrmSzr`, and `AbnNrmSzr` datasets. The TUH-EEG Seizure dataset was their weakness as their performance on that dataset was closer to the MD algorithm then than the GMM-UBM algorithm. However, when that dataset was not included its performance the I-Vectors tracked the performance of the GMM-UBMs. These results suggested I-Vectors were a viable alternative for subject verification, reaffirming what was previously reported, but now using larger and more diverse datasets [193].

Significantly, the I-Vector performance was strongest when built from TVMs of only 25 dimensions. Using only cosine distance to evaluate these vectors for each subject set, performance was comparable to two far more elaborate techniques. The MD technique is reliant on a pooled covariance matrix used to determine the distance between the averaged means of each subject [64]. It uses vectors with sizes equal to the number of features (26 to 40) and a similarly sized covariance matrix to evaluate the testing data. Its results occasionally bested the I-Vectors when using CEP features, failed to produce competitive results with COH features, and always trailed I-Vectors when using PSD features. When the MD CEP performance exceed that of the I-Vectors, the equivalent PSD features performance favored I-Vectors and exceeded the MD CEP performance.

While the MD performance represented a minimum performance threshold, the GMM-UBM performance represented a target threshold for the I-Vectors to meet or potentially exceed [42, 163, 200]. Unfortunately, the I-Vectors were unable to exceed the performance of the GMM-UBM when using the same sized UBM. On most datasets, the GMM-UBM performance using 4 mixture UBMs eclipsed the 0.75 threshold and often outperformed the MD results and the I-Vector for any sized UBM. However, as the UBM mixture size continued to increase its performance would begin to diminish when reaching the 1024 and 2048 mixture sizes.

When the GMM-UBM used a 4-mixture UBM it was effectively operating with 160 degrees of freedom (40 elements in a PSD feature vector times 4 mixtures) for each enrollment model and the target model. The MD classifier was a 40×40 pooled covariance matrix and 40-element enrollment and testing feature vectors. However, the I-Vectors were only 25-element vectors for each enrollment and testing subject, making them an order of magnitude smaller than the GMM-UBM approach and nearly half the size of the MD approach. The PSD and COH feature vectors contained the same number of elements, while the CEP feature vector was only 26 elements. The CEP features put the MD and I-Vectors on a very similar scale, but both retained an order of magnitude reduction over the GMM-UBM's 104 elements.

Taking the functional aspects of the algorithms into account, it appeared that I-Vectors were far more efficient at distilling the critical components of the datasets into a functional low dimensional space. The subject specific I-Vectors matched the order of vectors generated by the MD algorithm produced while producing performance scores comparable to the GMM-UBMs. If these techniques were expanded to produce channel classifications or epoch classifications, the GMM-UBM would produce significantly more data (an additional full set of UBM mixtures for each subject) and the MD would be forced to constrain even more subjects, (2,398 subject-channels from the 109 PhysioNet Database subjects) through its 40 × 40 mean centered covariance matrix.

This would increase the GMM-UBM training and computation time in conjunction with increasing its memory and disk needs, while the dimensionality constraints of the MD would likely continue to hinder its performance as the worst of the three classifiers. However, the I-Vector technique would only produce more 25 element feature vectors and require only a linear increase in computation cycles and no increase in memory or disk storage. Thus it is likely that I-Vectors represent the best balance within the context of these experiments and possess an extensible framework with a confined resource footprint.

6.1.3 Constraints

The results were limited by the datasets due to the differing number of subjects. It was beneficial to include diverse data in terms of quality and quantity, but additional insight would have been gained if each dataset was constrained to a common subject count. Adding in larger datasets that were known to be easier (`Mot`) or those known to be harder (`Szr`) to classify made them the dominant statistical components when paired with smaller datasets (`Abn` and `Nrm`). Thus, performance of `AbnMot` and `NrmMot` was an indirect evaluation on the `Mot` dataset as it represented over 2/3 of the aggregated dataset, 50 subjects to 109 subjects, and even moreso for the AbnSzr and NrmSzr with 411 versus 50 subjects.

When comparing the performance of the algorithms between the AbnNrm and the AbnSzr and NrmSzr datasets, the performance for the CEP and PSD features decreased, roughly 0.40 and 0.04, respectively. Conversely, the performance for the COH features increased slightly by 0.02 on average. When evaluated against the AbnMot and NrmMot, it was the PSD and COH feature sets that improved performance while the CEP features decreased for the GMM-UBM and I-Vector algorithms. The MD algorithm instead had improvements across all feature sets, nearly 0.50 for the COH features.

The wide range of the algorithms' performances were thought to be caused by the training data, which produced the UBMs and the pooled covariance matrices, being biased by overwhelming amount of Mot or Szr data in the aggregated datasets. This could have been the reason the AbnSzr and NrmSzr results were so similar to those of the AbnNrmSzr results, since the majority of modeled data comes from the Szr dataset.

Understanding the impact of these training data discrepancies would have required replicating these ten experiments using controlled datasets of 50 subjects each (the subject counts of the Nrm and Abn datasets). While balancing the weight of each source of data, this would have also reduced the total amount of data available for training the UBMs. Comparing only 150 subjects, or even 200 subjects if all four datasets were combined would also have been too small to test the desired TVMs dimensions, and would barely double the previously tested 109 subjects of the PhysioNet Database.

Given that some of best GMM-UBM performances required UBM mixtures of size 64 (see Figures 6.26 and 6.29) it was likely the complexity of the data was underestimated relative to the UBM mixture sweep. This was suggested in

Figures 6.4, 6.7, 6.19, 6.20, 6.25 and 6.28, where the larger TVM dimensions began to produce steadily increasing scores while the smaller TVMs' performances began to wane. These listed figures represented all tests using the `Szr` dataset. This supported the idea that the TUH-EEG Seizure data was very robust in that the GMM-UBM achieved maximum performance with just a 64 mixture UBM. As each mixture was on the order of 41 components, using 64 mixtures meant that there were 2,624 independent degrees of freedom to build the 570 subject models used in classification.

Comparatively, the I-Vector's best score relied on 570 vectors of 25 elements for the `AbnSzrMot` and `NrmSzrMot` datasets. In each experiment, larger dimension TVMs did not provide improved performance. In the larger datasets the TVMs frequently converged at the larger mixture sizes, Figures 6.23, 6.26 and 6.29. This suggested they were constrained by the smaller dimensional space produced by the previous UBMs. This was affirmed by the previously noted CEP feature plots when the `Szr` dataset was used.

6.2 LDA Enhanced Performance

The impact of LDA did not yield any performance improvements during the Parameter Sweeps of Section 5.2.1. This result disagreed with performance improvements documented by the speech recognition community, where reduction in the I-Vectors' dimensionality was met with improved classification performance [203, 208, 192]. The Algorithm Benchmarks provided another chance to test the impact of LDA on the TVM, since the datasets had increased in quality and quantity. As the LDA only apply to the I-Vectors, the optimal UBM mixtures were chosen and compared against the best scores of the GMM-UBM over the given UBM mixture range.

This allowed the continued use of the C Metric style plots with the x-axis modified to represent the LDA dimension. However, the LDA dimensions were updated to align better with the chosen TVM dimensions and to provide at least three LDA iterations for each TVM. This new alignment,Table 6.3, represented an improvement from the search space seen in Chapter 5, Table 5.1. This allowed for the unmodified TVM result to be shown (when the LDA dimension matches the TVM dimension) and all ensuing LDA dimensions linked by a common line style.

Table 6.3. Updated LDA Dimensions

UBM	TVM	LDA
32 64 128	25	20 15 5
	50	45 25 15
	75	70 50 25
	100	95 75 50 25
	200	195 100 75 50 25

6.2.1 Results

These experiments mirrored the ten experiments of the previous Native TVM Performance, which thus produced a total of 30 new figures. The initial experiment, Figures 6.31–6.33, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal and Normal datasets. This combined dataset served as a basline comparison point as it consisted of the smallest number of subjects, 100. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when using an LDA improved TVM for the CEP and PSD features. All three algorithms reported their best scores when using the CEP features.

Despite a TVM dimension of 25 being the strongest for the native TVM, LDA was unable to improve upon its performance. However, the next largest TVM dimensions, 50 and 75, showed significant performance gains when LDA reduced them to 45 and 70 dimensions, respectively. These improvements were enough to make the 45 dimension I-Vectors on par with the native 25 dimension I-Vectors, and the 70 dimension I-Vectors exceed the MD performance. The two largest dimensions, 100 and 200, did not improve their performance when paired with LDA.

The second experiment, Figures 6.34–6.36, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal and Seizure datasets. This combined dataset consisted of 461 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when using an LDA improved TVM for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95, respectively. The PSD and COH features saw improvements for the TVMs of dimension 50, 75, and 100. In both cases LDA only decreased the performance of TVMs of dimension 25.

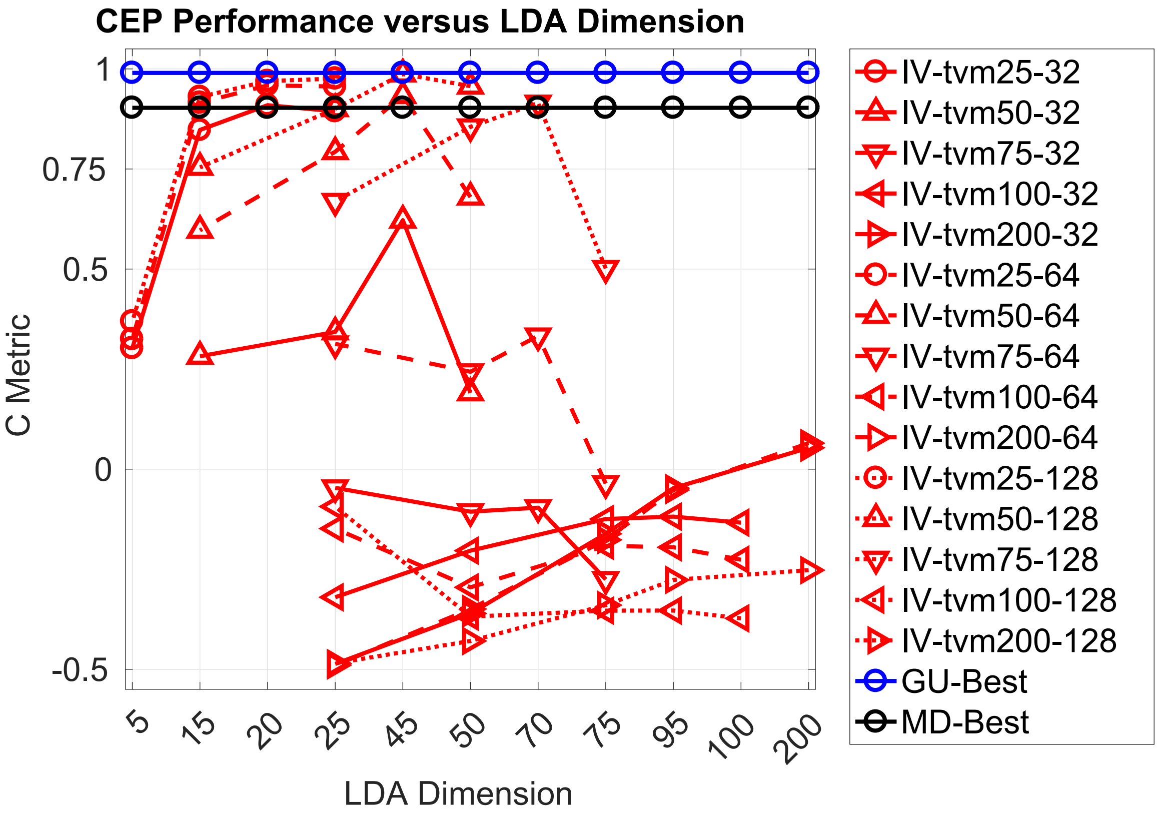


Figure 6.31. C Metric Plot of CEP AbnNrm with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal and Normal datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the 100 and 200 dimension TVMs to a dimension of 99.

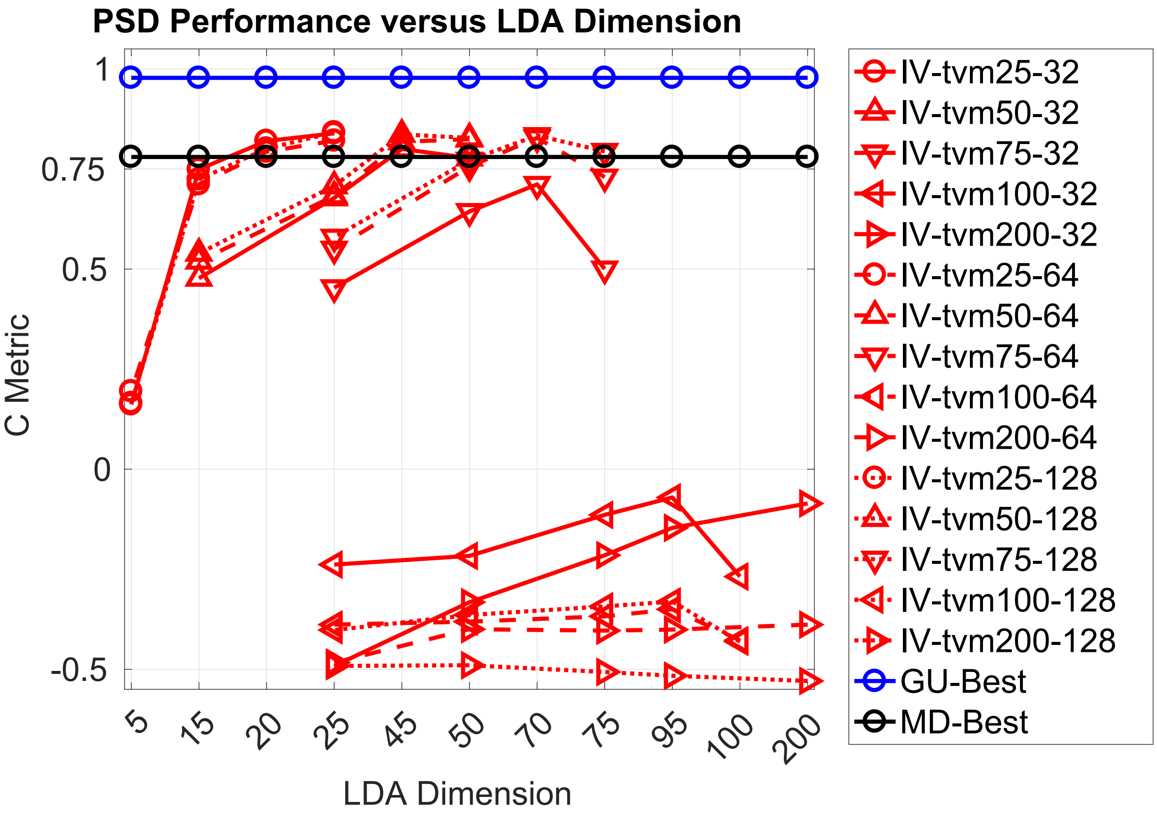


Figure 6.32. C Metric Plot of PSD `AbnNrm` with LDA. This C Metric plot shows the PSD based TUH-EEG Abnormal and Normal datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the 100 and 200 dimension TVMs to a dimension of 99.

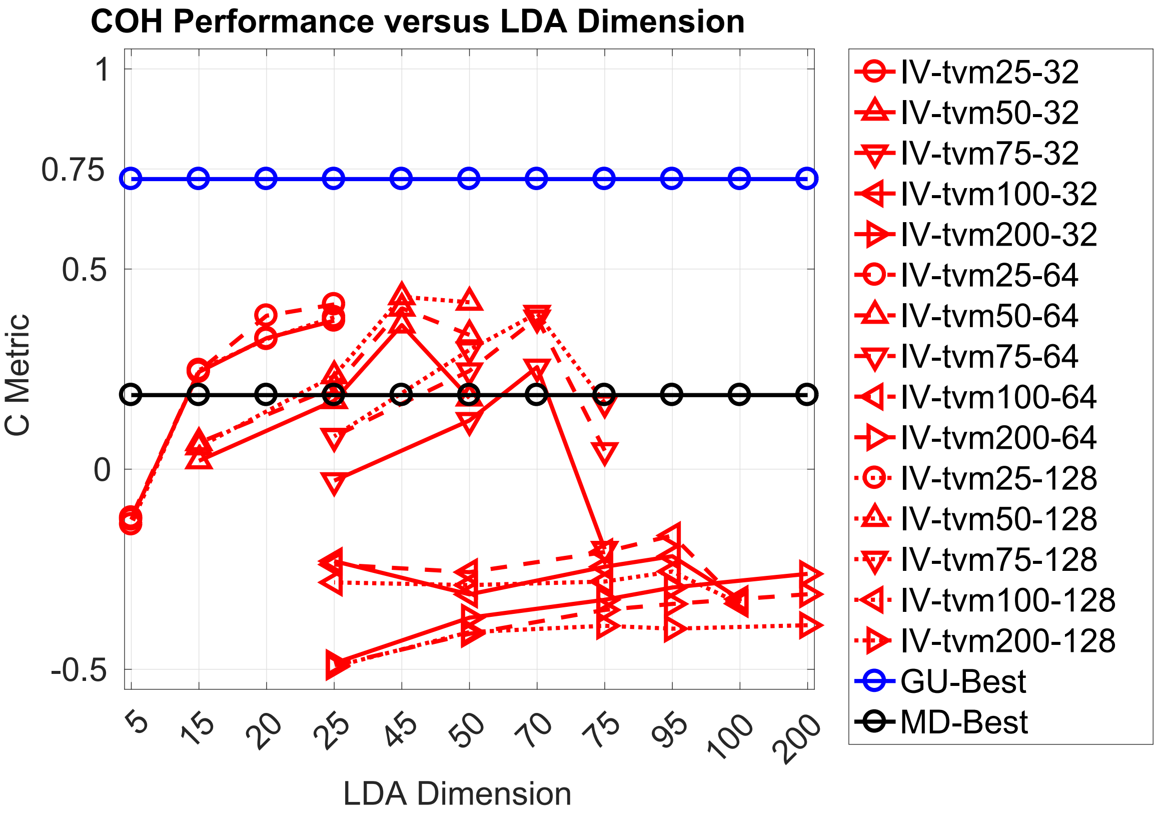


Figure 6.33. C Metric Plot of COH `AbnNrm` with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal and Normal datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the 100 and 200 dimension TVMs to a dimension of 99.

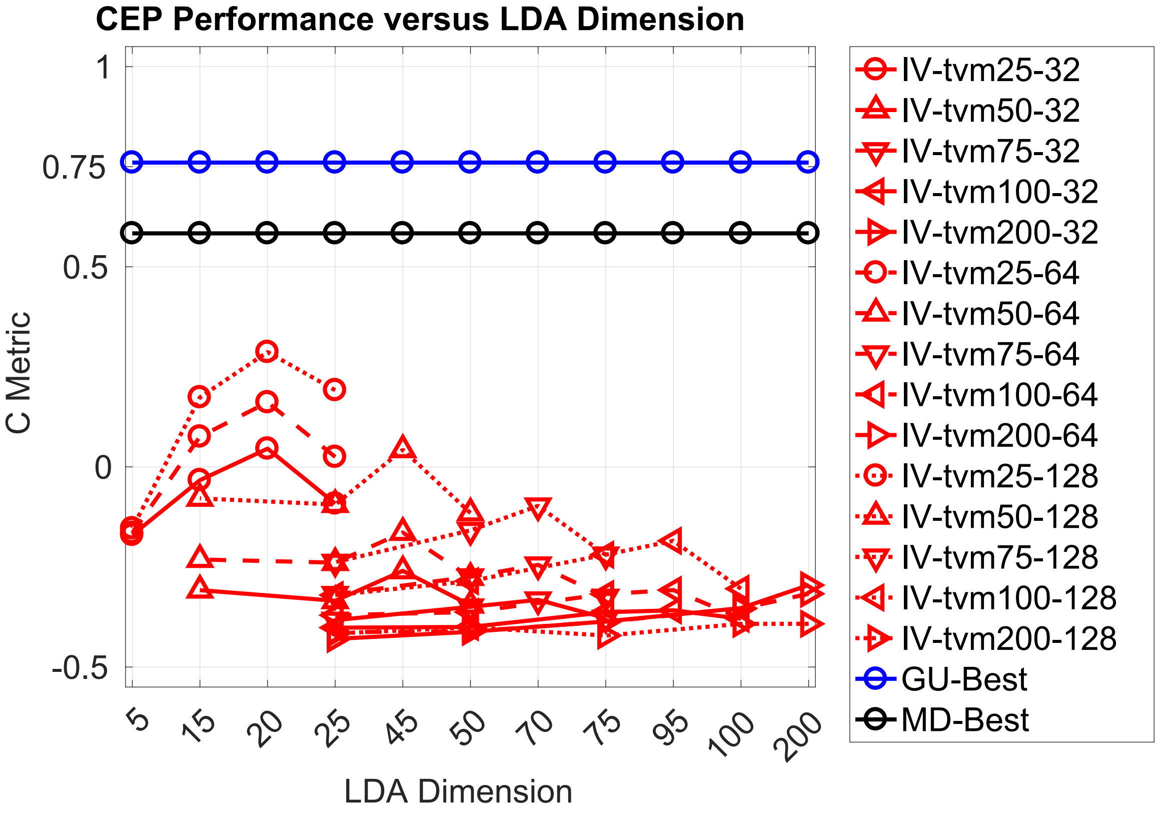


Figure 6.34. C Metric Plot of CEP `AbnSzr` with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

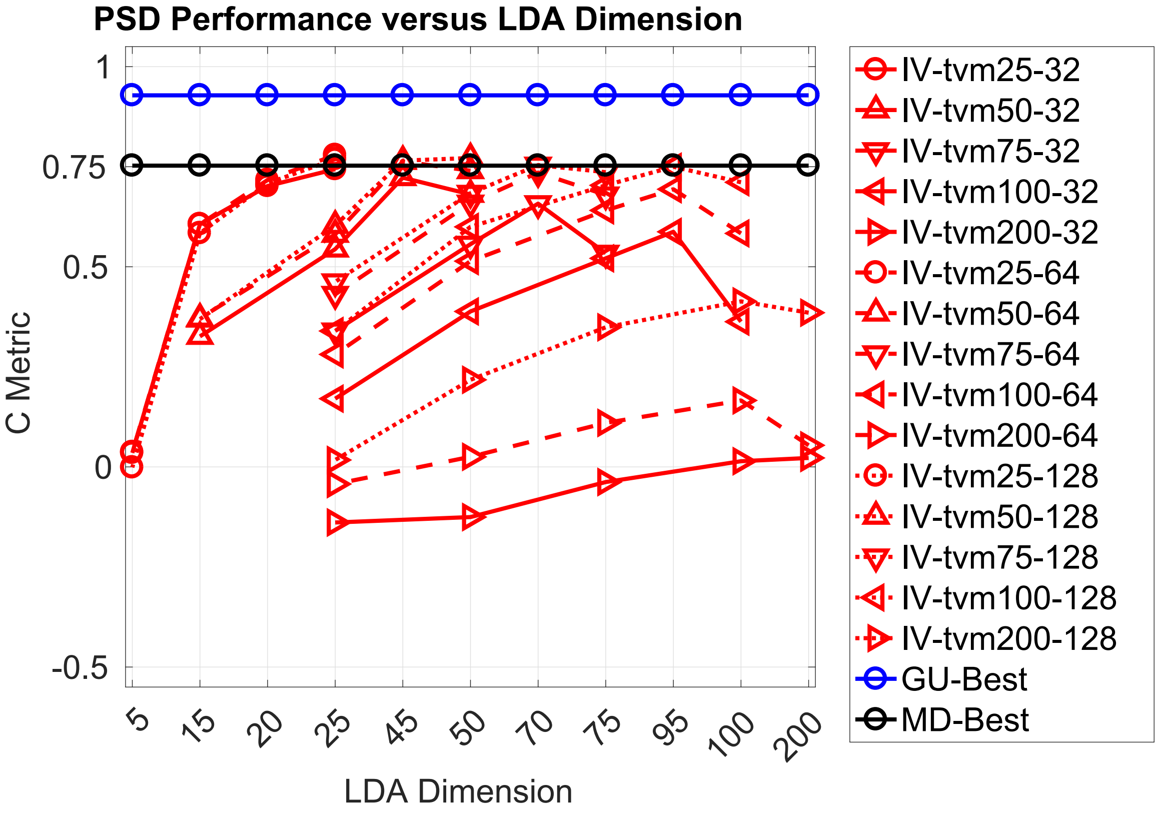


Figure 6.35. C Metric Plot of PSD AbnSzr with LDA. This C Metric plot shows the PSD based TUH-EEG Abnormal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

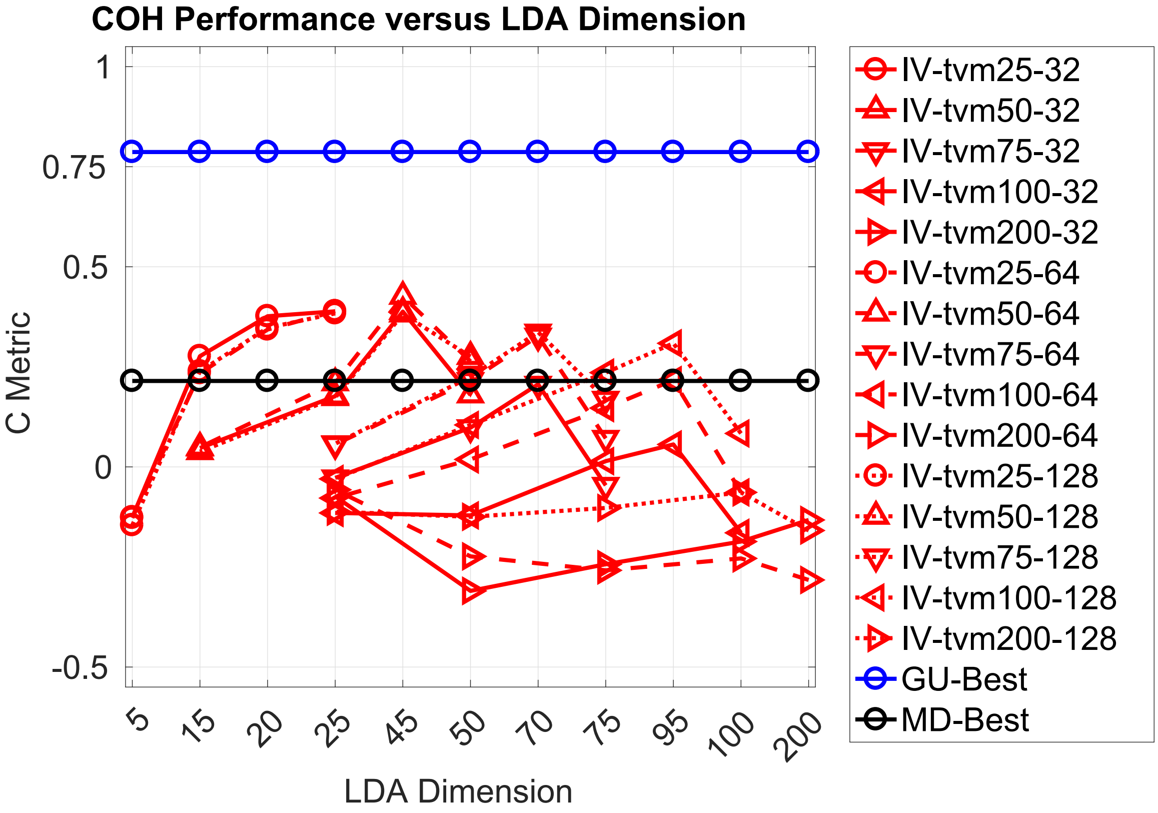


Figure 6.36. C Metric Plot of COH `AbnSzr` with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

The third experiment, Figures 6.37–6.39, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Normal and Seizure datasets. This combined dataset consisted of 461 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when using an LDA improved TVM for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95 respectively. The PSD and COH features saw improvement of the TVMs in dimensions 50, 75, 100, and 200 for each UBM mixture. In both cases LDA only decreased the performance of TVMs of dimension 25.

The fourth experiment, Figures 6.40–6.42, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal and PhysioNet Database Motion datasets. This combined dataset consisted of 159 subjects. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with the 64 and 128 mixtures UBMs when using an LDA improved TVM for all three features. The 32 mixture UBM only exceeded this threshold for the PSD and COH features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95 respectively. The PSD features experienced minimal if any

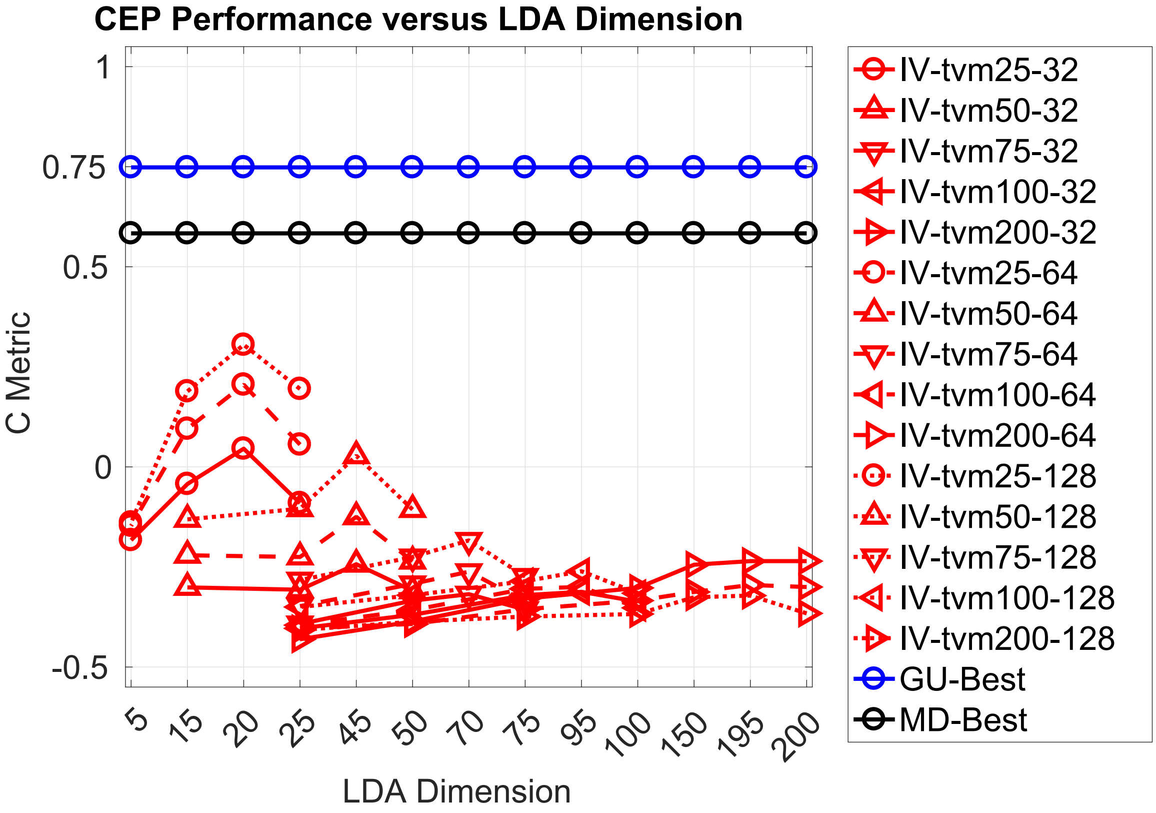


Figure 6.37. C Metric Plot of CEP `NrmSzr` with LDA. This C Metric plot shows the CEP based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

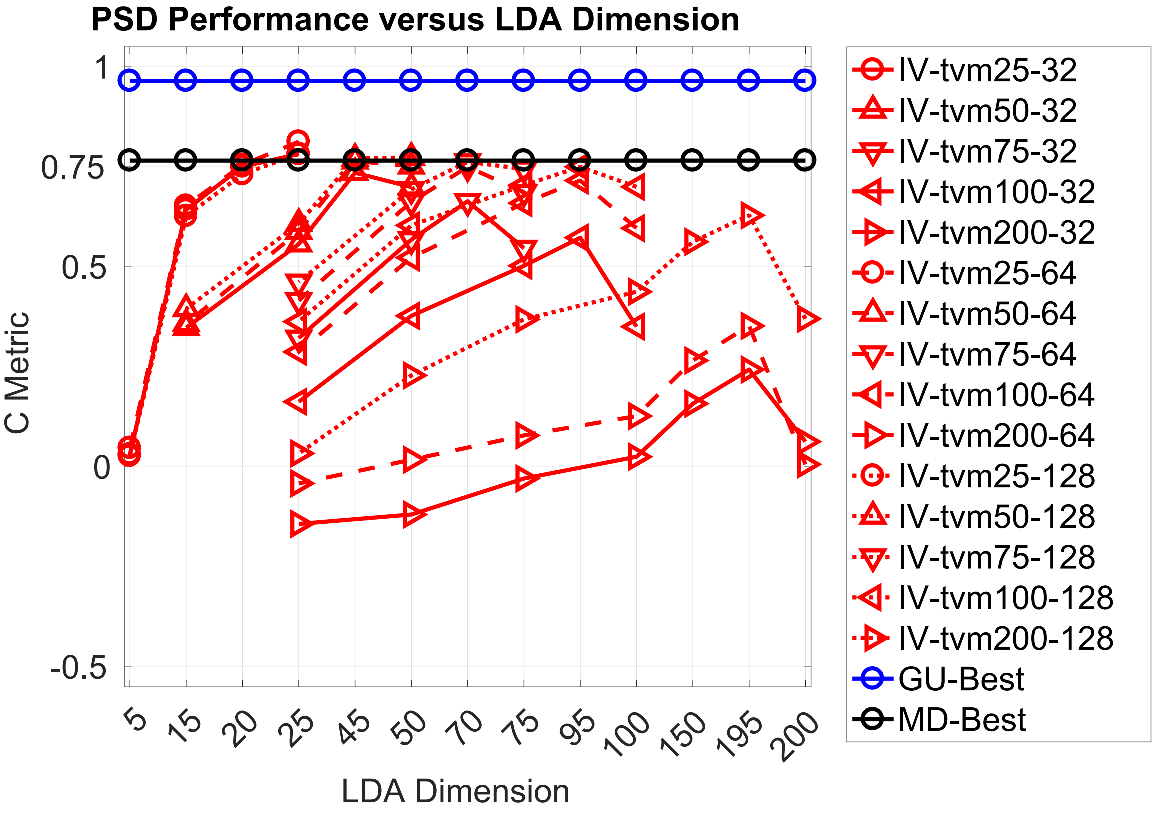


Figure 6.38. C Metric Plot of PSD `NrmSzr` with LDA. This C Metric plot shows the PSD based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

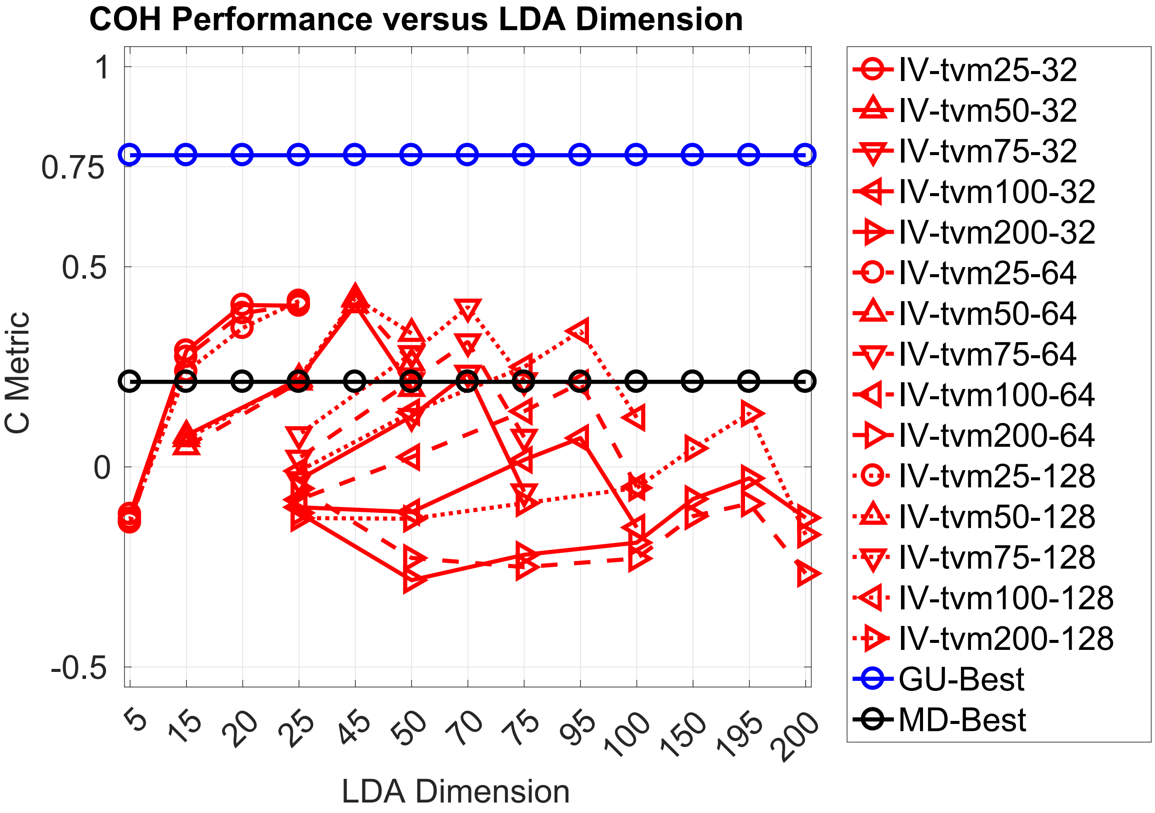


Figure 6.39. C Metric Plot of COH `NrmSzr` with LDA. This C Metric plot shows the COH based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

improvement for each TVM dimension. However, the COH features saw improvement of the TVMs in dimensions 50, 75, and 100 for the 32 and 64 mixture UBMs. In both cases LDA only decreased the performance of TVMs of dimension 25 and 200.

The fifth experiment, Figures 6.43–6.45, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Normal and Seizure datasets. This combined dataset consisted of 461 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for the PSD and COH features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when using an LDA improved TVM for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in all dimensions for the first iteration of LDA across all UBM mixtures. The PSD features for TVM dimensions of 75, 100, and 200 improved when LDA reduced them down to 70, 95, and 109. Continued used of LDA only decreased performance as was the case of TVM dimensions of 25 and 50. The COH features experience improvement with the initial use of LDA for all dimensions except 25. In many cases this reduction of dimensions drove I-Vector performance above that of the MD performance for the PSD and COH features.

The sixth experiment, Figures 6.46–6.48, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Seizure and PhysioNet Database Motion datasets. This combined dataset consisted of 520 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for the PSD and COH features. The I-Vectors were able to exceed the 0.75 score threshold with each of the

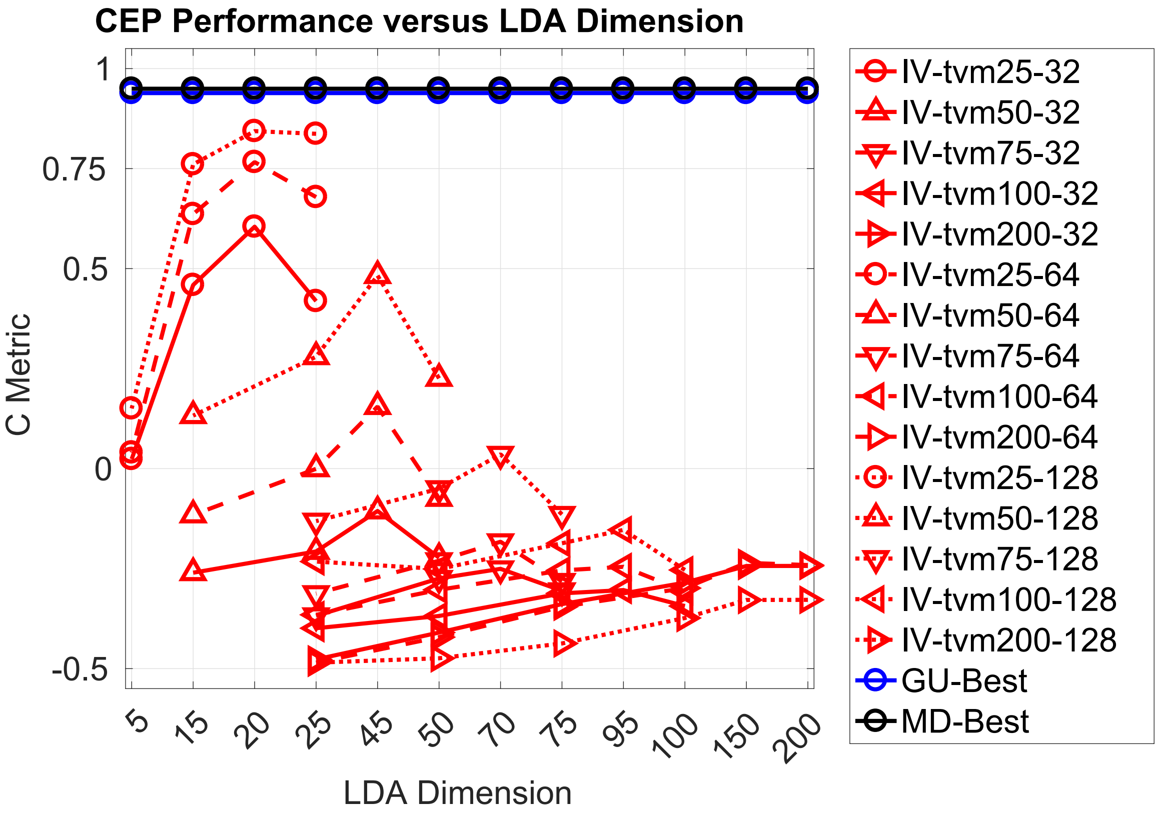


Figure 6.40. C Metric Plot of CEP `AbnMot` with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the 200 dimension TVM to 158 dimensions.

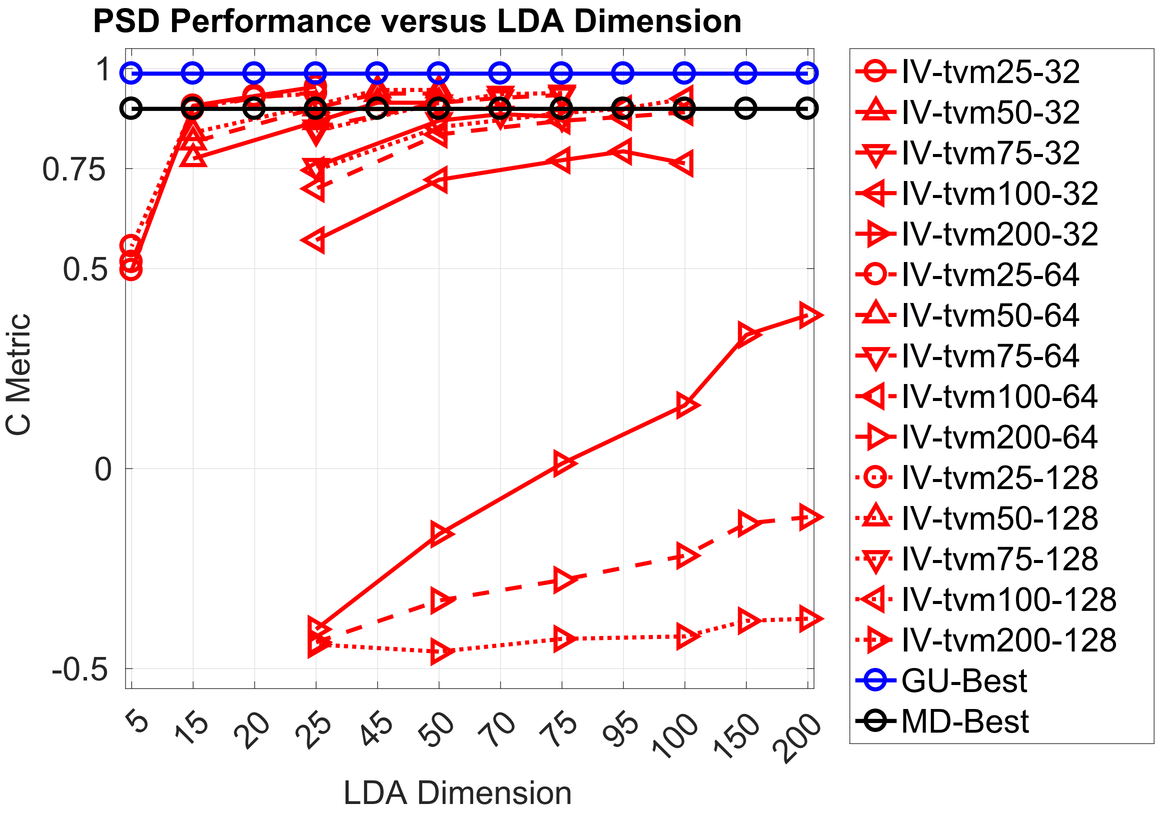


Figure 6.41. C Metric Plot of PSD AbnMot with LDA. This C Metric plot shows the PSD based TUH-EEG Abnormal and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the 200 dimension TVM to 158 dimensions.

