ADAPTIVE LEARNING OF NEURAL ACTIVITY DURING DEEP BRAIN STIMULATION

by

Arindam Dutta

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2015 by the
Graduate Supervisory Committee:

Antonia Papandreou-Suppappola, Chair
Keith Holbert
Daniel W. Bliss

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

Parkinsons disease is a neurodegenerative condition diagnosed on patients with clinical history and motor signs of tremor, rigidity and bradykinesia, and the estimated number of patients living with Parkinsons disease around the world is seven to ten million. Deep brain stimulation (DBS) provides substantial relief of the motor signs of Parkinsons disease patients. It is an advanced surgical technique that is used when drug therapy is no longer sufficient for Parkinsons disease patients. DBS alleviates the motor symptoms of Parkinsons disease by targeting the subthalamic nucleus using high-frequency electrical stimulation.

This work proposes a behavior recognition model for patients with Parkinson's disease. In particular, an adaptive learning method is proposed to classify behavioral tasks of Parkinsons disease patients using local field potential and electrocorticography signals that are collected during DBS implantation surgeries. Unique patterns exhibited between these signals in a matched feature space would lead to distinction between motor and language behavioral tasks. Unique features are first extracted from deep brain signals in the time-frequency space using the matching pursuit decomposition algorithm. The Dirichlet process Gaussian mixture model uses the extracted features to cluster the different behavioral signal patterns, without training or any *prior* information. The performance of the method is then compared with other machine learning methods and the advantages of each method is discussed under different conditions.

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank a few people who's knowledge and assistance have been invaluable to me. First and foremost, I would like to extend my gratitude to my advisor Professor Antonia for mentoring me and providing invaluable insight at every stage of my research.

I would also like to thank Professors Keith Holbert and Dan Bliss for taking the time to be members of my committee, and sharing their knowledge and expertise with me.

I would also like to thank my lab members especially Dr Narayan Kovvoli who helped me whenever I got stuck with something.

Last but not the least, I would like to thank my friends and family, especially my parents, for their constant love and support.

This project was a collaborative work between Arizona State University and Colorado Neurological Institute (CNI). The data with which this work has been done was provided my CNI. This work has been accepted and published as a conference paper by the Asilomar Conference on Signals, Systems, and Computers, 2014.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Table Page

Chapter 1

INTRODUCTION

Parkinson's disease (PD) which is also known as idiopathic or primary Parkinsonism, hypokinetic rigid syndrome, or paralysis agitans, is a degenerative disorder of the central nervous system [2]. It is recognized on the basis of clinical history and motor signs of tremor, rigidity and bradykinesia [3]. The death of dopamine-generating cells in the Substantia Nigra, a region of the mid-brain, causes the motor symptoms of Parkinson's disease. The reason of this cell death is unknown. As mentioned, the most perceptible symptoms in the early progression of the disease are movement-related; these include shaking, rigidity, and slowness of movement.

1.1 Motivation and Problem formulation

A large number of people suffer from PD in the United States and new patients with PD are diagnosed each year [4]. When drug therapy is no longer efficient for patients with PD, a state-of-the-art surgical technique named Deep Brain Stimulation (DBS) is applied which provides relief form the motor symptoms of PD [3] [5]. The current DBS therapy method is an open-loop experiment that has proven to be effective for treatment of PD as well as essential tremor which is another neurological disorder that causes a rhythmic shaking. By open loop, we mean that a unidirectional signal is generated from the device and delivered to the brain. However, the open-loop therapy DBS has various shortcomings as discussed in [5]. The design and advancement of a closed-loop implantable pulse generator (IPG) to sense and respond to physiologic signals within or outside the brain is considered the next frontier in brain stimulation research. By close-loop, we mean that bidirectional signals are sending

and responding in both directions, thus enabling feedback to the simulation process. As a consequence, advancements and accomplishments in diagnosing the behavior of PD patients by analyzing and classifying the electrical signals of their brain with and without DBS will aid the development of the next generation closed-loop DBS system, which is the major goal of this study.

1.2 Background

Several studies have been published on the classification of certain behaviors shown by patients with PD while performing certain tasks. Most of these studies were performed by processing of electroencephalography (EEG) recordings taken from the patients and very few using Local Field Potential (LFP) signals. Most classification methods used in the studies were based on integrating a feature extraction algorithm with a supervised classifier. In [6] different emotions such as happiness and sadness were classified using EEF signals by first filtering the signals into an optimal frequency band, using common spatial patterns as features, and the linear support vector machine (SVM) classifier. It was found that the gamma band (30-100 Hz) is suitable for emotion classification from EEG signals. In [7], emotional states in PD patients were compared to healthy controls using machine learning algorithms, taking into account the fact that PD patients are characterized by emotional deficits. The study involved the recording of EEG signals of PD patients and healthy controls while evoking emotions using multimodal stimulus (audio-visual aids), and the dynamic change in emotional state classified using different features. Four different types of EEG features were considered: bispectrum feature, power spectrum feature, wavelet feature and features extracted from non linear dynamical analysis. The study showed that the best classification result was obtained using the bispectrum feature and that higher frequency bands (alpha, beta and gamma) are more important in emotional

activities than lower frequency bands (delta and theta). The paper also reveals that the path of emotion changes can be visualized by reducing subject-independent features with manifold learning. In [8] the same data was used to distinguish emotional states such as happiness, sadness, fear, anger, surprise and disgust. In this case, features such as absolute and relative power, frequency and asymmetry measures were subjected to repeated ANOVA (a three-way repeated analysis of variance) in order to compare different groups and to discriminate their functionality as feature candidates in classification algorithm.

Of the few studies performed using LFP signals to classify patient behavior, the work in [9] presented a method to enable a single trial behavioral task recognition for random behavior, speech and motor. The approach was based on using wavelet coefficients as features and the SVM classifier.

In the aforementioned studies on the classification of PD patients behavioral tasks, most of the signals used are EEG signals, the features most commonly used are either time-based or frequency-based, and the classifiers are in general based on supervised classification methods. We thus want to consider LFP signals, which are collected using an electrode inside the brain, use more localized features and classifiers that do not require any training.

1.3 Thesis Contribution

As the number of people suffering from PD is increasing, the need of a better DBS design, and in particular a closed-loop DBS system design, has become necessary. For such a design, feedback signals in form of LFP signals need to be processed and features need to be extracted from these signals that will provide the best matched information on the patient behavior.

The use of time-frequency representations (TFRs) to analyze signals provides ac-

curate information about signals that vary both in time and frequency; TFRs have also been used to separate the time-varying signals from noise or interfering signals. Signals like LFP, EEG or electrocorticography (ECOG) time-varying since their frequency content changes with time. As a result, in this work, we use, time-frequency features to represent LFP signals as they provide unique patterns for classification.

The extracted feature vectors are used as input to the Dirichlet process Gaussian mixture model (DP-GMM) for unsupervised clustering. The approach allows for an unlimited number of mixture components. This number is learned adaptively from the provided features and does not need to be known as *a priori*. As a result, using DP-GMM, we do not require to train the data for classification.

In our work, we consider four different behavior tasks: simple motor task, language task, language with motor task and language without motor task. We use the matching pursuit decomposition algorithm to extract informative time-frequency features from the LFP signals corresponding to these for tasks. Clustering is then performed using DP-GMM. More specifically, we perform clustering between: (a) simple motor and language with motor (b) simple motor and language without motor (c) language with motor and language without motor

(d) simple motor and language with motor and language without motor

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 provides a background on DBS and explains the need for a closed loop DBS system for patients with PD. It also provides a background on LFP signals and how they are collected. Chapter 3 explains the MPD algorithm and how the LFP feature vectors are extracted using a Gaussian dictionary. Chapter 4 discuss the DP-GMM and its implementation using blocked Gibbs sampler, and it provides our proposed integrated MPD feature and DP-GMM classifier for the

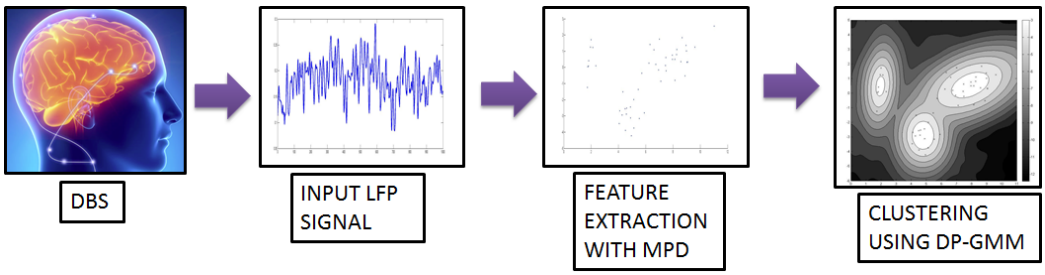


Figure 1.1: Block diagram summarizing the proposed method of adaptive clustering of LFP signals.

behavior tasks. Chapter 5 provides the simulation results and compares our approach with other methods. Finally, in Chapter 6, conclusions and extensions to future work are discussed.