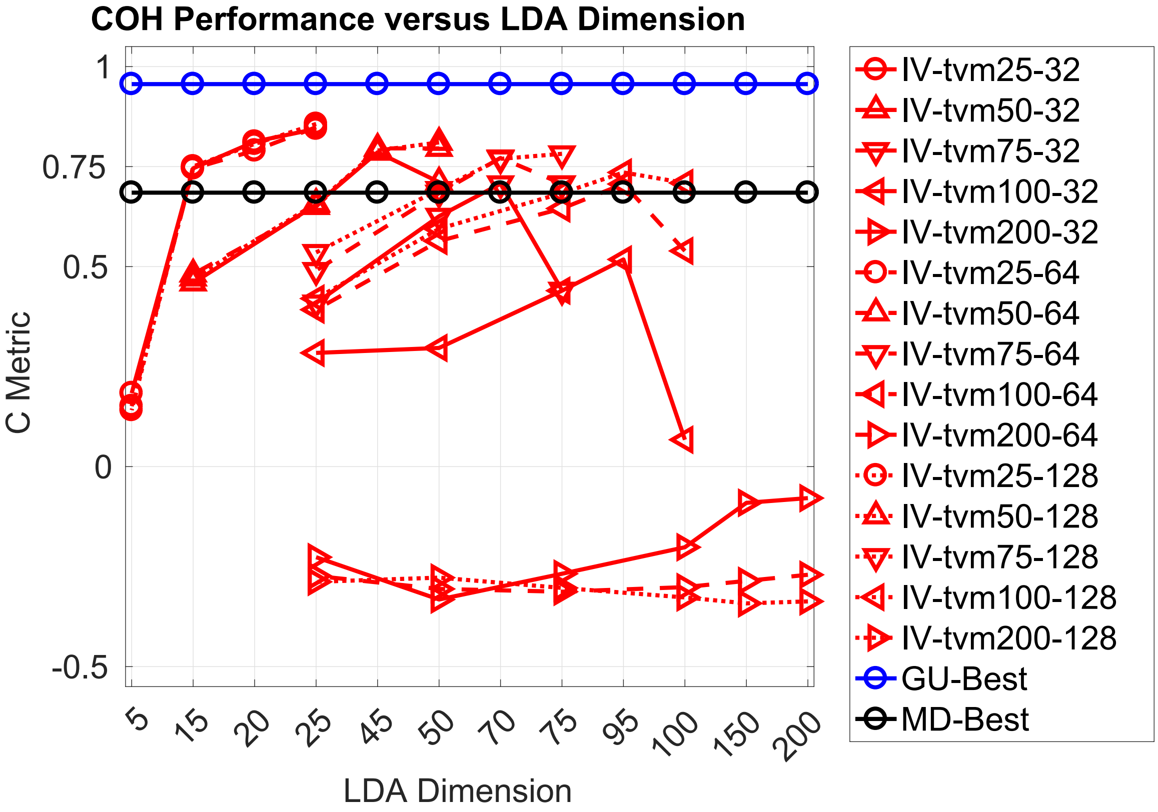


Figure 6.42. C Metric Plot of COH `AbnMot` with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 159 subjects, limiting the 200 dimension TVM to 158 dimensions.

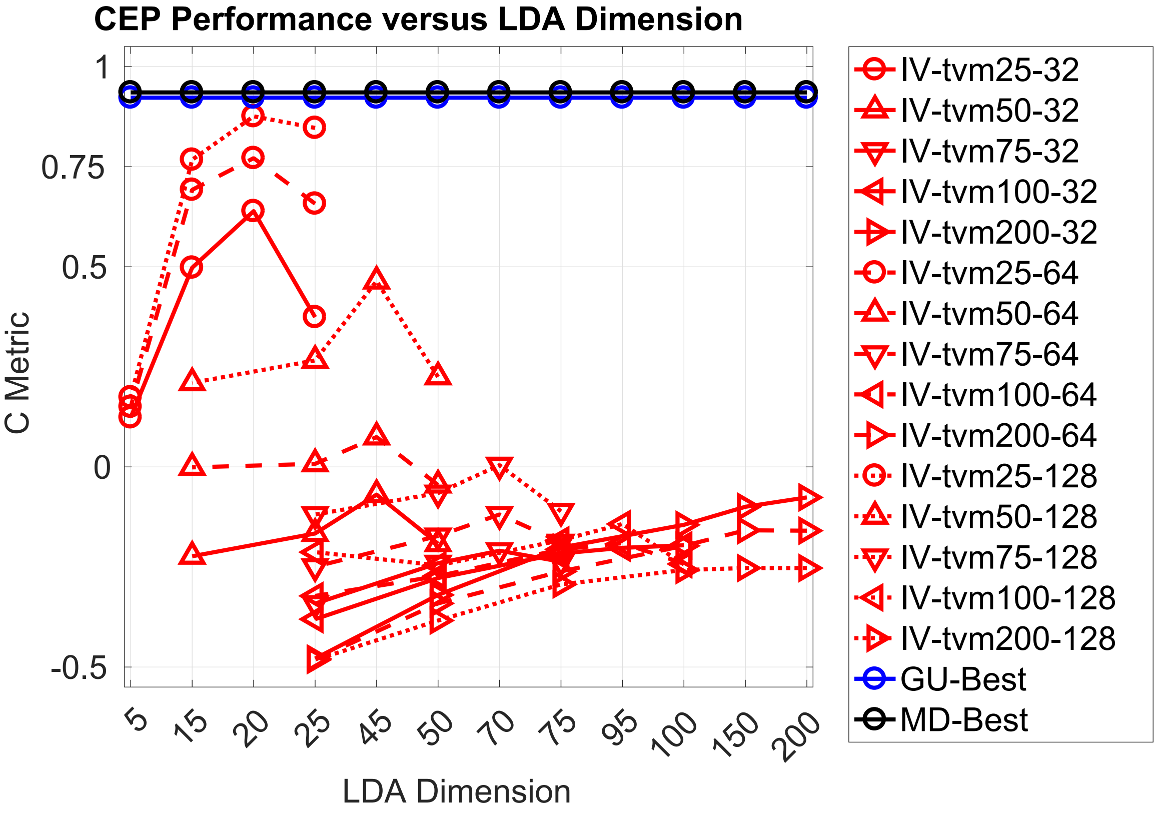


Figure 6.43. C Metric Plot of CEP `NrmSzr` with LDA. This C Metric plot shows the CEP based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

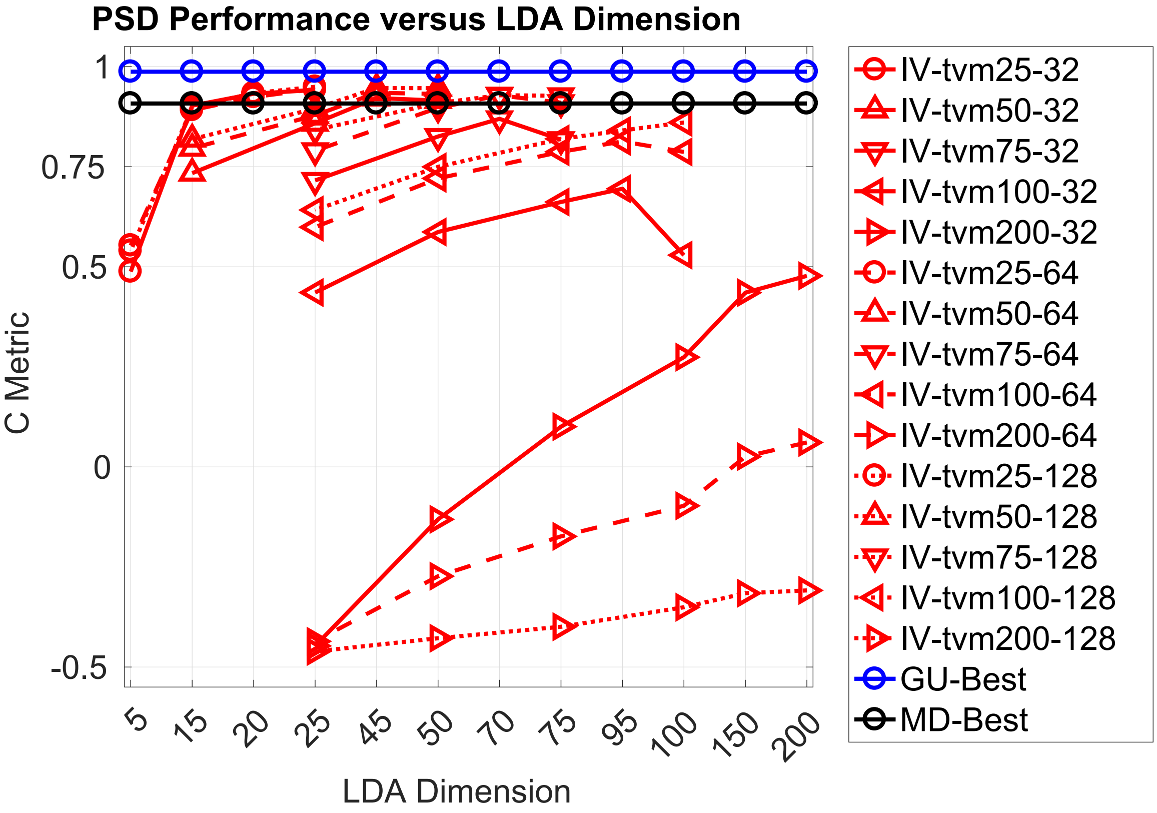


Figure 6.44. C Metric Plot of PSD NrmSzr with LDA. This C Metric plot shows the PSD based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

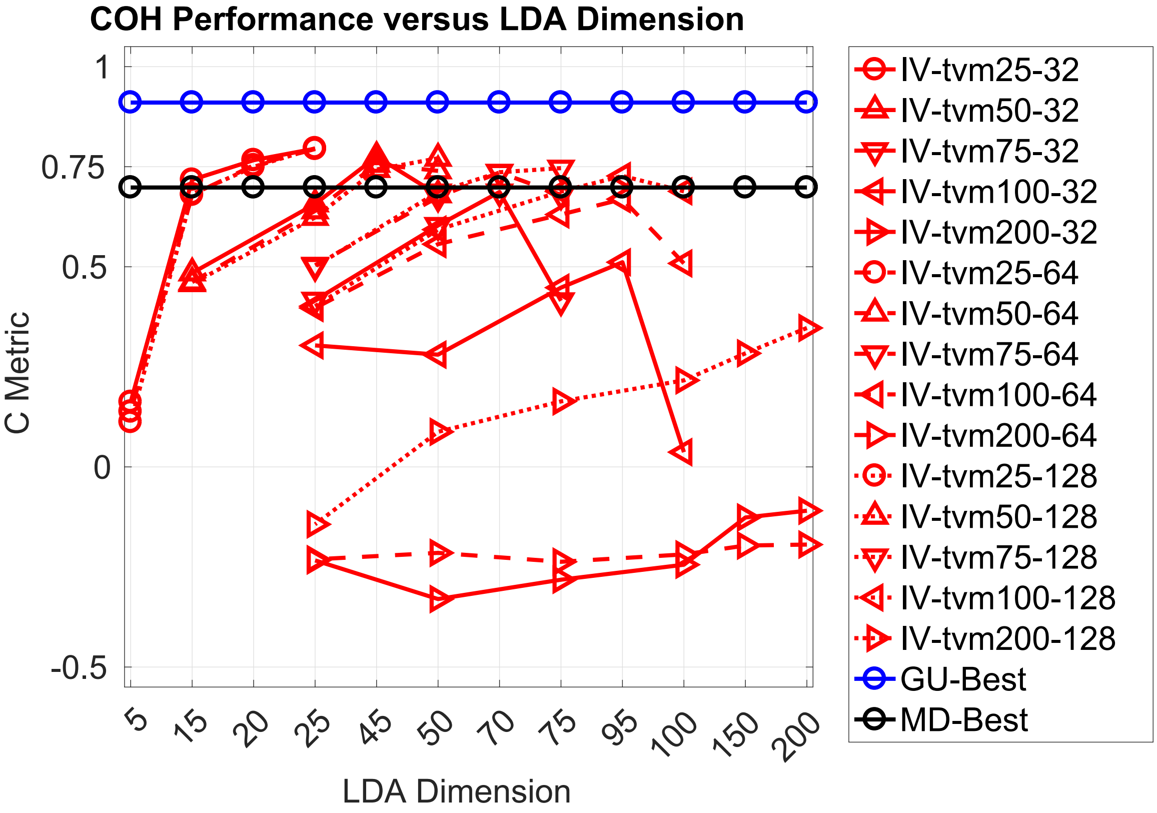


Figure 6.45. C Metric Plot of COH `NrmSzr` with LDA. This C Metric plot shows the COH based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

UBM mixtures when using an LDA improved TVM for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95 respectively. The 32 mixture UBM PSD feature TVMs of dimension 75, 100, and 200 were improved by LDA. The 64 and 128 mixture UBMs only experienced improvement for TVMs of dimension 100 and 200. The COH features were improved by LDA for every TVM dimension aside from 25 for each UBM mixture. Again, these improvements drove the I-Vector scores over that produced by the MD algorithm.

The seventh experiment, Figures 6.49–6.51, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal, Normal, and Seizure datasets. This combined dataset consisted of 511 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with using 64 and 129 mixture UBMs paired with native TVMs of dimension 50 and all mixtures when paired with native TVM of dimension 25 for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95 respectively. The PSD features saw improvement of the TVMs of dimension 75, 100, and 200 for all UBMs. The 32 mixture UBM also saw improvement for the TVM of dimension 50. The COH features saw improvement of the TVMs in dimensions 50, 75, 100, and 200 for each UBM mixture. For TVMs

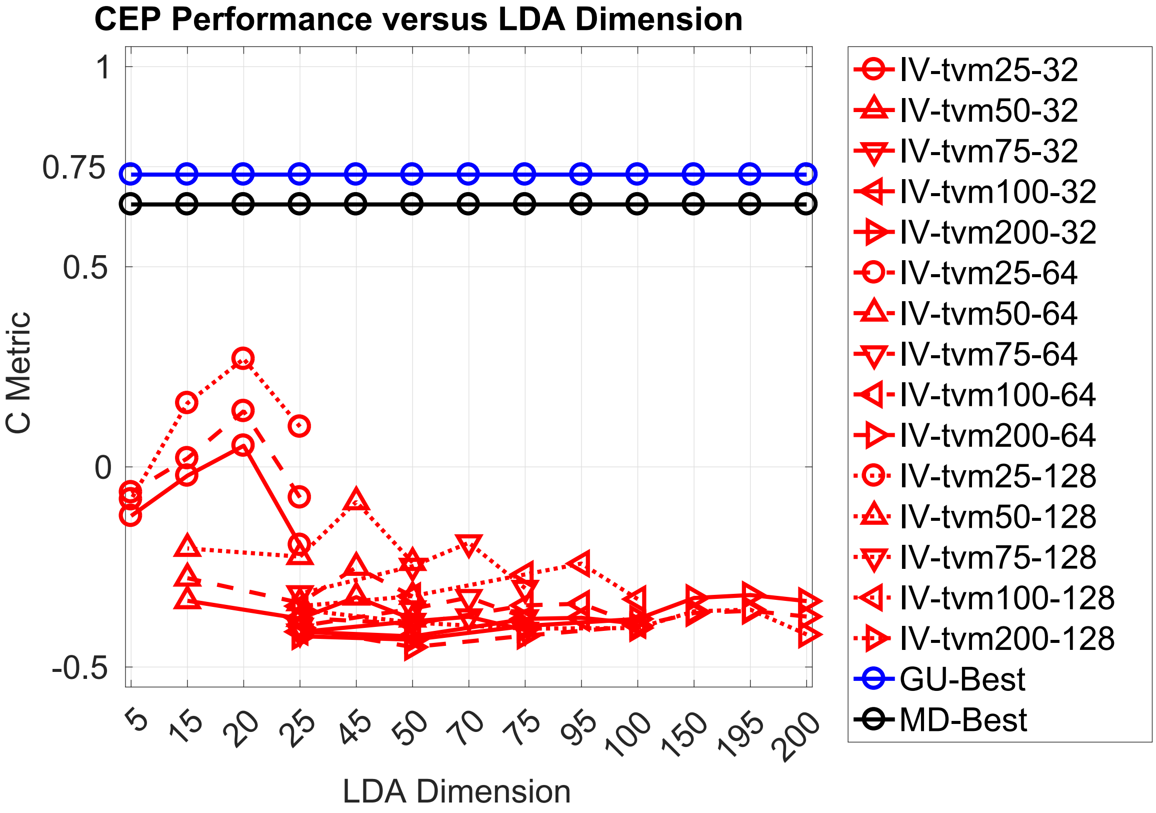


Figure 6.46. C Metric Plot of CEP `SzrMot` with LDA. This C Metric plot shows the CEP based TUH-EEG Seizure and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

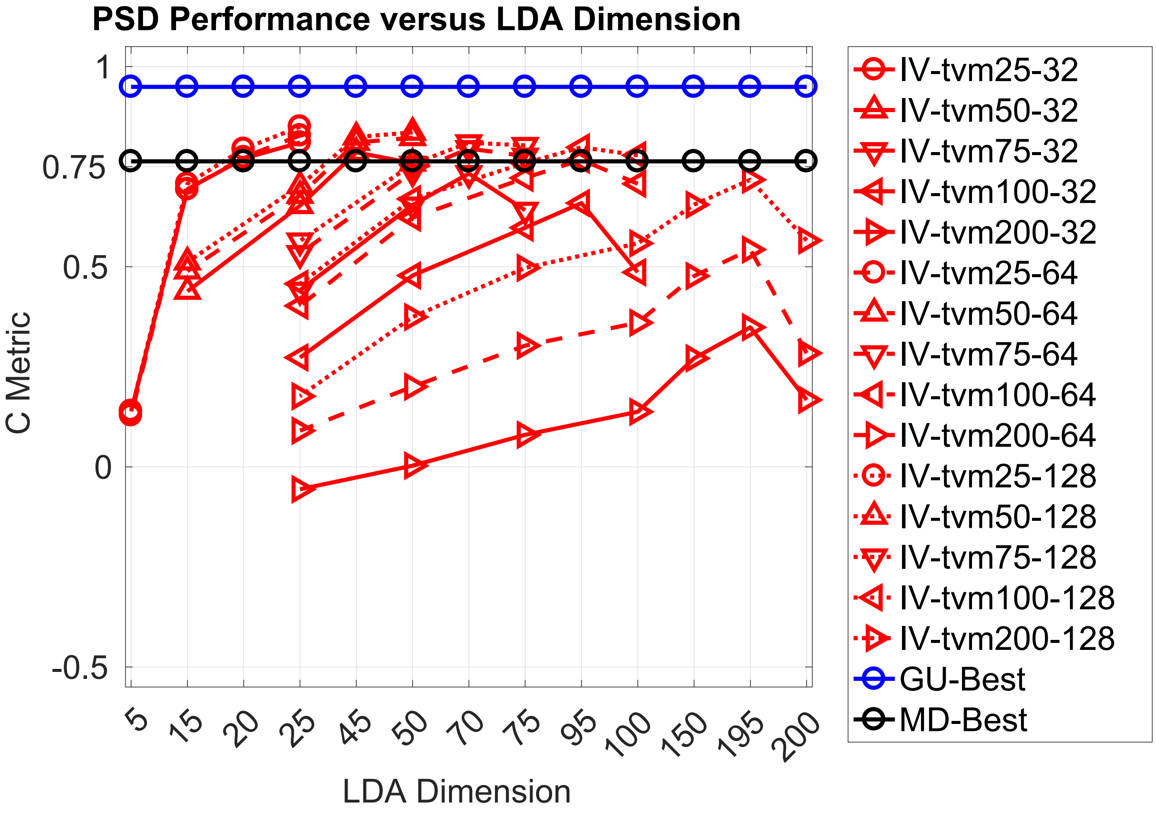


Figure 6.47. C Metric Plot of PSD `SzrMot` with LDA. This C Metric plot shows the PSD based TUH-EEG Seizure and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

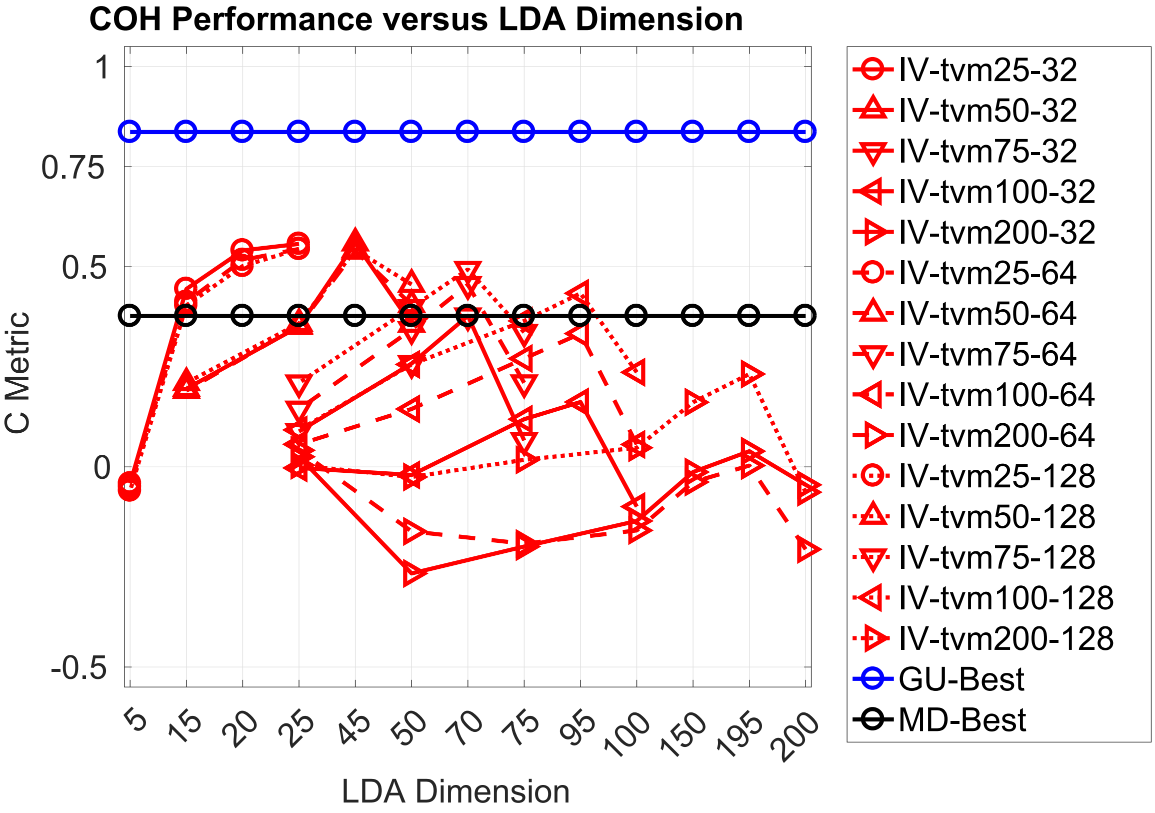


Figure 6.48. C Metric Plot of COH SzrMot with LDA. This C Metric plot shows the COH based TUH-EEG Seizure and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

of dimension 50, 75 and 100 the impact of LDA drove the I-Vector scores over that produced by the MD algorithm.

The eighth experiment, Figures 6.52–6.54, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets. This combined dataset consisted of 209 subjects. The MD algorithm exceeded the 0.75 score threshold for the CEP and PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when using a native or LDA improved TVM for the PSD features. For the CEP features only the 64 and 128 mixture UBMs using TVMs of dimension 25 exceeded the threshold. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of the TVMs in dimensions 25, 50, 75, and 100 for each UBM mixture when it reduced them to 20, 45, 70, and 95 respectively. The PSD features saw improvement of the 32 mixture UBM for TVM dimensions of 50, 75, and 100. The 64 and 128 mixture UBMs saw improvement for the TVM dimensions of 75, 100, and 200. The COH features saw improvement of all UBMs for the TVMs in dimensions 50, 75, 100, and 200. For TVMs of dimension 50, 75 and 100 the impact of LDA drove the I-Vector scores over that produced by the MD algorithm.

The ninth experiment, Figures 6.55–6.57, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets. This combined dataset consisted of 570 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with each of the

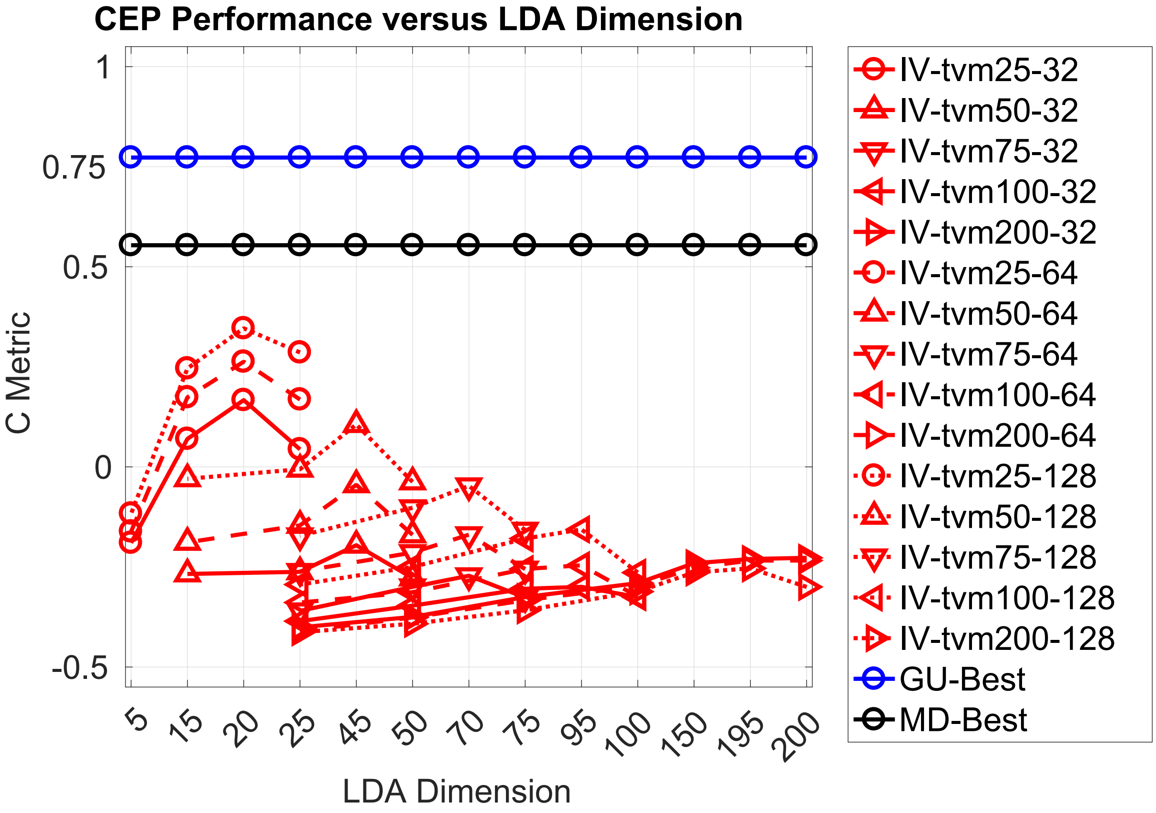


Figure 6.49. C Metric Plot of CEP `AbnNrmSzr` with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

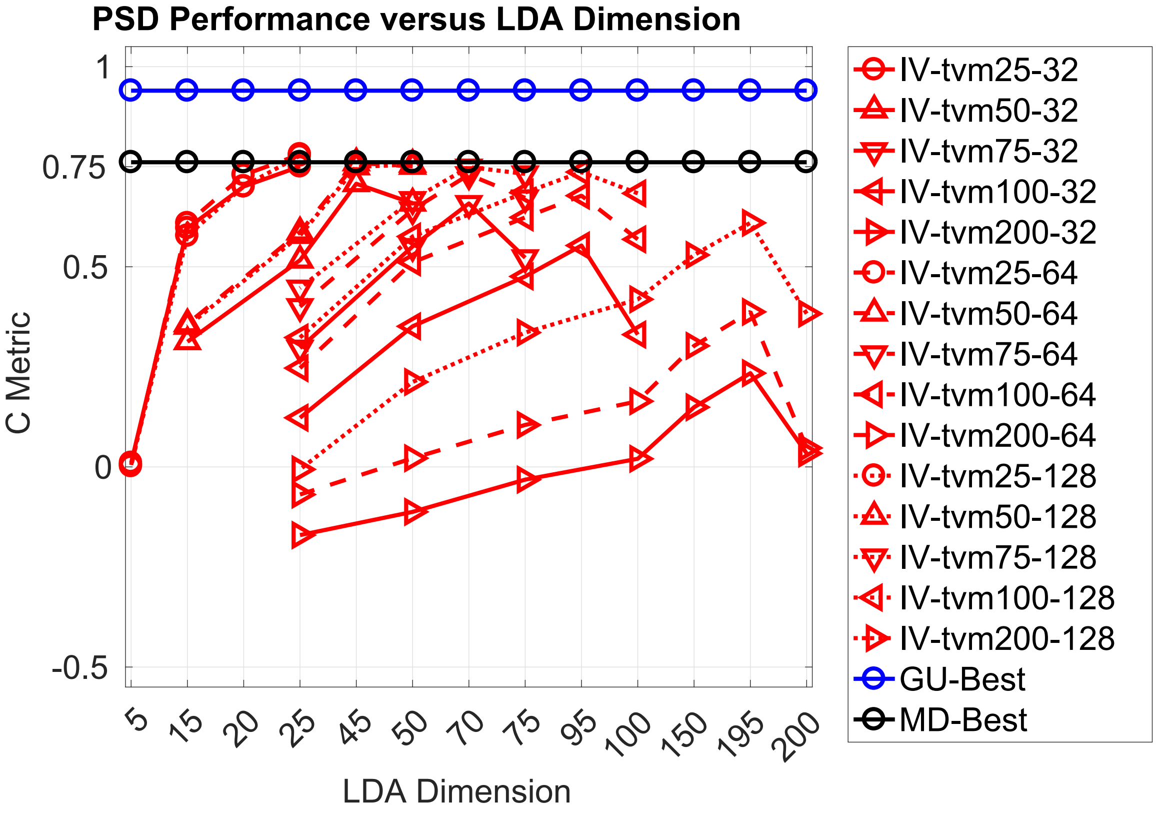


Figure 6.50. C Metric Plot of PSD `AbnNrmSzr` with LDA. This C Metric plot shows the PSD based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

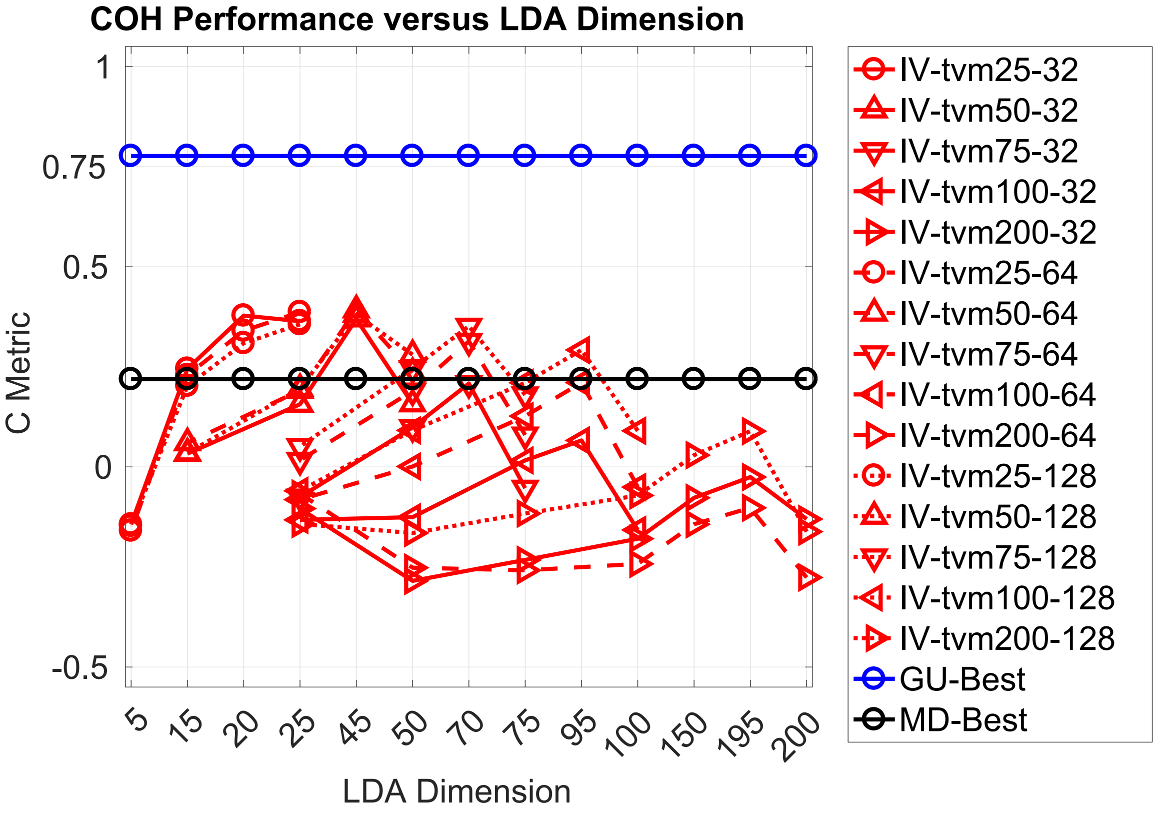


Figure 6.51. C Metric Plot of COH `AbnNrmSzr` with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

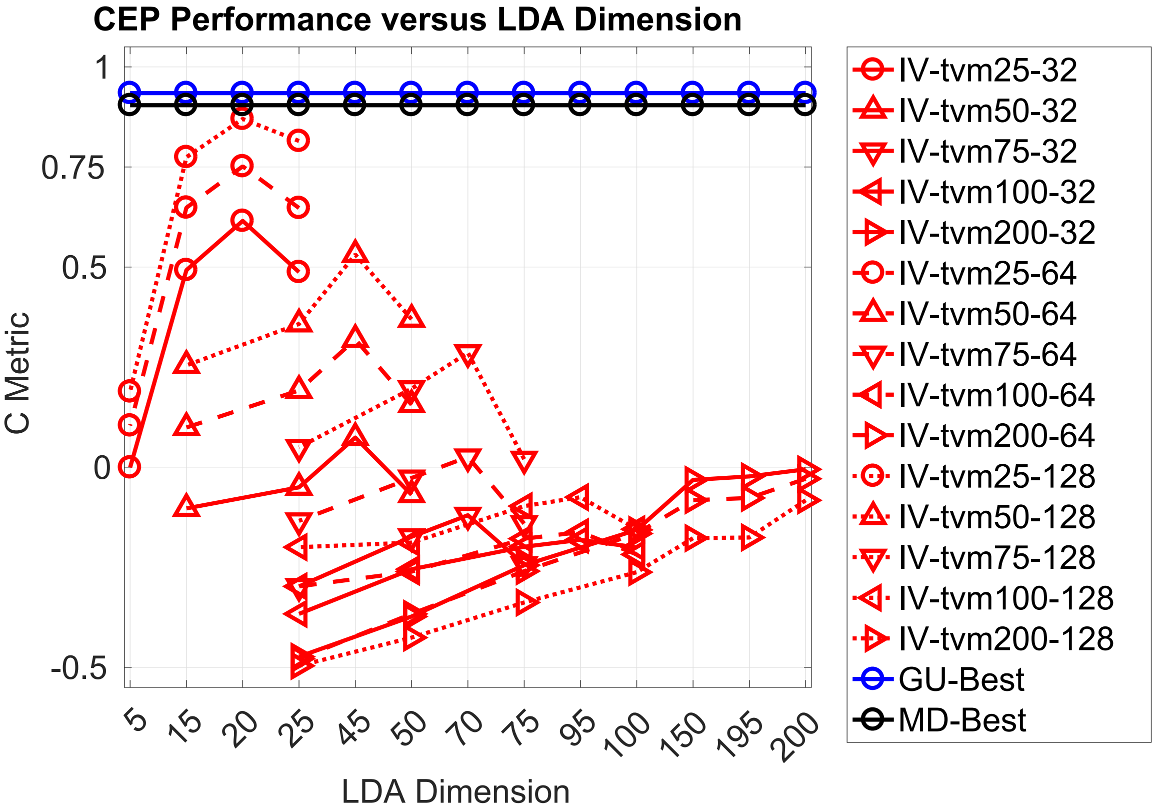


Figure 6.52. C Metric Plot of CEP `AbnNrmMot` with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

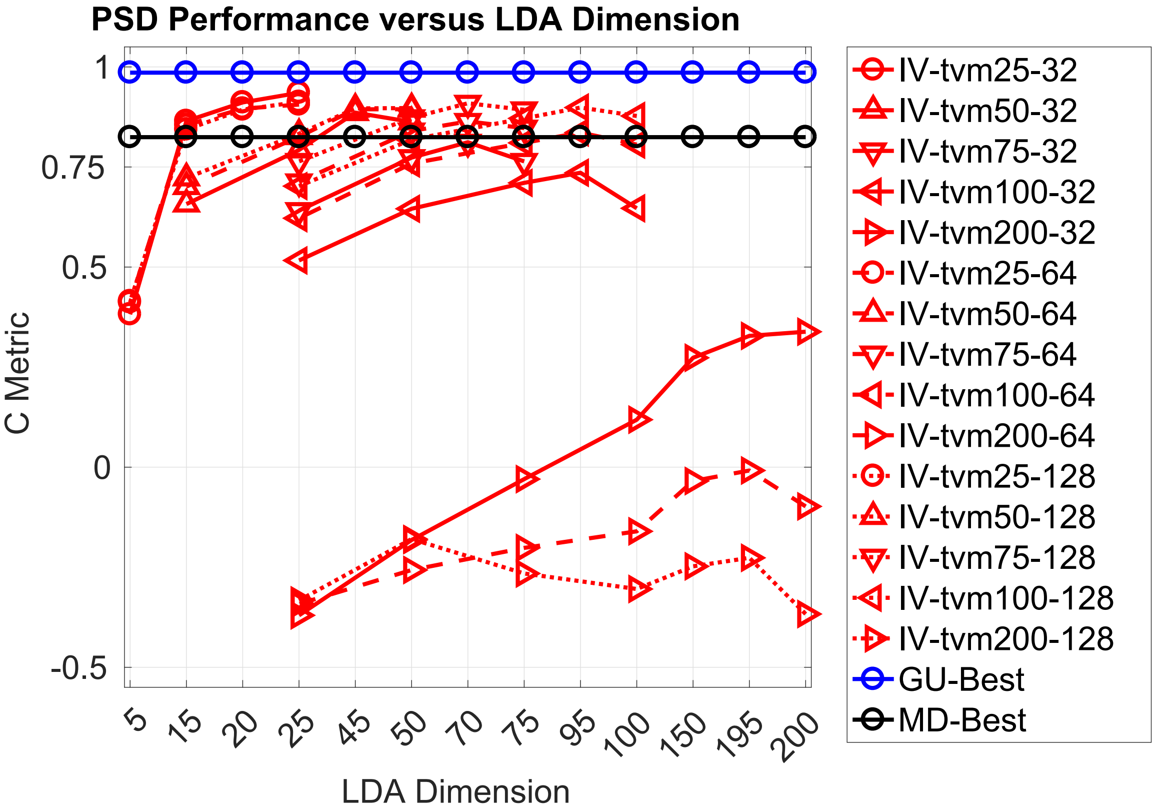


Figure 6.53. C Metric Plot of PSD `AbnNrmMot` with LDA. This C Metric plot shows the PSD based TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

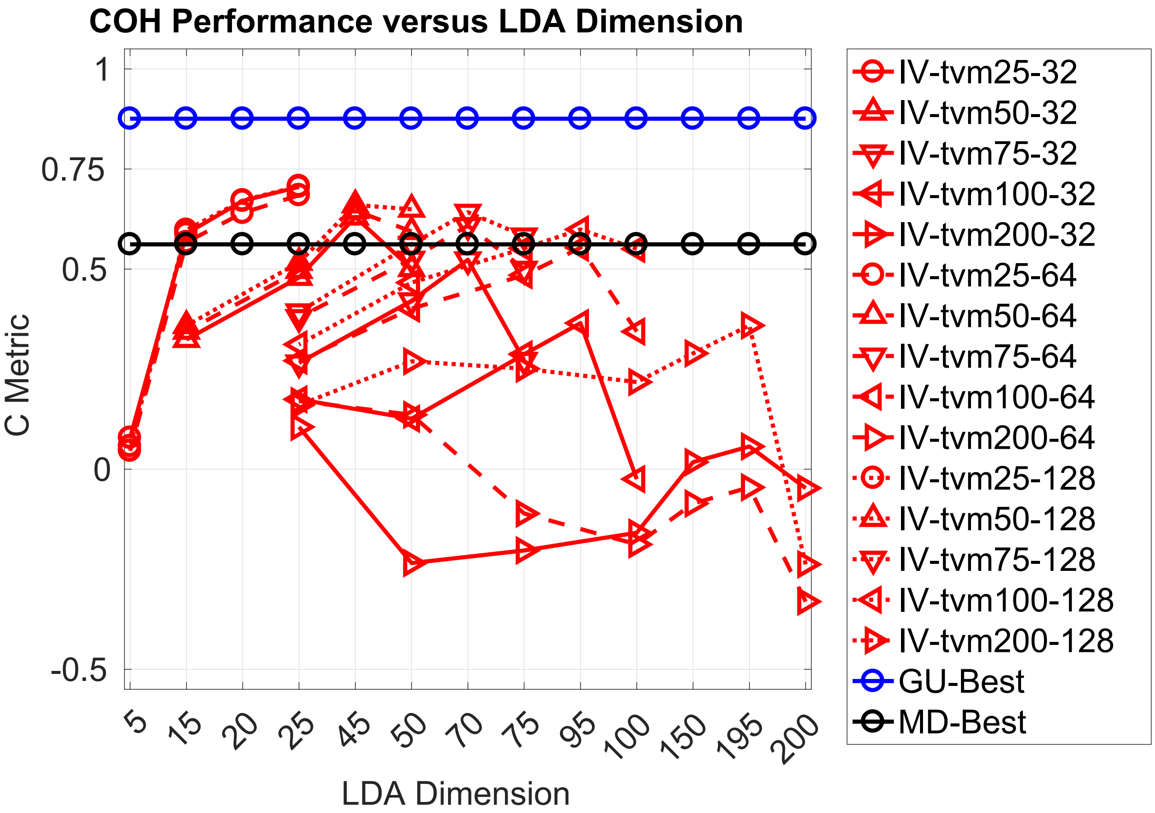


Figure 6.54. C Metric Plot of COH `AbnNrmMot` with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal, Normal, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

UBM mixtures when using an LDA improved TVM for the PSD features. All three algorithms reported their best scores when using the PSD features.

When paired with the CEP features, LDA was able to improve the performance of all TVMs built from the 128 mixture UBM. For the 64 mixture UBM, the TVMs of dimension 25, 50, 75, and 100 was improved by LDA. For the 32 mixture UBM, the TVMs of dimension 25, 50, 75, and 100 was improved by LDA. When paired with the PSD features, LDA was able to improve the performance of the TVMs of dimension 75, 100, and 200. For the 64 mixture UBM, the TVMs of dimension 75, 100, and 200 was improved by LDA. For the 32 mixture UBM, the TVMs of dimension 50, 75, 100, and 200 was improved by LDA. For the TVMs of dimension 50, 75, and 100 the impact of LDA drove the I-Vector score over that produced by the MD algorithm. When paired with the COH features, LDA was able to improve the performance of the TVMs of dimension 50, 75, 100, and 200 for all UBMs. These improvements drove the I-Vector score over that produced by the MD algorithm for the TVMs of dimension 50, 75, and 100.

The tenth experiment, Figures 6.58–6.60, tested the impact of LDA on the reduced UBM mixtures for each of the three feature sets when using the TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets. This combined dataset consisted of 570 subjects. The MD algorithm exceeded the 0.75 score threshold for the PSD features. The GMM-UBM algorithm exceeded the 0.75 score threshold for all three features. The I-Vectors were able to exceed the 0.75 score threshold with each of the UBM mixtures when paired with the native TVM of dimension 25. The 64 mixture UBM exceeded the threshold with a TVM of dimension 50 and the 128 mixture UBM exceeded the threshold with TVMs of dimension 50, 75, and 100. All three algorithms reported their best scores when using the PSD features.