A block diagram representing the main contribution of this work is shown in Figure 1.1. Also, the acronyms used throughout this thesis are summarized in Table 1.1

Table 1.1: Alphabetical List of Acronyms used in this Dissertation

Acronym	**Definition**
DBS	Deep Brain Stimulation
DHMM	Discrete Hidden Markov Model
DP	Dirichlet Process
DP-GMM	Dirichlet Process Gaussian Mixture Model
ECOG	Electrocorticography
EEG	Electroencephalography
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LFP	Local Field Potential
MCMC	Markov Chain Monte Carlo
MPD	Matching Pursuit Decomposition
PD	Parkinson's Disease
PDF	Probability Distribution Function
SVM	Support Vector Machine
IPG	Implantable Pulse Generator

Chapter 2

DEEP BRAIN STIMULATION AND LOCAL FIELD POTENTIAL

Parkinson's disease (PD) is a progressive neurological condition, resulting from the degeneration of neurons that produce dopamine in the substantia nigra located at the lower pat of the brain [3]. It effects functional activities like writing, typing, walking, speech and many other routine activities. Although the early treatments of managing the motor symptoms of this disease are effective, drugs eventually become ineffective as the disease progresses. When drugs no longer help PD patients, deep brain stimulation (DBS) treatment can be used to alleviate motor symptoms.

2.1 Deep Brain Stimulation

Deep brain stimulation is a surgical procedure used to treat several disabling neurological symptoms, and most commonly the debilitating motor symptoms of PD, such as tremor, rigidity, stiffness, slowed movement, and walking problems. Over the last two decades, the clinical success of DBS has contributed to a rapid expansion of DBS into a wide range of neurological disorders. In 1997 the first commercial DBS system was approved for the treatment of tremor [10]. DBS provides a train of stimulatory pulses of certain frequencies to the brain. So far an open loop DBS therapy has been used as an effective treatment of PD and essential tremor. An open loop DBS system basically involves a one-way flow of the signal generated by the DBS system to the brain. However, this open loop system has been shown to have some side effects like impaired cognition, speech and balance.

Figure 2.1 shows a DBS device with thin coated wires (leads) that transmit the electrical energy to the targeted portion of the brain, mostly to the subthalamic

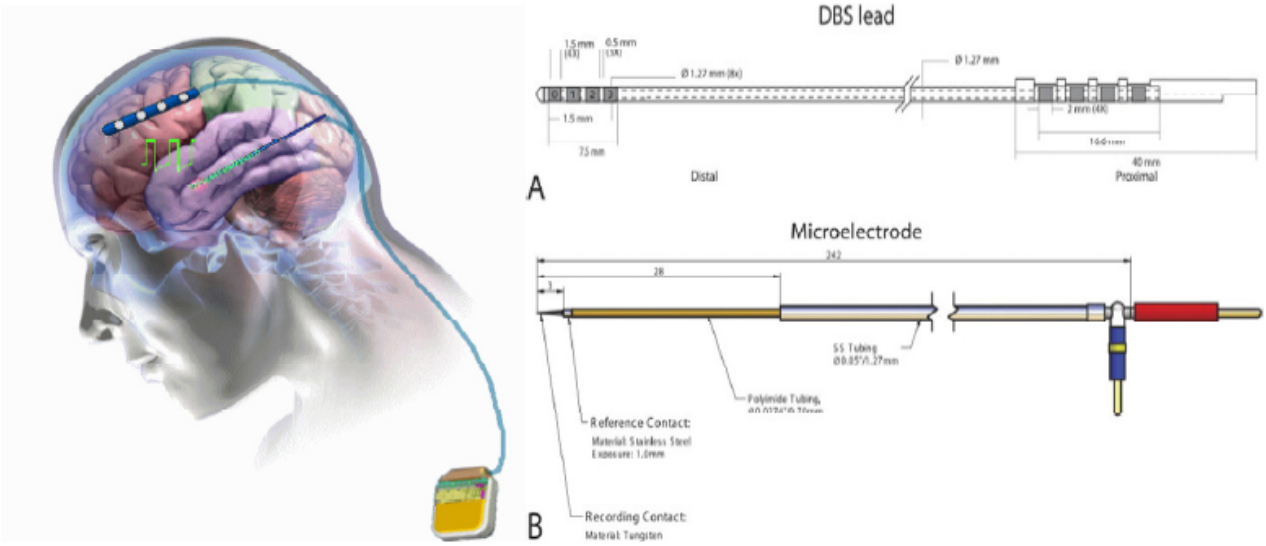


Figure 2.1: Electrodes to Record LFP Signals from the Subthalamic Nucleus.

nucleus for PD. The invasive microelectrodes record the LFP signals that reflect the oscillatory activity within the nuclei of the basal ganglia. Figure 2.2 shows the neurotransmitter which includes the computer chip that determines the waveform and electric impulses that are sent to the brain. The computer chip is individually programmable to *fine tune* the system to the patient.

The design of a closed loop implantable pulse generator (IPG) to sense and respond to physiologic signals within or outside the brain is considered to be the next big thing in brain stimulation research and will likely broaden the field to include new applications for neuromodulation. A closed-loop system involves bidirectional signals moving in both sensing and responding directions, allowing sensor signals to provide feedback based on the stimulation. The goal of the closed-loop DBS in PD is to restore the functionality of the targeted part of the brain. The DBS produces LFP signals which are sent to a specific area of the brain, based on the motor task. Knowledge

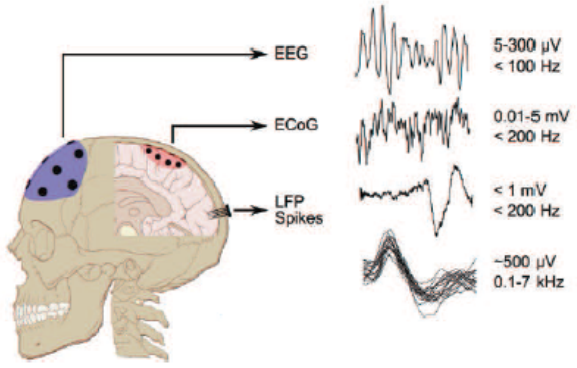
Figure 2.2: Neurostimulator that contains Battery and Micro-Electronic Circuitry [1]

of the LFP features of the PD patient while the patient is performing a specific behavior task under normal conditions (no tremors or other motor symptoms), then the DBS can be used to restore the LFP signals to the patient while the patient is having severe tremors. As a result, it is very important to be able to correctly cluster different behavior tasks under different conditions as this will be a stepping stone toward the success of a closed loop DBS system.

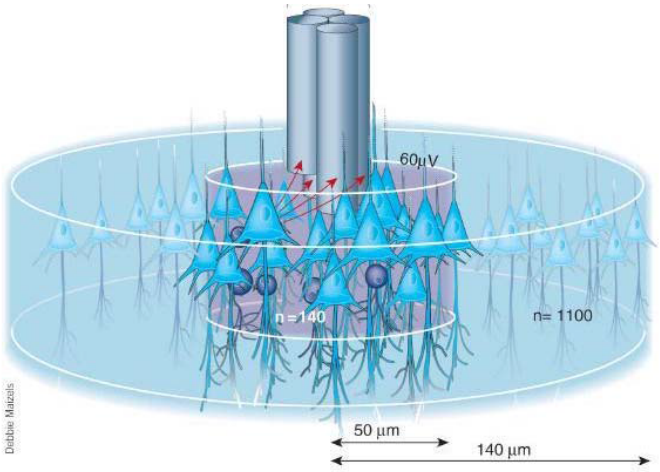
2.2 Local Field Potential

Electrical events at deeper locations in the brain which can be recorded by inserting metal or glass electrodes, or silicon probes into the brain are called LFP (also known as micro-EEG). Figure 2.3b shows an LFP signal while being recorded with a microelectrode. The LFP signal is the most informative brain signal as it contains action potentials and other membrane potentials-derived fluctuations in a small neuron volume [11]. The LFP differs from normal EEG or ECOG signals, and it ranges between 51,000 μV with frequency less than 200 Hz (see Figure 2.3a)

The LFP signals used to assess the performance of our proposed methods were obtained from a study involving twelve patients undergoing DBS implantation for



(a) Different Brain Signals: EEG, ECoG (macroscopic), LFP, Action Potentials or Spikes (microscopic).



(b) LFP Recordings with Microelectrodes [12]

Figure 2.3: Different types of brain signals

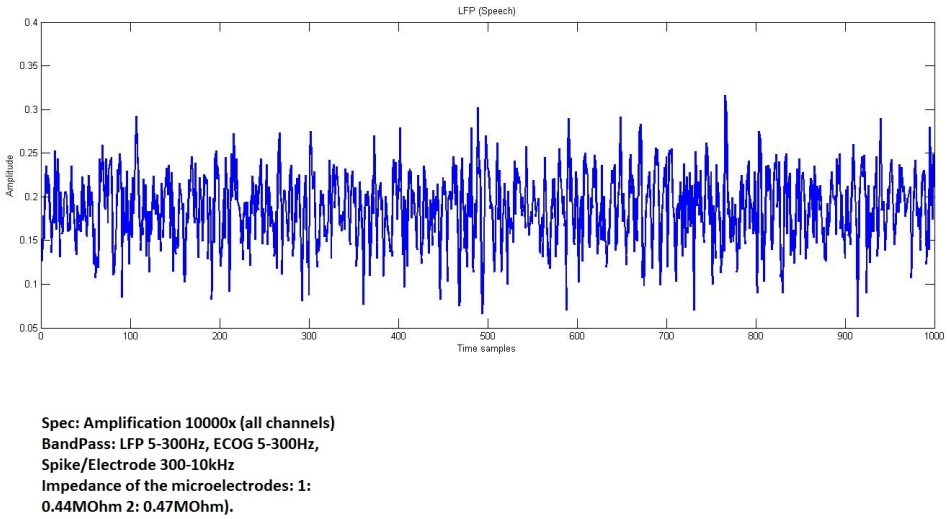


Figure 2.4: LFP of a PD Patient while performing a language task

treatment of idiopathic PD [3]. The signals were simultaneous LFP signals recorded during behavioral tasks (see Figure 1b). The tasks described four types of behaviors: simple motor task, language task, language with motor task, and language without motor task.