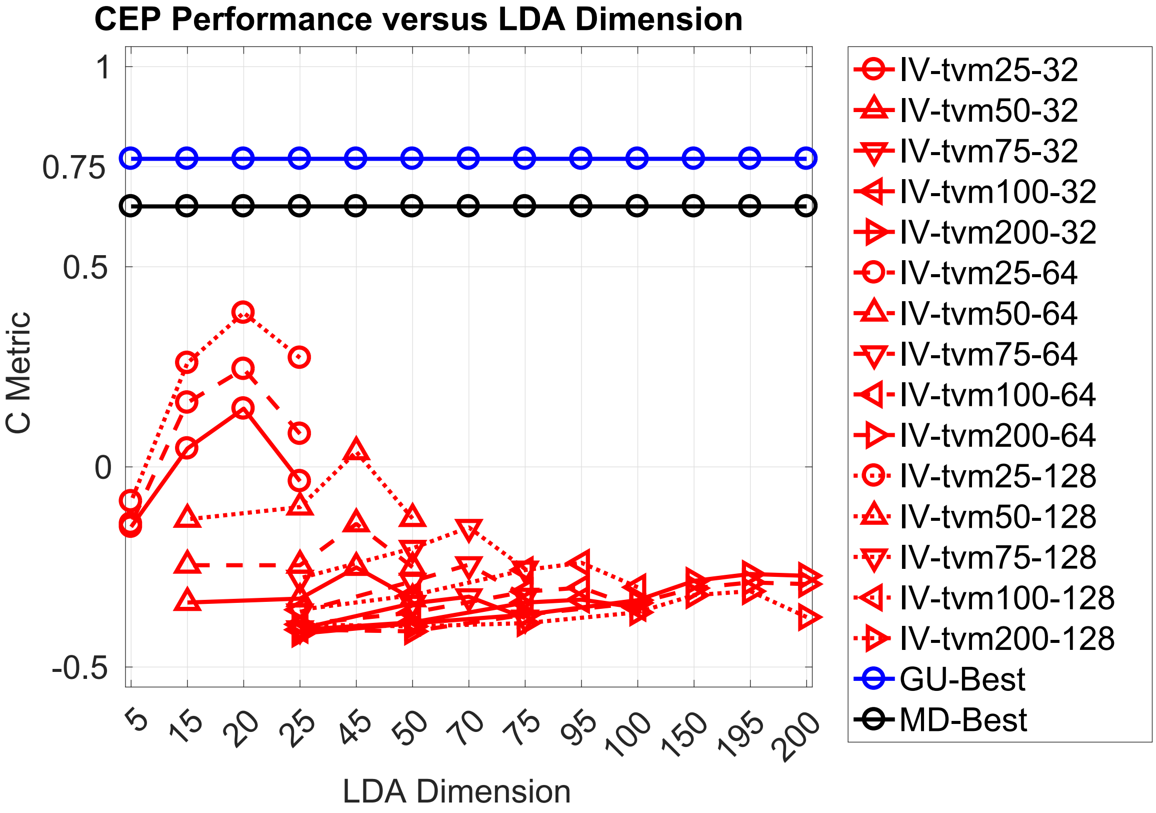


Figure 6.55. C Metric Plot of CEP `NrmSzrMot` with LDA. This C Metric plot shows the CEP based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

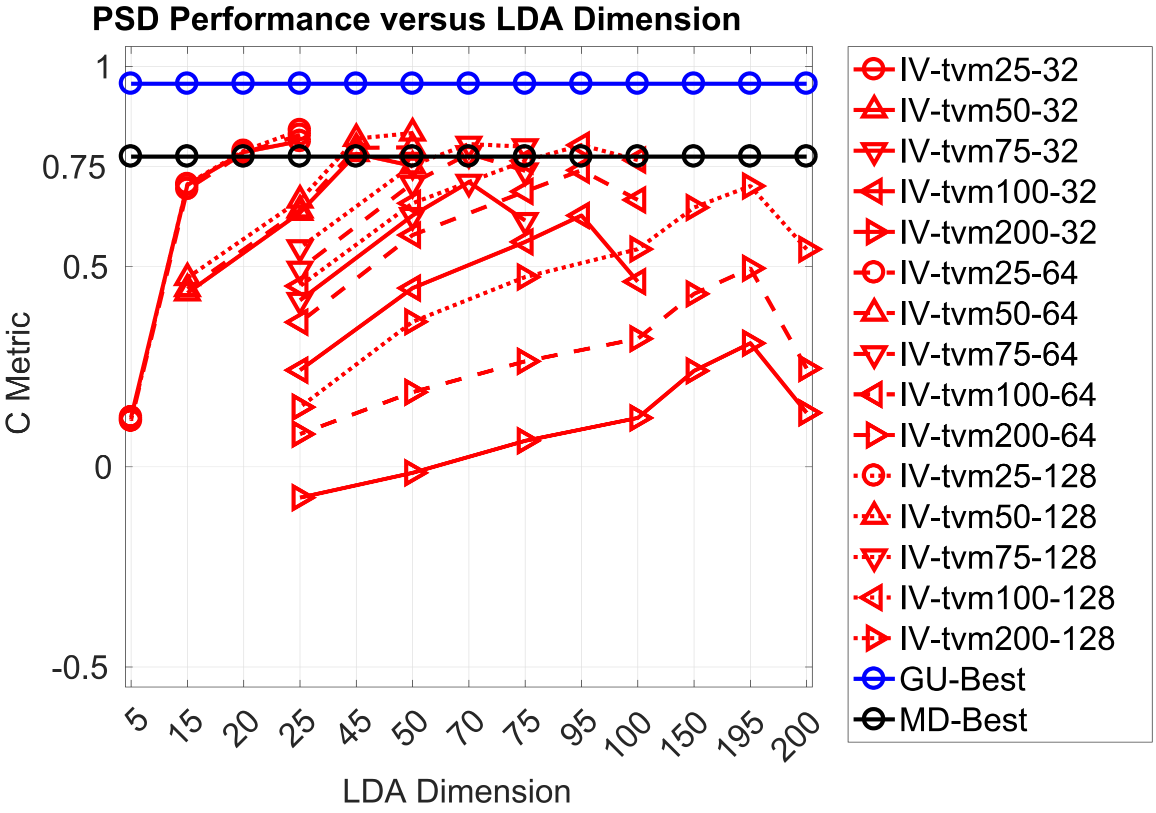


Figure 6.56. C Metric Plot of PSD `NrmSzrMot` with LDA. This C Metric plot shows the PSD based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

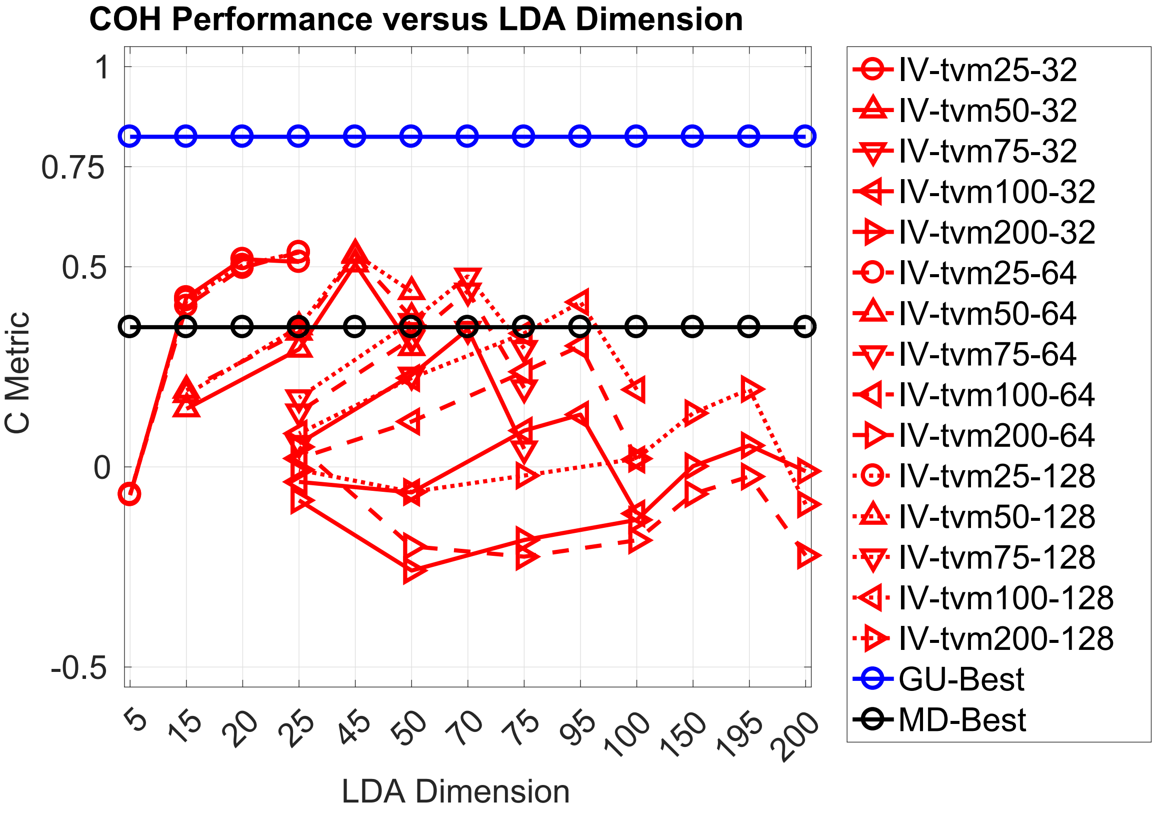


Figure 6.57. C Metric Plot of COH `NrmSzrMot` with LDA. This C Metric plot shows the COH based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

When paired with the CEP features, LDA was able to improve the performance of all TVMs built from the 128 mixture UBM. For the 64 mixture UBM, the TVMs of dimension 25, 50, 75, and 100 was improved by LDA. For the 32 mixture UBM, the TVMs of dimension 25, 50, 75, and 100 was improved by LDA. When paired with the PSD features, LDA was able to improve the performance of the TVMs of dimension 75, 100, and 200. For the 64 mixture UBM, the TVMs of dimension 75, 100, and 200 was improved by LDA. For the 32 mixture UBM, the TVMs of dimension 50, 75, 100, and 200 was improved by LDA. For the TVMs of dimension 50, 75, and 100 the impact of LDA drove the I-Vector score over that produced by the MD algorithm. When paired with the COH features, LDA was able to improve the performance of the TVMs of dimension 50, 75, 100, and 200 for all UBMs. These improvements drove the I-Vector score over that produced by the MD algorithm for the TVMs of dimension 50, 75, and 100.

6.2.2 Discussion

The top scores for each algorithm, dataset, and feature set pairing are given in Figure 6.61. These represent the peak performance of each system within the closed range of UBMs. The minimum acceptable score of 0.75 was not indicated in the table, instead of the top two scores were highlighted for each dataset. The GMM-UBM algorithm had the most high scores with the I-Vector producing the most second highest scores. The MD algorithm failed to produce a single score capable of a top two performance. Regardless of classifier, these scores were predominately tied to the PSD features just like the native results in Table 6.2. Unlike the algorithms' performances, each feature set found itself among the top two scores for a given dataset at least once.

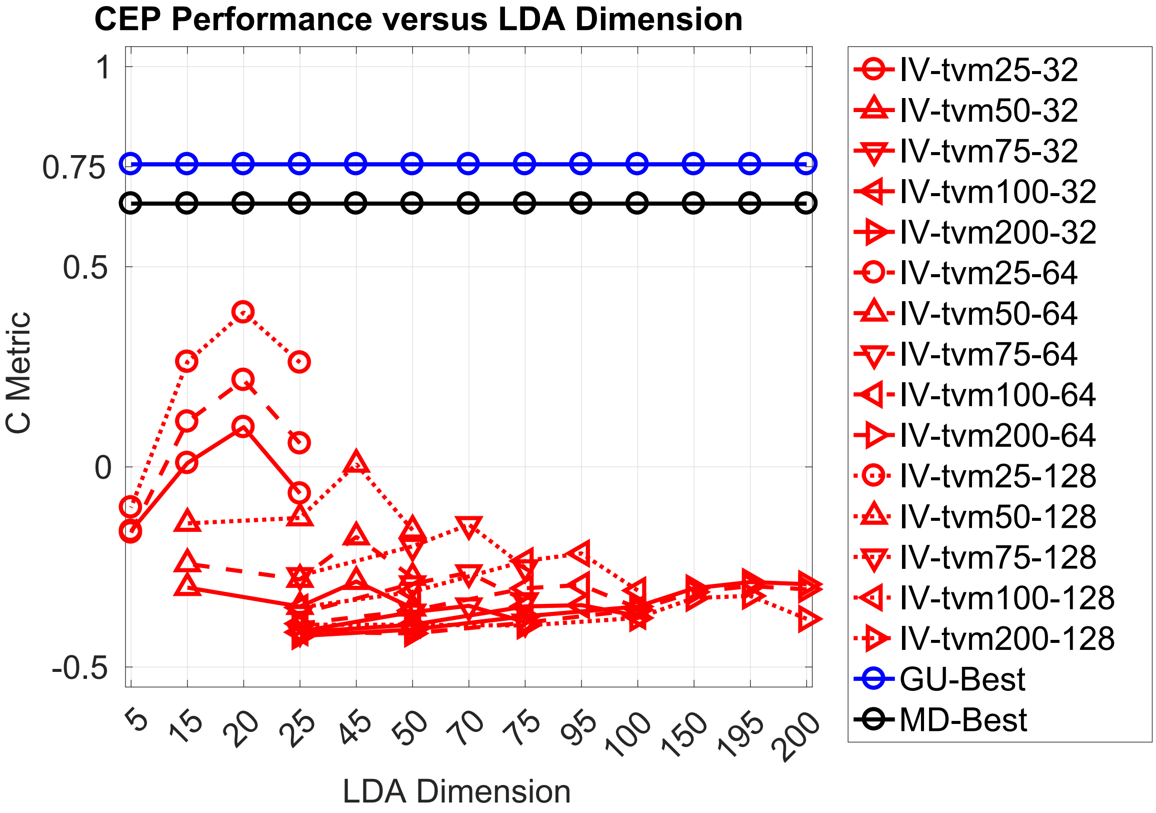


Figure 6.58. C Metric Plot of CEP `AbnSzrMot` with LDA. This C Metric plot shows the CEP based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

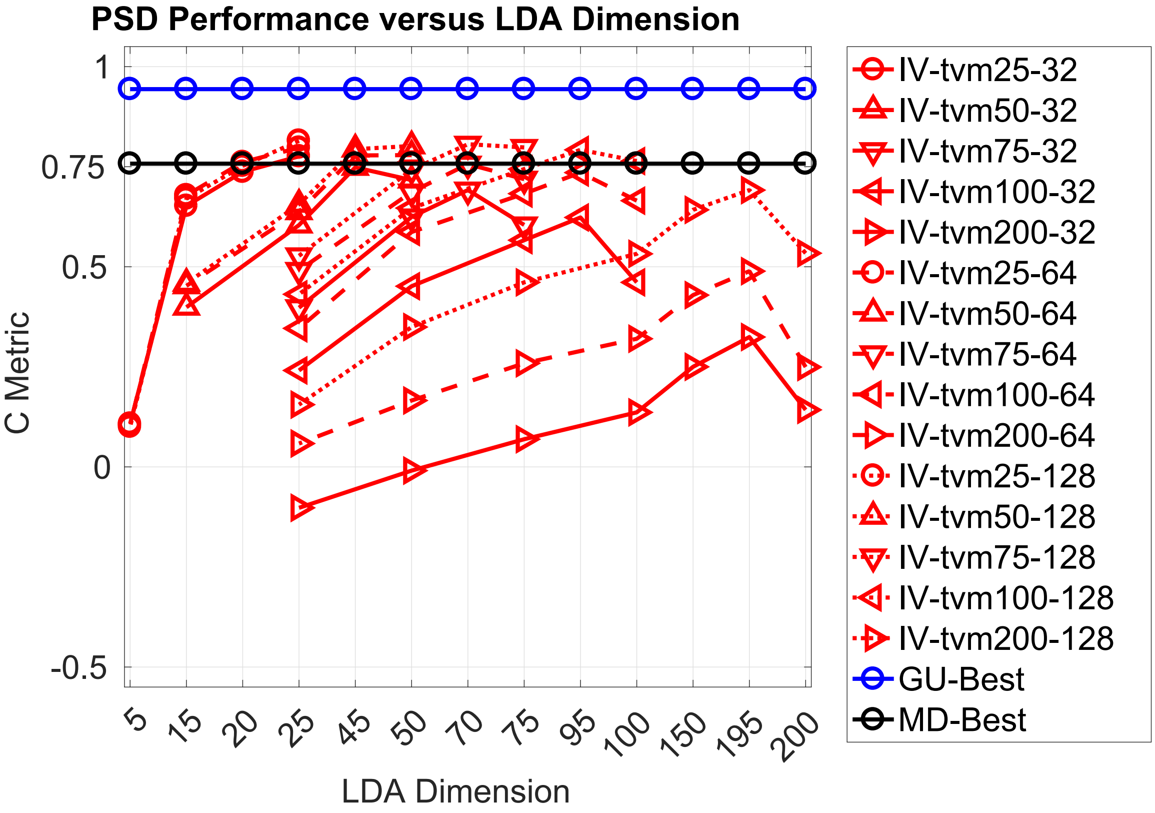


Figure 6.59. C Metric Plot of PSD AbnSzrMot with LDA. This C Metric plot shows the TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

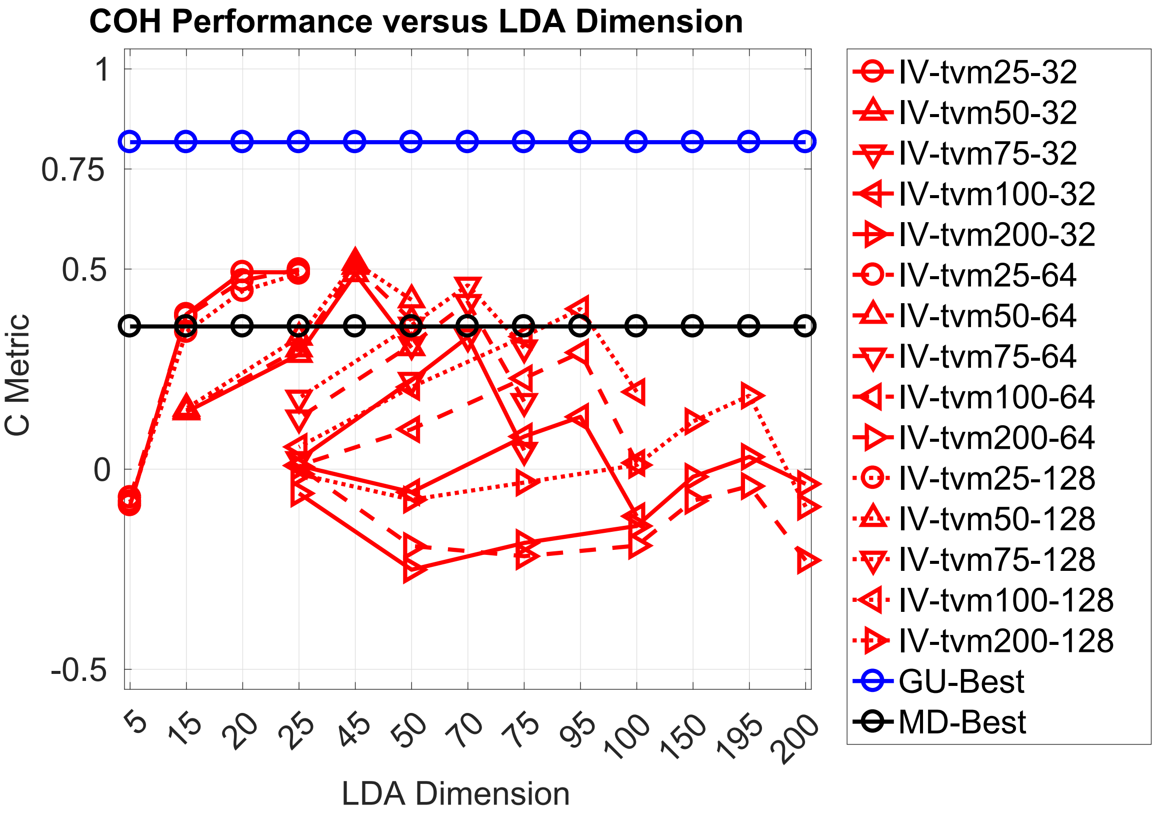


Figure 6.60. C Metric Plot of COH AbnSzrMot with LDA. This C Metric plot shows the COH based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

Figure 6.61. Top C Metric Performance with LDA. Top C Metric Performance with LDA

Dataset	Feature	GU	MD	IV
AbnNrm	CEP	**0.9900***	0.9028	*0.9872**
	PSD	0.9777	0.7800	0.8413
	COH	0.7247	0.1851	0.4306
AbnSzr	CEP	0.7604	0.5835	0.2871
	PSD	**0.9284**	0.7527	0.7791
	COH	*0.7867*	0.2148	0.4262
NrmSzr	CEP	0.7484	0.5835	0.3051
	PSD	**0.9651**	0.7657	*0.8130**
	COH	0.7787	0.2126	0.4221
AbnMot	CEP	0.9388	0.9496	0.8437
	PSD	**0.9874**	0.8994	0.9547
	COH	*0.9563*	0.6849	0.8567
NrmMot	CEP	0.9225	0.9357	0.8760
	PSD	**0.9874**	0.9080	*0.9497*
	COH	0.9104	0.6981	0.7960
SzrMot	CEP	0.7303	0.6558	0.2698
	PSD	**0.9496**	0.7635	*0.8501*
	COH	0.8367	0.3769	0.5594
AbnNrmSzr	CEP	0.7730	0.5583	0.3463
	PSD	**0.9345**	0.7613	*0.7808**
	COH	0.7768	0.2192	0.3921
AbnNrmMot	CEP	0.9349	0.9043	0.8700
	PSD	**0.9856**	0.8245	*0.9354*
	COH	0.8756	0.5616	0.7072
NrmSzrMot	CEP	0.7698	0.6509	0.3851
	PSD	**0.9574**	0.7754	*0.8412**
	COH	0.8248	0.3491	0.5360
AbnSzrMot	CEP	0.7561	0.6579	0.3859
	PSD	**0.9439**	0.7577	0.8149
	COH	*0.8169*	0.3565	0.5163

Changes in placement were marked with an *.

In the native TVM table Table 6.2, the I-Vectors produced 1 top score and 3 second place scores. Confined to the closed UBM set, the I-Vectors were unable to produce a top score, but managed seven second place scores. The majority of the I-Vector scores were produced by TVMs of dimension 25 using the 128 mixture UBM. There were only seven exceptions to this behavior, as shown in Figure 6.62.

Figure 6.62. Top C Metric I-Vector Exceptions. Top C Metric I-Vector Exceptions

Dataset	Feature	Score	TVM	LDA	UBM
AbnNrm	CEP	0.9872	50	45	128
	COH	0.4306	50	45	128
AbnSzr	COH	0.4262	50	45	64
NrmSzr	COH	0.4221	50	45	128
SzrMot	COH	0.5594	50	45	32
AbnNrmSzr	COH	0.3921	50	45	64
AbnSzrMot	COH	0.5163	50	45	128

These seven experiments were run again with an increased set of UBM mixtures given that I-Vectors had performed better under those conditions in the native TVM experiments. Unlike their smaller UBM mixture counterparts, these larger mixtures produced clusters of results based on their TVM dimension, Figures 6.63–6.69. The performance of larger TVMs was shown capable of matching and in one case exceeding the performance, Figure 6.68, of the 25 dimension TVMs. However, even with this increased UBM mixtures the impact of LDA was consistent in improving the performance of all TVMs larger than 25 dimensions regardless of mixture size.

These behaviors suggested that LDA was capable of improving the performance of a given TVM under the right circumstances, but that was insufficient for improving the performance of I-Vectors against the other classifiers. A distinct trend to this

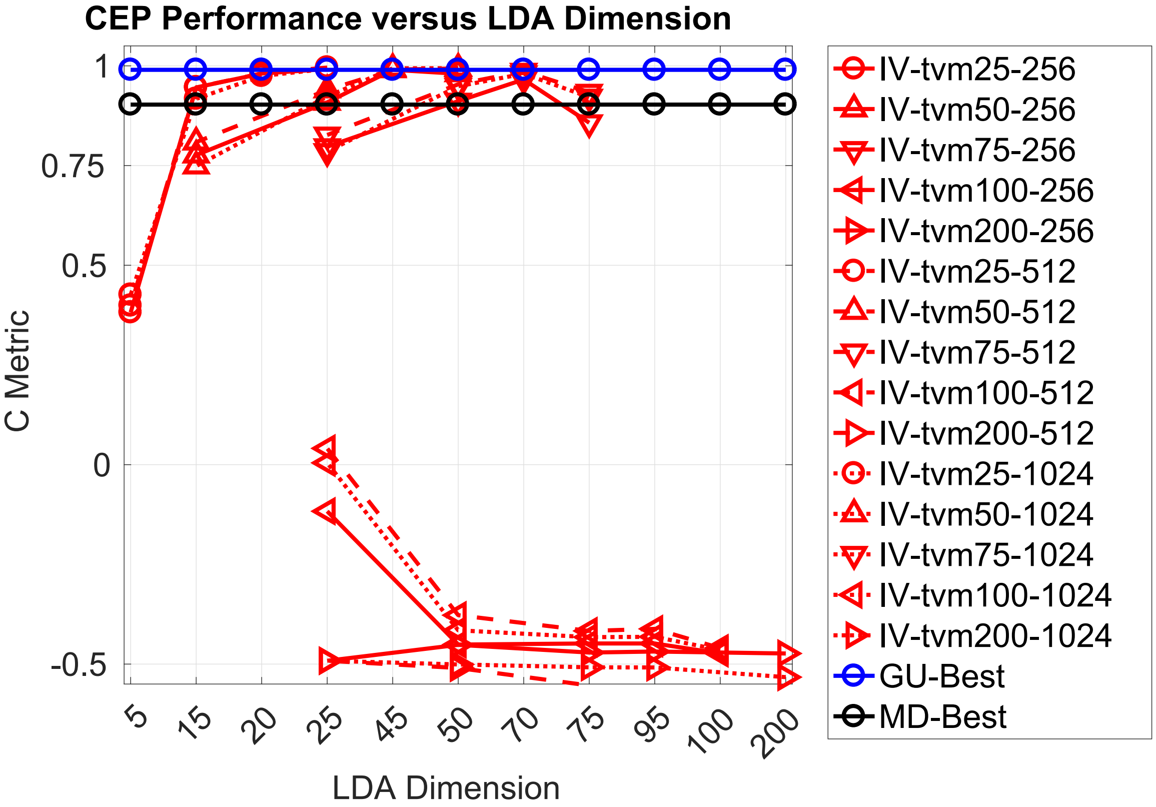


Figure 6.63. C Metric Plot of CEP `AbnNrm` with LDA, Larger UBMs. This C Metric plot shows the CEP based TUH-EEG Abnormal and Normal datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the 100 and 200 dimension TVMs to a dimension of 99.

p]

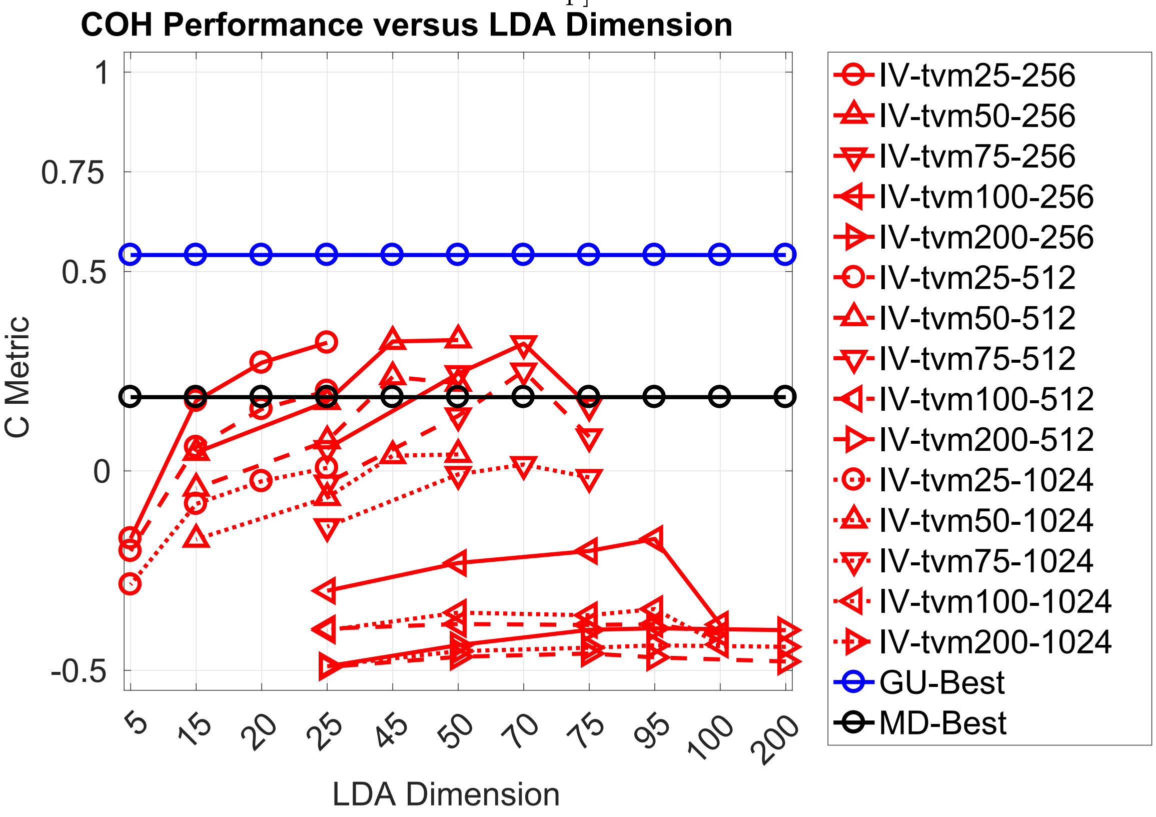


Figure 6.64. C Metric Plot of COH `AbnNrm` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Abnormal and Normal datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures. The dataset contained 100 subjects, limiting the 100 and 200 dimension TVMs to a dimension of 99.

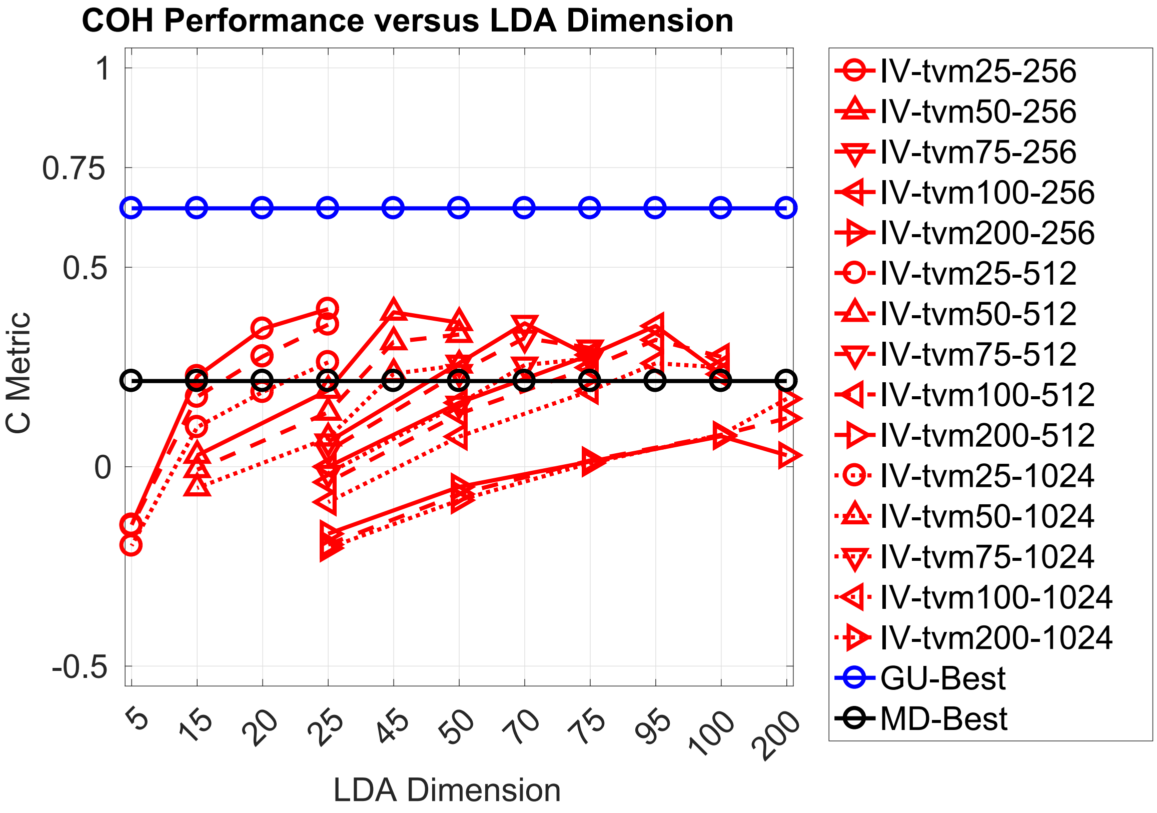


Figure 6.65. C Metric Plot of COH `AbnSzr` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Abnormal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

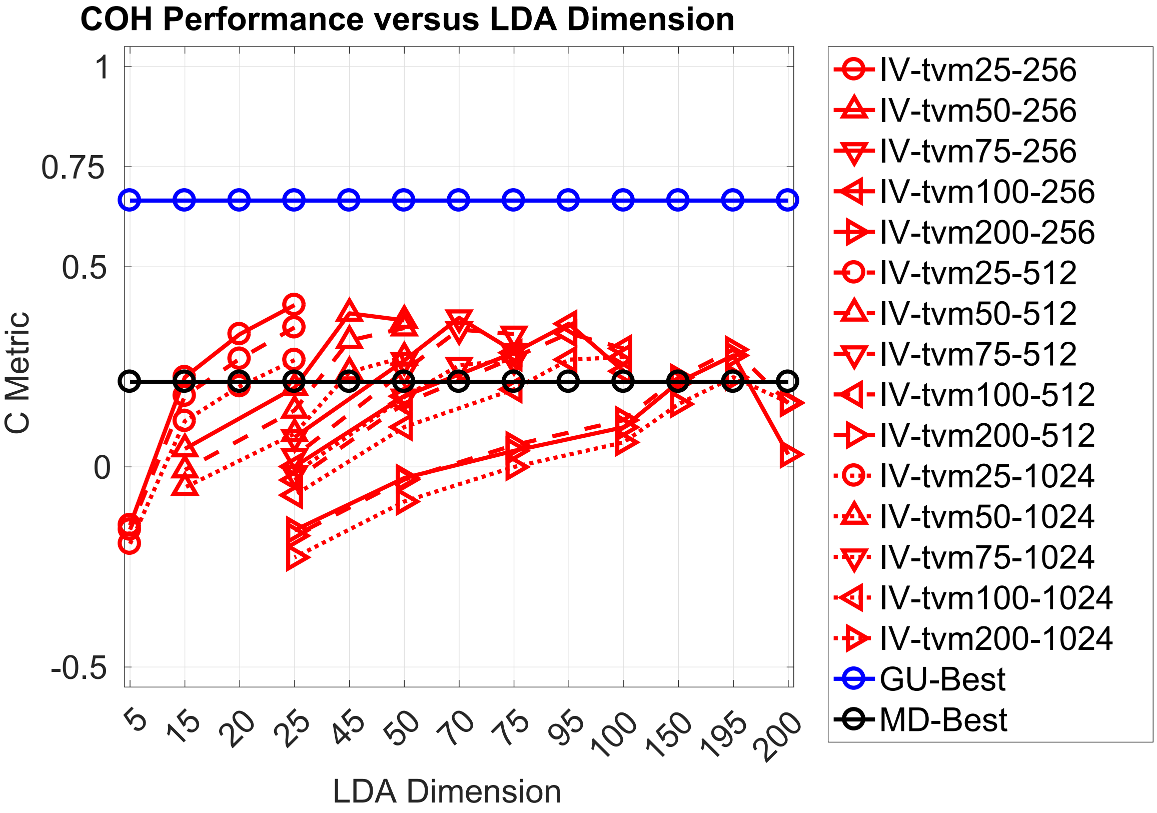


Figure 6.66. C Metric Plot of CEP `NrmSzr` with LDA, Large UBMs. This C Metric plot shows the COH based TUH-EEG Normal and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

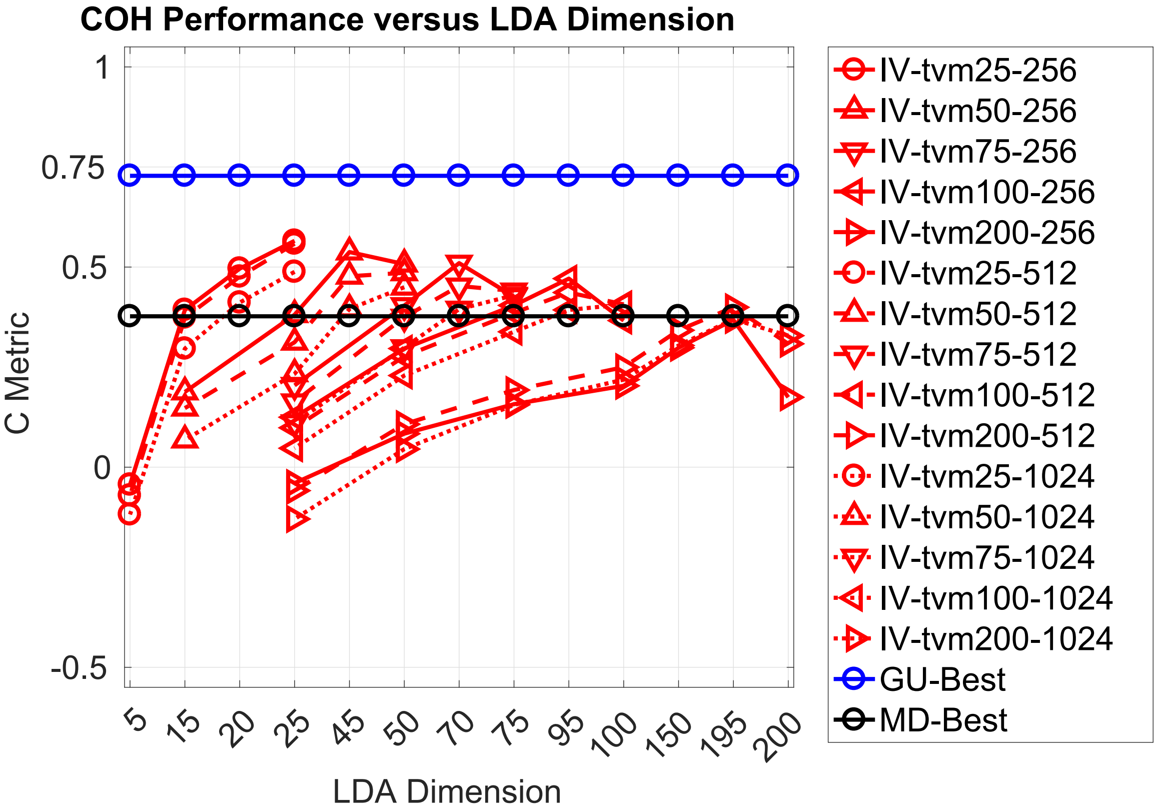


Figure 6.67. C Metric Plot of COH `SzrMot` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Seizure and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

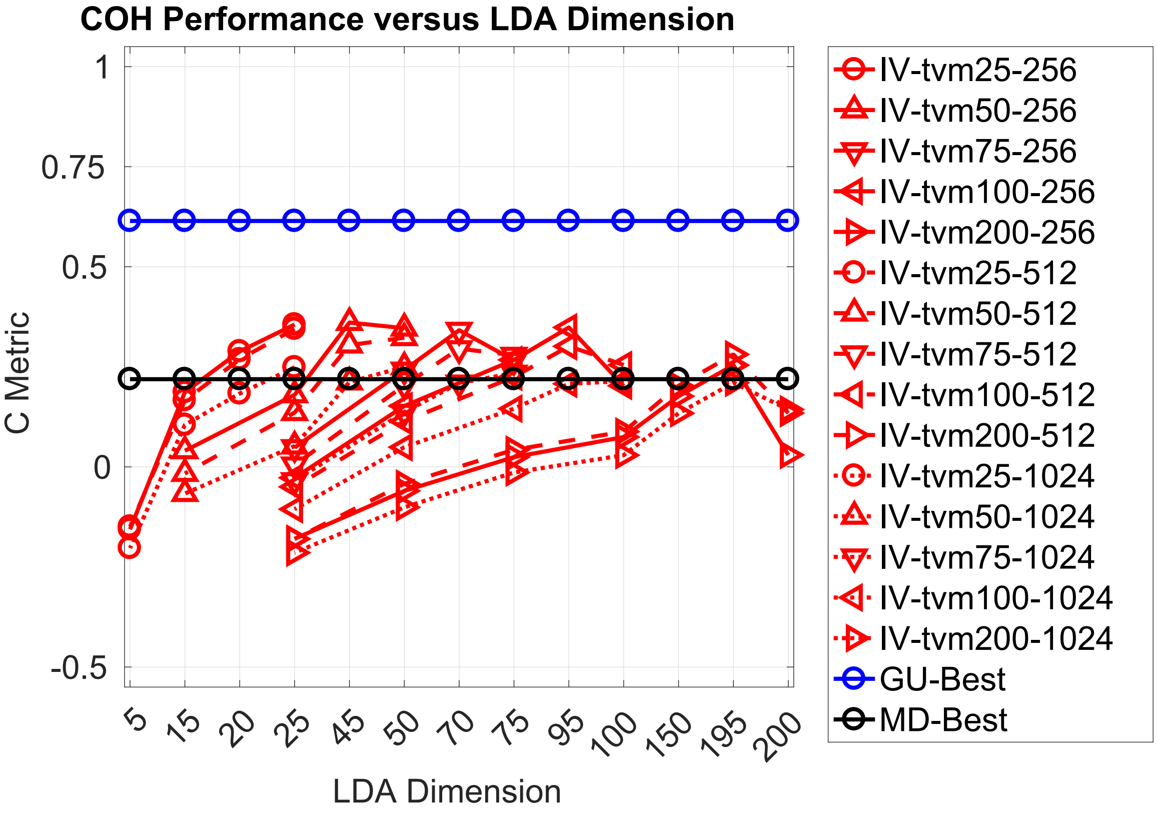


Figure 6.68. C Metric Plot of COH `AbnNrmSzr` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

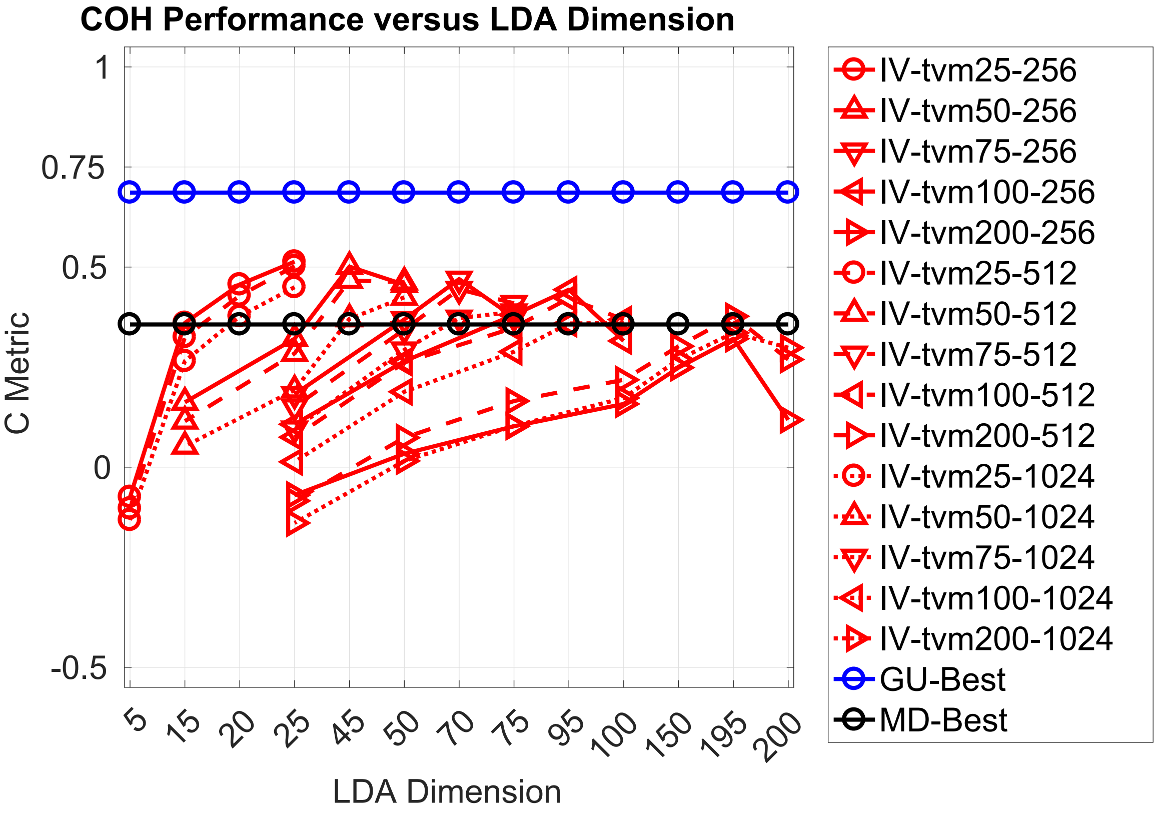


Figure 6.69. C Metric Plot of COH `AbnSzrMot` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Abnormal, Seizure and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

point was the fall off in performance of the smaller UBM mixtures as the TVM dimension was increased, as exemplified in Figures 6.49–6.51. For the smallest TVM dimension, all UBM mixtures produce equivalent performance on the PSD and COH features. Despite increasing the mixtures in the UBM, the larger TVM dimensions appeared capable of only replicating the scores achieved by the 25 dimension TVM for the PSD features. However, the strength of the larger 128 mixture UBM was apparent when using CEP features with the 25 dimension TVM.

The `AbnNrmSzr` was one of the larger and more complex datasets tested, which made it difficult to believe the best classification option would be one of the most dimensional constricting. However, this appeared directly related to the TVM dimension as the larger PSD UBMs, Figure 6.70, showed larger TVM dimensions and their LDA children producing similar levels of performance. This behavior was so consistent it nearly appeared as an artificial ceiling on I-Vector performance as neither the larger TVM dimensions nor larger UBMs produced a stronger C Metric score. This carried over to LDA only improving the 200 dimension TVMs and replicating the native TVM scores for the first LDA steps of the 100 and 75 dimension TVMs.