For the DBS lead recording design, recordings were obtained from each of the four contacts of the DBS lead (Medtronic 3389, see Figure 2.1). Although primarily designed for stimulation, these electrodes have been used for LFP recording in humans, as they do not require modification of standard surgical practice. The DBS lead contact is platinum/iridium, has a surface area of 6.0 mm^2 and impedance of 1.7 $k\Omega$. Signals were amplified,sampled using a sampling frequency of 5kHz, and combined with event markers and subject response signals.

A typical LFP signal taken from one of the subjects with Idiopathic PD is shown in Figure 2.4.

Chapter 3

MATCHING PURSUIT DECOMPOSITION AND FEATURES EXTRACTION

3.1 Matching Pursuit Decomposition Algorithm

As loacl field potentials (LFPs) are signals whose frequency content changes with time we apply the matching pursuit decomposition (MPD) algorithm to extract time-frequency based features [13], [14]. Using the MPD, a signal is decomposed into a linear expansion of Gaussian basis functions that are selected from a redundant basic dictionary. Each dictionary element is a Gaussian signal that is a time-shifted, frequency-shifted and scaled version of a basic Gaussian signal at the time-frequency origin. The feature obtained from each extracted dictionary element is a four-dimensional (4-D) vector consisting of the extraction weight coefficient, time shift, frequency shift and scale change parameters.

The MPD is a well known technique for sparse signal representations. It is a greedy algorithm that expands a signal into a linear approximation of basis functions by iteratively projecting the signal over a redundant, possibly non-orthogonal set of signals called dictionary. Since it is a greedy algorithm, the approximation may be sub-optimal. The dictionary functions are iteratively selected to best match the signal structure, resulting in a sub-optimal expansion. The main steps of the MPD algorithm are shown in Appendix A, and are discussed next in detail.

In general any basis function can be used as a dictionary to decompose the required signal, can be shown that the only signal that achieves the lower bound is the Gaussian signal. However, the Gaussian signal is most often selected since it attains Heisenberg's uncertainty principle [ref]. According to this principle, a signal cannot

simultaneously achieve high temporal resolution and high frequency resolution; the signal that achieves the best trade-off in both time and frequency resolution is the Gaussian signal. This can be shown by computing the time-bandwidth product of a Gaussian signal

$$x(t) = e^{-bt^2} \implies T_x F_x = \frac{1}{4\pi} \tag{3.1}$$

where T_x and F_x are the duration and bandwidth of the Gaussian signal, respectively. For all other signals it can be shown that $T_x F_x > (1/4\pi)$. The Gaussian dictionary element g_γ(t), $\gamma = 1, \ldots, \Gamma$ is given by

$$g_\gamma(t) = \kappa e^{2\pi\kappa(t-tau)^2} e^{j2\pi t\nu} \tag{3.2}$$

This forms a dictionary $\mathcal{D}$ with Γ independent Gaussian waveforms. The MPD algorithm begins by projecting the signal on each dictionary signal $g_\gamma(t)$ and computing the residue after every iteration. After P iterations, the MPD results in a linear weighted expansion of P selected Gaussian elements, $g_p(t)$ and their corresponding weight coefficients α_p, $p = 1, \ldots, P$. This is given by

$$x(t) = \sum_{p=1}^{P} \alpha_p g_p(t) + r_P(t) \tag{3.3}$$

where $r_P(t)$ is the residual signal after P iterations. The iterations start by setting $r_0(t) = x(t)$; at the Pth iteration, the best matched Gaussian signal $g_p(t)$ is selected that results in the maximum correlation with the remainder signal $r_p(t)$. In particular,

$$g_p(t) = \underset{g_\gamma(t)\in\mathcal{D}\gamma=1,\ldots,\Gamma}{\arg\max} \left| \int r_{p-1}(t) g_\gamma^*(t) dt \right| \tag{3.4}$$

The weight coefficient is then obtained using

$$\alpha_p = |\int r_{p-1}(t) g_p^*(t) dt| = |\langle r_{p-1}(t),\ g_p(t)\rangle \tag{3.5}$$