Therefore the previously assumed operating range of UBMs mixtures was incorrect. As the number of mixtures in the UBM approached and/or exceeded the necessary threshold for a given dataset the TVMs matrices converged their performance. Subtly, the use of LDA indicated this as well by no longer enhancing the performance of the TVM, but suppressing it. On the two largest datasets, Figures 6.71 and 6.72, containing both TUH-EEG Seizure and PhysioNet Database Motion datasets, their trends are nearly mirrored to each other and follow that of Figure 6.70. The C Metric scores appeared to flatten out for the native and first

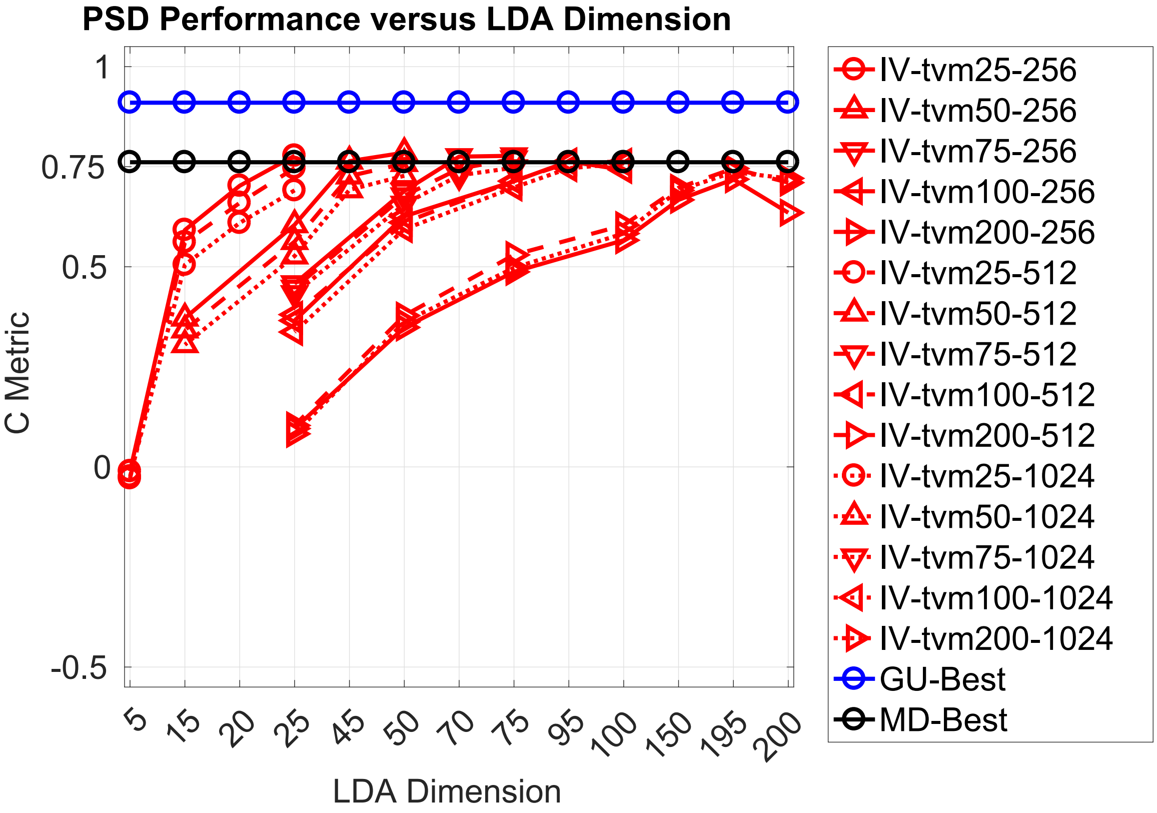


Figure 6.70. C Metric Plot of PSD AbnNrmSzr with LDA, Larger UBMs. This C Metric plot shows the PSD based TUH-EEG Abnormal, Normal, and Seizure datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

LDA reduction of the larger TVM dimensions, but are unable to improve overall performance.

Using LDA when the mixtures of the UBM were insufficiently sized for the given dataset improved performance by improving the independence of the TVM. Once the UBM had achieved sufficient size relative to the dataset, those using the larger sized seizure datasets were the only experiments to produce acceptable scores for TVMs of dimension 100 and 200. LDA's impact was negligible for properly sized UBMs with the initial reduction of 5 dimensions, but would then rapidly degrade performance. Overall, the largest TVMs did not outperform those of dimensions 25 and 50. The same was true of the larger UBMs failing to improve upon the results of the 128, 256 and 512 mixture UBMs utilizing TVMs of 25 and 50 dimensions.

6.2.3 Constraints

The impact that LDA had on the various TVMs was linked to the associated UBM mixture size. Unfortunately, for logistical reasons, the range of UBMs was limited to models up to 2048 mixtures as the single datasets showed performance falling off at the highest UBM mixture sizes. However, the larger aggregated datasets, specifically those paired with the TUH-EEG Seizure dataset, such as in Figures 6.71 and 6.72, could have benefited from UBMs with larger mixture sizes. This was seen in, Figures 6.73 and 6.74, where the TVM dimension and UBM pairings shifted from optimal performance being TVM dimension 25 with a 128 mixture UBM to a 200 dimension TVM with a 2048 mixture UBM.

To control these larger UBMs, the range of TVM dimensions would have needed to be larger as well. In fact, even without increasing the mixture sizes of the tested UBMs, larger TVMs dimensions may have improved the performance of the 1024 and 2048 mixture UBMs and helped define the trend of LDA on those larger matrices.

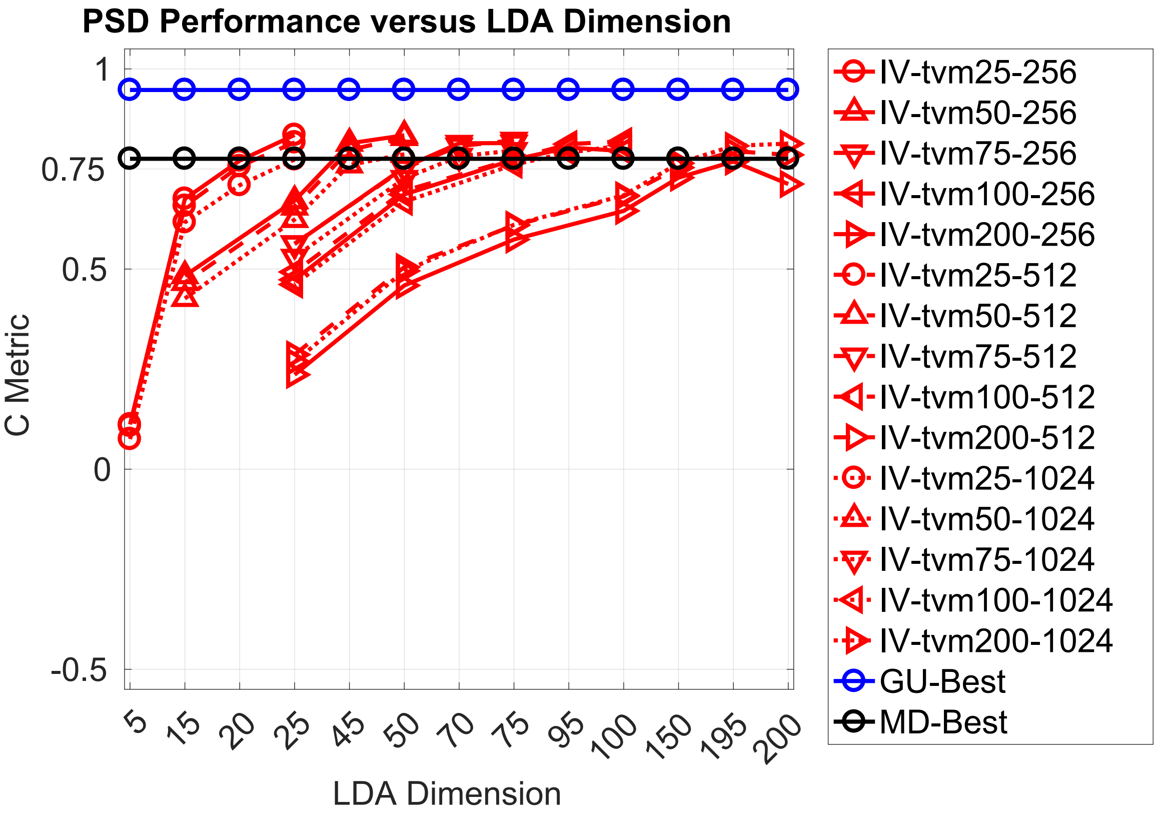


Figure 6.71. C Metric Plot of PSD `NrmSzrMot` with LDA, Larger UBMs. This C Metric plot shows the PSD based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

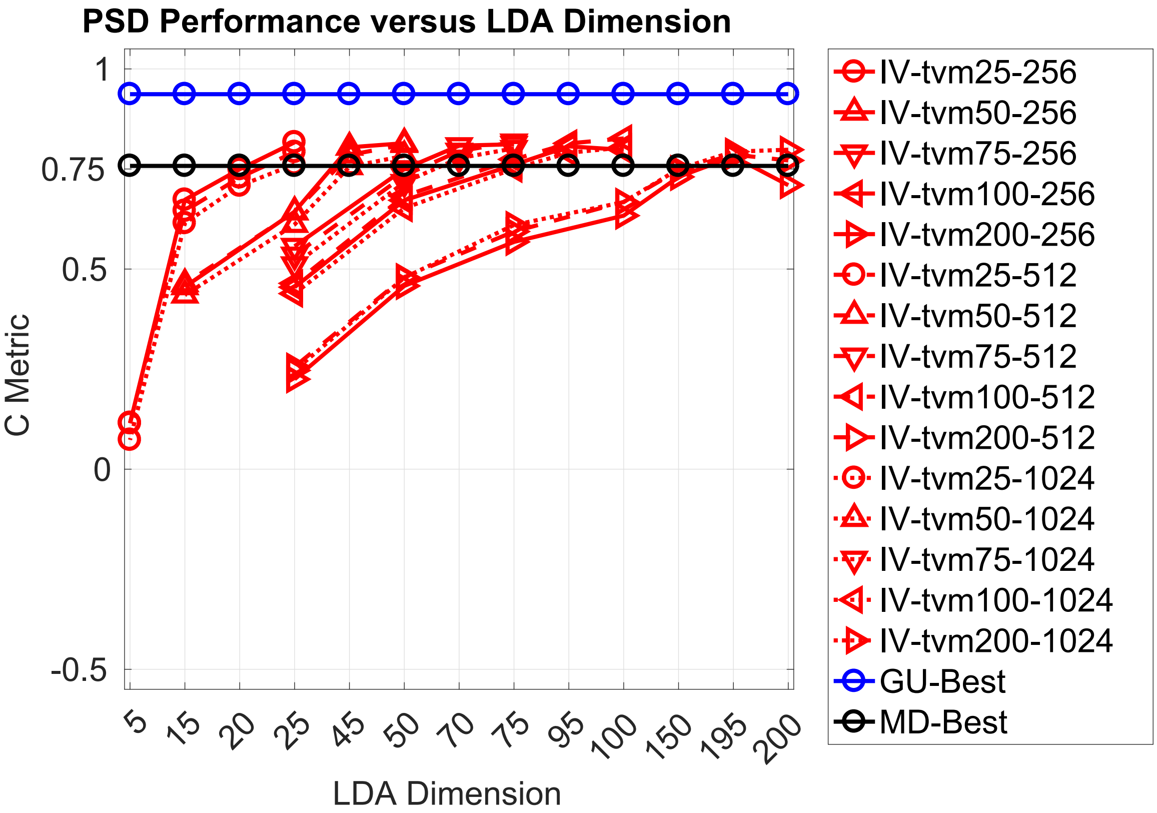


Figure 6.72. C Metric Plot of PSD `AbnSzrMot` with LDA, Larger UBMs. This C Metric plot shows the COH based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

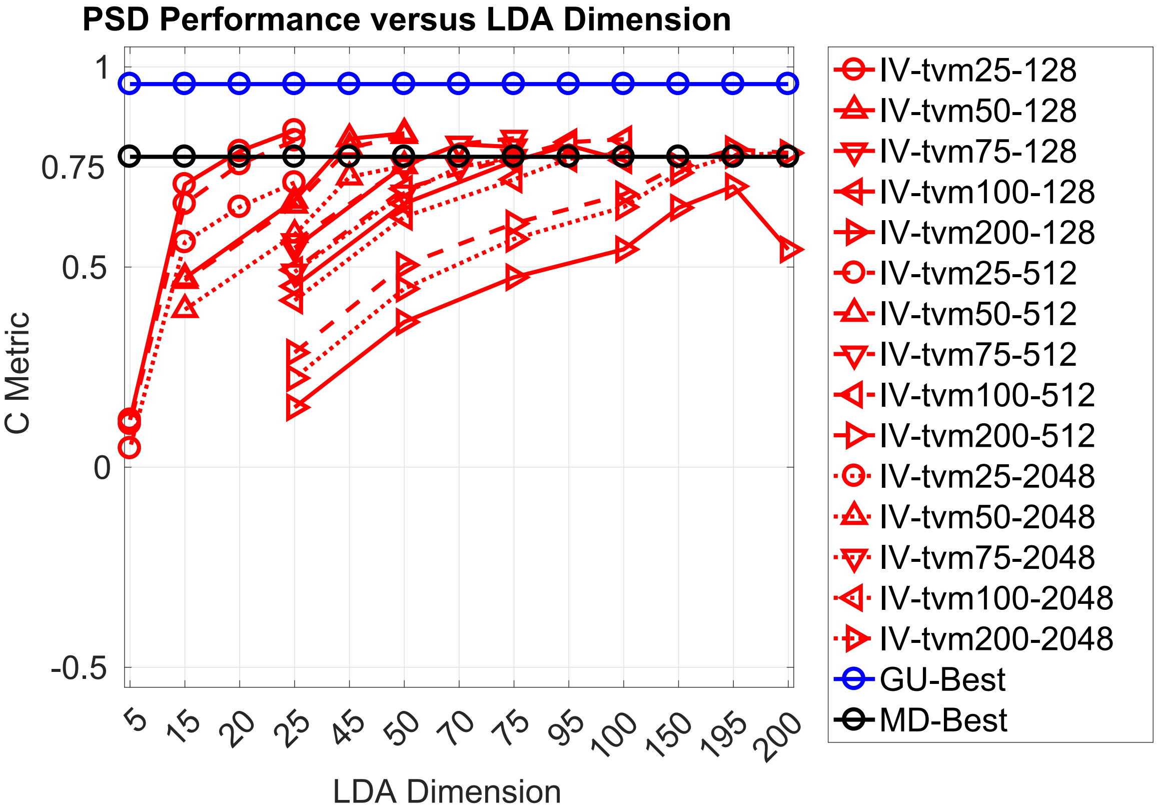


Figure 6.73. C Metric Plot of PSD `NrmSzrMot` with LDA, Largest UBMs. This C Metric plot shows the PSD based TUH-EEG Normal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

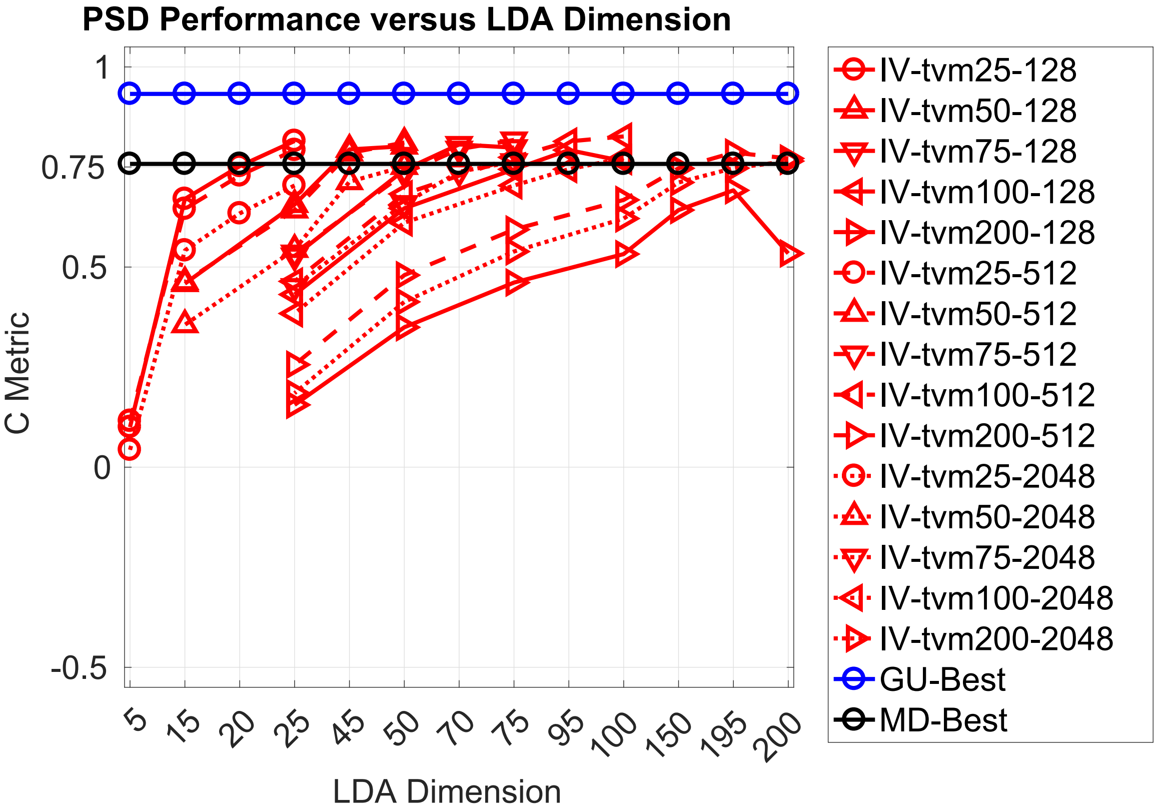


Figure 6.74. C Metric Plot of PSD AbnSzrMot with LDA, Largest UBMs. This C Metric plot shows the PSD based TUH-EEG Abnormal, Seizure, and PhysioNet Database Motion datasets performance as a function of LDA dimension. The UBM mixture sizes are given for the I-Vector and GMM-UBM results. The MD results were not dependent on UBM mixtures.

However, if larger TVMs or UBMs decreased performance that would have been insightful as well. Given that, the results suggested, for the chosen datasets and epoch configurations, I-Vectors encountered a performance ceiling due to the interaction of the TVM dimensions and UBM mixture sizes. Ultimately this limited their C Metric scores to above the MD algorithm, but unable to encroach on the GMM-UBM algorithm.

Resolving when this trend subsided would potentially inform a better understanding of the relationship between TVMs dimensions, UBMs mixture size, and classification performance. This was why using a closed set of UBMs turned out to be ill-informed for the LDA experiments. The native TVMs outperformed their LDA modified counterparts because they used non-optimal UBMs. However, finding the optimal UBM through performance scores alone masked the true nature of the TVM. As LDA was capable of improving the performance of an otherwise deficient native TVM score of 0.50 to 0.92 by reducing its dimensions from 75 to 70, Figure 6.31.

6.3 Conclusion

In these Algorithm Benchmarks, the number of tested subjects was substantially increased from other related bio-metric research, which has frequently relied on the PhysioNet Database as the largest dataset [68, 104, 86, 162]. In addition to their size, the datasets chosen for this work were diverse, covering abnormal, normal, motion, and seizure activity. Across all of these parameters, I-Vectors performed equivalently with the established EEG classification techniques, MD and GMM-UBM. I-Vector performance, when using the native TVM and TVMs enhanced by LDA, showed a remarkable stability across feature sets. Critically, the I-Vectors of dimension 25, 20, and 15 frequently outperformed MD and performed equivalently to GMM-UBM despite operating in the lowest possible dimensional space.

However, the I-Vector technique's ability to classify was exceed by the GMM-UBMs which had the advantage of building complete UBMs for each enrollment and testing subject. Understandably, I-Vectors were not able to exceed the GMM-UBM performance given the increased articulation available via the mixtures of the UBMs. I-Vectors were able to leverage those exact same UBMs to produce similar performance by leveraging the dependencies built into the TVM. This was the actual goal of the experiments: to fully prove the efficacy of I-Vectors on EEGs given claims of their performance in the realm of speech recognition [121, 174, 170, 207].

The MD algorithm performed well when given CEP and PSD features, but this failed to occur when paired with the COH features. Its inclusion provided a lower bound on performance given the simplicity of each enrollment and testing subject was constrained to a feature vector comprised of the mean values of its epochs. This vector was on the order of 26/40 elements for the CEP and PSD/COH features and relied on a pooled covariance matrix built from all the enrollment data. The drawback of this technique was that it provided minimal insight into the nature of the data beyond the models built for each subject.

Everything was tied to the enrollment and testing subjects, which made subject verification possible, but insight into EEGs was limited to the dependencies found in the covariance matrix. Therefore, the selection of data segments used to build the 'subject' data dictated what the algorithm would learn and use to differentiate the subjects. This required *a priori* knowledge to learn content at a deeper level, such as seizure versus non-seizure, whereas individual subjects were easier to generalize.

Conversely, the UBMs used by the GMM-UBM and I-Vector were provided with an abundance of models that may or may not have been based upon the discrimination field (subject, age, disorder, etc). This means that the partitioning of the enrollment and testing 'subjects' required no *a priori* knowledge. Providing datasets based on subject or seizure versus motion did not change what was learned by the UBMs, but rather only altered those mixtures most closely related to those conditions. Mixtures in the UBM would likely pertain specifically to motion, seizure, or subject qualities.

As the GMM-UBM built models for each subject, it was able to specifically target the mixtures contained in each enrollment and testing dataset. There was no mechanism to constrain the creation of these mixtures which led to incredible performance. The drawback was that the number of UBM mixtures could exceed the natural modes of the datasets, turning it into an over-fitted system. This was most apparent at larger mixture sizes as GMM-UBM performance began to roll off seen in Figures 6.28–6.30. With too many equations for the available unknowns, instability was introduced which caused a drop in performance. [1]

This behavior was mitigated by the I-Vector algorithm because the TVM dimensions were limited to one less than the number of subjects. However, this was likely unnecessary given the I-Vector dimensions were constrained by the TVM attempting to model a far larger mixture space. In essence, the dimensions of the I-Vector were not laid out to map to a specific subject, but rather to control changes in the means of the UBM. Attempts to map a single element of the I-Vector to control a single subject would fail because it was necessary to account for all the other elements doing the same through the TVM.

The dependency of the TVM forced the I-Vector elements to work in concert by cooperating to define each subject. An assumption was made that this forced the I-

[1] This is likely the reason for the increased EER from Chapter 5's published epoch sweep results.

Vectors to link mixtures that were related for the given subject. Individual control of the mixtures was otherwise impossible as each additional mixture introduced another feature's worth of elements, roughly 20, to control. The I-Vector of dimension 25 using a 512 mixture UBM that bested the GMM-UBM score in Figure 6.1 was controlling over 10,000 elements in the TVM. As indicated in the LDA experiments' discussion, the TVM dimensions were likely too small given the task asked of them as the UBM grew in size.

Ultimately, I-Vectors worked for the purpose of subject verification using EEG data. The configuration used in these experiments was straightforward, relying on the basic introductory approach proposed by the pioneers of the technique. Since their introduction, various advancements have been made to enhance performance in terms of UBM and TVM generation [169, 209, 210] and I-Vector production [180, 211, 212] and evaluation [213]. These were considered beyond the scope of current work and were therefore not addressed here. However, it was likely these techniques could have improved performance by possibly pushing I-Vector performance past that of their GMM-UBM counterparts.

Chapter 7

UBM-TVM RELATIONSHIP

I-Vector performance was predicated on the ability of the TVM to cull information from the UBM. This process distilled any relationships between the mixtures of the UBM into the TVM while simultaneously reducing the dimensionality to that of the TVM dimension. Therefore it was impossible to discern what each row of the TVM represented. For each I-Vector, TVM rows worked in concert to modify the UBM in the feature space for classification.

The other algorithms, MD and GMM-UBM, were more transparent about how they modified the enrollment and testing datasets to perform subject verification. This made it possible to determine how those algorithms made their decisions, but the information was not insightful (a pooled subject covariance matrix and a multitude of subject specific UBMs). The difficulty with I-Vectors was that the information contained within the TVM was insightful, there was not direct method to extract or interpret it. In fact, this is one of the drawbacks of the technique in that resultant I-Vector elements are unidentifiable components which cannot be uniquely identified [209]. This drawback was less problematic in the context of speech processing because a substantial knowledge base existed related to the details of sound production and human phonemes [123, 137]. With the success of this technique on speech data, and its now proven ability to process EEG data, it therefore became necessary to explore techniques capable of highlighting the logic of the TVM and I-Vector algorithm.

Critically, we have shown that I-Vectors worked well for all feature sets, but performed best with the PSD features, exceeding a C Metric score of 0.75 for many of the aggregated datasets. These performances suggested that the TVMs were capable of distilling the dominant mixtures within a given UBM in a dependent fashion through the lower dimensional I-Vector space. Thus, the focus of this chapter was on determining where those mixture dependencies were occurring in the TVM and how those relationships could be used to enhance classification performance. The working hypothesis was that mixtures from similar sources (normal, abnormal, motion, and seizure) would have various levels of affinity for other mixtures within their native dataset, and likewise for mixtures constructed from other datasets.

An immediate outcome was that it would be possible to identify components of the individual datasets that acted in opposition to each other. These decision surfaces manifested because of the mixture modeling used to produce the UBM and not because of the scope of classification. This was why the resting trials were removed from the PhysioNet Database dataset, to ensure each dataset possessed unique characteristics from the others. Otherwise the `Mot` trials would have been more aligned with the the `Abn` and `Nrm` data that were included (despite being very similar) as controls. Iterating through the 10 aggregated datasets produced distinct edges from which it was possible to learn what mixtures were unique to a given dataset's UBM and then, by association, which mixtures were caused by abnormal, normal, motion or seizure phenomena. Taken to its end, the learning paradigm was limited only by the data available for testing and its ability to produce reasonable classification results.

7.1 Example

The crux of the newly developed approach was to map two distinct UBMs to each other. This was made easier by using a small number of datasets containing common data subsets. This ensured there was a common pool of data within the training dataset and thus a common thread between the UBMs for the mapping algorithm to identify. As an example, the process of matching the *target* PSD `AbnMot` dataset, Figure 7.2 into the *base* PSD `AbnNrm` dataset, Figure 7.1, started with the construction of their 8-mixture UBMs. The visualization of larger mixture UBMs would be fruitless given the overlap between mixtures as they increase in number within a confined feature space. Yet, an 8-mixture UBM example helps illustrate the intent of the proposed technique by showing that the UBMs (a) produced unique mixtures and (b) appeared distinct from one another.

Naturally, the mixtures within each UBM that appeared most unique would be of interest. However, the UBM attempted to build its models to cover the entire dimensional space of the dataset, so “unique” does not necessarily imply “important”. Recall that none of the datasets were pre-processed to correct or remove artifacts. In the base, Figure 7.1, and target, Figure 7.2, distributions, the means and variances of the mixtures are similar making it difficult to know if they represent artifacts or underlying trends of their dataset.

In addition, the mixtures’ weights vary, deviating quite far from the anticipated average weight of 12.5% with the light blue mixtures accounting for upwards of 20% of the dataset. As more mixtures are added, they begin to resolve into smaller subspaces originally outlined by the 8, 16, and 32-mixture UBMs. Thus even the mixture weights offer less insight for larger mixture UBMs as they drop below 2% by the 6th UBM iteration.

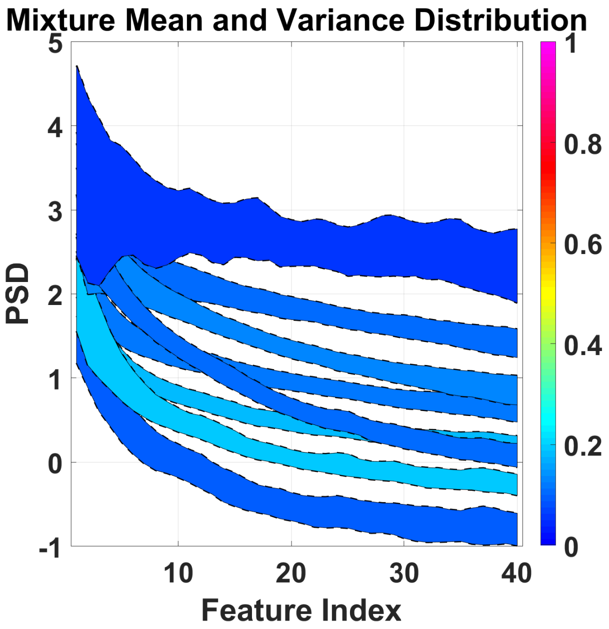


Figure 7.1. The PSD AbnNrm UBM Mixture Distribution. Distribution of the 8-mixture PSD UBM AbnNrm base. Coloration indicates mixture weight. Area of a mixture is +/- one standard deviation.

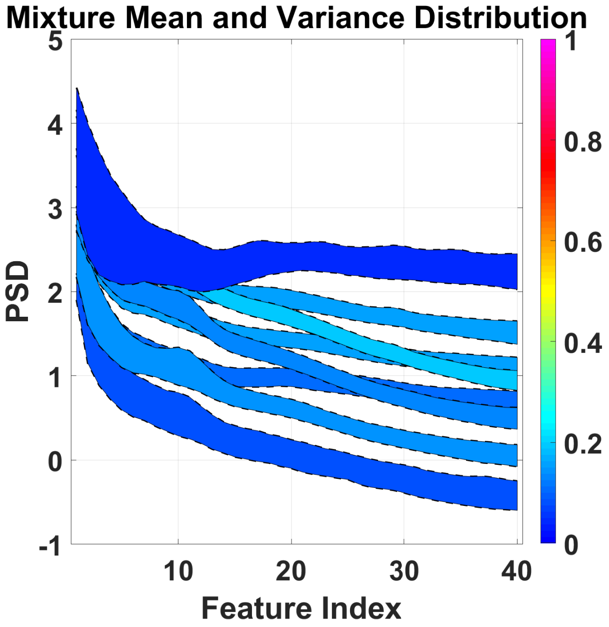


Figure 7.2. The PSD AbnMot UBM Mixture Distribution. Distribution of the 8-mixture PSD UBM AbnMot target. Coloration indicates mixture weight. Area of a mixture is +/- one standard deviation.

Given the difficulty of conceptualizing the differences between the UBM distributions in this manner, they were instead reduced down to their inter-mixture distances. Directly comparing the two UBMs against each other via their Fréchet Distance (D^F) provided a way to understand the similarity between the UBMs, see Figure 7.3. The Fréchet Distance is based upon two GMMs' means ($\overrightarrow{\mu}$) and variances ($\boldsymbol{\sigma}$) and is given by:

$$D^F_{1,2} = \text{norm}(\overrightarrow{\mu_1} - \overrightarrow{\mu_2})^2 + \text{trace}(\boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 - 2(\boldsymbol{\sigma}_1 * \boldsymbol{\sigma}_2)^{\frac{1}{2}}) \qquad (7.1\text{-}1)$$

The distances were reported as log values given the range of distance produce during testing.

When evaluating these distances for the UBMs of the "base" and "target" in Figure 7.3, the main interests are the mixtures that are closest and furthest apart from each other . Under this premise, it is seen that base mixture 2 and target mixture 3 were the closest mixtures between the two UBMs. Conversely, base mixture 1 and target mixture 8 and base mixture 8 and target mixture 1 were the furthest apart. In addition base mixtures 3, 4, 5, and 6 were close to target mixtures 5, 4, 1, and 2, respectively. This indicated that both UBMs had modeled a similar feature space despite operating on distinct datasets.

Attention was also given to the weight of each UBM's mixture, shown as the bar plots on the top and right axis of the confusion matrix. These indicated that base mixture 8 and target mixture 8 were derived from the smallest subset of the original dataset. However, target mixture 3 represented only 10% of the `AbnMot` dataset while being the strongest link to the the base UBM via base mixture 2 which represented 12% of the `AbnNrm` dataset. This helped establish that the UBMs shared a feature space despite being built on different datasets.

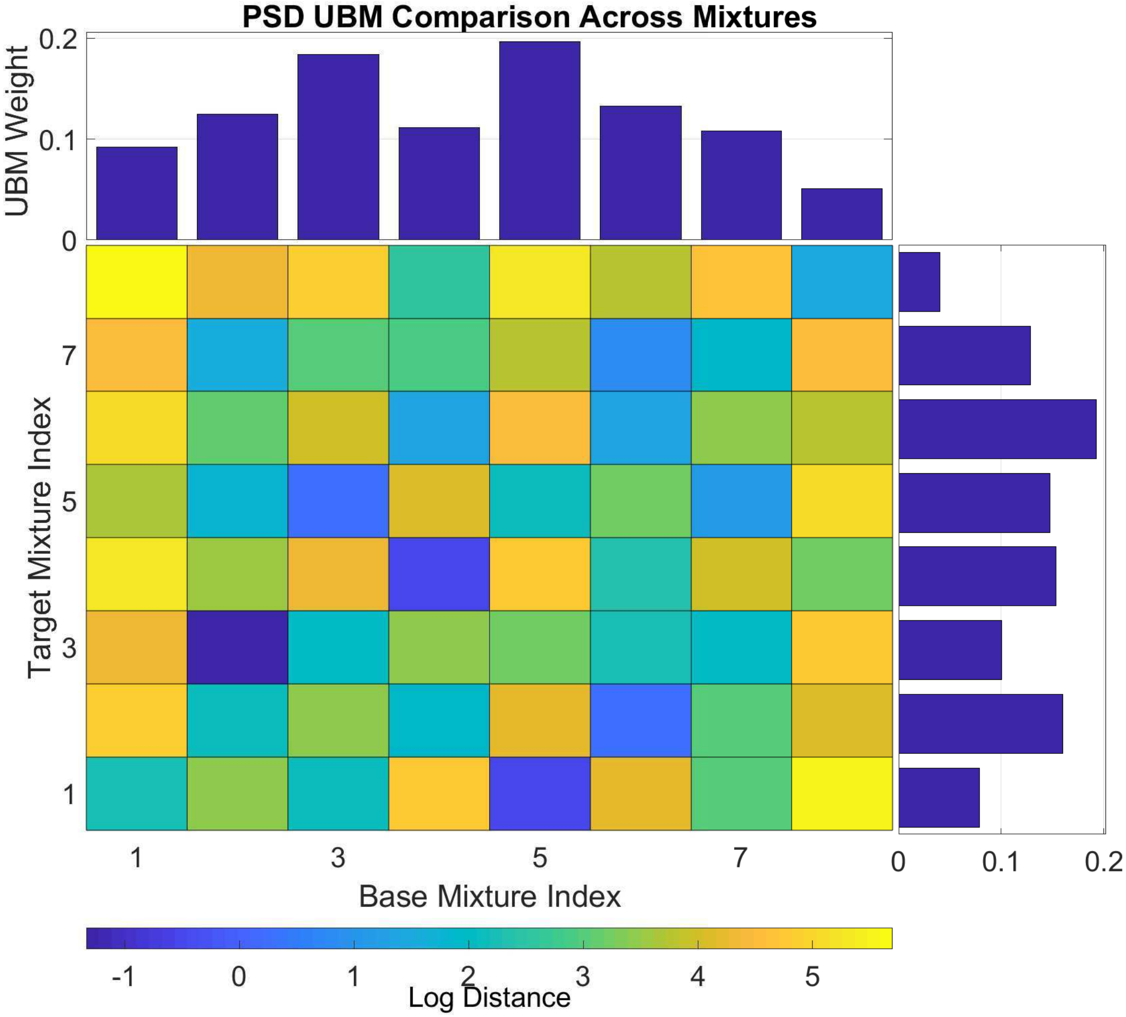


Figure 7.3. The PSD AbnMot UBM Confusion Matrix. The relationship between the PSD based AbnNrm and AbnMot datasets presented as a confusion matrix of the differences between each UBM's mixtures. The bar plots on the edges maintain the weights from their native datasets providing insight into the prevalence of a given mixture.

A similar figure was produced for the TVMs of these two datasets, Figure 7.4, comparing the coefficients produced for each mixture. As a TVM acts on the means of the UBM mixtures, it was necessary to find a way to compare the alignment of these coefficients against each other that didn't consider variances, as with the Fréchet Distance. This was achieved by generating an "impulse response" from the TVM by means of a 'unit' I-Vector.

$$\begin{aligned}
\boldsymbol{T} &= \begin{bmatrix} t_{1,1} & \cdots & t_{1,CF} \\ \vdots & \ddots & \vdots \\ t_{L,1} & \cdots & t_{L,CF} \end{bmatrix} \\
\overrightarrow{T}^{IR} &= \boldsymbol{T} * \begin{bmatrix} 1_1 \\ \vdots \\ 1_L \end{bmatrix} \\
\boldsymbol{T}^{IR} &= \begin{bmatrix} \overrightarrow{T}^{IR}_{1,1} & \cdots & \overrightarrow{T}^{IR}_{1,C} \\ \vdots & \ddots & \cdots \\ \overrightarrow{T}^{IR}_{F,1} & \cdots & \overrightarrow{T}^{IR}_{F,C} \end{bmatrix} \\
D^{IR}_{1,2} &= -\cos(\Theta_{\boldsymbol{T}_{:,1},\boldsymbol{T}_{:,2}})
\end{aligned} \tag{7.1-2}$$

This approached evaluated the relationship between the coefficients over a range of -1 to 1 in terms of a scaled CD. The resultant values corresponded to coefficients being perfectly aligned at -1 or entirely divergent at 1 within the F-dimensional feature space. These values were used to convey the relationship between the TVMs that corresponded to the same two UBMs shown in Figure 7.4. The coefficients were normalized by the variance of their associated UBM mixture making the CD based upon proportional changes of the mean and not its magnitude, and the results were presented as a confusion matrix.

Weights, like those derived by the UBM indicated the model's coverage of the dataset, were instead a representation of which TVM mixture generated the largest shifts, based upon the scaled standard deviation of the associated UBM determined variance. In this system, each feature was evaluated across the set of mixtures, thereby limiting the number of weight votes (largest change for the given feature) to the number of features. This meant that for larger UBMs many mixtures would not be assigned any weight. The purpose of this system was to provide insight into which TVM mixture coefficients exerted the most impact on their associated UBM mixture. This allowed the resultant TVM confusion plot to mirror the layout and content of its UBM counterpart.

The hypothesized premise was that if two mixtures existed in the same dimensional space (the UBMs) and represented the same component of EEG morphology, then the resultant changes to these mixtures would be consistent given the common feature space. While the UBMs could provide insight into position and shape of these mixtures within the feature space, it offered no insight into their actual use. The TVMs were able provide direct links between how each mixture's means were updated. In turn, these updates offered insight based upon the magnitude and direction of their adjustments to drive the subject classification process. The UBMs' mixtures could be aligned or divergent within the TVM space, providing another level of discrimination beyond their location and shape in the feature space, thereby giving insight into their classification purpose.

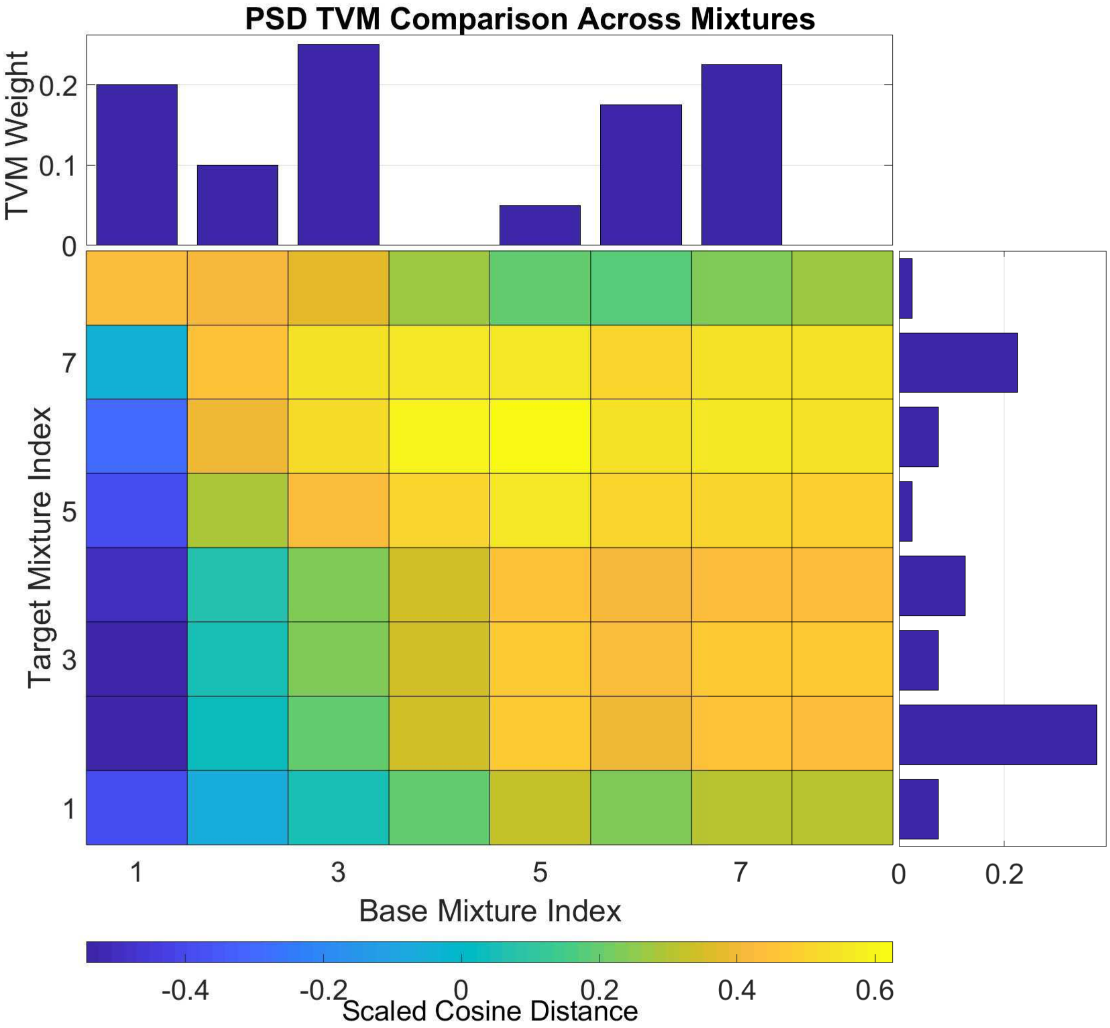


Figure 7.4. The PSD `AbnMot` TVM Confusion Matrix. The relationship between the PSD based `AbnNrm` and `AbnMot` datasets presented as a confusion matrix of the differences between each TVM's mixtures. The bar plots on the edges maintain the weights from their native datasets providing insight into the prevalence of a given mixture.

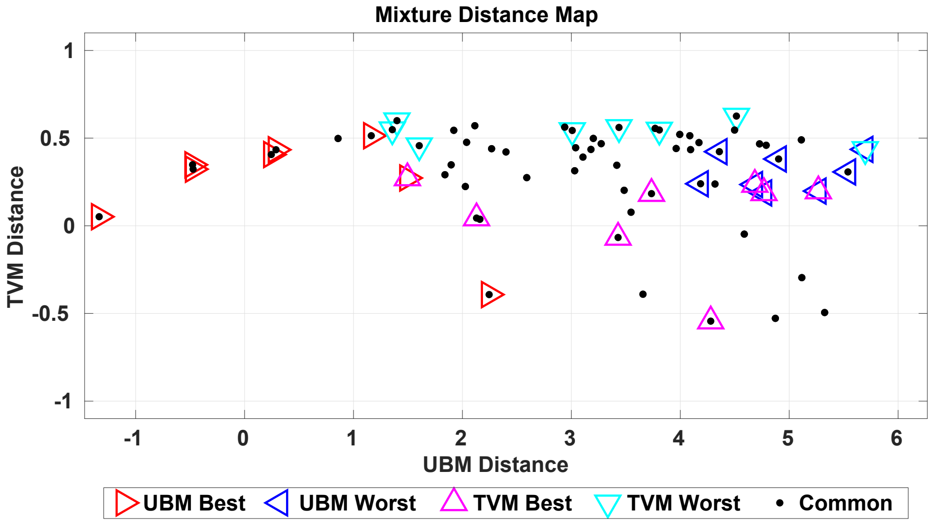


Figure 7.5. The PSD `AbnMot` Flagged Distance Map. The relationship between the PSD based `AbnNrm` and `AbnMot` datasets presented using the UBM distances on the x-axis and the TVM distances on the y-axis. The lowest `AbnNrm` distances are indicated by a caret away from their axis and the largest distances by a caret toward their axis.

The UBM and TVM distance mappings were then compared to determine where the best and worst distances aligned in their shared space. The x-axis was used for the UBM distance because it served as the independent control whereas the the TVM scaled CD was the dependent parameter. The distances between the mixtures for the UBMs and TVMs were plotted and flagged as either best, worst, or common (eg. “other”) distances. The best and worst distances were presented as carets pointing toward each other while the commons were dots. Separate colors were used for the UBM and TVM mixtures resulting in Figure 7.5. By splitting the axis in half ($\gtrless 0$ for TVM distance and $\gtrless 2$ for UBM distance) each quadrant could be idealized in terms of the UBM and TVM characteristics of furthest-divergent (quadrant I), closest-divergent (quadrant II), closest-aligned (quadrant III), and furthest-aligned (quadrant IV).

The AbnNrm-AbnMot dataset comparison resulted in a number of UBM-TVM distance pairings: 1 closest-aligned, 3 furthest-aligned, and 1 furthest-divergent.

Table 7.1. Example Mixture Distances

Base Mix	Target Mix	Match	UBM	TVM
8	8	C-A	1.495	0.273
-	-	-	-	-
4	1	F-A	4.769	0.188
5	8	F-A	5.268	0.198
7	8	F-A	4.685	0.235
1	8	F-D	5.698	0.436

These represented the points of interest between the UBM modeling process and the learned TVM model. These overlapping occurrences are recorded in Table 7.1 showing how the UBM and TVM interpret the same mixture pairings. The values of the associated distances are also given to provide context for the match labels. Using this mapping, it is possible to extract and compare the mixtures from the base and target datasets using a similar style figure as Figure 7.2 where the colors are set to blue for the base and pink for the target mixtures.

The closest-aligned mixture pairing in (Figure 7.6) shows what the UBMs and TVMs classified as being similar. While these mixtures occupied a similar feature space, their shape in that feature space is distinct and they were unique to their datasets, Figures 7.1 and 7.2. By being aligned in the TVM space, the mixtures were the most similar from the view of the `AbnNrm` base, Figure 7.4, which meant that the other mixtures were more divergent. Despite not being strongly aligned, given the near 0 TVM distance, the importance of their relationship was that they resulted in lowest amount of differentiation compared to the other possible base mixture to target mixture pairings. Thus changes in a given I-Vector would be less apparent through this mixture relationship than others.

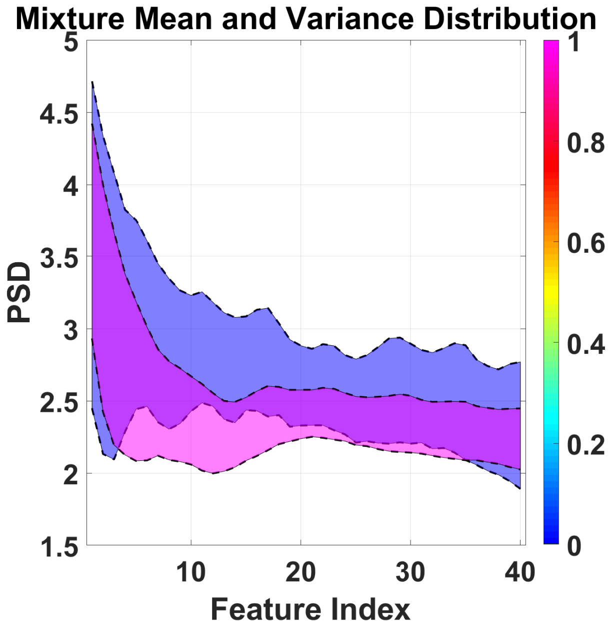


Figure 7.6. The 8 Mixture PSD `AbnNrm` Closest-Aligned. The PSD `AbnNrm` (blue base) and `AbnMot` (pink target) **Closest-Aligned** mixtures drawn from the 8-mixture UBM. Area of a mixture is +/- one standard deviation. The base mixture is 8 and the target mixture is 8.

The 3 furthest-aligned mixtures were split, with base mixture 4 matching to target mixture 1 (Figure 7.7) and base mixtures 5 and 7 matching to target mixture 8 (Figure 7.8). The alignment in the TVM space of these mixture pairings was slightly greater than 0, indicating that these mixtures were the least altered by a given I-Vector compared to the other possible pairings. It is clear that the mixtures have unique means and variances, but the TVM modifies their means in a similar fashion as their coefficients are aligned. Thus instead of focusing on further separating these already distinct mixtures, the TVM generally maintains their UBM distance by grouping them together. This is of particular interest when multiple mixtures, Figure 7.8, exhibit the same behavior in relation to a single target mixture. In this case, target mixture 8 is again used in the pairing considering it is the least weighted mixture from its UBM, Figure 7.3.

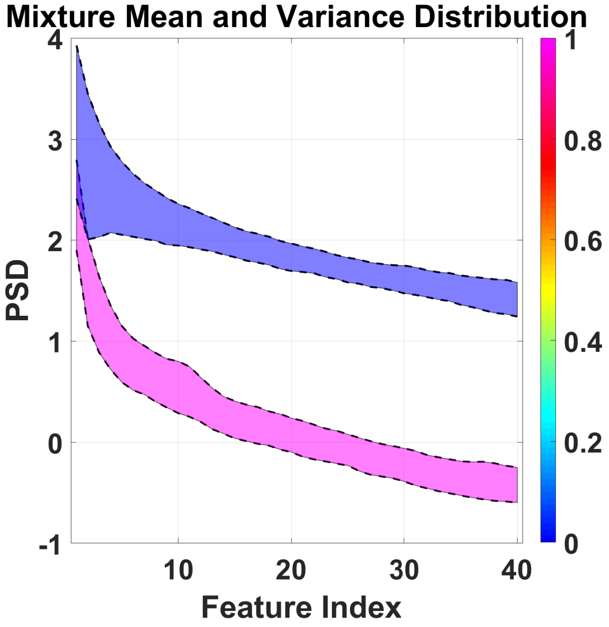


Figure 7.7. The 8 Mixture PSD AbnNrm Furthest-Aligned (1). The PSD AbnNrm (blue base) and AbnMot (pink target) **furthest-aligned** mixtures drawn from the 8-mixture UBM. Area of a mixture is +/- one standard deviation. The base mixture is 4 and the target mixture is 1.

The final pairing (furthest-divergent, see Figure 7.9) contained base mixture 1 and target mixture 8. While the previous figures contained aligned TVM distances, this divergent classification indicates that the TVM wants to further separate these mixtures in the feature space. Again target mixture 8 was paired with a previously unpaired base mixture. However, base mixture 1 was far from the majority of the target mixtures, Figure 7.3, while also being aligned with the majority of them in the TVMs space, Figure 7.4. Thus despite already being separable for the majority of target mixtures, the relationship with target mixture 8 had to be augmented. Given the previous pairings, it was clear that target mixture 8 was important to not only

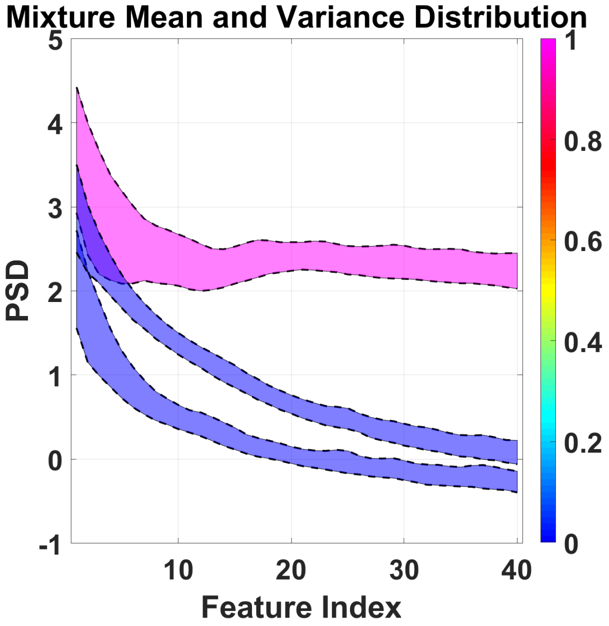


Figure 7.8. The 8 Mixture PSD AbnNrm Furthest-Aligned (2). The PSD AbnNrm (blue base) and AbnMot (pink target) **furthest-aligned** mixtures drawn from the 8-mixture UBM. Area of a mixture is +/- one standard deviation. The base mixtures are 5 and 7 and the target mixture is 8.

its dataset, but also its classification performance. This made all of its relationships critical to classification which is why it appeared frequently in the analysis.

Lacking from this example were classifications of the closest-divergent. This classification which would have suggested the presence of a decision surface between the flagged mixtures as the UBM and TVM evaluations would have been in conflict. The concept of occupying a similar feature space and being shifted in opposing directions by the TVMs is suggestive of mixtures with unique properties. Therefore classifications of divergent TVM distances were thought to be indicative of decisions surfaces between the two datasets, while those classified as aligned would suggest the feature space was more important than the means and variances of the mixtures.

When addressing the TVM aligned mixtures, the assumption was made that those mixtures represented linked signal phenomena. The feature space of EEGs

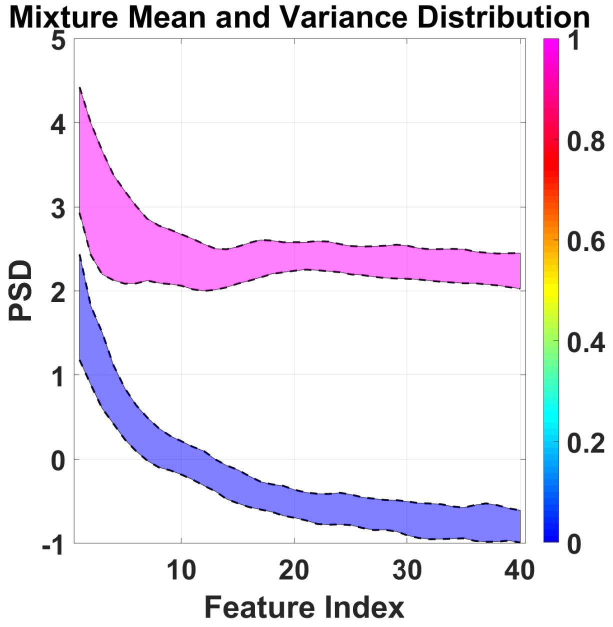


Figure 7.9. The PSD AbnNrm UBM Mixture Distribution. The PSD AbnNrm (blue base) and AbnMot (pink target) **furthest-divergent** mixtures drawn from the 8-mixture UBM. Area of a mixture is +/- one standard deviation. The base mixture is 1 and the target mixture is 8.

represented a large but closed dimensional space. Therefore, specific EEG events were likely to contain unique characteristics which these algorithms were attempting to model. An artifact, low frequency event or as in this example motor control manifested as a mixture with somewhat unique means, but distinct variances in Figure 7.7. Here the target mixture of Figure 7.7 resides in a similar position to the base mixtures of Figure 7.8. These mixtures from the AbnMot dataset were contrasted with base mixtures that they overlapped with in the feature space. This was seen again in Figure 7.9, but now the TVM was separating the target mixture 8 and base mixture 1, despite base mixture 1's similarity to the base mixtures in Figure 7.8.

This indicated that mixtures could exist in the same feature space, but their relationships could be distinct. Such behavior was beyond the ability of the UBM metric to quantify and was thus only discovered through the TVM feature space. While all of these insights may not have been critical for subject verification, they did point toward the unique qualities of each dataset. For example, while sharing the `Abn` dataset, the inclusion/exclusion of the `Nrm` and `Mot` dataset was apparent in the visual UBM distributions given the similar means but distinct variances of Figure 7.6. This difference was masked when the two UBMs were evaluated in terms of distance and weights, but the visual aspect was only useful given the limited mixtures used.

This process was carried out for each feature set, but the PSD features were the least abstract of those tested. Using them was intuitive for human interpretation, which helped in troubleshooting and understanding the technique. However, the approach was not limited by feature set, as these overlapping conditions existed for all feature sets as seen in Figures 7.10 and 7.11.

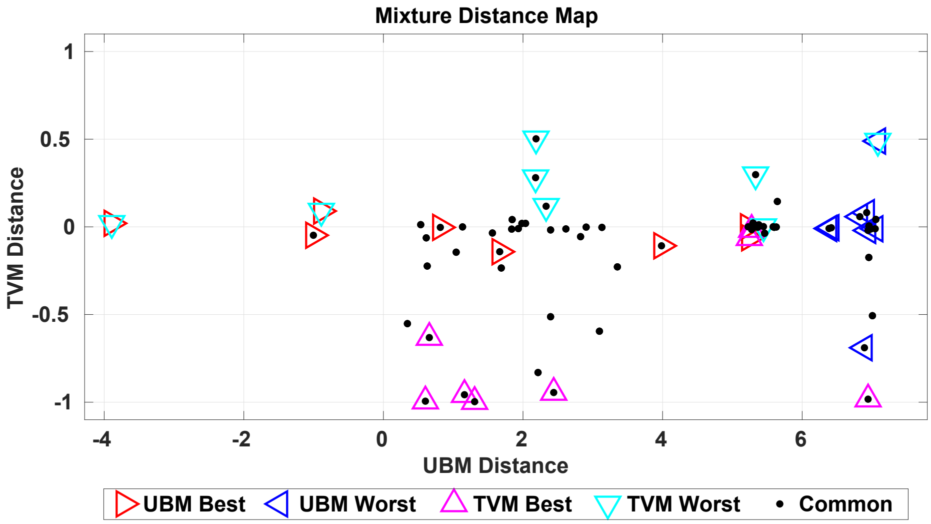


Figure 7.10. The CEP `AbnMot` Flagged Distance Map. The relationship between the CEP based `AbnNrm` and `AbnMot` datasets presented using the UBM distances on the x-axis and the TVM distances on the y-axis. The lowest AbnRnm distances are indicated by a caret away from their axis and the largest distances by a caret toward their axis.

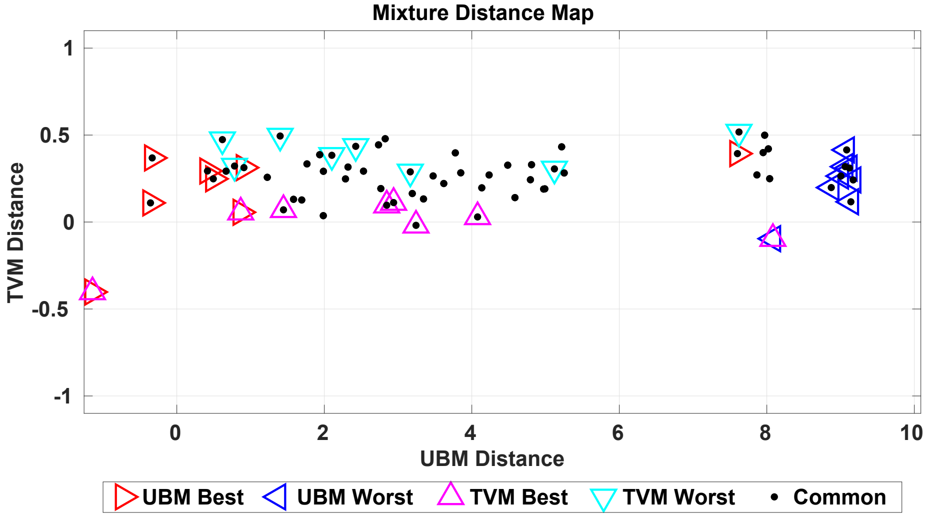


Figure 7.11. The COH AbnMot Flagged Distance Map. The relationship between the COH based AbnNrm and AbnMot datasets presented using the UBM distances on the x-axis and the TVM distances on the y-axis. The lowest AbnRnm distances are indicated by a caret away from their axis and the largest distances by a caret toward their axis.

7.2 Mixture Scaling

Two experiments, comparing AbnMot to NrmMot in Table 7.2 and AbnSzr to NrmSzr in Table 7.3, were used as validate the technique across multiple UBM sizes. The Mot based results had 159 subjects and the Szr based results had 461 subjects.

The AbnMot and NrmMot results produced four or more pairings of furthest-divergent labels at each mixture size, as seen in Table 7.2. While the TVM distance values fluctuated at each mixture size, they remained relatively consistent around a value of 5. However, other pairings did not exhibit this behavior, as the furthest-aligned was only produced for the 32-mixture UBMs. No closest-divergent or furthest-aligned were produced for the 8-mixture UBMs.

Meanwhile, the AbnSzr and NrmSzr results produced 6 or more pairings of closest-aligned at each mixture size (Table 7.3). These did not have one single target mixture matching to multiple base mixtures, but instead exhibited a one-to-one pairing between the datasets for every mixture size. For these pairings, the UBM distance was inconsistent, ranging from -2 to 1, while the TVM distances were consistently at or below -0.5. This suggested that the mixtures were more aligned than occupying similar feature spaces despite already being in very close proximity to each other.

The Mot based results appeared to continually pair the same target mixture against the same base mixtures regardless of initial UBM, Figures 7.12–7.14. These furthest-divergent pairings were indicative of a decision surface involving the feature space occupied by the target mixture and base mixtures.

The majority of results in Table 7.3 were closest-aligned pairings, which made sense given the bias toward the large amount of Szr data compared to the amount of Abn and Nrm data. This made the lone furthest-divergent mixture when using 32-mixture UBMs of particular interest. Examining it showed two very distinct mixtures, Figure 7.16. This occurred in a similar fashion using the 16-mixture UBMs, which produced the closest-divergent pairing shown in Figure 7.17. The base mixture in Figure 7.16 appears to occupy a similar position in the feature space as the two mixtures of Figure 7.17.

This likely explains the shift in classification, as the 8-mixture UBMs classified a pair of mixtures as closest-aligned, Figure 7.18, which are clearly those in Figure 7.17. The mixtures were all in the same feature space, making discrimination difficult based on UBM distance alone. However the TVM adjusted correctly at the 16 mixture level while the UBM discovered the mixture's existence when increased to 32 mixtures.

At this larger mixture size, the distribution of paired mixture distances begins to converge, Figure 7.15. The improved mixture differentiation means there are less overlaps between the best and worst pairings of UBM and TVM mixtures. This makes the outliers more pronounced, but masks potential decision surfaces by pairing

Table 7.2. `NrmMot` to `AbnMot` Evaluation with 8 Mixtures

Mixture	Base Mix	Target Mix	Match	UBM	TVM
8	1	1	C-A	1.047	-0.127
	-	-	-	-	-
	-	-	-	-	-
	2	1	F-D	4.608	0.063
	4	1	F-D	5.120	0.153
	6	1	F-D	4.888	0.403
	7	1	F-D	4.287	0.357
16	1	1	C-A	1.157	-0.478
	13	11	C-A	-3.689	-0.580
	15	15	C-D	1.120	0.031
	-	-	-	-	-
	4	1	F-D	5.195	0.362
	6	1	F-D	5.173	0.430
	7	1	F-D	4.670	0.344
	8	1	F-D	5.506	0.384
	10	1	F-D	4.932	0.040
	12	1	F-D	5.456	-0.003
32	17	17	C-A	-0.113	-0.583
	11	7	C-D	-2.110	0.396
	1	32	F-A	6.475	-0.498
	13	32	F-A	5.471	-0.488
	24	1	F-A	5.781	-0.400
	26	1	F-A	5.276	-0.343
	14	1	F-D	5.478	0.199
	15	1	F-D	5.006	0.303
	20	1	F-D	5.466	0.409
	22	1	F-D	5.442	0.264

Table 7.3. NrmSzr to AbnSzr Evaluation with 8 Mixtures

Mixture	Base Mix	Target Mix	Match	UBM	TVM
8	1	1	C-A	0.016	-0.993
	2	2	C-A	-1.047	-0.975
	3	3	C-A	-0.762	-0.944
	4	4	C-A	-3.431	-0.836
	5	5	C-A	-0.400	-0.695
	8	8	C-A	1.091	-0.513
	-	-	-	-	-
	7	8	F-A	5.479	-0.611
	1	8	F-D	6.287	-0.317
	2	8	F-D	5.324	-0.421
	3	8	F-D	5.695	-0.434
	5	8	F-D	5.924	-0.206
	8	1	F-D	6.211	-0.022
16	3	3	C-A	-0.389	-0.496
	8	8	C-A	-1.563	-0.744
	10	10	C-A	-1.689	-0.755
	11	11	C-A	-0.507	-0.679
	15	15	C-A	-3.578	-0.805
	16	16	C-A	-4.159	-0.609
	1	9	C-D	-0.920	0.088
	-	-	-	-	-
	-	-	-	-	-
32	2	2	C-A	0.365	-0.820
	5	5	C-A	-0.211	-0.644
	10	10	C-A	-1.284	-0.860
	11	11	C-A	-0.743	-0.835
	20	20	C-A	0.998	-0.430
	22	22	C-A	-2.109	-0.792
	30	30	C-A	0.010	-0.933
	-	-	-	-	-
	-	-	-	-	-
	1	32	F-D	7.303	0.170

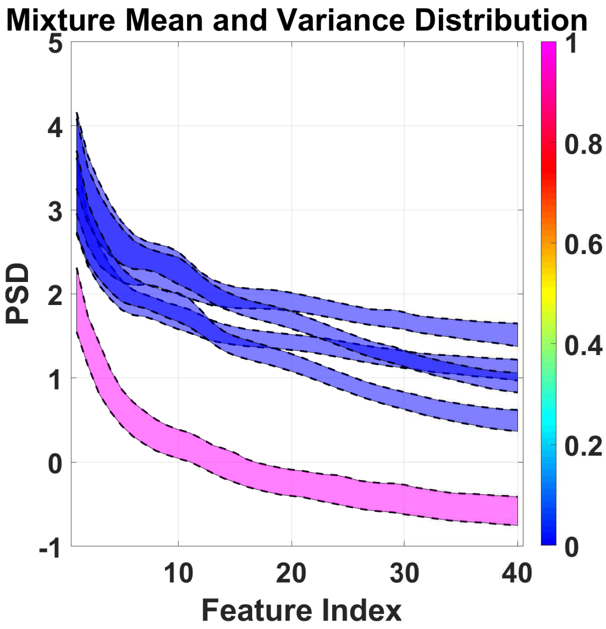


Figure 7.12. The 8 Mixture PSD `AbnMot` `NrmMot` Furthest-Divergent. The mixtures associated with the 8 Mixture PSD `AbnMot` (target) and `NrmMot` (base) **Furthest-Divergent** pairing. Area of a mixture is +/- one standard deviation. The blue indicates the base mixture (2, 4, 6, and 7) and the pink indicates the target mixture (1).

the best and worst of one model with a common pairing of another. This worked focused on only mixtures were the best and worst overlapped, but the UBM and TVM continued to produce mixtures of interest. For example, the best and worst carets located within the cluster of common results present as a new set of outliers relative to the positions of their UBM/TVM cohort.

The results of these brief mixture scaling experiments presented similar insights to that of the example Section 7.1. In addition to accurately mapping between UBMs and TVMs, when given diverse data and increasingly larger UBMs, performance was consistent. The ability to maintain a decision surface across mixture sizes was critical, as that was the underlying mechanism use by the I-Vectors. However, the ability of the TVM to discern mixtures in close proximity was not indicative that the same

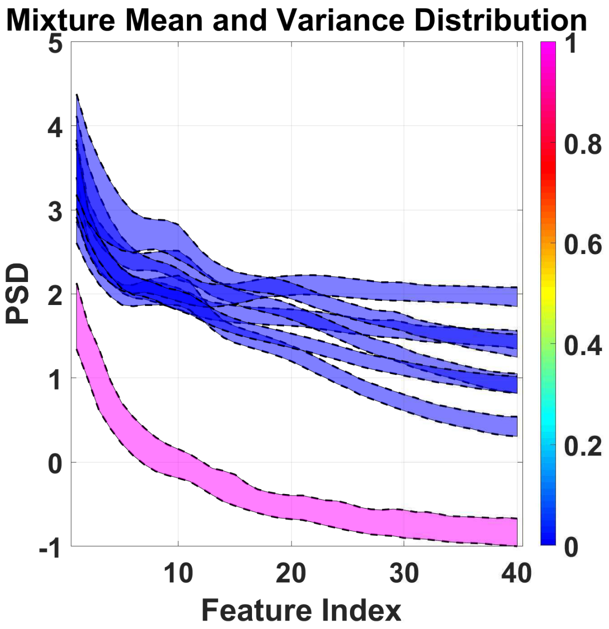


Figure 7.13. The 16 Mixture PSD `AbnMot` `NrmMot` Furthest-Divergent. The mixtures associated with the 16 Mixture PSD `AbnMot` (target) and `NrmMot` (base) **Furthest-Divergent** pairing. Area of a mixture is +/- one standard deviation. The blue indicates the base mixture (4, 6, 7, 8, 10, and 12) and the pink indicates the target mixture (1).

behavior was more important. This suggested the TVM was capable of growth and interpretation on data alone in a way that was not feasible for the UBM.

7.3 Results

The previous example and mixture scaling experiments walked through the analysis steps of the proposed UBM-TVM mapping technique. Just as in those sections, the result's figures focus on differences between the mixtures in each dataset via their classification (closest-aligned, closest-divergent, furthest-aligned, furthest-divergent) and their evolution within the incremental UBM mixture sizes. As such any pairings involving those mixtures classified as common, (see figure 7.5) were not used which

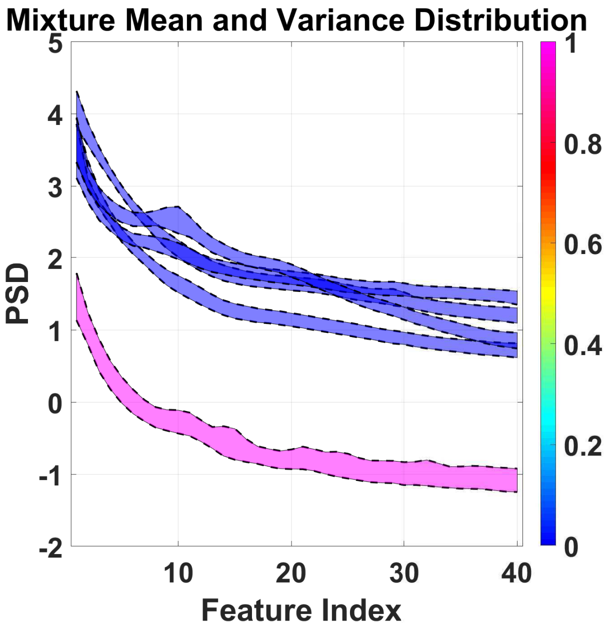


Figure 7.14. The 32 Mixture PSD `AbnMot` `NrmMot` Furthest-Divergent. The mixtures associated with the 32 Mixture PSD `AbnMot` (target) and `NrmMot` (base) **Furthest-Divergent** pairing. Area of a mixture is +/- one standard deviation. The blue indicates the base mixture (14, 15, 20, and 22) and the pink indicates the target mixture (1).

simplified the analysis and was necessary given the potential complexity of parsing the large UBMs.

To further simplify the analysis, only TVMs of dimension 25 were used when paired with the 8, 16, and 32-mixture UBMs, which were at the lower end of the optimal UBM mixture sizes identified in Chapter 5. Despite knowing that the aggregated datasets required larger UBM mixtures for acceptable I-Vector performance, the GMM-UBM performance peaked earlier, indicating that the UBMs had likely mapped the feature space to find the dominant components of the datasets.

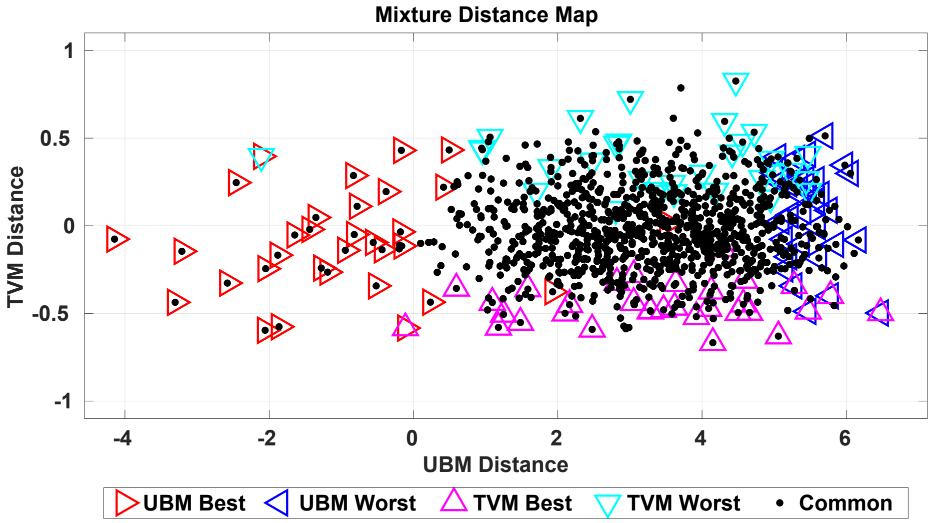


Figure 7.15. The PSD AbnMot 32-Mixture Distance Map. The relationship between the PSD based AbnMot (target) and NrmMot (base) datasets presented using the UBM distances on the x-axis and the TVM distances on the y-axis. The worst distances are indicated by a caret pointing in a positive direction and the best distances are indicated by a caret point in a negative direction.

7.3.1 Base: AbnNrm

In these instances, the AbnNrm dataset was used as the base mixture with each target mixture containing one or both of the Abn or Nrm datasets. This anchored the results against a single dataset in an effort to control the numbers of degrees of freedom. The target datasets were therefore limited to AbnMot, AbnSzr, NrmMot, NrmSzr, AbnNrmSzr, and AbnNrmMot.

7.3.1.1 Target: Motion

The first experiment compared the base dataset against the AbnMot, NrmMot, and AbnNrmMot datasets. The AbnMot and NrmMot datasets contained 159 subjects and overlapped with roughly 33% of the base AbnNrm dataset. The AbnNrmMot dataset contained 209 subjects and overlapped with roughly 48% of the base AbnNrm

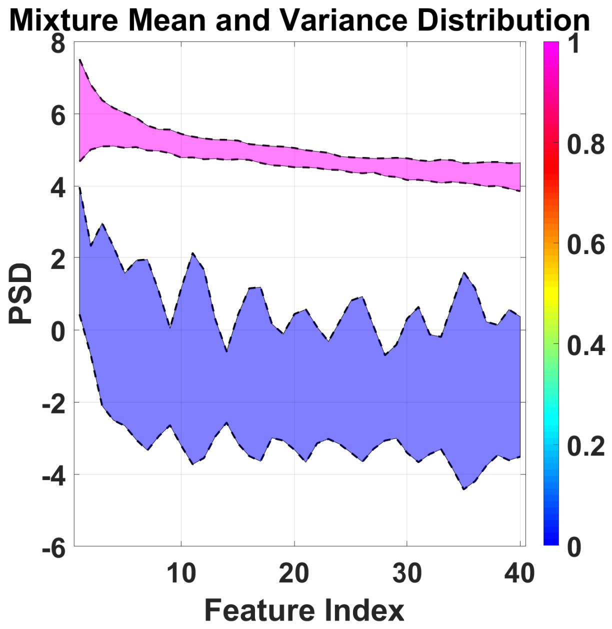


Figure 7.16. The 32 Mixture PSD `AbnSzr` `NrmSzr` Furthest-Divergent. The mixtures associated with the 32 Mixture PSD `AbnSzr` (target) and `NrmSzr` (base) **Furthest-Divergent** pairing. Area of a mixture is +/- one standard deviation.

dataset. The `Mot` dataset contained rest tasks between the scripted motion tasks, likely shifting this percentage closer in terms of content. However, the subjects and recording environments were unique between the datasets.

The sets of the 8, 16, and 32-mixture UBMs in Tables 7.4–7.6 showed the effect of increasing the mixture size on the datasets. Only when using the 16-mixture UBM were all four types found.

Within the 8-mixture base `AbnNrm` UBM, Table 7.4, mixtures 2 and 3 were not given any classification. Mixture 1 was used for all three targets. Mixtures 4 and 8 were used for `NrmMot` and `AbnMot`. Aligned (C-A and/or F-A) and divergent (C-D and/or F-D) classifications were found in all target datasets.

Within the 16-mixture base `AbnNrm` UBM, Table 7.5, only 10 mixtures (1, 3, 5, 6, 8, 9, 10, 11, 13, and 15) were given a classification. Those without a classification

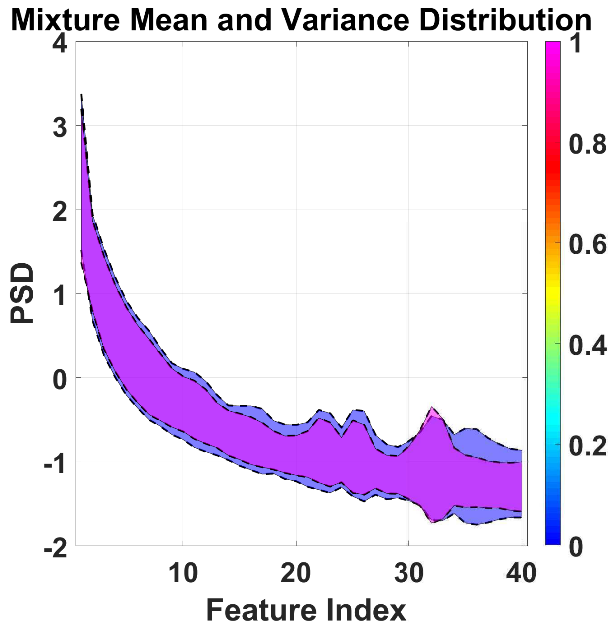


Figure 7.17. The 16 Mixture PSD `AbnSzr` `NrmSzr` Closest-Divergent. The mixtures associated with the 16 Mixture PSD `AbnSzr` (target) and `NrmSzr` (base) **Closest-Divergent** pairing. Area of a mixture is +/- one standard deviation.

were 2, 4, 7, 12, 14, and 16. Mixture 1 was used for all three datasets. Mixtures 5, 8, 9, 10, and 11 were used for `AbnMot` and `AbnNrmMot`. Divergent TVM scores were not found for the `NrmMot` dataset.

Within the 32-mixture base `AbnNrm` UBMs, Table 7.6 only 11 mixtures (4, 7, 8, 10, 16, 18, 20, 25, 26, 28, and 32) were given a classification. Mixtures 4 and 8 were used for `NrmMot` and `AbnMot`. Mixture 10 was used for `AbnMot` and `AbnNrmMot`. Divergent TVM scores were found in all target datasets.

7.3.1.2 Target: Seizure

In this set of experiments all of the target datasets contained the `Szr` dataset. This limited the target datasets to `AbnSzr`, `NrmSzr`, and `AbnNrmSzr`. The `AbnSzr` and `NrmSzr` contained 461 subjects meaning the `AbnNrm` had an overlap of 22%. The

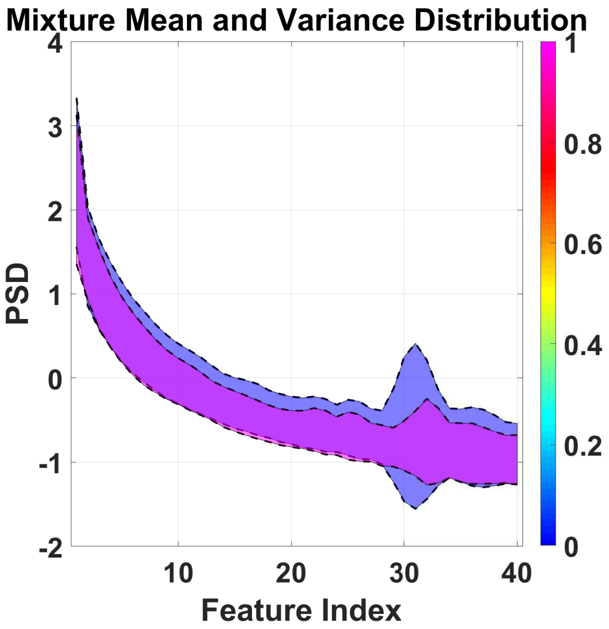


Figure 7.18. The 8 Mixture PSD `AbnSzr` `NrmSzr` Closest-Aligned. The mixtures associated with the 8 Mixture PSD `AbnSzr` (target) and `NrmSzr` (base) **Closest-Aligned** pairing. Area of a mixture is +/- one standard deviation.

`AbnNrmMot` contained 511 subjects which meant the `AbnNrm` had an overlap of 20%. The `Szr` contained subjects that experienced a seizure, but that did not mean that all recorded data was of seizures. This likely meant the data was a mix of seizure activity and/or abnormal and normal brain activity as well. This was similar to the `Mot` containing resting states between the motion tasks.

The sets of the 8, 16, and 32-mixture UBMs in Tables 7.7–7.9 showed the effect of increasing the mixture size on the datasets. Only when using the 16-mixture UBM were all four types found.

Within the 8-mixture base `AbnNrm` UBMs, Table 7.7, all mixtures were given a classification. Mixtures 1, 6, and 7 were used for all three datasets. Mixture 5 was used for `NrmSzr` and `AbnSzr`. Mixtures 2, 3, and 8 were used for `AbnSzr` and `AbnNrmSzr`. Divergent TVM scores were found in all target datasets.

Table 7.4. AbnNrm Base Evaluation with 8 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	-	-	-	-	-
	1	1	C-D	0.778	-0.002
	6	6	C-D	0.761	0.604
	4	1	F-A	5.045	-0.240
	6	1	F-A	4.550	-0.024
	8	1	F-A	5.731	0.053
	-	-	-	-	-
AbnMot	8	8	C-A	1.495	0.273
	-	-	-	-	-
	4	1	F-A	4.769	0.188
	5	8	F-A	5.268	0.198
	7	8	F-A	4.685	0.235
	1	8	F-D	5.698	0.436
AbnNrmMot	-	-	-	-	-
	1	1	C-D	0.842	-0.375
	1	8	F-A	5.694	-0.579
	-	-	-	-	-

Table 7.5. AbnNrm Base Evaluation with 16 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	1	1	C-A	1.182	-0.754
	8	16	C-A	0.764	-0.139
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
AbnMot	1	1	C-A	2.359	-0.475
	3	5	C-A	0.091	-0.308
	-	-	-	-	-
	-	-	-	-	-
	5	16	F-D	5.549	0.386
	9	16	F-D	5.761	0.559
	10	16	F-D	4.519	0.584
	11	16	F-D	5.133	0.586
AbnNrmMot	15	15	C-A	0.644	-0.050
	1	1	C-D	1.003	0.763
	5	9	C-D	-0.516	0.355
	5	16	F-A	5.601	-0.339
	9	16	F-A	5.807	-0.297
	10	16	F-A	4.608	-0.480
	11	16	F-A	5.199	-0.465
	13	16	F-A	5.461	-0.146
	6	1	F-D	4.879	0.318
	8	1	F-D	5.868	0.431

Table 7.6. AbnNrm Base Evaluation with 32 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	-	-	-	-	-
	8	16	C-D	2.676	0.236
	4	1	F-A	5.402	-0.412
	32	1	F-A	6.732	-0.358
	16	1	F-D	6.192	0.101
	18	32	F-D	5.403	0.298
AbnMot	17	17	C-A	2.489	-0.430
	28	28	C-A	0.857	-0.702
	-	-	-	-	-
	-	-	-	-	-
	4	1	F-D	5.711	0.399
	8	1	F-D	6.378	0.249
	10	1	F-D	5.204	0.202
	26	1	F-D	5.368	0.138
AbnNrmMot	-	-	-	-	-
	7	13	C-D	0.327	0.282
	10	7	C-D	-0.608	0.099
	25	1	C-D	0.947	0.299
	20	17	F-A	5.694	-0.235
	-	-	-	-	-

Table 7.7. AbnNrm Base Evaluation with 8 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmSzr	1	1	C-A	-0.080	-0.769
	5	5	C-D	-0.659	0.648
	5	8	F-A	6.018	0.404
	6	8	F-A	5.083	0.353
	7	8	F-A	5.601	0.382
	-	-	-	-	-
AbnSzr	1	1	C-A	0.819	-0.750
	8	8	C-A	2.559	-0.089
	3	3	C-D	-0.255	0.539
	5	5	C-D	0.353	0.622
	2	8	F-A	5.248	0.024
	3	8	F-A	5.609	0.047
	5	8	F-A	5.859	-0.012
	6	8	F-A	4.831	-0.009
	7	8	F-A	5.414	-0.105
	-	-	-	-	-
AbnNrmSzr	8	8	C-A	2.489	0.221
	3	3	C-D	-1.138	0.689
	4	4	C-D	-0.731	0.635
	6	6	C-D	-1.564	0.566
	1	8	F-A	6.167	-0.347
	2	8	F-A	5.227	0.162
	3	8	F-A	5.592	0.281
	7	8	F-A	5.395	0.272
	-	-	-	-	-

Within the 16-mixture base AbnNrm UBMs, Table 7.8, seven mixtures (5, 6, 7, 10, 13, 14, and 15) were not given a classification. Mixture 1 was used for AbnSzr and AbnNrmSzr. Mixture 8 was used for NrmSzr and AbnSzr. Mixtures 12 and 16 were used for NrmSzr and AbnNrmSzr. Divergent TVM scores were not found for the AbnNrm and AbnNrmSzr.

Within the 32-mixture base AbnNrm UBMs, Table 7.9, 18 mixtures (1, 4, 6, 8, 10, 13, 14, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, and 30) were not given a classification. Mixture 31 was used for NrmSzr and AbnNrmSzr. Divergent TVM scores were found for all target datasets.

7.3.2 Base: AbnNrmMot

The 209 subject AbnNrmMot was used as a base against the 100 subject AbnNrm, 159 subject AbnMot, and 159 subject NrmMot. In this series of experiments, the larger dataset was used as the base with the targets containing combinations of the included mixtures.

Within the 8-mixture base AbnNrmMot UBMs, Table 7.10, only Mixture 6 was not given a classification. Mixtures 2 and 7 were used for all three datasets. Mixtures 4 and 8 were used for AbnNrm and AbnMot. Mixtures 1 and 7 were used for AbnNrm and NrmMot. Mixtures 3 and 5 were used for AbnMot and NrmMot. Divergent TVM scores were found for all target datasets.

Within the 16-mixture base AbnNrmMot UBMs, Table 7.11, 4 mixtures (8, 10, 13, and 14) were not given a classification. Mixtures 5 and 16 were used for all three datasets. Mixtures 3, and 7 were used for AbnNrm and AbnMot. Mixture 4 was used for AbnNrm and NrmMot. Mixtures 1, 2, and 12 were used for AbnMot and NrmMot. Divergent TVM scores were found for all target datasets.

Table 7.8. AbnNrm Base Evaluation with 16 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmSzr	9	1	C-A	-1.643	-0.511
	11	3	C-A	-0.034	-0.329
	2	7	C-D	-0.196	0.553
	4	9	F-A	5.521	-0.019
	3	16	F-D	6.364	0.386
	8	9	F-D	6.066	0.101
	12	9	F-D	5.768	0.304
	16	9	F-D	6.623	0.213
AbnSzr	1	1	C-A	1.006	-0.805
	-	-	-	-	-
	8	1	F-A	6.055	-0.386
	-	-	-	-	-
AbnNrmSzr	1	1	C-A	1.742	-0.792
	-	-	-	-	-
	12	9	F-A	5.702	-0.393
	16	9	F-A	6.559	-0.558
	-	-	-	-	-

Table 7.9. AbnNrm Base Evaluation with 32 Mixtures

AbnNrm	Base Mix	Target Mix	Match	UBM	TVM
NrmSzr	32	16	C-A	0.839	-0.447
	3	3	C-D	-0.031	0.464
	9	1	C-D	-1.181	0.532
	16	24	C-D	0.368	0.079
	31	31	C-D	0.126	0.414
	7	32	F-A	6.573	-0.060
	9	32	F-A	6.932	-0.074
	11	32	F-A	6.578	-0.122
	-	-	-	-	-
AbnSzr	-	-	-	-	-
	2	3	C-D	0.476	0.661
	18	19	C-D	0.153	0.438
	28	28	C-D	0.189	0.691
	-	-	-	-	-
	12	1	F-D	6.187	0.438
AbnNrmSzr	15	15	C-A	-0.588	-0.233
	21	21	C-A	0.233	-0.596
	29	29	C-D	-0.569	0.502
	-	-	-	-	-
	5	32	F-D	6.729	0.378
	31	32	F-D	6.040	0.580

Table 7.10. AbnNrmMot Base Evaluation with 8 Mixtures

AbnNrmMot	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	-	-	-	-	-
	2	6	C-D	0.951	0.016
	8	8	C-D	1.441	-0.514
	2	1	F-A	4.781	-0.434
	4	1	F-A	5.309	-0.575
	1	8	F-D	5.721	-0.215
	7	1	F-D	4.548	-0.520
AbnMot	4	4	C-A	-2.495	-0.070
	8	8	C-D	-3.052	0.333
	-	-	-	-	-
	2	1	F-D	4.143	0.239
	3	8	F-D	4.632	0.241
	5	8	F-D	5.063	0.409
	7	8	F-D	4.157	0.405
NrmMot	5	5	C-A	-4.239	-0.083
	7	7	C-D	-2.608	0.325
	-	-	-	-	-
	1	8	F-D	5.422	0.107
	2	1	F-D	4.501	0.374
	3	8	F-D	4.479	0.452
	5	8	F-D	4.939	0.439

Table 7.11. AbnNrmMot Base Evaluation with 16 Mixtures

AbnNrmMot	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	11	7	C-A	0.198	-0.305
	16	8	C-D	0.447	0.148
	4	1	F-A	5.428	-0.283
	6	1	F-A	5.384	-0.297
	3	16	F-D	5.800	0.276
	5	16	F-D	6.006	0.351
	7	16	F-D	5.580	0.223
AbnMot	-	-	-	-	-
	1	1	C-D	1.096	0.494
	9	9	C-D	0.890	0.425
	16	16	C-D	-1.147	0.141
	3	16	F-A	4.852	-0.659
	5	16	F-A	5.192	-0.560
	7	16	F-A	4.542	-0.397
	2	1	F-D	4.484	0.168
	12	1	F-D	5.226	0.096
NrmMot	12	12	C-A	-2.700	-0.367
	1	1	C-D	-1.522	0.751
	5	5	C-D	-4.640	0.444
	4	1	F-A	5.188	-0.298
	2	1	F-D	4.780	0.024
	15	1	F-D	4.725	0.007
	16	1	F-D	5.926	0.217

Within the 32-mixture base AbnNrmMot UBMs, Table 7.12, nine mixtures (3, 8, 9, 12, 15, 17, 19, 24, and 25) were given a classification. Mixture 24 was used for AbnNrm and AbnMot. Mixture 15 was used for AbnNrm and NrmMot. Divergent TVM scores were found for all target datasets.

7.3.3 Base: AbnNrmSzr

The 511 subject AbnNrmSzr was used as a base against the 100 subject AbnNrm, 461 subject AbnSzr, and 461 subject NrmSzr. In this series of experiments the larger dataset was used as the base with the targets containing combinations of the included mixtures.

Within the 8-mixture base AbnNrmSzr UBMs, Table 7.13, all the mixtures were given a classification. Mixtures 1, and 5 were used for all three datasets. Mixture 8 was used for AbnNrm and AbnSzr. Mixture 7 was used for AbnSzr and NrmSzr. Divergent TVM scores were found for all target datasets.

Within the 16-mixture base AbnNrmSzr UBMs, Table 7.14, seven mixtures (2, 4, 5, 8, 12, 14, and 15) were given a classification. Mixture 2 was used for AbnNrm and NrmSzr. Divergent TVM scores were found for all target datasets.

Within the 32-mixture base AbnNrmSzr UBMs, Table 7.15, 15 mixtures (2, 4, 12, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, and 29) were given a classification. Mixture 2 was used for all three datasets. Mixture 17 was used for AbnNrm and NrmSzr. Divergent TVM scores were found for all target datasets.