With this choice $r_{p-1}(t)$ is projected onto $g_p(t)$ and decomposed as follows:

$$r_{p-1}(t) = r_p(t) + \alpha_p g_p(t), \tag{3.6}$$

From Equation (3.6), we can see that the decomposition of $x(t)$ is given by (3.3) where α_p is given by (3.5). It can be shown that $r_P(t)$ converges exponentially to 0 when p tends to ∞

$$\lim_{p\to\infty} \|r_P(t)\| = 0; \tag{3.7}$$

Hence

$$x(t) = \sum_{k=0}^{\infty} |\langle r_k(t),\ g_k(t)\rangle| g_k(t), \tag{3.8}$$

and

$$|x(t)|^2 = \sum_{k=0}^{\infty} |\langle r_k(t), g_k(t)\rangle|^2, \tag{3.9}$$

Thus, the original vector $x(t)$ is decomposed into a sum of the dictionary signals that matches best the signal and its residuals at each iterations. It can also be seen that the decomposition preserves the signal energy assymptotically.

In this case the algorithm was repeated for 15 iterations on LFP signals of length 1000 samples each. The dictionary used was of the size 1000×40000. Fig. 3.1 shows the raw LFP signal and also the sum of 15 Gaussian waves from the dictionary that matches the signal. The MPD algorithm is shown in Appendix A.

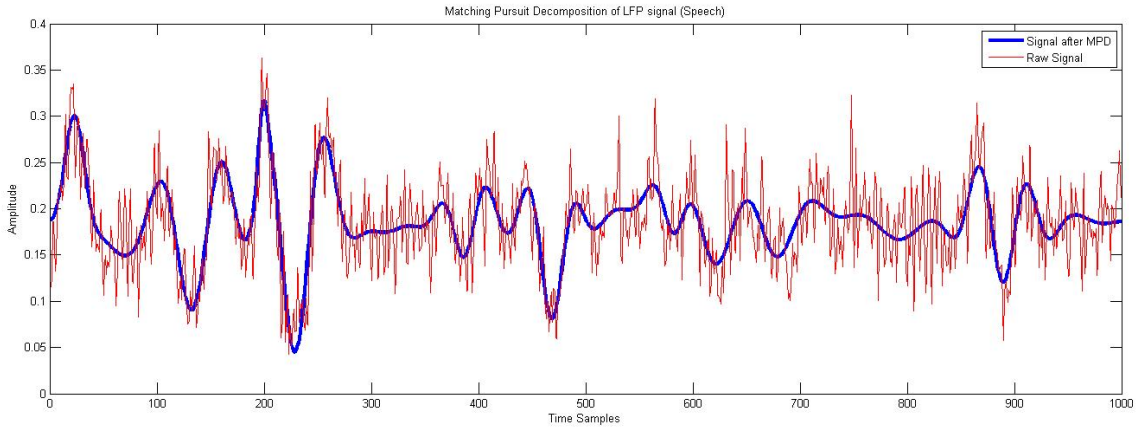


Figure 3.1: Language task LFP signal after 15 Iterations of MPD

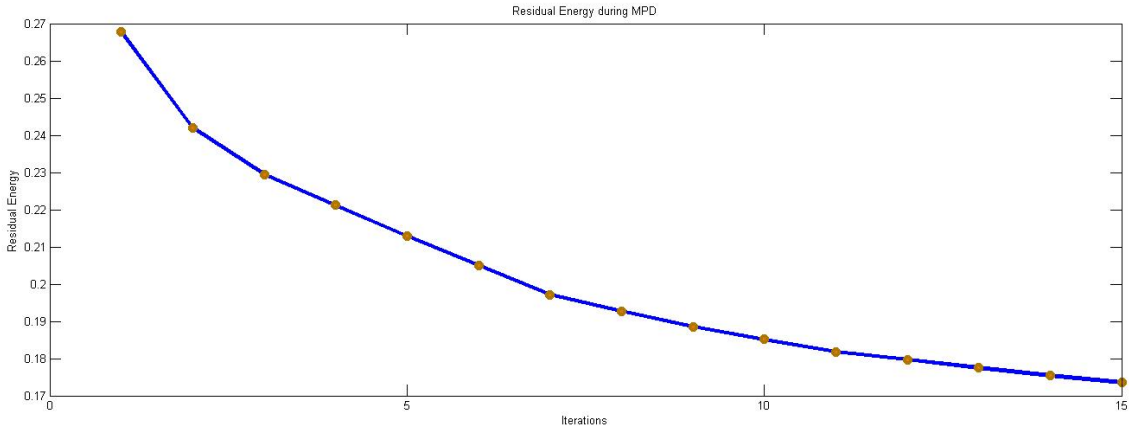


Figure 3.2: Residual energy at each MPD iterations

3.2 Matching Pursuit Decomposition of LFP signals

We applied the MPD algorithm to the LFP signals of behavior tasks that were collected from PD patients during DBS implantation surgeries. We had LFP signals from J experiments and $M = 4$ behavior tasks. We denote the LFP signal vector from the jth experiment, $j = 1, \ldots, J$, corresponding to the mth task, $m = 1, 2, 3, 4$, as $\mathbf{s}_i^m$. After discretization, the signal has N samples, thus the signal vector is given by $\mathbf{s}_j^m = [s_j^m[0]\, s_j^m[1]\, \ldots\, s_j^m[N-1]]^T$, where T denotes vector transpose. Task $m = 1$ task