Table 7.12. AbnNrmMot Base Evaluation with 32 Mixtures

AbnNrmMot	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	17	17	C-A	1.323	-0.291
	15	15	C-D	0.489	0.246
	12	17	F-A	5.605	-0.260
	24	17	F-A	5.847	-0.292
	-	-	-	-	-
AbnMot	3	3	C-A	-0.570	-0.544
	24	24	C-D	-0.570	0.261
	-	-	-	-	-
	-	-	-	-	-
NrmMot	9	9	C-A	-2.386	-0.258
	19	19	C-A	-4.974	-0.289
	25	25	C-A	-4.081	-0.736
	15	15	C-D	-3.156	0.595
	8	1	F-A	5.638	-0.776
	-	-	-	-	-

Table 7.13. AbnNrmSzr Base Evaluation with 8 Mixtures

AbnNrmSzr	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	1	1	C-A	0.344	-0.079
	2	2	C-D	-0.191	0.651
	3	3	C-D	-1.138	0.689
	5	5	C-D	-0.782	0.610
	2	8	F-A	4.572	0.278
	4	1	F-A	5.310	0.191
	6	1	F-A	4.867	-0.058
	8	1	F-A	6.167	-0.347
	-	-	-	-	-
AbnSzr	5	5	C-A	-1.158	-0.712
	8	8	C-D	-3.607	-0.279
	-	-	-	-	-
	1	8	F-D	6.179	-0.125
	5	8	F-D	5.813	-0.019
	7	8	F-D	5.305	-0.231
NrmSzr	5	5	C-A	-1.715	-0.565
	7	7	C-D	-5.949	-0.355
	-	-	-	-	-
	1	8	F-D	6.314	-0.364
	4	1	F-D	5.371	-0.302
	5	8	F-D	5.974	-0.195
	6	8	F-D	5.048	-0.238

Table 7.14. AbnNrmSzr Base Evaluation with 16 Mixtures

AbnNrmSzr	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	8	8	C-A	0.635	-0.444
	15	15	C-A	-0.923	-0.234
	2	10	C-D	1.262	0.110
	5	5	C-D	-1.618	0.257
	-	-	-	-	-
	-	-	-	-	-
AbnSzr	-	-	-	-	-
	14	14	C-D	-1.184	0.011
	-	-	-	-	-
	-	-	-	-	-
NrmSzr	12	12	C-A	-2.929	-0.657
	2	2	C-D	-3.345	0.026
	-	-	-	-	-
	4	9	F-D	5.499	0.136

Table 7.15. AbnNrmSzr Base Evaluation with 32 Mixtures

AbnNrmSzr	Base Mix	Target Mix	Match	UBM	TVM
AbnNrm	-	-	-	-	-
	17	25	C-D	-0.372	0.061
	2	32	F-A	5.939	-0.276
	16	17	F-A	6.494	-0.395
	20	17	F-A	5.593	-0.510
	-	-	-	-	-
AbnSzr	25	25	C-A	2.399	-0.266
	2	2	C-D	0.015	0.403
	18	18	C-D	-1.495	0.654
	23	23	C-D	-0.144	0.553
	24	24	C-D	-2.380	0.531
	-	-	-	-	-
	-	-	-	-	-
NrmSzr	17	17	C-A	-1.881	-0.321
	29	27	C-A	0.639	-0.630
	2	2	C-D	-1.973	0.340
	12	22	C-D	0.355	0.155
	19	19	C-D	-0.204	0.613
	22	22	C-D	-1.176	0.481
	26	32	F-A	6.209	-0.372
	27	32	F-A	6.454	-0.564
	4	32	F-D	6.056	0.234

7.3.4 Base: AbnSzrMot

The 570 subject `AbnSzrMot` was used as a base against the 159 subject `AbnMot`, 461 subject `AbnSzr`, and 520 subject `SzrMot`. In this series of experiments, the larger dataset was used as the base with the targets containing combinations of the included mixtures.

Within the 8-mixture base `AbnSzrMot` UBMs, Table 7.16, all mixtures were given a classification. Mixtures 3, 5, 7, and 8 were used for all three datasets. Mixture 6 was used for `AbnMot` and `SzrMot`. Mixtures 1 and 2 were used for `AbnSzr` and `SzrMot`. Divergent TVM scores were found for `AbnMot` and `SzrMot` target datasets.

Within the 16-mixture base `AbnSzrMot` UBMs, Table 7.17, nine mixtures (1, 2, 8, 9, 10, 13, 14, 15, and 16) were given a classification. Mixture 1 was used for all three datasets. Mixtures 13 and 15 were used for `AbnMot` and `AbnSzr`. Mixture 9 was used for `AbnMot` and `SzrMot`. Mixture 14 was used for `AbnSzr` and `SzrMot`. Divergent TVM scores were found for `AbnMot` and `SzrMot` target datasets.

Within the 32-mixture base `AbnSzrMot` UBMs, Table 7.18, ten mixtures (15, 8, 10, 17, 21, 22, 25, 28, 30, and 32) were given a classification. Mixture 28 was used for all three datasets. Mixture 21 was used for `AbnMot` and `AbnSzr`. Divergent TVM scores were found for `AbnSzr` and `SzrMot` target datasets.

Table 7.16. AbnSzrMot Base Evaluation with 8 Mixtures

AbnSzrMot	Base Mix	Target Mix	Match	UBM	TVM
AbnMot	8	8	C-A	2.629	0.156
	6	6	C-D	-1.696	0.536
	3	8	F-A	4.632	-0.108
	5	8	F-A	5.049	0.222
	7	8	F-A	4.299	0.193
	-	-	-	-	-
AbnSzr	1	1	C-A	-0.085	0.012
	8	8	C-A	2.028	-0.179
	-	-	-	-	-
	2	8	F-A	4.935	0.083
	3	8	F-A	5.423	0.118
	5	8	F-A	5.690	0.033
	7	8	F-A	5.161	-0.081
	-	-	-	-	-
SzrMot	5	5	C-A	-6.438	-0.484
	6	6	C-A	-5.886	-0.569
	8	8	C-D	-1.759	0.102
	2	1	F-A	4.534	-0.615
	4	1	F-A	5.220	-0.520
	8	1	F-A	5.814	-0.449
	1	8	F-D	5.783	0.243
	3	8	F-D	5.026	0.136
	5	8	F-D	5.345	-0.128
	7	8	F-D	4.736	-0.119

Table 7.17. AbnSzrMot Base Evaluation with 16 Mixtures

AbnSzrMot	Base Mix	Target Mix	Match	UBM	TVM
AbnMot	16	16	C-A	4.305	-0.379
	1	9	C-D	0.964	0.644
	13	9	C-D	0.779	0.115
	2	1	F-A	4.483	-0.365
	15	16	F-A	4.526	-0.369
	9	16	F-D	5.929	0.374
AbnSzr	13	13	C-A	-0.946	-0.204
	14	14	C-A	0.371	-0.324
	15	15	C-A	0.382	-0.487
	1	9	C-D	0.541	0.014
	10	10	C-D	-0.190	0.587
	-	-	-	-	-
	-	-	-	-	-
SzrMot	14	14	C-A	-1.809	-0.302
	1	1	C-D	-3.574	0.688
	9	9	C-D	-2.687	0.944
	-	-	-	-	-
	8	9	F-D	5.751	0.644

Table 7.18. AbnSzrMot Base Evaluation with 32 Mixtures

AbnSzrMot	Base Mix	Target Mix	Match	UBM	TVM
AbnMot	8	8	C-A	0.947	-0.295
	28	28	C-A	-0.724	-0.828
	-	-	-	-	-
	21	32	F-A	5.310	-0.290
	-	-	-	-	-
AbnSzr	10	18	C-A	0.082	-0.724
	21	21	C-D	-0.109	0.482
	28	28	C-D	1.199	0.772
	-	-	-	-	-
	-	-	-	-	-
SzrMot	5	5	C-A	-0.956	-0.761
	22	22	C-A	-0.293	-0.395
	28	28	C-A	-3.524	-0.753
	25	25	C-D	-1.392	0.273
	-	-	-	-	-
	17	32	F-D	6.535	0.091
	22	9	F-D	5.587	0.519
	30	9	F-D	5.599	0.102
	32	9	F-D	6.961	0.225

7.3.5 Base: NrmSzrMot

The 570 subject `AbnSzrMot` was used as a base against the 159 subject `NrmMot`, 461 subject `NrmSzr`, and 520 subject `SzrMot`. In this series of experiments the larger dataset was used as the base with the targets containing combinations of the included mixtures.

Within the 8-mixture base `NrmSzrMot` UBMs, Table 7.19, all mixtures were given a classification. Mixtures 1, 4, 7, and 8 were used for all three datasets. Mixture 5 was used for `NrmSzr` and `SzrMot`. Divergent TVM scores were found for `AbnSzr` and `SzrMot` target datasets.

Within the 16-mixture base `NrmSzrMot` UBMs, Table 7.20, 13 mixtures (1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 15, and 16) were given a classification. Mixtures 16 was used for all three datasets. Mixtures 2, 5, and 15 were used for `NrmMot` and `NrmSzr`. Mixture 9 was used for `NrmMot` and `SzrMot`. Divergent TVM scores were found for all target datasets.

Within the 32-mixture base `NrmSzrMot` UBMs, Table 7.21, 16 mixtures (2, 4, 5, 6, 9, 10, 12, 15, 18, 19, 20, 22, 26, 30, 31, and 32) were given a classification. No mixtures overlapped within the target datasets. Divergent TVM scores were found for the `NrmMot` and `NrmSzr` target datasets.

7.3.6 Discussion

The experiments were reviewed based upon their base mixture. The larger base mixtures, `AbnNrmMot` and `AbnNrmSzr` and `AbnSzrMot` and `NrmSzrMot`, were grouped into two sections while the `AbnNrm` was reviewed by itself.

Table 7.19. NrmSzrMot Base Evaluation with 8 Mixtures

NrmSzrMot	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	1	1	C-A	0.437	-0.317
	7	7	C-D	-1.113	0.461
	2	1	F-A	4.376	-0.231
	4	1	F-A	5.146	-0.225
	6	1	F-A	4.809	-0.006
	8	1	F-A	5.794	-0.027
	-	-	-	-	-
NrmSzr	8	8	C-A	2.952	-0.097
	-	-	-	-	-
	1	8	F-A	6.259	-0.003
	5	8	F-A	5.904	0.131
	7	8	F-A	5.394	0.047
	4	1	F-D	5.409	0.741
SzrMot	5	5	C-A	-1.510	-0.585
	6	6	C-A	-4.657	-0.609
	8	8	C-D	-1.924	-0.150
	4	1	F-A	5.205	-0.606
	1	8	F-D	5.831	-0.101
	2	8	F-D	4.489	-0.021
	3	8	F-D	5.076	-0.016
	5	8	F-D	5.400	-0.233
	7	8	F-D	4.762	-0.281

Table 7.20. NrmSzrMot Base Evaluation with 16 Mixtures

NrmSzrMot	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	-	-	-	-	-
	3	3	C-D	0.068	0.598
	9	1	C-D	1.129	0.200
	2	1	F-A	4.672	-0.604
	5	16	F-A	5.299	-0.041
	15	1	F-A	4.626	-0.242
	16	1	F-A	6.561	-0.409
	6	1	F-D	5.088	0.533
	8	1	F-D	5.645	0.473
	10	1	F-D	4.810	0.088
	12	1	F-D	5.413	0.333
NrmSzr	5	5	C-A	-0.779	-0.377
	15	15	C-A	0.280	-0.548
	16	16	C-A	1.653	-0.564
	7	7	C-D	0.023	0.346
	2	16	F-A	5.884	-0.595
	-	-	-	-	-
SzrMot	-	-	-	-	-
	1	1	C-D	-1.146	0.613
	9	9	C-D	-0.112	0.933
	16	9	F-A	6.567	-0.703
	13	16	F-D	6.104	0.614

Table 7.21. NrmSzrMot Base Evaluation with 32 Mixtures

NrmSzrMot	Base Mix	Target Mix	Match	UBM	TVM
NrmMot	-	-	-	-	-
	18	18	C-D	-1.235	0.550
	22	14	C-D	0.055	0.470
	20	1	F-A	5.529	-0.355
	22	1	F-A	5.400	-0.288
	30	1	F-A	5.516	-0.459
	31	1	F-A	5.092	-0.413
	32	1	F-A	6.880	-0.352
	-	-	-	-	-
NrmSzr	-	-	-	-	-
	9	9	C-D	-0.069	0.327
	10	10	C-D	-0.050	0.261
	19	19	C-D	-2.238	0.428
	26	26	C-D	-0.149	0.432
	2	32	F-A	6.332	-0.146
	4	32	F-A	6.064	-0.142
	5	32	F-A	6.676	-0.169
	6	32	F-A	6.029	-0.148
	15	32	F-A	6.279	-0.249
	-	-	-	-	-
SzrMot	-	-	-	-	-
	-	-	-	-	-
	12	9	F-A	5.716	-0.363
	-	-	-	-	-

7.3.6.1 AbnNrm

The mixture mappings of the AbnNrm UBMs were collected into Table 7.22. This allowed trends of the base mixtures to be compared across the six target datasets. As noticed in the example and calibration, not all classification pairings were generated for each experiment configuration. The F-D classification failed to generate pairings in 10 instances, the F-A in 6, and the C-D and C-A in 5. This indicated that the chosen 'best/worst' case classifications did not overlap frequently for all experiments. Increasing the pairings of interest to include those shifting between the common labels and the 'best/worst' case labels could have been used to better track changes between the mixture sizes. Despite this, a number of mixtures were clearly dominant based upon mixture size or datatset.

For the 8-mixture UBM, base mixture 1 appeared in 7 pairings. In 6 of those pairings, the 16-mixture UBM produced pairings as well. Examining that relationship, it was discovered that that base mixture 1 of the 8-mixture UBM had become base mixtures 1 and 9 of the 16-mixture UBM, see Figures 7.19–7.22. This was not unique to the Szr associated data, as the Mot datast showed a similar behavior in Figures 7.24–7.27.

The treatment of these mixtures across UBM mixture sizes indicated that their use was driven by the dataset. When paired with the seizure data, the base mixtures were closest to and aligned with mixtures representative of abnormal and/or seizure phenomena. However, they were set in opposition to mixtures that were produced by the motion data in their UBM and TVM distances. Thus this nondescript base mixture was used by multiple TVMs as a discrimination surface against the unique mixtures from the target datasets. In the case of the 3- mixture UBM AbnNrm base

Table 7.22. AbnNrm Base Mixture Matches

Matches		AbnNrm					
		NrmMot	AbnMot	AbnNrmMot	NrmSzr	AbnSzr	AbnNrmSzr
C-A	8	-	8	-	1	1,8	8
	16	1,8	1,3	15	9,11	1	1
	32	-	17,28	-	32	-	15,21
C-D	8	1,6	-	1	5	3,5	3,4,6
	16	-	-	1,5	2	-	-
	32	8	-	7,10,25	3,9,11,16,31	2,18,28	29
F-A	8	4,6,8	4,5,7	1	5,6,7	2,3,5,6,7	1,2,3,7
	16	-	-	5,9,10,11,13	4	8	12,16
	32	4,32	-	20	7,9,11	-	-
F-D	8	-	1	-	-	-	-
	16	-	5,9,10,11	6,8	3,8,12,16	-	-
	32	16,18	4,8,10,26	-	-	12	5,31

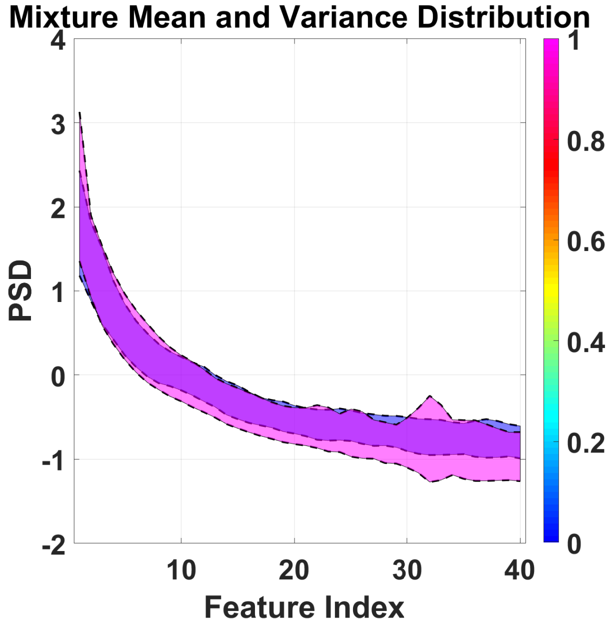


Figure 7.19. The 8 Mixture PSD AbnNrm NrmSzr Closest-Aligned. Pairing of AbnNrm (base) 8-mixture UBM's mixture 1 with its C-A pair from the NrmSzr (target).

and NrmSzr target, the TVMs resolved that this feature space was occupied by both datasets but represented distinct phenomena, see Figure 7.23.

The only other fully populated row in Table 7.22 belonged to the F-A values of the 8-mixture UBMs. Here again, repeated clusters of base mixtures were seen, although the trend did not continue for the 16-mixture UBMs. While nearly every mixture was used in this row, a small subset of mixtures (5, 6, and 7) appeared dominant, Figures 7.28–7.31. The base mixtures are clustered together and paired against a unique mixture from the target dataset.

In three of these instances, Figures 7.29–7.31, the target mixture sits in a similar position in the feature space. This again indicated a bias based upon the dataset, with the Szr data producing a distinct target mixture potentially modified by the presence of the abnormal data in Figure 7.31. And again the Mot dataset's characteristic mixture was used in Figure 7.29.

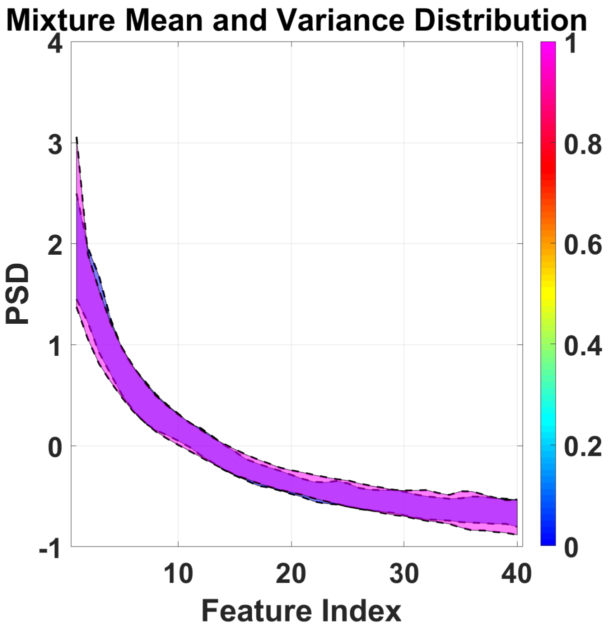


Figure 7.20. The 16 Mixture PSD AbnNrm NrmSzr Closest-Aligned. Pairing of AbnNrm 16-mixture UBM's mixture 1 with its C-A pair from the NrmSzr.

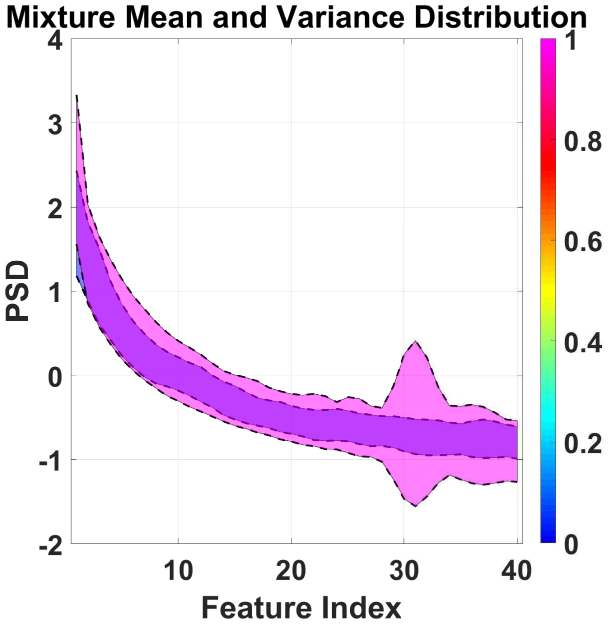


Figure 7.21. The 8 Mixture PSD AbnNrm NrmSzr Closest-Aligned. Pairing of AbnNrm 8-mixture UBM's mixture 1 with its C-A pair from the AbnSzr.

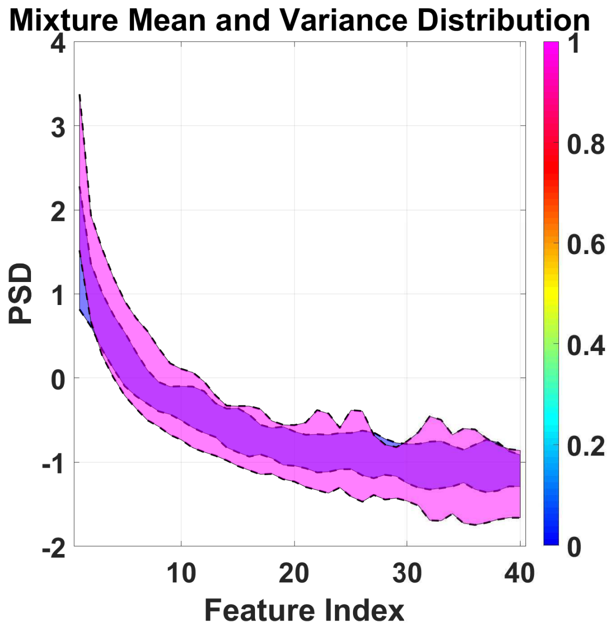


Figure 7.22. The 16 Mixture PSD `AbnNrm` `NrmSzr` Closest-Aligned. Pairing of `AbnNrm` 16-mixture UBM's mixture 1 with its C-A pair from the `AbnSzr`.

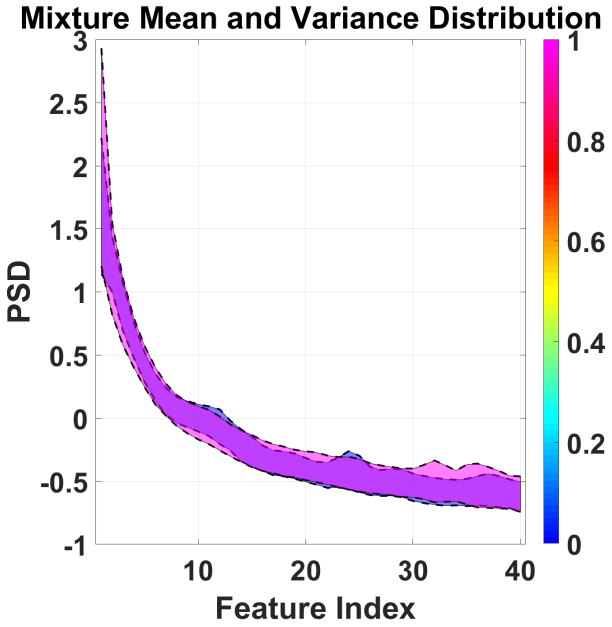


Figure 7.23. The 32 Mixture PSD `AbnNrm` `NrmSzr` Closest-Divergent. Pairing of `AbnNrm` 32-mixture UBM mixture with its C-D pair from the `NrmSzr`.

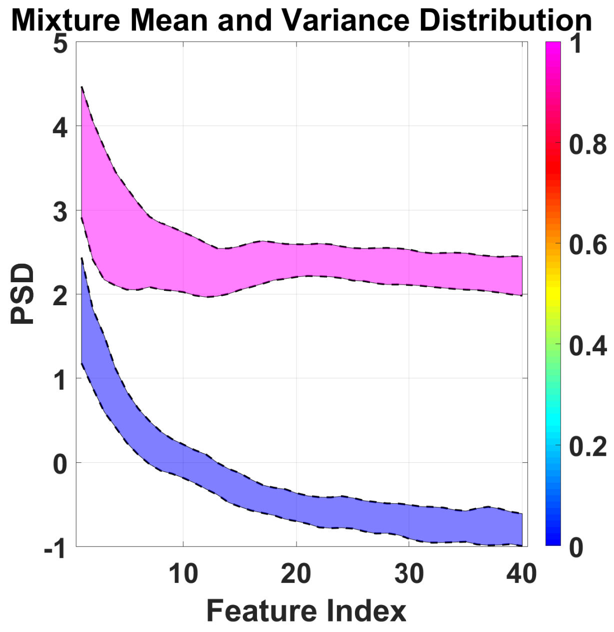


Figure 7.24. The 8 Mixture PSD `AbnNrm` `AbnNrmMot` Furthest-Aligned. Pairing of `AbnNrm` 8-mixture UBM's mixture 1 with its F-A pair from the `AbnNrmMot`.

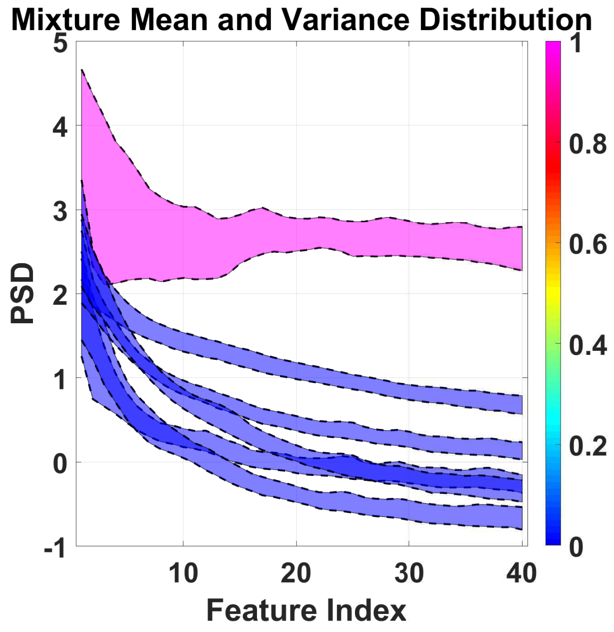


Figure 7.25. The 16 Mixture PSD `AbnNrm` `AbnNrmMot` Furthest-Aligned. Pairing of `AbnNrm` 16-mixture UBM's mixture 1 with its F-A pair from the `AbnNrmMot`.

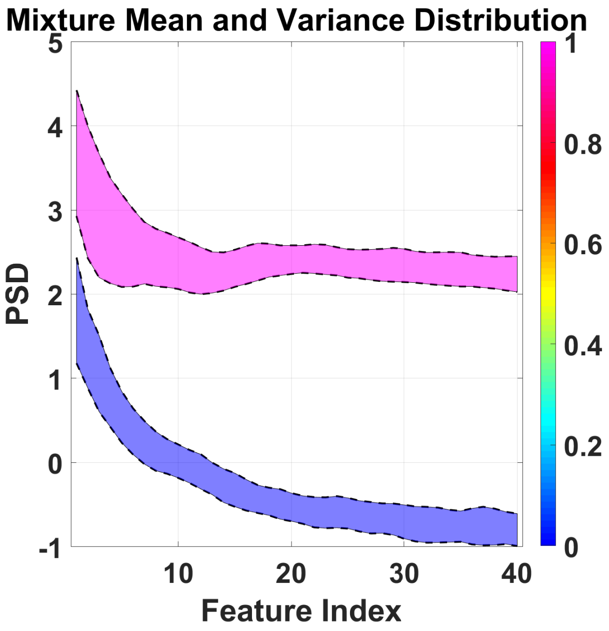


Figure 7.26. The 8 Mixture PSD AbnNrm AbnMot Furthest-Divergent. Pairing of AbnNrm 8-mixture UBM's mixture 1 with its F-D pair from the AbnMot.

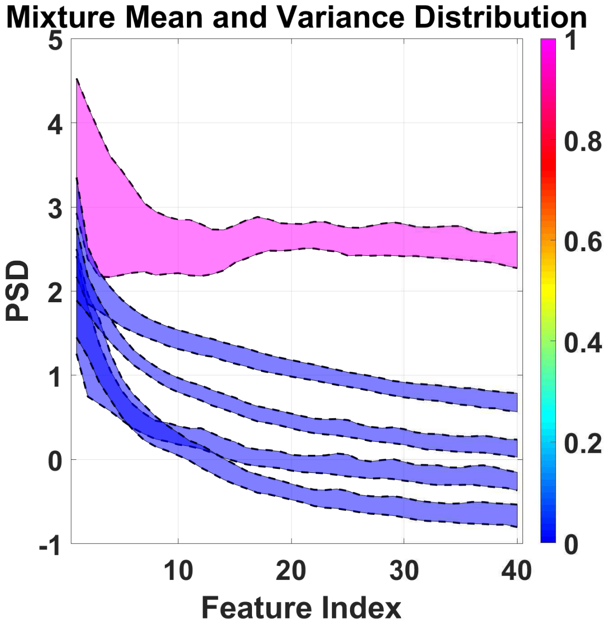


Figure 7.27. The 16 Mixture PSD AbnNrm AbnMot Furthest-Divergent. Pairing of AbnNrm 16-mixture UBM's mixture 1 with its F-D pair from the AbnMot.

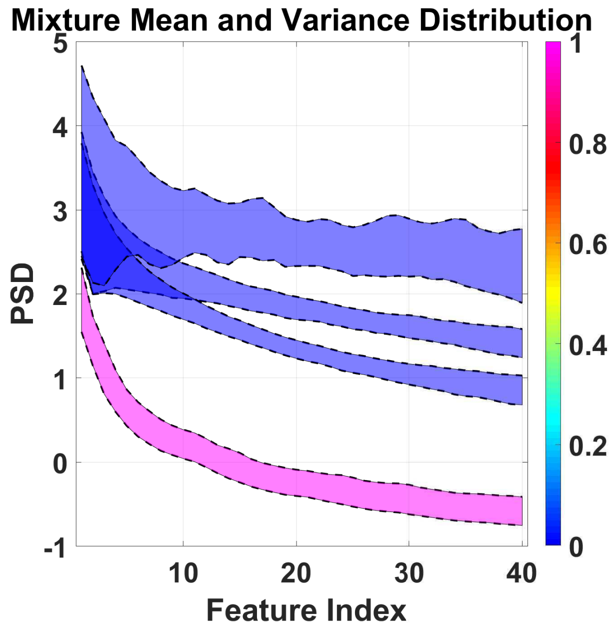


Figure 7.28. The 8 Mixture PSD AbnNrm NrmMot Furthest-Aligned. Pairing of AbnNrm 8-mixture UBM with its F-A pair from the NrmMot.

In both cases, characteristic mixtures of the base dataset were paired against unique mixtures from the target datasets. These Furthest-Aligned classifications did not stand out when viewed through their UBM distances, 7.32 as the mixtures all appear well-matched across the datasets. Despite the overwhelming contribution of the seizure data, the TVMs determined that the mixtures should be treated as wholly divergent from each other, Figure 7.33.

7.3.6.2 AbnNrmMot and AbnNrmSzr

Strictly using larger datasets as the base provided more consistency in the modeling process, Table 7.23. This time, the more well-defined mixtures were matched back into their fundamental forms. The AbnNrm data was a control against each of the larger datasets allowing comparisons between them and also against their specific motion and seizure datasets. The F-D classification failed to generate 9 pairings, the

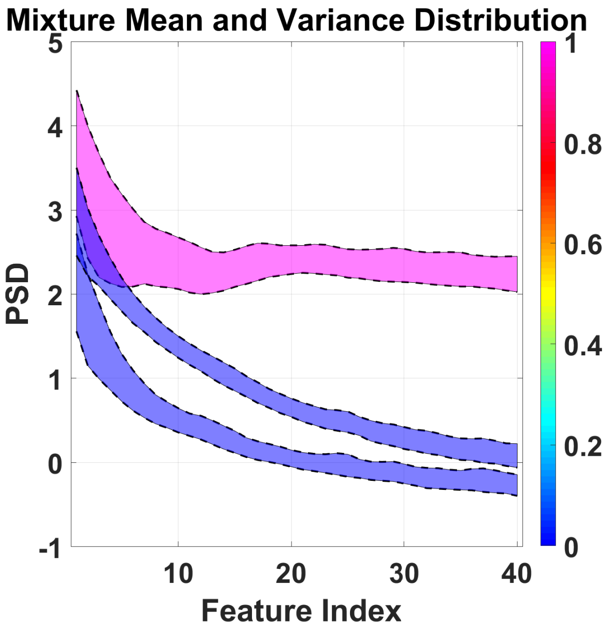


Figure 7.29. The 8 Mixture PSD `AbnNrm AbnMot` Furthest-Aligned. Pairing of `AbnNrm` 8-mixture UBM with its F-A pair from the `AbnMot`.

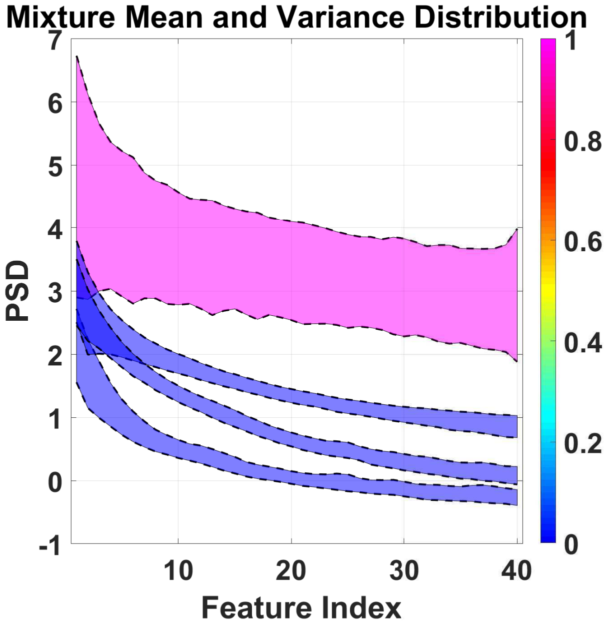


Figure 7.30. The 8 Mixture PSD `AbnNrm NrmSzr` Furthest-Aligned. Pairing of `AbnNrm` 8-mixture UBM with its F-A pair from the `NrmSzr`.

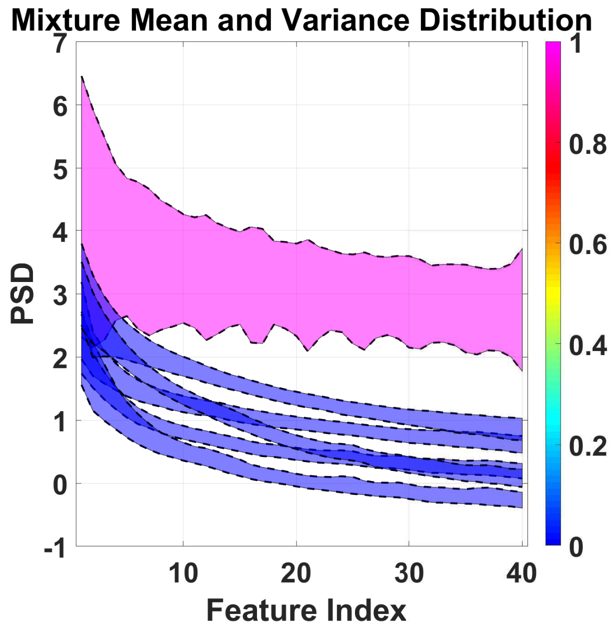


Figure 7.31. The 8 Mixture PSD AbnNrm AbnSzr Furthest-Aligned. Pairing of AbnNrm 8-mixture UBM with its F-A pair from the AbnSzr.

F-A failed to generate 7, and the C-A failed to generate 4. The C-D classification generated pairings for all of its experimental configurations. These 20 missing pairings were an improvement from the AbnNrm base missing 26 pairings from the previous experiment, Table 7.22.

The most used mixture from the 8-mixture UBMs when using the AbnNrm target data was mixture 2. This mixture was used by both datasets for their C-D (Figures 7.34 and 7.35) and F-A (Figures 7.36 and 7.37) pairings. Despite the similar mixture number, their mixtures were, of course, distinct as each base was constructed from a different dataset. However, the base mixtures were linked across the classifications Figures 7.34 and 7.36 and Figures 7.35 and 7.37.

From the C-D classifications, although the mixtures occupy similar feature spaces, the TVM finds them to be distinct. Naturally, one might make the same assumptions upon viewing them, but recall that these are the best distance matches according to

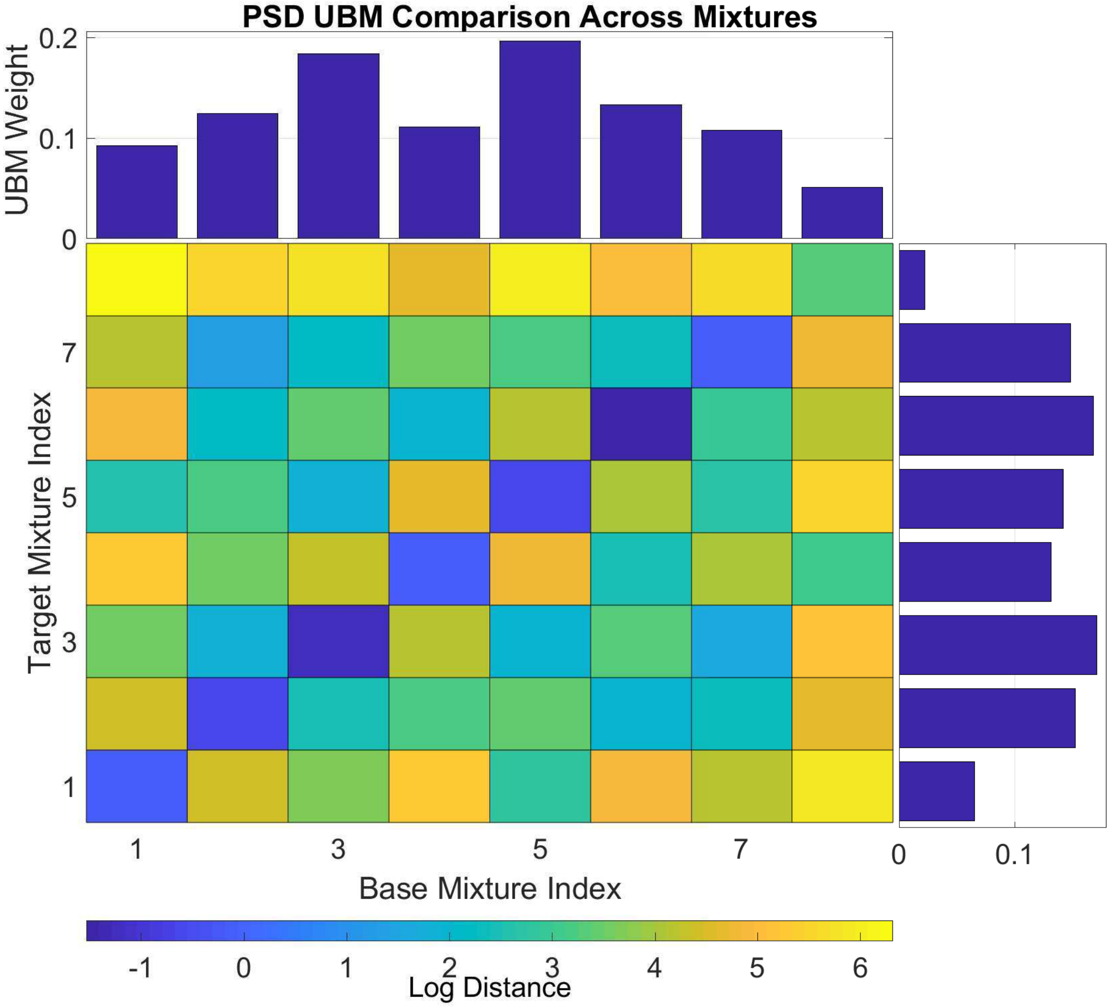


Figure 7.32. The 8 Mixture PSD AbnNrm NrmSzr UBM Confusion Matrix. UBM distance measurements between the AbnNrm base and NrmSzr target.

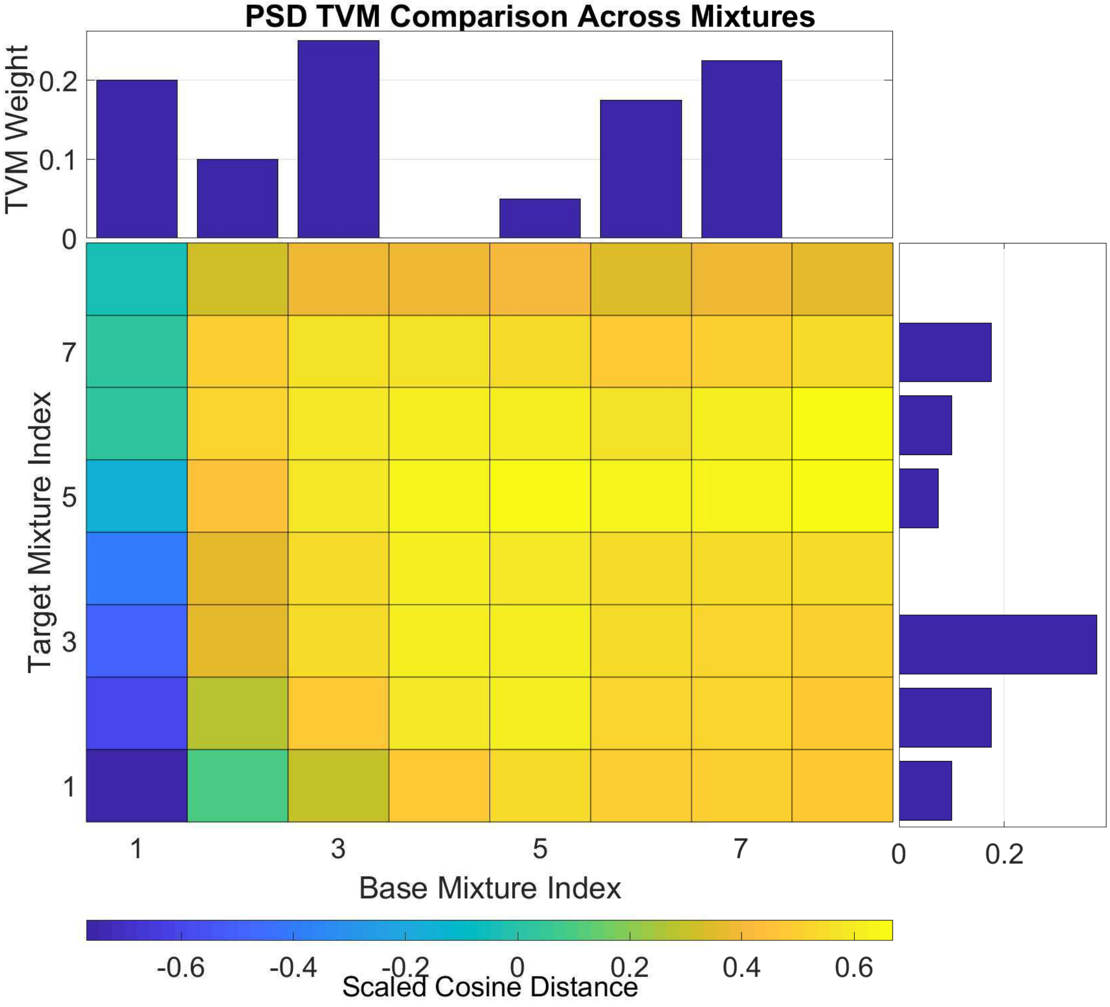


Figure 7.33. The 8 Mixture PSD AbnNrm NrmSzr TVM Confusion Matrix. TVM distance measurements between the AbnNrm base and NrmSzr target.

Table 7.23. `AbnNrmMot` and `AbnNrmSzr` Base Mixture Matches

Matches		`AbnNrmMot`			`AbnNrmSzr`		
		`AbnNrm`	`NrmMot`	`AbnMot`	`AbnNrm`	`NrmSzr`	`AbnSzr`
C-A	8	-	5	4	1	5	5
	16	11	12	-	8,15	12	-
	32	17	9, 19,25	3	-	17,29	25
C-D	8	2,8	7	8	2,3, 5	7	8
	16	16	1,5	1, 9,16	2,5	2	14
	32	15	15	24	17	2,12, 19,22	2,18 23,24
F-A	8	2,4	-	-	2,4, 6,8	1,4, 5,6	-
	16	4,6	4	3, 5,7	2,5	-	-
	32	12,24	8	-	2, 16,20	26,27	-
F-D	8	1,7	1,2, 3,5	2,3, 5,7	-	1,4, 5,6	1,5 7
	16	-	5,9, 10,11	6,8	-	4	-
	32	-	-	-	-	4	-

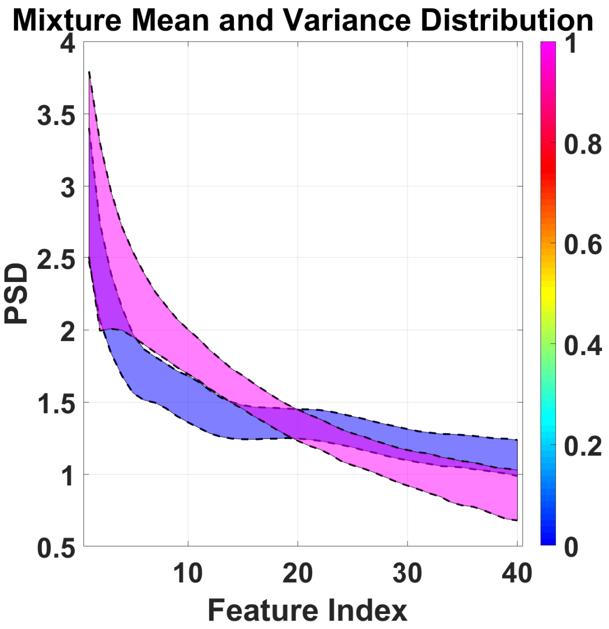


Figure 7.34. The 8 Mixture PSD AbnNrmMot AbnNrm Closest-Divergent. Pairing of AbnNrmMot (base) 8-mixture UBM with its C-D pair from the AbnNrm (target).

the UBMs. More interesting is the presence of previously uncovered dataset-specific mixtures in the F-A pairings. Both larger datasets appeared to pick the same target mixture from the AbnNrm UBM and pair it against mixtures native to motion and seizure data.

Figures 7.34–7.37 suggested that decision surfaces were easy to find between the two datasets given the repeated use of the same base mixtures. This was similar to the AbnNrm results in Section 7.3.6.1 which also used 8-mixture UBMs. This was more likely due to their locations in the feature space rather than the mixtures themselves, as repeated mixtures occurred less frequently as the UBM mixture size was increased, seen in Tables 7.22 and 7.23.

Even with success using smaller UBMs, the larger 32-mixture UBMs still offered valuable insight when the TVM found divergent distances in Figures 7.38 and 7.39. The presence of motion data in Figure 7.38 caused a variation in mixtures around the

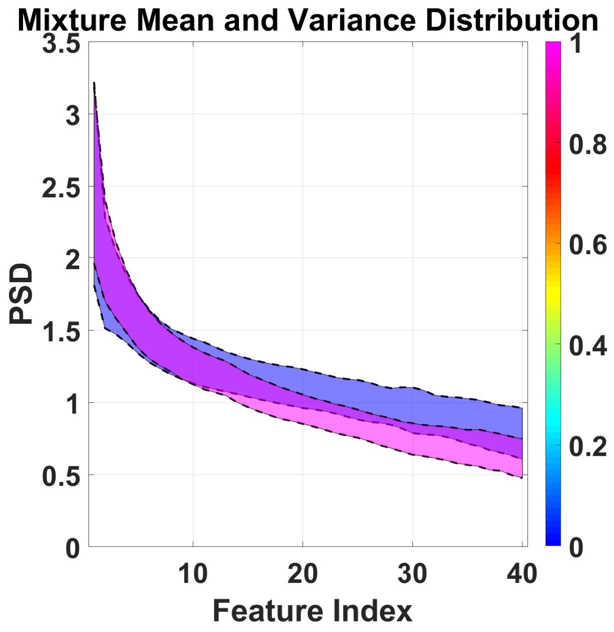


Figure 7.35. The 8 Mixture PSD `AbnNrmSzr AbnNrm` Closest-Divergent. Pairing of `AbnNrmSzr` 8-mixture UBM with its C-D pair from the `AbnNrm`.

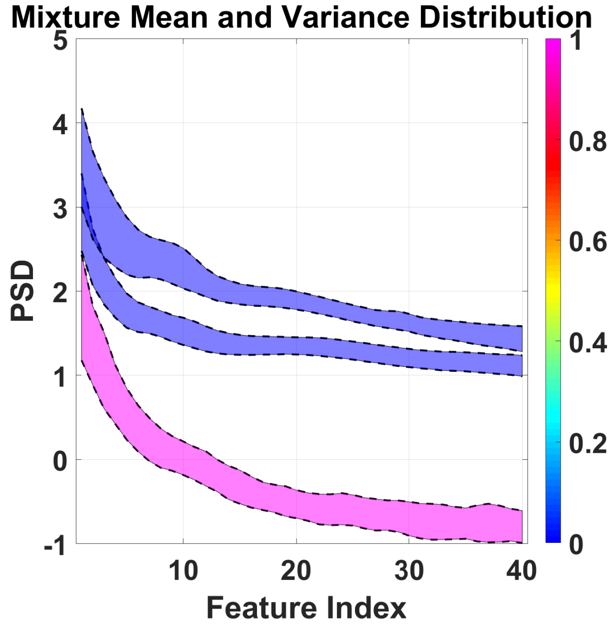


Figure 7.36. The 8 Mixture PSD `AbnNrmMot AbnNrm` Furthest-Aligned. Pairing of `AbnNrmMot` 8-mixture UBM with its F-A pair from the `AbnNrm`.

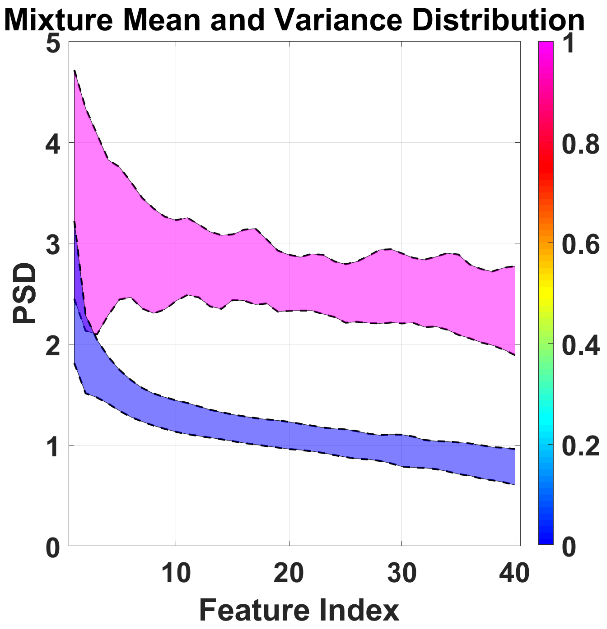


Figure 7.37. The 8 Mixture PSD `AbnNrmSzr` `AbnNrm` Furthest-Aligned. Pairing of `AbnNrmSzr` 8-mixture UBM with its F-A pair from the `AbnNrm`.

alpha band (8-12Hz or 8-12 feature index). This generated a decision surface between the base and target datasets. However, the C-D classification for the `AbnNrmSzr` base and `AbnNrm` target in Figure 7.39 had no such behavior. Its pairing was wholly different in terms of means and variances for the base and target mixture as compared to the motion mixture.

Reviewing the most active rows, as done previously, the F-D classifications of `AbnNrmMot` and `AbnNrmSzr` presented with similar mixture groupings. In the case of the `AbnNrmMot`, some of these mixtures mapped into the `AbnNrm` dataset as well. The F-D classification had previously provided insight into mixtures unique to both datasets, making it an ideal classification for finding decision surfaces.

The `AbnNrmMot` base produced Figure 7.40 for target `NrmMot` and Figure 7.41 for target `AbnMot`. In both instances, similar base and target mixtures were found. The

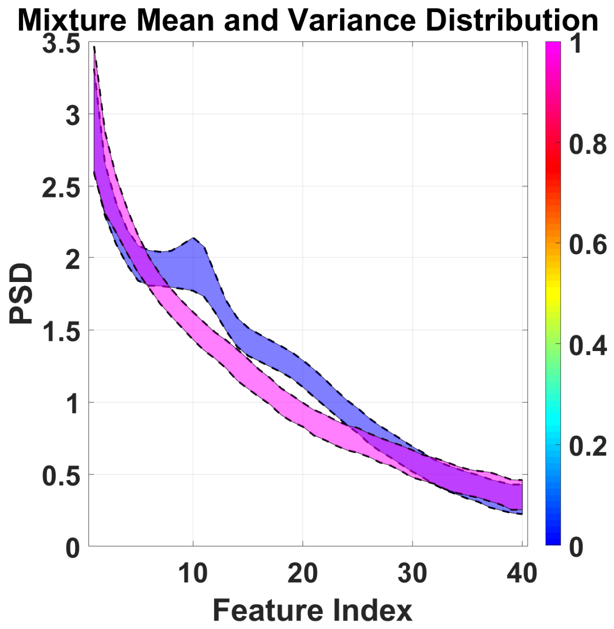


Figure 7.38. The 32 Mixture PSD `AbnNrmMot AbnNrm` Closest-Divergent. Pairing of `AbnNrmMot` 32-mixture UBM with its C-D pair from the `AbnNrm` dataset.

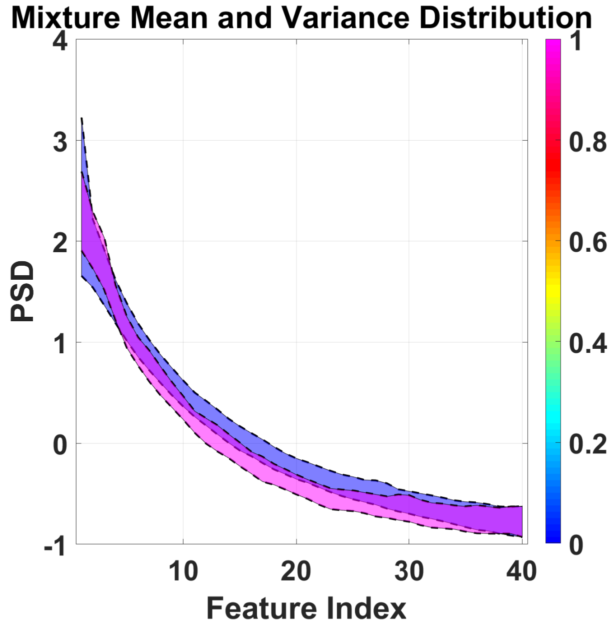


Figure 7.39. The 32 Mixture PSD `AbnNrmSzr AbnNrm` Closest-Divergent. Pairing of `AbnNrmSzr` 32-mixture UBM with its C-D pair from the `AbnNrm` dataset.

AbnNrmSzr base produced Figure 7.42 for target NrmSzr and Figure 7.43 for target AbnSzr. These too contained similar base and target mixtures.

These mixtures appeared to contain a mixture that was previously associated with the seizure dataset (in Figures 7.19 and 7.21), specifically with the nodule between feature elements 30 and 40. Similar mixtures were prevalent in both the base mixtures, Figures 7.42 and 7.43, and target mixture, Figure 7.44. Multiple datasets produced similar mixtures, but in each instance, their use was determined by the TVM to the point that even when both base and target contained them, they were still used to separate the datasets.

Results such as these indicated that the relationship between UBM and TVM had the ability to identify mixtures that were similar not only in their location in the feature space, but also their ability to drive classification. The progression of mixtures through the UBM mixture sizes was yet another way to understand the relationship between the UBMs and TVMs. As the mixtures became better articulated with each subsequent UBM, the TVM reevaluated how to link the mixtures for classification resulting in an unexpected parent-child relationship between the mixture sizes. While some instances of this were presented in this work, a more thorough analysis would require building upon the mixture pairings tables and a wider range of target datasets to track the lineage of all such mixtures across their base datasets. This would bring the work closer to the overarching goal of a clinician free annotation system for EEG data.

7.3.6.3 AbnSzrMot and NrmSzrMot

The final experiments used the largest and most diverse datasets from Chapter 6, the AbnSzrMot and NrmSzrMot. Given the small size and similarity of the Abn and Nrm datasets, it hoped that the results would be well correlated across the base datasets.

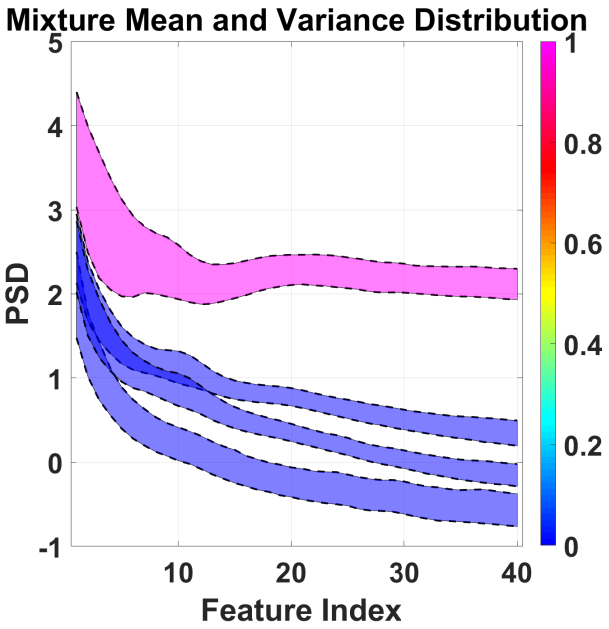


Figure 7.40. The 8 Mixture PSD AbnNrmMot NrmMot Furthest-Divergent. Pairing of AbnNrmMot 8-mixture UBM with its F-D pair from the NrmMot.

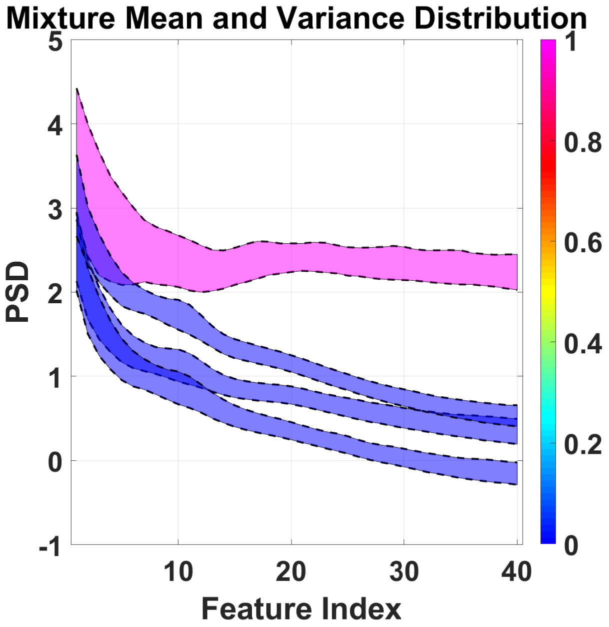


Figure 7.41. The 8 Mixture PSD AbnNrmMot AbnMot Furthest-Divergent. Pairing of AbnNrmMot 8-mixture UBM with its F-D pair from the AbnMot.

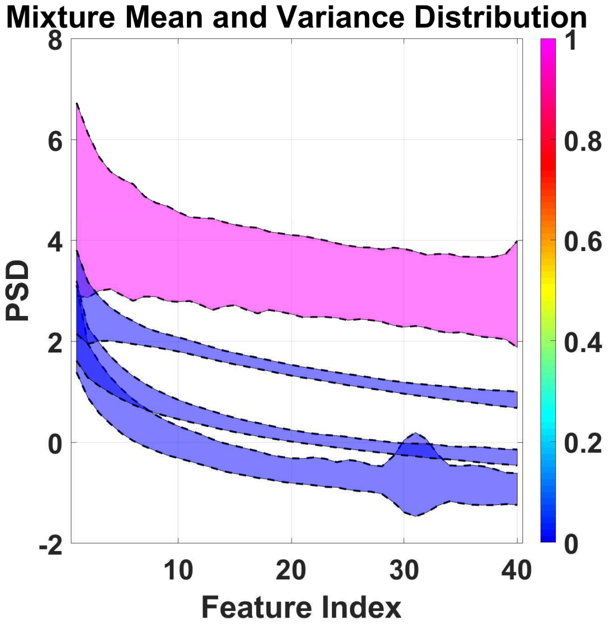


Figure 7.42. The 8 Mixture PSD `AbnNrmMot` `NrmSzr` Furthest-Divergent. Pairing of `AbnNrmSzr` 8-mixture UBM with its F-D pair from the `NrmSzr`.

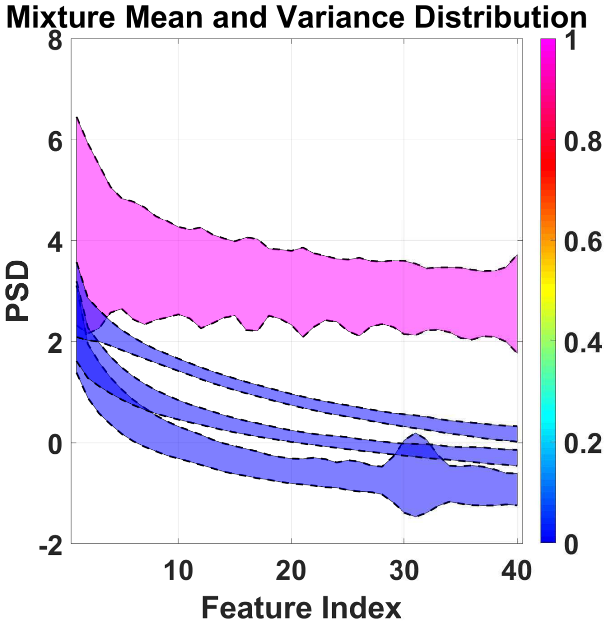


Figure 7.43. The 8 Mixture PSD `AbnNrmSzr` `AbnSzr` Furthest-Divergent. Pairing of `AbnNrmSzr` 8-mixture UBM with its F-D pair from the `AbnSzr`.

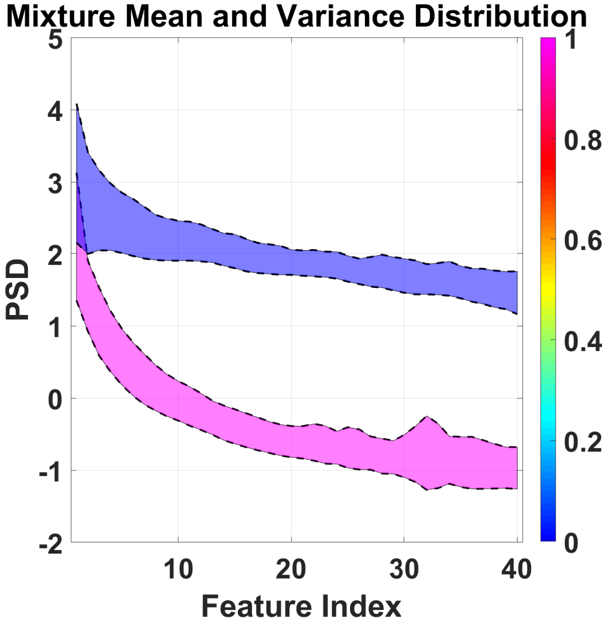


Figure 7.44. The 8 Mixture PSD AbnNrmSzr AbnSzr Furthest-Divergent. Pairing of AbnNrmSzr 8-mixture UBM with another F-D pair from the NrmSzr.

The control dataset this time was SzrMot which was the largest dual dataset tested. The Abn base further tested against the AbnMot and AbnSzr, while the Nrm base used NrmMot and NrmSzr.

These test configurations produced 23 missing pairings, Table 7.24. This was worse than the previous large base datsets, and a slight improvement over the AbnNrm results. The F-D classification failed to generate 10 pairings, the F-A failed to generate 4, the C-D failed to generate 4, and and the C-A failed to generate 5 pairings. The failures of the C-A pairings occurred only within the NrmSzrMot dataset and the AbnSzrMot accounted for those in the F-A pairings. As in the prior section with Table 7.23, there was a distinct cluster of mixtures in one of the rows of Table 7.24. This time it was the C-D 16-mixture UBM pairings consisting of mixtures 1 and 9

When these repeated pairings were examined in Figure 7.45 (AbnSzrMot to SzrMot) they were similar to those of the mixtures in Figure 7.46 (NrmSzrMot to

Table 7.24. `AbnSzrMot` and `NrmSzrMot` Base Mixture Matches

Matches		`AbnSzrMot`			`NrmSzrMot`		
		`AbnMot`	`AbnSzr`	`SzrMot`	`NrmMot`	`NrmSzr`	`SzrMot`
C-A	8	8	1,8	15	1	8	5,6
	16	16	13, 14,15	14	-	5,15, 16	-
	32	8,28	10	5,22, 28	-	-	-
C-D	8	6	-	8	7	-	8
	16	1,13	1,10	1,9	3,9	7	1,9
	32	-	21,28	25	18,22	9,10, 19,26	-
F-A	8	3, 5,7	2,3, 5,7	2,4,8	2,4, 6,8	1, ,5,7	4
	16	2,15	-	-	2,5, 15,16	2	16
	32	21	-	-	20,22, 30,31, 32	2,4, 5,6, 16	12
F-D	8	-	-	1,3, 5,7	-	4	1,2 3,5, 7
	16	9	-	8	6,8, 10,12	-	13
	32	-	-	17,22, 30,32	-	-	-

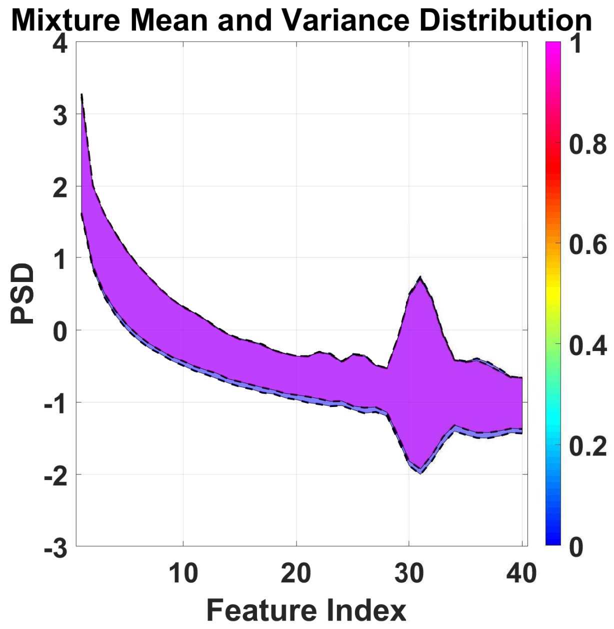


Figure 7.45. The 16 Mixture PSD AbnSzrMot SzrMot Closest-Divergent. Pairing of AbnSzrMot 16-mixture UBM with its C-D pair from the SzrMot dataset.

SzrMot). These pairings reinforced the influence of what was thought to be an established seizure mixture regardless of what datasets were paired with the seizure dataset.

Given the previous results played out similarly, the 8-mixture F-D pairings were examined as well despite being sparse outside of the SzrMot dataset, Figures 7.47 and 7.48. These mixtures again reflective of previous mixtures and groupings, given their appearance in the previous results of Figures 7.30 and 7.31. The continued model and discovery of the same influential UBM mixtures across datasets and UBM sizes suggested that the proposed analysis had worked.

Using large complex base and target datasets appeared to be the trigger to activating the 32-mixture UBMs. Within the NrmSzrMot experiments, the F-A 32-mixture UBM pairings produced 5 mixtures for the NrmMot, Figure 7.49, and NrmSzr, Figure 7.50, datasets. In these two instances, the two pairings fit perfectly

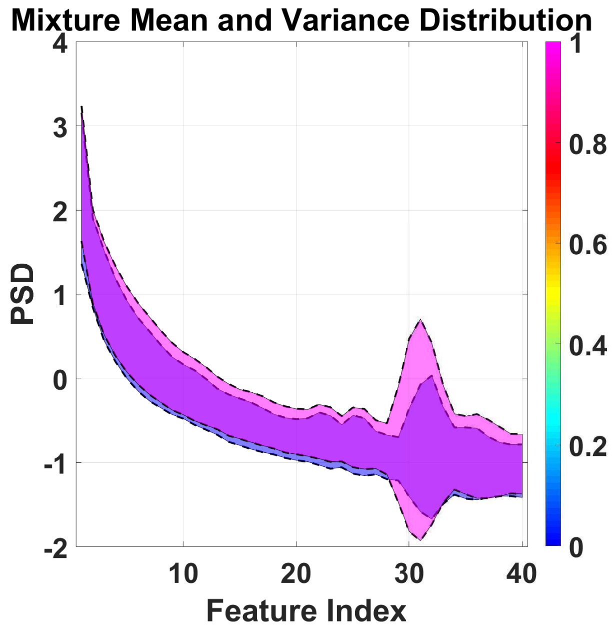


Figure 7.46. The 16 Mixture PSD `NrmSzrMot` `SzrMot` Closest-Divergent. Pairing of `NrmSzrMot` 16-mixture UBM with its C-D pair from the `SzrMot` dataset.

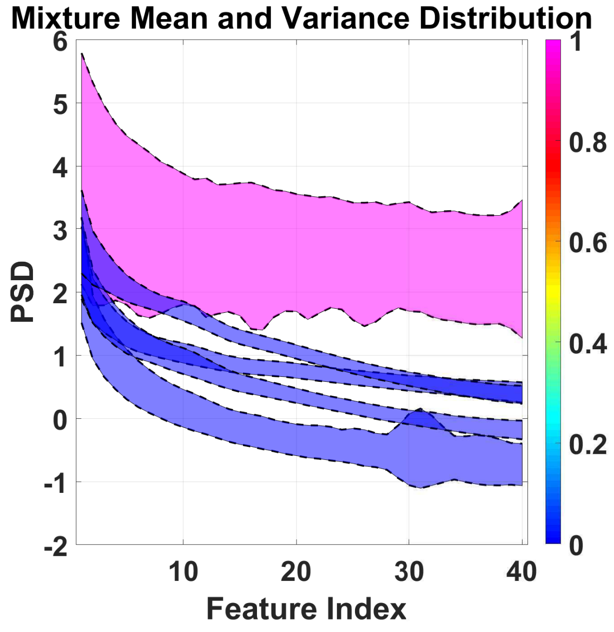


Figure 7.47. The 8 Mixture PSD `AbnSzrMot` `SzrMot` Furthest-Divergent. Pairing of `AbnSzrMot` 8-mixture UBM with its F-D pair from the `SzrMot` dataset.

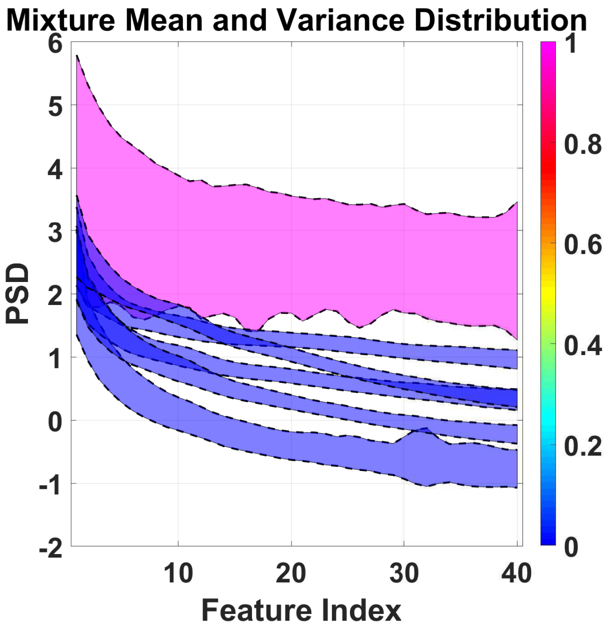


Figure 7.48. The 8 Mixture PSD `NrmSzrMot` `SzrMot` Furthest-Divergent. Pairing of `NrmSzrMot` 8-mixture UBM with its F-D pair from the `SzrMot` dataset.

around each other in the feature space. The exception was that the target mixture of Figure 7.49 appeared to match within the base mixtures of Figure 7.50 suggesting it would also pair with its target mixture.

A similar pattern occurred again at the 16-mixture UBM F-A pairings as well, Figures 7.51 and 7.52. Here, the target one of pairing overlaps area occupied by the base of the other pairing and vice versa, confirming the link between these mixtures and suggesting these two feature spaces are integral to their datasets. This is in a way a self evaluation, if the base and target datasets were identical, which was omitted from analysis to prevent such recursive relationships. The novel ability to track similar mixtures across different base and target datasets to identify such a recursive relationship confirmed robustness of the proposed technique.

However, if operating on a single dataset, the results may have been ignored as artifacts due to the small weight assigned to the mixtures by their UBMs. In both

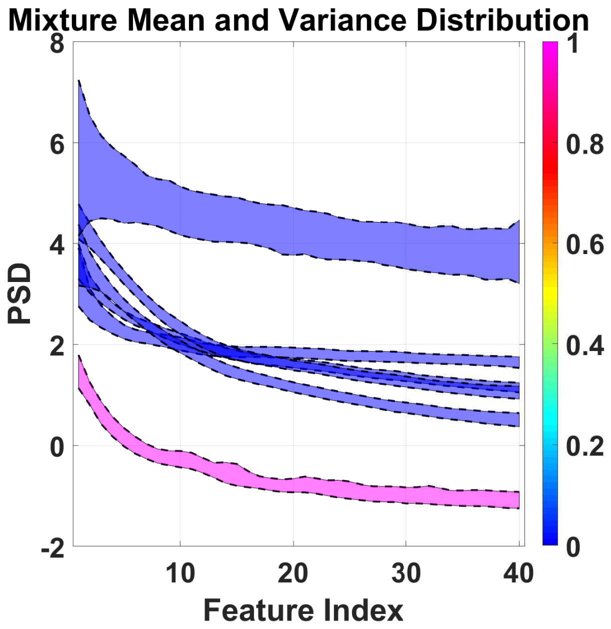


Figure 7.49. The 32 Mixture PSD NrmSzrMot NrmMot Furthest-Aligned. Pairing of NrmSzrMot 32-mixture UBM with its F-A pair from the NrmMot dataset.

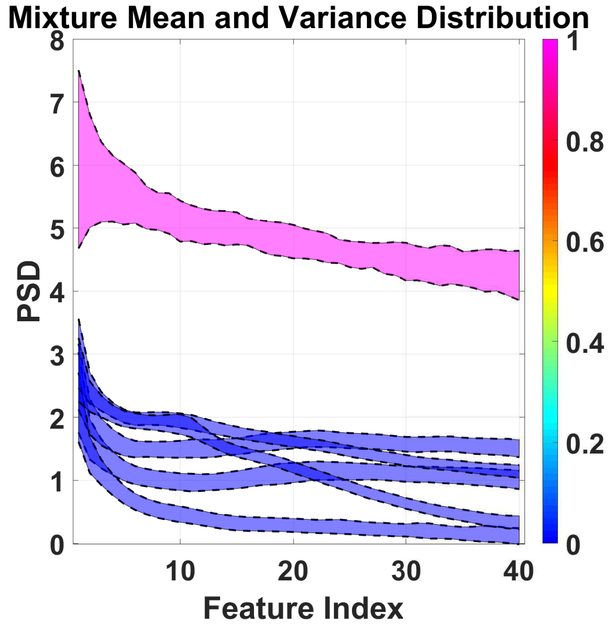


Figure 7.50. The 16 Mixture PSD NrmSzrMot NrmSzr Furthest-Aligned. Pairing of NrmSzrMot 32-mixture UBM with its F-A pair from the NrmSzr dataset.

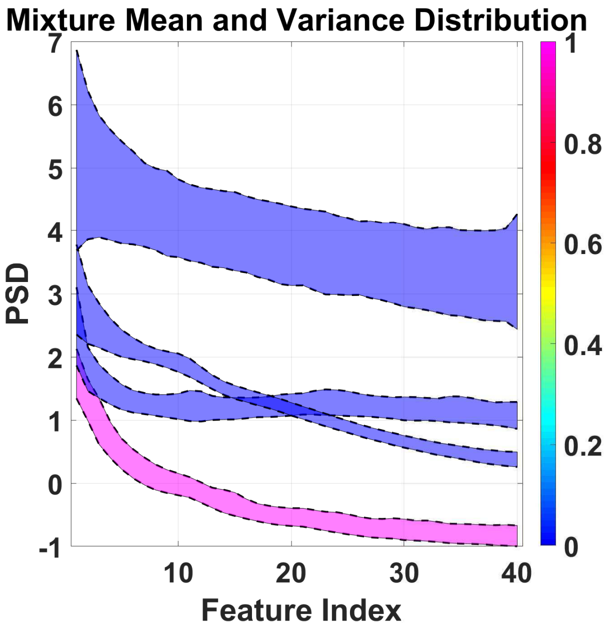


Figure 7.51. The 16 Mixture PSD `NrmSzrMot` `NrmMot` Furthest-Aligned. Pairing of `NrmSzrMot` 16-mixture UBM with its F-A pair from the `NrmMot` dataset.

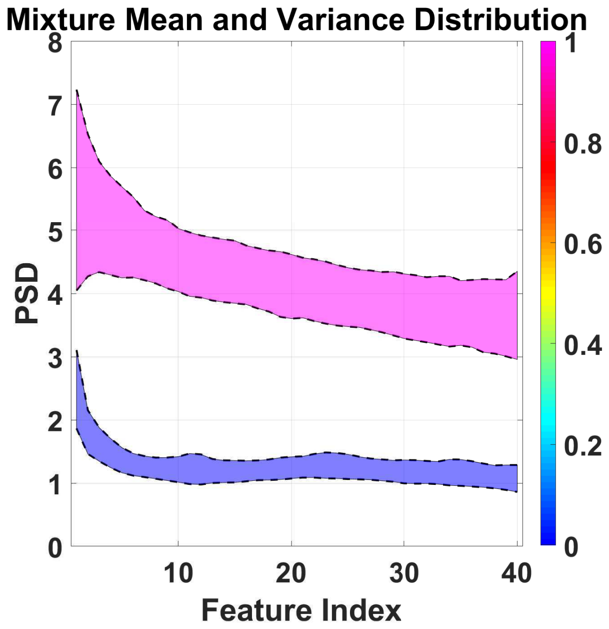


Figure 7.52. The 16 Mixture PSD `NrmSzrMot` `NrmSzr` Furthest-Aligned. Pairing of `NrmSzrMot` 16-mixture UBM with its F-A pair from the `NrmSzr` dataset.

sets of pairings, the UBM distance of the targeted mixtures were the lowest of the 16 and both TVMs gave that same mixture zero weight as well. This would have suggested the mixtures in question were artifacts, but given the TVMs had aligned them with other mixtures in the base it was possible they represented authentic data at which point it would be necessary to compare the single dataset with a unique target to determine the true nature of the mixtures. This is precisely the problem with unsupervised techniques is that they are entirely dependent on the data given to them.

Validating such claims is difficult without a clinician, but this us yet another way the technique serves to provide context without them. A true artifact would likely be an outlier in any dataset. However if the underlying UBM mixture produces minimal distances to other mixtures that are (a) heavily weighted and (b) flagged as one of the four UBM-TVM relationships, it would be reasonable to assume the mixture is not an artifact. This was beyond the scope of the work, but should be addressed if the work is continued.

The potential outlier mixture was found again by the F-A pairings of the 8-mixture UBMs across both feature sets in Figures 7.53–7.56. In these instances, the target and base were considered aligned by the TVMs indicating that the area occupied by the target feature space in Figures 7.53 and 7.55 was critical for modeling specific datasets. Therefore artifact mixtures in this region could have been harder to diagnosis given its importance across datasets.

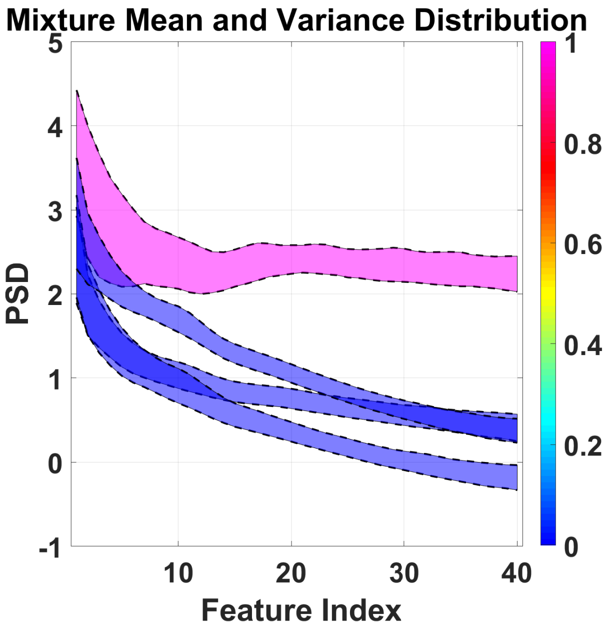


Figure 7.53. The 8 Mixture PSD AbnSzrMot AbnMot Furthest-Aligned. Pairing of AbnSzrMot 8-mixture UBM with its F-A pair from the AbnMot dataset.

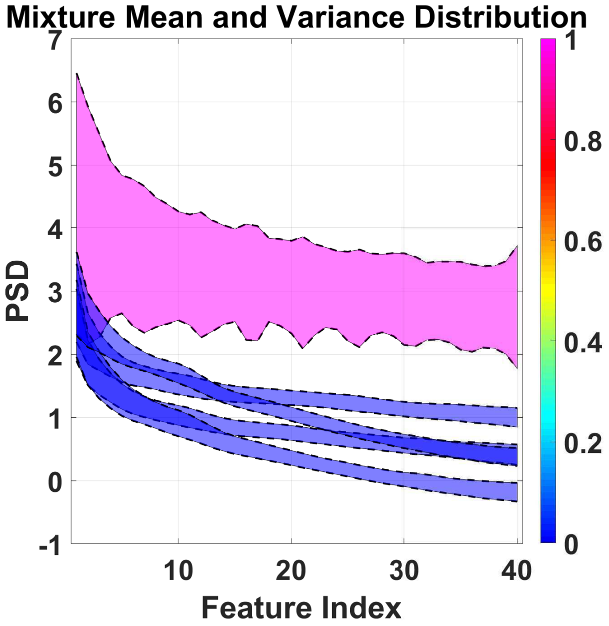


Figure 7.54. The 8 Mixture PSD AbnSzrMot AbnSzr Furthest-Aligned. Pairing of AbnSzrMot 8-mixture UBM with its F-A pair from the AbnSzr dataset.

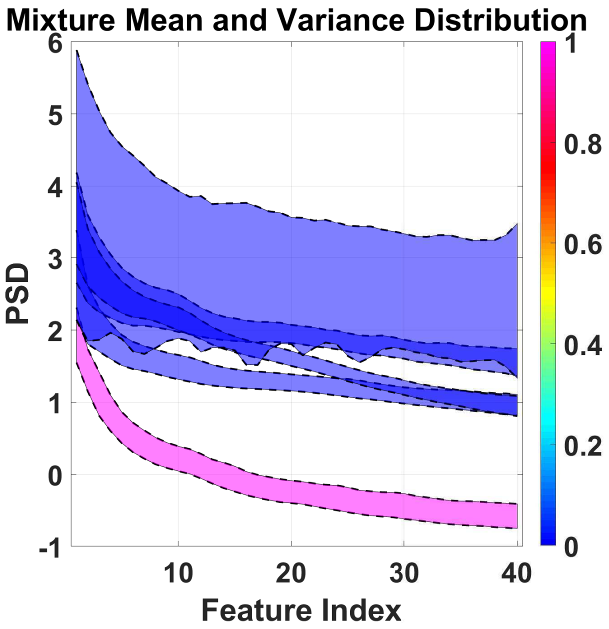


Figure 7.55. The 16 Mixture PSD NrmSzrMot NrmMot Furthest-Aligned. Pairing of NrmSzrMot 8-mixture UBM with its F-A pair from the NrmMot dataset.

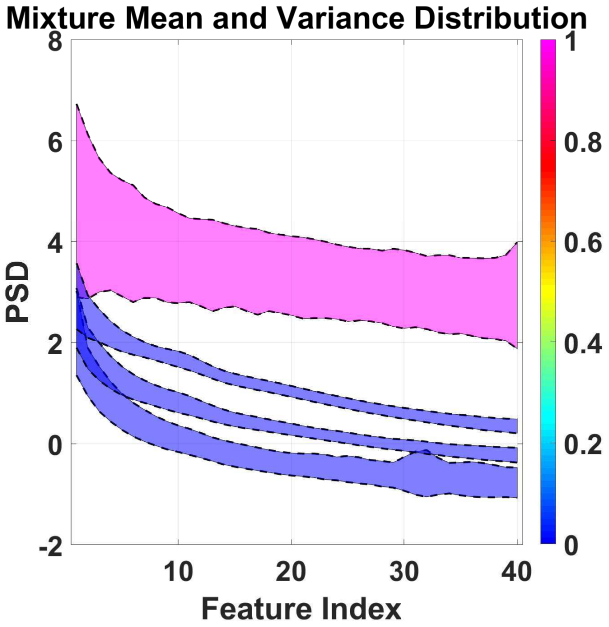


Figure 7.56. The 16 Mixture PSD NrmSzrMot NrmSzr Furthest-Aligned. Pairing of NrmSzrMot 8-mixture UBM with its F-A pair from the NrmSzr dataset.

7.3.7 Constraints

The proposed analysis technique was new to the field of EEG analysis, which led to a number of limitations. Chiefly among these was the lack of diverse data that limited previous work to parse the fundamental components of EEG waveforms. Despite operating on over 500 subjects drawn from four distinct data sources, this was not enough breadth or depth either. The failure to produce classification pairings for each experimental test point was a major drawback. As mentioned previously, this should be mitigated by including the common to aligned/divergent and common to closest/furthest label pairings.

Increasing the amount of data and comparison points would have improved the understanding of where artifacts, dataset specific mixtures, and generalized EEG mixtures appeared within the feature space. However, achieving this would have required exhaustive testing using the individual seizure, motion, normal and abnormal datasets from Chapter 5. Adding all of these into this chapter was not feasible as it was beyond the scope of the work. This was due to the need to present results using datasets that shared natural modes which were to allow the UBM to be linked in the given feature space.

It should have been possible to link across the feature sets as well, but again that would have required extensive simulations that were beyond the scope of work. Tracking the results of this work required presenting concise visualizations which proved difficult to maintain for larger UBMs. Operating on metrics and equations alone would have provided minimal tools to debug and validate the technique during it's development. Now that the process has been validated this is no longer a concern, but increasing the analysis to link through the feature sets or included more datasets will still compounded the difficulties in interpreting the results.

Linking the TVMs via their cosine similarity proved to be effective, but was no ideal. Understanding the alignment of the 'impulse response' from each TVM mixture should have been addressed as determining the direction of lines in the dimensions of the feature space. This type of problem was beyond the scope of this work, but generating not only alignment but a significant understanding of the directions in the feature space would have enhanced the mapping between the TVMs.

7.4 Conclusion

The results of the example, calibration, and various dataset specific experiments showed that the proposed I-Vector technique was capable of determining which mixtures that the TVM had used to inform its decision surfaces. While the results were not perfect, they did show the ability to resolve to the same closed set of solutions across varied datasets and UBMs of three sizes. These were promising results, despite clear setbacks in terms of missing pairings and datasets with known strong overlapping features.

The impact of UBM mixture size was difficult to judge as the 8-mixture labels did an acceptable job of mapping the feature space, despite not capturing all of the nuance of the mixtures. This was addressed by structuring the experiments to draw from the larger datasets as their base, there was direct improvement in classification pairings in terms of quantity and quality. However, it was intriguing to see how much information a 8-mixture UBM could contain given the amount of data used in training it. This was not entirely unexpected given the strength of the GMM-UBMs classifier in Chapter 5 for the 4, 8, and 16-mixture UBMs. Yet, even that technique struggled to perform subject verification when operating on the `Szr` dataset.

This UBM-TVM relationship analysis technique could be improved with additional testing, as noted in the Constraints section(s), but also by expanding to data that does not come from EEGs. acpIV were an adaptation of JFA, made an in effort to serve the speech recognition community's needs. The headway made here by an unsupervised machine learning approach suggested the potential for this technique in assisting the EEG processing communities growing need of labeled data.

Chapter 8

CONCLUSION

At the time of this writing there was minimal if any work on the use of I-Vectors on EEG for the purpose of bio-metric subject verification. Work had previously been carried out using UBMs for subject verification [103, 163] and ADHD classification [42] and MD based subject verification [64, 65]. The deployment of these three algorithms over the PhysioNet Database and TUH-EEG datasets represented not only the formal introduction of I-Vectors for use on EEG data, but also a substantial set of benchmark experiments for the algorithms and datasets involved.

The experiments titled "Protocol Replication" and "Parameter Sweeps" were carried out in Chapters 4–6 following a protocol based on the work of La Rocca [64]. These addressed not only how to select the appropriate epochs and tune the I-Vector parameters, but also sweeps across CEP, PSD, and COH feature sets. This laid the groundwork for further development of I-Vectors beyond the subject verification experiments, based upon the UBM-TVM relationship analysis as presented in Chapter 7.

8.1 Parameter Sweeps

The Parameter Sweeps addressed the operational parameters of all the algorithms by testing feature set, epoch duration, number of epochs, and different combinations of datasets. Specific to the I-Vectors were tests focused on the construction of the UBM mixture size, TVM dimension, and LDA dimension.

Based upon epoch duration and UBM mixture sweeps using the PhysioNet Database dataset and all three algorithms, it was determined that the range of tested UBM mixtures should be 2 to 2048 and the range of epoch duration should be 10s and 2s. Further testing across each independent Abnormal, Normal, Motion, and Seizure dataset using each of the three features indicated that the optimal epoch duration was 2s with 12 epochs removed from the training/enrollment dataset to be used for the testing dataset. This configuration was developed by testing I-Vectors alone but applied to all algorithms for the remainder of the experiments. However, the ideal UBM range was found to be 32 to 128 mixtures which was shared by the GMM-UBM and I-Vector algorithms. The I-Vector specific components of TVM and LDA dimension were revealed to be optimal at a 25 dimension TVM and no LDA. These results necessitated the continued testing of each feature set, and the LDA dimensions into the Algorithm Benchmarks.

A shortcoming of these results was that the number of epochs in the training/enrollment dataset was not controlled. While the number of testing epochs was specified, the remaining epochs were used for training/enrollment which was dependent on the epoch duration of 2s. Thus the balance of training to testing was set at 48 training epochs for every 12 testing epochs, corresponding to a 80/20 split. This required required the cross validation to produce 6 randomized sets instead of the 6 exhaustive sets used in the Protocol Replication to ensure an unbiased accounting of the algorithms' performances.

Furthermore, the overall amount of data was a clear factor in the performance of the larger UBM mixtures and TVM dimensions. The smaller (50 subject) Normal and Abnormal datasets were unable to utilize the larger components as well as the 411 subject Seizure dataset. This indicated that reaching a data threshold, in terms of the total number of training epochs, was necessary before enabling the larger TVM dimensions, despite the number of epochs being well in excess of the maximum number of UBM mixtures generated. The datasets offered three subject counts (50, 109, and 411) but these values were not enough to accurately map the epoch threshold level required for all the TVM dimensions.

These constraints were minor, as the overall performance of I-Vectors across the feature and dataset combinations suggested they would be competitive with GMM-UBMs on the aggregated datasets. In fact, their performance during the Protocol Replication was able to exceed that of the GMM-UBM and MD algorithms. This proved the efficacy of the protocols developed and provided a substantial reduction in degrees of freedom leading to more controlled experiments for the Algorithm Benchmarks of Chapter 6 and the UBM-TVM relationship of Chapter 7.

8.2 Algorithm Benchmarks

The Algorithm Benchmarks addressed the ability of the I-Vectors to compete with with the MD and GMM-UBM algorithms for each feature type over the aggregated datasets. These datasets provided a wider range of subject counts (100, 159, 461, 511, 520, and 570) which enabled the impact of LDA on I-Vector generation to be completed as part of the Parameter Sweep experiments.

Overall, I-Vector performance for subject verification was found to be on par with similarly reported techniques following the Parameter Sweeps. The initial results of the Protocol Replication showed I-Vectors matched the results of La Rocca's work [64], but the Algorithm Benchmarks showed the technique produced results on par with other GMM and/or GMM-UBM classification tasks [42, 86, 98, 103, 163]. These results indicated that I-Vectors were competitive with other unsupervised learning tools for EEG subject verification, despite being unable to surpass the performance of the GMM-UBM algorithm. However, the trade off was that of the three algorithms the I-Vectors used the smallest feature vector for classification making it an acceptable classifier while also acting as a strong dimensional reduction technique (even before applying LDA).

Each algorithms' performance was effected by feature and dataset, which was most stark when using the TUH-EEG seizure dataset. While the GMM-UBM continued to be the top performer, it frequently failed to exceed the desired minimum 0.75 C Metric score. This was anticipated, since the classification of seizure EEG remains an active research topic for clinicians and data scientists alike [43]. In general, each variant of EEG data finds itself operating on a new novel dataset tailored for the given condition such as seizures [108], sleep [81] or motion tasks [86]. These experiments frequently use one of any number of classifiers, many of which were outlined in Chapter 2 such as SVMs, DPs, and the rapidly advancing field of NNs, while also searching for an optimal feature set (entropies [152], raw data [14] or fractal features [214]). Despite the novelty of such experiments, the work presented here was far more ambitious because it combined multiple datasets together and analyzed them using multiple feature sets. This tested the ability of the classification algorithms and the feature sets in ways not previously documented on publicly sourced datasets.