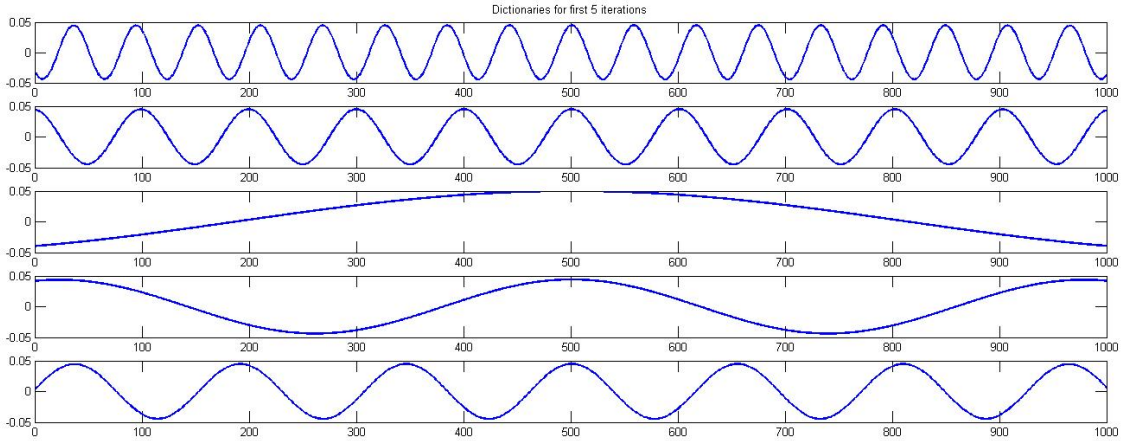


Figure 3.3: Gaussian waveforms for the first 5 MPD iterations

corresponds to the simple motor task, Task $m = 2$ is the language task, Task $m = 3$ is the the language with motor task, and Task $m = 4$ is the language without motor task. The MPD is applied to the mth behavior task signal from the jth experiment, $\mathbf{s}_j^m$, in Appendix A. If we assume that the MPD performed P iterations, then the extracted feature matrix corresponding to $\mathbf{s}_j^m$ is given by the $4 \times P$ matrix $\mathbf{F}_j^m$ whose pth column is given by $[\mathbf{F}_j^m]_p = [\alpha_p \, \tau_p \, \nu_p \, \kappa_p]^T$, $p = 1, \ldots, P$. Here, α_p is the MPD weight coefficient, τ_p is the time shift parameter, ν_p is the frequency shift parameter, and κ_p is the scaling parameter.

Using the actual experimental LFP signals, we run the MPD using $P = 15$ iterations, and the length of the sampled LFP signals was $N = 1,000$ samples. We used an MPD dictionary with $\Gamma = 100,040,000$ Gaussian signal atoms. Figure 3.1 shows an example of an LFP signal superimposed with its MPD linear expansion signal after 15 iterations. Table 3.1 shows the resulting feature vectors $[\mathbf{F}_j^m]_p$, $p = 1, \ldots, 15$ (for all 15 MPD iterations). For this same example, the normalized residual energy is shown to be decreasing monotonically at each iteration in Figure 3.2 and the first 5 extracted Gaussian signals are shown in time in Figure 3.3.

Table 3.1: MPD feature vector of the LFP signal in Figure 3.1

iterations	**amplitude** (α)	**time-shift** (τ)	**frequency-shift** (ν)	**scale change** (σ)
1	5.853	1	1	533
2	-0.695	144	173	272
3	0.466	118	142	405
4	-0.370	148	0	612
5	-0.364	153	198	33
6	0.349	73	12	20
7	0.342	1	64	785
8	-0.251	127	0	367
9	0.243	136	149	849
10	0.214	168	0	478
11	-0.211	157	0	114
12	-0.170	161	0	919
13	0.168	201	0	302
14	-0.168	1	45	840
15	-0.151	1	165	771

The normalized residual energy with each iterations is shown in Figure. 3.2. It can be seen that the energy is decreasing monotonously. Figure. 3.3 shows the gaussian waveforms that matches the signal for the first 5 iterations.

Using the extracted features, these feature vectors are set at the input to DP-GMM and the best feature sets are evaluated based on successful classification of the behavior tasks.