In doing so, it was discovered that the previous optimal UBM mixture range as determined in Chapter 5 was poorly understood, as I-Vector performance peaked with larger UBMs on the aggregated datasets. Even the GMM-UBM algorithm performance required larger UBMs for their performance to peak. This indicated that the Parameter Sweeps were lacking. To address this, the aggregated datasets of Chapter 6 should have been used in the Parameter Sweeps until an inflection point in performance due to UBM size was observed. Multiple plots in Chapter 5 indicated that larger UBMs could have improved I-Vector performance, particularly on the datasets containing seizure data.

Within the tested experiment protocol, LDA showed the ability to provide C Metric score gains of 10% to 30% depending on the feature set and dataset as seen in Chapter 6. In some instances this was enough to push the I-Vector performance beyond that of the MD algorithm. LDA's impact was more prevalent on the CEP and COH features as they were unable to match the performance of the PSD features. This may have been an indication that CEP and COH features are for niche applications as COH features are frequently used on ERP classification tasks given the emphasis on synchronicity [162]. However, for all three features, LDA's ability to improve the performance of the 25 dimension TVMs was minimal if any at all. This suggested the TVMs are all modeling different facets of the data, as their performance when reduced to the same dimension via LDA was never similar.

The shortfall of the Algorithm Benchmarks was that at the 1024 and 2048 mixture UBMs the GMM-UBM performance appeared to be trailing off just as the larger dimension TVMs were improving their performance. It is likely that both the number of UBM mixtures and TVM dimensions were undersized for the largest of datasets. Adding the larger datasets to the Parameter Sweeps might have addressed this, but the experimental protocol relied on 2 minutes of data from each subject which clearly was too little in hindsight. Scaling up the amount of data from each subject while maintain 2s duration epochs would likely require far larger UBMs and might have shown I-Vector to be superior to GMM-UBM on larger datsets given their performance trends.

8.3 UBM-TVM Relationship

The UBM-TVM relationship experiments were the most pioneering aspect of this work. As useful as ML algorithms have become, much of their progress has remained obscured in their multiple layers of weighted coefficient matrices and nonlinear activation functions. While they rapidly advance our ability to use EEG data, they often fail to provide insight into the data itself and how they made their classification decisions. Articulating why a decision is made requires understanding how a decision is made, which was the goal of understanding how the TVM adapts itself to allow for such low-dimension feature vectors to produce near state of the art subject verification performance.

By directly comparing the relationships produced by the UBMs and their TVM counterparts, the basic mechanics of the I-Vector generation process were brought into focus. The results of Chapter 7 indicated that by contrasting the mixture positions in the EEG feature space with the nominal mixture-specific TVM mean shifts, a relationship between position and application was modeled for each experiment's dataset. By pairing these datasets against each other as a known 'base' and unknown 'target', mixtures that acted as decision surfaces between the datasets were discovered. Given the scale of these experiments, these mixtures were not related directly to the subjects, but rather to the qualities of the datasets (abnormal, normal, motion, and seizure) themselves.

This initial step was possible because of the strong performance of I-Vectors and GMM-UBMs on the PSD feature sets. The entire technique relied on the UBMs to have accurately modeled the data, which was affirmed by the strong classification performance of GMM-UBMs using the smaller mixture sizes. Likewise, the TVMs had to be capable of providing classification on par with their GMM-UBM counterparts, which was evidenced in the 32, 64, 128, and 256 mixture UBMs. Thus, analyzing the modeling done prior to the strong subject verification, using the 8, 16, and 32 mixture UBMs would provide insight into the four type of datasets.

The results showed that the largest aggregated datasets (built from three of the abnormal, normal, motion, and seizure datasets) used the same mixtures to generate consistent relationships against the target datasets across each size of the UBMs. When comparing the mixture pairings across datasets for an 8 or 16-mixture UBM, the base dataset used the same mixtures to produce the closest-divergent and furthest-divergent decision surfaces. It was also able to confirm that low weighted closest-aligned mixtures were not artifacts but rather distinct phenomena related to a particular dataset, seizure or motion. Thus the UBMs were able to link the mixtures within a given feature space and the TVMs provided insight into how those mixtures were used in relation to each other for classification purposes.

These initial classifications were broad and would need to be refined, as the number of mixtures and dataset contents are expanded. Not all pairings labeled as "aligned" were truly aligned with a TVM distance score below 0.5, just as not all "divergent" labels were truly divergent with a distance score greater than 0.5. Thus the middle range (-0.5 to 0.5), wherein two dimensional vectors would be orthogonal to each other, represents a space potentially just as important as the fully aligned or divergent areas. To accurately map this space it likely would have been necessary to account for all label transitions, and not just those between the best and worst overlaps between the UBM and TVM. However, organizing a system to classify beyond just the outliers was beyond the scope of this work.

There are a number of ways to expand on this technique now that it has shown its ability to explain how decisions are made by the I-Vectors. If traced back to and evaluated against known annotations it should be possible to identify certain mixtures with a known annotation and to leverage this to then build groups of related mixtures under a given annotation label.

8.4 Going Forward

The three research aims were fulfilled and laid claim to a strong reference for future experiments concerning the tested feature sets, the tested algorithms, and the tested datasets. Linking all of these components together by using publicly available datasets provides a strong benchmark for further the development of algorithms and feature sets within the EEG processing community. Historically, subject verification experiments have been reliant on the PhysioNet Database which limited the possible types of experiments given its limited recording duration and low subject count. Now experiments that previously used that dataset have an alternative dataset against which to reevaluate their work and compare to the results presented in this work.

However, a hurdle remains in helping those researchers find data of interest for their studies. Too often, protocols have required that experiments gather their own data, which helped them with their specific research but failed to add to the collective pool of common experimental results. The NEDC's efforts to collect and publicly release EEG data was an attempt to address this failure. Yet this is only part of the problem, as like the PhysioNet Database dataset which lacked any information about its subjects, the data is often only as useful as the medical information associated with it. Data that cannot be organized on the highest of levels (subject, age, gender, condition) makes the process of understanding the components of those factors harder.

Thus, the impetus for this work stemmed from questions about how one would search and organize such a larger amount of EEG recordings. There was going to be a need for a tool that could map all of this data to itself while also being capable of placing new novel recordings into an appropriate place within the existing mapping. The cohort retrieval problem was thus born and I-Vectors were seen as a possible solution.

If that conversation were to happen again today, it could be said that I-Vectors are a potential solution given the results presented herein. They were competitive with an algorithm that used feature vectors nearly double their order of magnitude and afforded the opportunity to analyze the relationships being created that allowed such performance and dimension reduction. While this did not make them the superior classifier, it did make them the superior tool for research and analysis of EEG data without the need for clinically sourced annotations. The creation of such a tool, that previously did not exist within the EEG community, should lend itself to helping addressing questions about feature selection and dataset composition while providing a potential alternative source for generating informed annotations.

The use of I-Vectors outside of speech recognition has found success in text recognition [215], acoustic detection [216], and image recognition [217] and now EEG subject verification. This is likely to continue into other areas of multi-modal or multi-source signal process problems if not supplanted by the use of x-vectors. However, the fact that I-Vectors can operate successfully without labeled to produce the various UBM-TVM relationships and is less computational intensive may imbue it with some longevity until x-vectors mature further.

Depending on the level of abstract in the DNN statistics pooling it may be possible to recreate a similar analysis to the I-Vector analysis used to compare the TVM and UBM spaces. Yet this would likely be occurring on the subject level as the statistical transform of the TVM would be equivalent to the processes that generate the first embedding. Thus it may not be possible to constrain the parameters of the DNN in a such way replicate the TVM impulse response via the two embeddings. At the very least this would be an interesting area of research that melds the presented work with what is likely the future of speaker recognition.

BIBLIOGRAPHY

[1] O. N. Markand, “Pearls, Perils, and Pitfalls in the Use of the Electroencephalogram,” *Semin. Neurol.*, vol. 23, no. 1, pp. 007–046, 2003.

[2] T. W. Picton, “The P300 Wave of the Human Event-Related Potential,” *J. Clin. Neurophysiol.*, vol. 9, no. 4, pp. 456–479, oct 1992.

[3] P. Khanna *et al.*, “Modeling distinct sources of neural variability driving neuroprosthetic control,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 3068–3071, 2016.

[4] Lun-De Liao *et al.*, “Biosensor Technologies for Augmented Brain-Computer Interfaces in the Next Decades,” *Proc. IEEE*, vol. 100, no. SPL CONTENT, pp. 1553–1566, may 2012.

[5] B. J. Lance *et al.*, “Brain–Computer Interface Technologies in the Coming Decades,” *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1585–1599, may 2012.

[6] S. Ramgopal *et al.*, “Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy,” *Epilepsy Behav.*, vol. 37, pp. 291–307, 2014.

[7] S. Lopez *et al.*, “Automated identification of abnormal adult EEGs,” in *2015 IEEE Signal Process. Med. Biol. Symp.*, vol. 37, no. 6. IEEE, dec 2015, pp. 1–5.

[8] H. Nolan, R. Whelan, and R. B. Reilly, “FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection,” *J. Neurosci. Methods*, vol. 192, no. 1, pp. 152–162, sep 2010.

[9] E. Schulz *et al.*, “Decoding an individual’s sensitivity to pain from the multivariate analysis of EEG data,” *Cereb. Cortex*, vol. 22, no. 5, pp. 1118–1123, 2012.

[10] N. Kannathal, M. L. Choo, U. R. Acharya, and P. K. Sadasivan, “Entropies for detection of epilepsy in EEG,” *Comput. Methods Programs Biomed.*, vol. 80, no. 3, pp. 187–194, 2005.

[11] V. Lawhern, D. Slayback, D. Wu, and M. Kass, “Efficient Labeling of EEG Signal Artifacts Using Active Learning,” *Proc. - 2015 IEEE Int. Conf. Syst. Man, Cybern. SMC 2015*, pp. 3217–3222, 2016.

[12] F. Lotte and C. Guan, “Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms.” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–62, feb 2011.

[13] H. Chu, C. K. Chung, W. Jeong, and K.-H. Cho, "Predicting epileptic seizures from scalp EEG based on attractor state analysis," *Comput. Methods Programs Biomed.*, vol. 143, pp. 75–87, may 2017.

[14] D. F. Wulsin *et al.*, "Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement." *J. Neural Eng.*, vol. 8, no. 3, p. 036015, jun 2011.

[15] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Machine-Learning-Based Coadaptive Calibration for Brain-Computer Interfaces," *Neural Comput.*, vol. 23, no. 3, pp. 791–816, 2011.

[16] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep Feature Learning for EEG Recordings," *Arxiv*, pp. 1–24, 2015.

[17] A. J. Izenman, *Modern Multivariate Statistical Techniques*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2008.

[18] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus." *Front. Neurosci.*, vol. 10, no. MAY, p. 196, may 2016.

[19] P. W. Kaplan and S. R. Benbadis, "How to write an EEG report: Dos and don'ts," *Neurology*, vol. 80, no. Issue 1, Supplement 1, pp. S43–S46, jan 2013.

[20] K. M. Tsiouris *et al.*, "An unsupervised methodology for the detection of epileptic seizures in long-term EEG signals," in *2015 IEEE 15th Int. Conf. Bioinforma. Bioeng.* IEEE, nov 2015, pp. 1–4.

[21] A. C. Grant *et al.*, "EEG interpretation reliability and interpreter confidence: A large single-center study," *Epilepsy Behav.*, vol. 32, pp. 102–107, mar 2014.

[22] N. Gaspard *et al.*, "Interrater agreement for Critical Care EEG Terminology," *Epilepsia*, vol. 55, no. 9, pp. 1366–1373, sep 2014.

[23] J. Halford *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings," *Clin. Neurophysiol.*, vol. 126, no. 9, pp. 1661–1669, sep 2015.

[24] S. C. Warby *et al.*, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods." *Nat. Methods*, vol. 11, no. 4, pp. 385–92, 2014.

[25] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr, "Mixed-Band Wavelet-Chaos-Neural Network Methodology for Epilepsy and Epileptic Seizure Detection," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 9, pp. 1545–1551, sep 2007.

[26] J. J. Halford *et al.*, “Characteristics of EEG interpreters associated with higher interrater agreement.” *J. Clin. Neurophysiol.*, vol. 34, no. 2, pp. 168–173, 2017.

[27] C. M. Epstein, “Guideline 7: Guidelines for Writing EEG Reports,” *J. Clin. Neurophysiol.*, vol. 23, no. 2, pp. 118–121, apr 2006.

[28] T. Banaschewski and D. Brandeis, “Annotation: What electrical brain activity tells us about brain function that other techniques cannot tell us - A child psychiatric perspective,” *J. Child Psychol. Psychiatry Allied Discip.*, vol. 48, no. 5, pp. 415–435, 2007.

[29] E. Westhall *et al.*, “Interrater variability of EEG interpretation in comatose cardiac arrest patients,” *Clin. Neurophysiol.*, vol. 126, no. 12, pp. 2397–2404, dec 2015.

[30] K. Gwet, “Kappa Statistic is not satisfactory for assessing the extent of agreement between raters,” *Stat. Methods Inter-Rater Reliab. Assessmen*, no. 1, pp. 1–5, 2002.

[31] P. A. Gerber *et al.*, “Interobserver Agreement in the Interpretation of EEG Patterns in Critically Ill Adults,” *J. Clin. Neurophysiol.*, vol. 25, no. 5, pp. 241–249, oct 2008.

[32] Z. Z. Wang *et al.*, “Cross-subject workload classification with a hierarchical Bayes model,” *Neuroimage*, vol. 59, no. 1, pp. 64–69, jan 2012.

[33] D. La Rocca, P. Campisi, and G. Scarano, “EEG Biometrics for Individual Recognition in Resting State with Closed Eyes,” *Int. Conf. Biometrics Spec. Interes. Gr.*, no. Figure 1, pp. 1–12, 2012.

[34] S. Makeig *et al.*, “Evolving signal processing for brain-computer interfaces,” in *Proc. IEEE*, vol. 100, no. SPL CONTENT, aug 2012, pp. 1567–1584.

[35] T. Schluter and S. Conrad, “An Approach for Automatic Sleep Stage Scoring and Apnea-Hypopnea Detection,” in *2010 IEEE Int. Conf. Data Min.*, vol. 6, no. 2. IEEE, dec 2010, pp. 1007–1012.

[36] A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, “Age and sex effects in the EEG: Differences in two subtypes of attention-deficit/hyperactivity disorder,” *Clin. Neurophysiol.*, vol. 112, no. 5, pp. 815–26, may 2001.

[37] H. Begleiter and B. Porjesz, “Genetics of human brain oscillations,” *Int. J. Psychophysiol.*, vol. 60, no. 2, pp. 162–171, 2006.

[38] J. A. Urigüen and B. Garcia-Zapirain, “EEG artifact removal - State-of-the-art and guidelines,” *J. Neural Eng.*, vol. 12, no. 3, 2015.

[39] Q. Gui, Z. Jin, and W. Xu, "Exploring EEG-based biometrics for user identification and authentication," *2014 IEEE Signal Process. Med. Biol. Symp. IEEE SPMB 2014 - Proc.*, 2015.

[40] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[41] A. Harati *et al.*, "Improved EEG event classification using differential energy," in *2015 IEEE Signal Process. Med. Biol. Symp. - Proc.*, no. December 2015. IEEE, dec 2016, pp. 1–4.

[42] J. L. Marcano, M. A. Bell, and A. L. Beex, "Classification of ADHD and non-ADHD subjects using a universal background model," *Biomed. Signal Process. Control*, vol. 39, pp. 204–212, 2018.

[43] U. R. Acharya *et al.*, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, no. July 2017, pp. 270–278, 2018.

[44] T. Rakthanmanon *et al.*, "Searching and mining trillions of time series subsequences under dynamic time warping," *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 262–270, 2012.

[45] P. Campisi, D. La Rocca, and D. L. Rocca, "Brain waves for automatic biometric-based user recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 5, pp. 782–800, 2014.

[46] B. Porjesz *et al.*, "The utility of neurophysiological markers in the study of alcoholism," *Clin. Neurophysiol.*, vol. 116, no. 5, pp. 993–1018, 2005.

[47] J. A. Coan and J. J. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biol. Psychol.*, vol. 67, no. 1-2, pp. 7–49, 2004.

[48] M. Schultze-Kraft *et al.*, "Unsupervised classification of operator workload from brain signals." *J. Neural Eng.*, vol. 13, no. 3, p. 036008, jun 2016.

[49] A. B. Gardner *et al.*, "Human and automated detection of high-frequency oscillations in clinical intracranial EEG recordings," *Clin. Neurophysiol.*, vol. 118, no. 5, pp. 1134–1143, may 2007.

[50] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, 2007.

[51] D. F. Wulsin, S. Jensen, and B. Litt, "A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling," *Proc. 29th Int. Conf. Mach. Learn.*, pp. 57–64, 2012.

[52] J. G. Bogaarts *et al.*, "Optimal training dataset composition for SVM-based, age-independent, automated epileptic seizure detection," *Med. Biol. Eng. Comput.*, vol. 54, no. 8, pp. 1285–1293, aug 2016.

[53] J. A. Blanco *et al.*, "Data mining neocortical high-frequency oscillations in epilepsy and controls," *Brain*, vol. 134, no. 10, pp. 2948–2959, oct 2011.

[54] V. Bajaj and R. Pachori, "Classification of seizure and nonseizure EEG signals using empirical mode decomposition," *Inf. Technol. Biomed. . . .*, vol. 16, no. 6, pp. 1135–1142, 2012.

[55] W. O. Tatum and W. O. Tatum, IV, *Handbook of EEG Interpretation*, 2nd ed. New York: Demos Medical, 2014.

[56] J. Buckelmüller, H.-P. Landolt, H. H. Stassen, and P. Achermann, "Trait-like individual differences in the human sleep electroencephalogram," *Neuroscience*, vol. 138, pp. 351–356, 2006.

[57] S. L. Wendt *et al.*, "Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 4250–4253, 2012.

[58] J. Zygierewicz *et al.*, "High resolution study of sleep spindles." *Clin. Neurophysiol.*, vol. 110, no. 12, pp. 2136–2147, 1999.

[59] M. Del Pozo-Banos, J. B. Alonso, J. R. Ticay-Rivas, and C. M. Travieso, "Electroencephalogram subject identification: A review," *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6537–6554, 2014.

[60] P. Tangkraingkij, C. Lursinsap, S. Sanguansintukul, and T. Desudchit, "Personal identification by EEG using ICA and neural network," *Comput. Sci. Its Appl. 2010*, pp. 419–430, 2010.

[61] H. H. Stassen, D. T. Lykken, P. Propping, and G. Bomben, "Genetic determination of the human EEG. Survey of recent results on twins reared together and apart." *Hum. Genet.*, vol. 80, no. 2, pp. 165–76, 1988.

[62] M. Doppelmayr, W. Klimesch, T. Pachinger, and B. Ripper, "Individual differences in brain dynamics: important implications for the calculation of event-related band power." *Biol. Cybern.*, vol. 79, no. 1, pp. 49–57, 1998.

[63] C. E. M. Van Beijsterveldt and G. C. M. Van Baal, “Twin and family studies of the human electroencephalogram: A review and a meta-analysis,” *Biol. Psychol.*, vol. 61, no. 1-2, pp. 111–138, 2002.

[64] D. La Rocca *et al.*, “Human brain distinctiveness based on EEG spectral coherence connectivity,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 9, pp. 2406–2412, 2014.

[65] D. La Rocca, P. Campisi, and J. Sole-Casals, “EEG based user recognition using BUMP modelling,” *Biometrics Spec. Interes. Gr. (BIOSIG), 2013 Int. Conf.*, pp. 1–12, 2013.

[66] K. Brigham and B. V. K. V. Kumar, “Subject identification from electroencephalogram (EEG) signals during imagined speech,” in *2010 Fourth IEEE Int. Conf. Biometrics Theory, Appl. Syst.* IEEE, sep 2010, pp. 1–8.

[67] L. De Gennaro *et al.*, “The electroencephalographic fingerprint of sleep is genetically determined: A twin study,” *Ann. Neurol.*, vol. 64, no. 4, pp. 455–460, 2008.

[68] M. Fraschini *et al.*, “An EEG-based biometric system using eigenvector centrality in resting state brain networks,” *IEEE Signal Process. Lett.*, vol. 22, no. 6, pp. 666–670, 2015.

[69] M. Näpflin, M. Wildi, and J. Sarnthein, “Test-retest reliability of resting EEG spectra validates a statistical signature of persons,” *Clin. Neurophysiol.*, vol. 118, no. 11, pp. 2519–2524, 2007.

[70] A. B. Ajiboye *et al.*, “Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration,” *Lancet*, vol. 389, no. 10081, pp. 1821–1830, 2017.

[71] C. Gouy-Pailler *et al.*, “Nonstationary Brain Source Separation for Multiclass Motor Imagery,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 2, pp. 469–478, feb 2010.

[72] P.-J. Kindermans, M. Tangermann, K.-R. Müller, and B. Schrauwen, “Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller,” *J. Neural Eng.*, vol. 11, no. 3, p. 035005, jun 2014.

[73] B. Blankertz *et al.*, “Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing,” *Adv. Neural Inf. Process. Syst.*, pp. 1–8, 2007.

[74] F. Lotte and C. Guan, "Learning from other subjects helps reducing Brain-Computer Interface calibration time," in *2010 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, no. 2. IEEE, 2010, pp. 614–617.

[75] P.-J. Kindermans *et al.*, "True zero-training brain-computer interfacing–an online study." *PLoS One*, vol. 9, no. 7, p. e102504, jul 2014.

[76] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. 6, pp. 510–523, 1988.

[77] N. Karamzadeh *et al.*, "Capturing dynamic patterns of task-based functional connectivity with EEG," *Neuroimage*, vol. 66, pp. 311–317, 2013.

[78] J. Jeong, "EEG dynamics in patients with Alzheimer's disease," *Clin. Neurophysiol.*, vol. 115, no. 7, pp. 1490–1505, 2004.

[79] E. Ba←(s) ar and B. Güntekin, "A review of brain oscillations in cognitive disorders and the role of neurotransmitters," *Brain Res.*, vol. 1235, pp. 172–193, 2008.

[80] S. J. Lupien *et al.*, "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition," *Brain Cogn.*, vol. 65, no. 3, pp. 209–237, 2007.

[81] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, 2016.

[82] B. Güntekin, E. Ba, and E. Ba←(s) ar, "Review of evoked and event-related delta responses in the human brain," *Int. J. Psychophysiol.*, vol. 103, pp. 43–52, 2016.

[83] C. Vidaurre *et al.*, "Toward unsupervised adaptation of LDA for brain-computer interfaces." *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 587–97, mar 2011.

[84] R. B. Paranjape, J. Mahovsky, L. Benedicenti, and Z. Koles, "THE ELECTROENCEPHALOGRAM AS A BIOMETRIC," in *Can. Conf. Electr. Comput. Eng.*, vol. 2, 2001, pp. 1363–1366.

[85] R. Palaniappan and D. P. Mandic, "Biometrics from Brain Electrical Activity: A Machine Learning Approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 738–742, 2007.

[86] S. Yang, F. Deravi, and S. Hoque, "Task sensitivity in EEG biometric recognition," *Pattern Anal. Appl.*, pp. 1–13, 2016.

[87] R. Mahajan and B. I. Morshed, "Unsupervised eye blink artifact denoising of EEG data with modified multiscale sample entropy, Kurtosis, and wavelet-ICA." *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 1, pp. 158–65, jan 2015.

[88] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomed. Signal Process. Control*, vol. 31, pp. 407–418, 2017.

[89] K. min Su, W. D. Hairston, and K. Robbins, "EEG-Annotate: Automated identification and labeling of events in continuous signals with applications to EEG," *J. Neurosci. Methods*, vol. 293, pp. 359–374, 2018.

[90] J. Gross, "Analytical methods and experimental approaches for electrophysiological studies of brain oscillations," *J. Neurosci. Methods*, vol. 228, pp. 57–66, may 2014.

[91] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, jun 2007.

[92] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[93] H. Wang and C.-s. Choy, "Automatic seizure detection using correlation integral with nonlinear adaptive denoising and Kalman filter," in *2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, aug 2016, pp. 1002–1005.

[94] G. Schalk *et al.*, "BCI2000: a general-purpose brain-computer interface (BCI) system." *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–43, jun 2004.

[95] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Networks*, vol. 57, pp. 39–50, sep 2014.

[96] A. Page *et al.*, "A Flexible Multichannel EEG Feature Extractor and Classifier for Seizure Detection," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 62, no. 2, pp. 109–113, feb 2015.

[97] B. C. Armstrong *et al.*, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," *Neurocomputing*, vol. 166, pp. 59–67, 2015.

[98] E. Maiorana, D. La Rocca, and P. Campisi, "On the Permanence of EEG Signals for Biometric Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 1, pp. 163–175, 2016.

[99] R. Bódizs, J. Körmendi, P. Rigó, and A. S. Lázár, "The individual adjustment method of sleep spindle analysis: Methodological improvements and roots in the fingerprint paradigm," *J. Neurosci. Methods*, vol. 178, no. 1, pp. 205–213, 2009.

[100] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." *Circulation*, vol. 101, no. 23, pp. E215–20, jun 2000.

[101] B. Blankertz *et al.*, "The BCI Competition III: Validating Alternative Approaches to Actual BCI Problems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, jun 2006.

[102] M. Radha, G. Garcia-Molina, M. Poel, and G. Tononi, "Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2014, pp. 1876–1880, 2014.

[103] S. Marcel, J. D. R. Millán, and J. d. R. Millan, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 743–748, 2007.

[104] D. Rodrigues *et al.*, "EEG-based person identification through Binary Flower Pollination Algorithm," *Expert Syst. Appl.*, vol. 62, pp. 81–90, nov 2016.

[105] M. Delpozo-Banos, C. M. Travieso, C. T. Weidemann, and J. B. Alonso, "EEG biometric identification: A thorough exploration of the time-frequency domain," *J. Neural Eng.*, vol. 12, no. 5, 2015.

[106] C. Vidaurre and B. Blankertz, "Towards a Cure for BCI Illiteracy," *Brain Topogr.*, vol. 23, no. 2, pp. 194–198, jun 2010.

[107] M. Spezialetti, L. Cinque, J. M. R. S. Tavares, and G. Placidi, "Towards EEG-based BCI driven by emotions for addressing BCI-Illiteracy: a meta-analytic review," *Behav. Inf. Technol.*, vol. 37, no. 8, pp. 855–871, aug 2018.

[108] J. Martinez-del Rincon *et al.*, "Non-linear classifiers applied to EEG analysis for epilepsy seizure detection," *Expert Syst. Appl.*, vol. 86, pp. 99–112, 2017.

[109] J. A. Blanco *et al.*, "Unsupervised Classification of High-Frequency Oscillations in Human Neocortical Epilepsy and Control Patients." *J. Neurophysiol.*, vol. 104, no. 5, pp. jn.01 082.2009–, 2010.

[110] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *J. Neural Eng.*, vol. 15, no. 3, p. 031005, 2018.

[111] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, 2017.

[112] K. A. I. Aboalayon, H. T. Ocbagabir, and M. Faezipour, "Efficient sleep stage classification based on EEG signals," *2014 IEEE Long Isl. Syst. Appl. Technol. Conf. LISAT 2014*, pp. 1–6, 2014.

[113] Shijian Lu *et al.*, "Unsupervised Brain Computer Interface Based on Intersubject Information and Online Adaptation," *Neural Syst. Rehabil. Eng. IEEE Trans.*, vol. 17, no. 2, pp. 135–145, apr 2009.

[114] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition," *Neural Networks*, vol. 41, no. 1995, pp. 188–201, 2013.

[115] M. K. Abdullah, K. S. Subari, J. L. C. Loong, and N. N. Ahmad, "Analysis of effective channel placement for an EEG-based biometric system," *Proc. 2010 IEEE EMBS Conf. Biomed. Eng. Sci. IECBES 2010*, no. December, pp. 303–306, 2010.

[116] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," pp. 257–286, 1989.

[117] P. Kenny *et al.*, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 5, pp. 980–988, jul 2008.

[118] H. Behravan *et al.*, "I-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition," *IEEE/ACM Trans. Speech Lang. Process.*, vol. 24, no. 1, pp. 29–41, 2016.

[119] D. A. Reynolds, "Gaussian Mixture Models," *Encycl. Biometrics*, no. 2, pp. 659–663, 2009.

[120] T. Hasan and J. H. L. Hansen, "A Study on Universal Background Model Training in Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 1890–1899, sep 2011.

[121] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, no. August, 2011, pp. 857–860.

[122] P. Kenny, T. Stafylakis, J. Alam, and M. Kockmann, "An I-vector backend for speaker verification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, 2015, pp. 2307–2311.

[123] H. Li, B. Ma, and K.-A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, may 2013.

[124] M. H. Bahari, M. McLaren, H. Van Hamme, and D. A. Van Leeuwen, "Speaker age estimation using i-vectors," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, sep 2012, pp. 506–509.

[125] N. Kasabov and E. Capecci, "Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes," *Inf. Sci. (Ny).*, vol. 294, pp. 565–575, feb 2015.

[126] M. H. Silber *et al.*, "The visual scoring of sleep in adults," *J. Clin. Sleep Med.*, vol. 3, no. 2, pp. 121–131, 2007.

[127] S. K. Loo and S. L. Smalley, "Preliminary report of familial clustering of EEG measures in ADHD," *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, vol. 147, no. 1, pp. 107–109, 2008.

[128] S. J. Segalowitz, D. L. Santesso, and M. K. Jetha, "Electrophysiological changes during adolescence: A review," *Brain Cogn.*, vol. 72, no. 1, pp. 86–100, 2010.

[129] S. Fazli, M. Danóczy, J. Schelldorfer, and K.-R. Müller, "L1-penalized linear mixed-effects models for high dimensional data with application to BCI," *Neuroimage*, vol. 56, no. 4, pp. 2100–2108, 2011.

[130] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems." *Neuroimage*, vol. 34, no. 4, pp. 1600–11, feb 2007.

[131] S.-Y. Cheng and H.-T. Hsu, "Mental Fatigue Measurement Using EEG," in *Risk Manag. Trends.* InTech, jul 2011.

[132] S. Dähne *et al.*, "SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters," *Neuroimage*, vol. 86, pp. 111–122, feb 2014.

[133] K. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 3rd ed. Maryland, USA: Advanced Analytics, LLC, 2012.

[134] A. Page, J. Turner, T. Mohsenin, and T. Oates, "Comparing Raw Data and Feature Extraction for Seizure Detection with Deep Learning Methods," *Twenty-Seventh Int. . . .*, pp. 284–287, 2014.

[135] V. Gandhi *et al.*, “Quantum neural network-based EEG filtering for a brain-computer interface.” *IEEE Trans. neural networks Learn. Syst.*, vol. 25, no. 2, pp. 278–88, feb 2014.

[136] B. Blankertz *et al.*, “Optimizing Spatial filters for Robust EEG Single-Trial Analysis,” *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[137] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, aug 1980.

[138] C. Guger *et al.*, “How many people are able to control a P300-based brain-computer interface (BCI)?” *Neurosci. Lett.*, vol. 462, no. 1, pp. 94–98, 2009.

[139] V. J. Monastra, J. F. Lubar, and M. Linden, “The development of a quantitative electroencephalographic scanning process for attention deficit–hyperactivity disorder: Reliability and validity studies.” *Neuropsychology*, vol. 15, no. 1, pp. 136–144, 2001.

[140] M. Scheffer *et al.*, “Early-warning signals for critical transitions.” *Nature*, vol. 461, no. September, pp. 53–59, 2009.

[141] N. Martin *et al.*, “Topography of age-related changes in sleep spindles,” *Neurobiol. Aging*, vol. 34, no. 2, pp. 468–476, 2013.

[142] F. Ferrarelli *et al.*, “Reduced sleep spindle activity in schizophrenia patients,” *Am. J. Psychiatry*, vol. 164, no. 3, pp. 483–492, 2007.

[143] E. J. Wamsley *et al.*, “Reduced sleep spindles and spindle coherence in schizophrenia: Mechanisms of impaired memory consolidation?” *Biol. Psychiatry*, vol. 71, no. 2, pp. 154–161, 2012.

[144] M. Mölle, L. Marshall, S. Gais, and J. Born, “Grouping of spindle activity during slow oscillations in human non-rapid eye movement sleep.” *J. Neurosci.*, vol. 22, no. 24, pp. 10 941–10 947, 2002.

[145] C. Huang *et al.*, “Discrimination of Alzheimer’s disease and mild cognitive impairment by equivalent EEG sources: a cross-sectional and longitudinal study,” *Clin. Neurophysiol.*, vol. 111, no. 11, pp. 1961–1967, nov 2000.

[146] A. Lenartowicz and S. K. Loo, “Use of EEG to Diagnose ADHD,” *Curr. Psychiatry Rep.*, vol. 16, no. 11, 2014.

[147] I. Buyck and J. R. Wiersema, “Resting electroencephalogram in attention deficit hyperactivity disorder: developmental course and diagnostic value.” *Psychiatry Res.*, vol. 216, no. 3, pp. 391–7, may 2014.

[148] S. K. Loo and S. Makeig, “Clinical Utility of EEG in Attention-Deficit/Hyperactivity Disorder: A Research Update,” *Neurotherapeutics*, vol. 9, no. 3, pp. 569–587, jul 2012.

[149] T. P. Jung *et al.*, “Removing electroencephalographic artifacts by blind source separation.” *Psychophysiology*, vol. 37, no. 2, pp. 163–78, mar 2000.

[150] A. Delorme, T. J. Sejnowski, and S. Makeig, “Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis,” *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[151] P. J. Kindermans *et al.*, “True zero-training brain-computer interfacing - An online study,” *PLoS One*, vol. 9, no. 7, 2014.

[152] U. R. Acharya *et al.*, “Automated diagnosis of epileptic EEG using entropies,” *Biomed. Signal Process. Control*, vol. 7, no. 4, pp. 401–408, jul 2012.

[153] A. Subasi, “Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients,” *Expert Syst. Appl.*, vol. 28, no. 4, pp. 701–711, 2005.

[154] N. Huang *et al.*, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” *Proc. R. Soc. Lond. A*, vol. 454, pp. 903–995, 1998.

[155] S. M. Pincus, “Approximate entropy as a measure of system complexity.” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 6, pp. 2297–301, mar 1991.

[156] J. S. Richman and J. R. Moorman, “Physiological time-series analysis using approximate entropy and sample entropy.” *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, no. 6, pp. H2039–49, jun 2000.

[157] C. L. Nikias and A. P. Petropulu, *Higher-order spectra analysis. a nonlinear signal processing framework.* Englewood Cliffs, New Jersey: PTR Prentice Hall, 1993.

[158] A. J. Gabor, R. R. Leach, and F. U. Dowla, “Automated seizure detection using a self-organizing neural network,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 99, no. 3, pp. 257–266, 1996.

[159] T. Kohonen, “The self-organizing map,” *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[160] H. Van Dis *et al.*, “Individual differences in the human electroencephalogram during quiet wakefulness,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 47, no. 1, pp. 87–94, 1979.

[161] H. H. Stassen, “Computerized recognition of persons by EEG spectral patterns,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 49, no. 1-2, pp. 190–194, 1980.

[162] M. V. Ruiz-blondet, Z. Jin, and S. Laszlo, “CEREBRE: A Novel Method for Very High Accuracy Event-Related Potential Biometric Identification,” *IEEE Trans. Inf. Forensics Secur.*, vol. 6013, no. c, pp. 1–13, jul 2016.

[163] P. Nguyen *et al.*, “EEG-Based person verification using Multi-Sphere SVDD and UBM,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7818 LNAI, no. PART 1, pp. 289–300, 2013.

[164] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and Session Variability in GMM-Based Speaker Verification,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, may 2007.

[165] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1-3, pp. 19–41, jan 2000.

[166] N. Dehak *et al.*, “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, may 2011.

[167] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Commun.*, vol. 52, no. 1, pp. 12–40, jan 2010.

[168] O. Glembek *et al.*, “Simplification and optimization of i-vector extraction,” in *2011 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2011, pp. 4516–4519.

[169] R. McClanahan and P. L. De Leon, “Reducing computation in an i-vector speaker recognition system using a tree-structured universal background model,” *Speech Commun.*, vol. 66, no. 1, pp. 36–46, 2015.

[170] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of I-vector Length Normalization in Speaker Recognition Systems,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 249–252, 2011.

[171] C. S. C. Greenberg *et al.*, “The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge,” *Proc. Speak. Lang. Recognit. Work.*, no. June, pp. 224–230, 2014.

[172] C.-H. Lee *et al.*, “Improved acoustic modeling for large vocabulary continuous speech recognition,” *Comput. Speech Lang.*, vol. 6, no. 2, pp. 103–127, apr 1992.

[173] A. Hyvärinen, J. Karhunen, and E. Oja, “Independent Component Analysis,” *Analysis*, vol. 26, no. 1, p. 481, 2001.

[174] H. Behravan, V. Hautamäki, and T. Kinnunen, “Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish,” *Speech Commun.*, vol. 66, pp. 118–129, feb 2015.

[175] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms,” *CRIM, Montr. CRIM-06/08-13*, pp. 1–17, 2005.

[176] M. Senoussaoui *et al.*, “An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech,” in *Odyssey Speak. Lang. Recognit. Work.*, 2010.

[177] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1-3, pp. 37–52, aug 1987.

[178] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[179] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[180] S. Cumani and P. Laface, “e-vectors: JFA and i-vectors revisited,” in *2017 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, mar 2017, pp. 5435–5439.

[181] S. J. Gershman and D. M. Blei, “A tutorial on Bayesian nonparametric models,” *J. Math. Psychol.*, vol. 56, no. 1, pp. 1–12, 2012.

[182] Y. W. Y. Y. Teh, M. I. M. M. I. Jordan, M. J. M. Beal, and D. M. Blei, “Hierarchical Dirichlet Processes,” *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, dec 2006.

[183] J. Mueller and A. Thyagarajan, “Siamese Recurrent Architectures for Learning Sentence Similarity,” in *Proc. 30th Conf. Artif. Intell. (AAAI 2016)*, no. 2012, 2016, pp. 2786–2792.

[184] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Interspeech 2017*, vol. 52, no. 2. ISCA: ISCA, aug 2017, pp. 999–1003.

[185] D. Snyder *et al.*, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5329–5333, 2018.

[186] B. Reuderink, J. Farquhar, M. Poel, and A. Nijholt, “A subject-independent brain-computer interface based on smoothed, second-order baselining,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 4600–4604, 2011.

[187] K. Su and K. A. Robbins, “A Framework for Content-based Retrieval of EEG with Applications to Neuroscience and Beyond.” *Proc. ... Int. Jt. Conf. Neural Networks. Int. Jt. Conf. Neural Networks*, pp. 1–8, 2013.

[188] A. M. Dymond, R. W. Coger, and E. A. Serafetinides, “Preprocessing by factor analysis of centro-occipital EEG power and asymmetry from three subject groups,” *Ann. Biomed. Eng.*, vol. 6, no. 2, pp. 108–116, 1978.

[189] C. Berka *et al.*, “EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks.” *Aviat. Space. Environ. Med.*, vol. 78, no. 5 Suppl, pp. B231–44, may 2007.

[190] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” *IEEE Trans. Acoust.*, vol. 29, no. 2, pp. 254–272, 1981.

[191] H. Behravan, V. Hautamäki, and T. Kinnunen, “Foreign Accent Detection from Spoken Finnish Using i-Vectors,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 79–83, 2013.

[192] M. McLaren and D. van Leeuwen, “Improved speaker recognition when using i-vectors from multiple speech sources,” *2011 IEEE Int. Conf. Acoust. Speech Signal Process.*, no. 1, pp. 5460–5463, may 2011.

[193] C. Ward and I. Obeid, “Application of identity vectors for EEG classification,” *J. Neurosci. Methods*, vol. 311, pp. 338–350, jan 2019.

[194] E. Khoury, L. E. Shafey, and S. Marcel, “Spear: An open source toolbox for speaker recognition based on Bob,” in *2014 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2014, pp. 1655–1659.

[195] D. Povey *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Work. Autom. speech Recognit. Underst.*, Cambridge, 2011.

[196] S. O. Sadjadi, M. Slaney, and L. Heck, “MSR identity toolbox v1. 0: a matlab toolbox for speaker-recognition research,” *Speech Lang. Process. Tech. Comm. Newsl.*, no. 1, pp. 4–7, 2013.

[197] N. Dehak *et al.*, “Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 1559–1562, 2009.

[198] O. Glembek *et al.*, “Comparison of scoring methods used in speaker recognition with joint factor analysis,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 4057–4060, 2009.

[199] N. V. Thakor and S. Tong, “Advances in quantitative EEG analysis methods,” *Annu. Rev. Biomed. Eng.*, vol. 6, no. 1, pp. 453–495, 2004.

[200] C. Ward and I. Obeid, "Feasibility of Identity Vectors for use as subject verification and cohort retrieval of electroencephalograms," in *2016 IEEE Signal Process. Med. Biol. Symp.*, vol. 1. IEEE, dec 2016, pp. 1–5.

[201] G. Saon, T. Sercu, S. Rennie, and H.-k. K. J. Kuo, "The IBM 2016 English conversational telephone speech recognition system," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 7–11, sep 2016.

[202] C. R. Ward, J. Picone, and I. Obeid, "Applications of UBMs and I-vectors in EEG subject verification," in *2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, aug 2016, pp. 748–751.

[203] M. McLaren and D. Van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors from Multiple Speech Sources," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 3, pp. 755–766, 2012.

[204] D. F. Wulsin and B. Litt, "An unsupervised method for identifying regions that initiate seizures on intracranial EEG," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3091–3094, 2011.

[205] M. Ahn and S. C. Jun, "Performance variation in motor imagery brain-computer interface: A brief review," *J. Neurosci. Methods*, vol. 243, pp. 103–110, 2015.

[206] M. Schröder *et al.*, "Robust EEG Channel Selection across Subjects for Brain-Computer Interfaces," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 19, pp. 3103–3112, 2005.

[207] S. Safavi, M. Russell, and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," *Comput. Speech Lang.*, vol. 50, pp. 141–156, jul 2018.

[208] E. Khoury, L. E. Shafey, and M. Ferras, "Hierarchical speaker clustering methods for the NIST i-vector Challenge," *Proc. Odyssey 2014 - Speak. Lang. Recognit. Work.*, no. 1, pp. 254–259, 2014.

[209] W. Rao, M.-W. Mak, and K.-A. Lee, "Normalization of total variability matrix for i-vector/PLDA speaker verification," in *2015 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, apr 2015, pp. 4180–4184.

[210] S. E. Shepstone *et al.*, "Total variability modeling using source-specific priors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 504–517, 2016.

[211] M. Li and S. Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Comput. Speech Lang.*, vol. 28, no. 4, pp. 940–958, jul 2014.

[212] A. Sizov *et al.*, "Joint Speaker Verification and Antispoofing in the <inline-formula> <tex-math notation="LaTeX">i </tex-math></inline-formula>-Vector Space," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 821–832, apr 2015.

[213] S. Seshadri, U. Remes, and O. Räsänen, "DIRICHLET PROCESS MIXTURE MODELS FOR CLUSTERING I-VECTOR DATA," in *2017 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2017, pp. 5470–5474.

[214] W.-Y. Y. Hsu, "Fuzzy Hopfield neural network clustering for single-trial motor imagery EEG classification," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 1055–1061, jan 2012.

[215] F. Troncoso-Pastoriza, C. García-Mateo, and M. Fairhurst, "Introducing an approach for writer recognition based on the i -vector paradigm," *IET Biometrics*, vol. 6, no. 3, pp. 191–199, 2016.

[216] B. Elizalde, H. Lei, and G. Friedland, "An i-Vector representation of acoustic environments for audio-based video event detection on user generated content," *Proc. - 2013 IEEE Int. Symp. Multimedia, ISM 2013*, pp. 114–117, 2013.

[217] R. Wallace and M. McLaren, "Total variability modelling for face verification," *IET Biometrics*, vol. 1, no. 4, pp. 188–199, 2012.

[218] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529310," *Society*, vol. 33, no. 1, pp. 159–174, 2008.

APPENDIX A

EQUATIONS

Relevant topics discussed in each sections that may not be familiar to all readers are briefly outlined here.

A.1 Background

A.1.1 Defining Similarity via Cohen's Kappa

It is difficult to produce annotated sets of EEG recordings without clinical support. To ensure the accuracy of these sets it is necessary to have multiple clinicians annotate the same data to build a consensus-based annotation. This process invites each clinician's bias into the annotation process which must be tracked and controlled in terms of intra-rater and inter-rater similarity scores. These scores provide a sense of strength of a clinician's ability and robustness of a dataset as a function of *agreement* evaluated as Cohen's Kappa (κ).

Given two raters and their tallies for class A or B in Table **??**, their inter-rater agreement κ is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{A.1}$$

Table A.1. Table of Cohen's Kappa

		S1	
		A	B
S2	A	q	w
	B	z	x

$$\begin{aligned} p_o &= \frac{q+x}{q+w+z+x} \\ p_e &= \frac{q+w}{q+w+z+x} * \frac{q+z}{q+w+z+x} + \frac{z+x}{q+w+z+x} * \frac{w+x}{q+w+z+x} \end{aligned} \tag{A.2}$$

In the above equation, p_o finds the percentage of agreement between the two raters[1]. Then p_e finds the percentage the raters chose the same label, how often S1 chose A and S2 chose A. The calculated expectation of similarity, p_e, is used to control for the outcome of similarity, p_o. The grades of agreement are quantified as follows: $\{ < 0$ poor; $0 - 0.20$ slight; $0.21 - 0.40$ fair; $0.41 - 0.60$ moderate; $0.61 - 0.80$ substantial; $0.81 - 1.00$ almost perfect$\}$ [218].

[1] In the event the two raters are the same clinician, the agreement represents intra-rater agreement instead of inter-rater agreement.