Chapter 4

CLUSTERING BEHAVIOR TASKS USING MPD FEATURES AND DIRICHLET PROCESS GAUSSIAN MIXTURE MODELLING

4.1 Integrated Clustering Algorithm Using MPD Features

As discussed in Chapter 3, use use the matching pursuit decomposition (MPD) algorithm to obtain time-frequency based feature vectors for the local field potential (LFP) behavior task signals. The feature vectors are then used as input to the Dirichlet process Gaussian mixture model (DP-GMM). A Gaussian mixture model (GMM) is a probabilistic model that assumes that all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The DP-GMM can be thought of as a GMM with a variable number of components modeled using the Dirichlet process. Dirichlet processes are often used in nonparametric Bayesian statistics, where the number of statistical representations can grow as more data are observed. They are thus specifically useful for unsupervised learning and clustering applications. As a result, the DP-GMM allows for an unlimited number of mixture components, the actual number of clusters does not need to be known a priori [15, 16, 17, 18].

To model a given set of data to DP-GMM and get clusters from it we first need to assume a prolific probabilistic model [19]. Let us consider the following joint distribution over the data, in this case the feature vectors $X = [\alpha\ \tau\ \nu\ \kappa]^T$ which we obtain from the local field potential (LFP)

$$p(X, Z, \mu, \Sigma, \pi) = p(X|Z, \mu, \Sigma)p(Z|\pi)p(\pi|\alpha)p(\mu, \Sigma|\lambda) \tag{4.1}$$

where Z is the latent variable which are correspondent variables between clusters and data points X, $\theta_k = \{\mu, \Sigma\}$ are the parameters of the Gaussian distributions that fit the data X and is sampled from a prior distribution which we assume here to be a Normal-Wishart distribution with parameter λ, π is the parameter that specifies the latent variable Z which is sampled from a Dirichlet distribution of parameter α, $p(X|Z, \mu, \Sigma)$ is the data likelihood probability distribution which is thought of as Gaussian in nature. $p(Z|\pi)$ is the correspondence probability which is a multinomial distribution that specifies the latent variable Z and $p(\pi|\alpha)$ is a mixture prior probability which follows a Dirichlet distribution. Taking the Dirichlet distribution as the conjugate prior of any multinomial distribution, if we multiply them together we always get a Dirichlet distribution, so we can compute the statistics in a closed form. Lastly $p(\mu, \Sigma|\lambda)$ is the parameter prior probability where Normal-Wishart distribution is chosen, for it is the conjugate prior of the Gaussian distribution.

In this problem we consider the same model as (4.1),but we let the latent variable Z to be defined as $\bar{\theta}_i$ sampled from

$$p(\bar{\theta}_i|\pi, \theta_k) = \sum_{k=1}^{K} \pi_k \delta(\theta_k, \bar{\theta}_i) \tag{4.2}$$

where δ is the Kronecker delta. The problem can be realized from the diagram given in Figure 4.1 and is explained more clearly in 4.4. So we can say the following about the parameters we need to estimate

$$\begin{aligned}
&\pi \sim \mathrm{Dir}(\alpha) \\
&\theta_k \sim H(\lambda) \\
&\bar{\theta}_i \sim G(\pi, \theta_k) \\
&G(\pi, \theta_k) = \sum_{k=1}^{K} \pi_k \delta(\theta_k, \bar{\theta}_i), K \to \infty
\end{aligned}$$

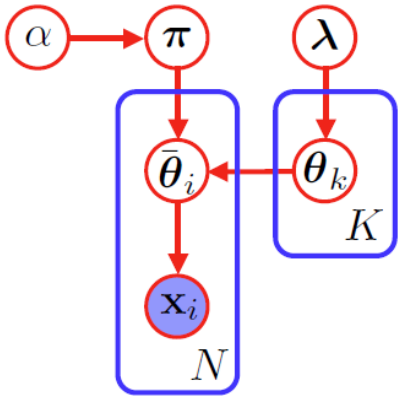


Figure 4.1: DP-GMM Block Diagram

Before describing the entire algorithm some of the definitions need to be explained.

4.2 Conjugate Prior

The concept, as well as the term *conjugate prior*, was introduced by Howard Raiffa and Robert Schlaifer in their work on Bayesian decision theory [20]. In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. Consider the general problem of inferring a distribution for a parameter θ given some datum or data x. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function og θ, $p(x|\theta)$ and prior $p(\theta)$, normalized by the probability of data $p(x)$:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')d\theta'} \tag{4.3}$$

Let the likelihood function be considered fixed; the likelihood function is usually well-determined from the data-generating process. It is clear that different choices of the prior distribution $p(\theta)$ may make the integral more or less difficult to calculate,

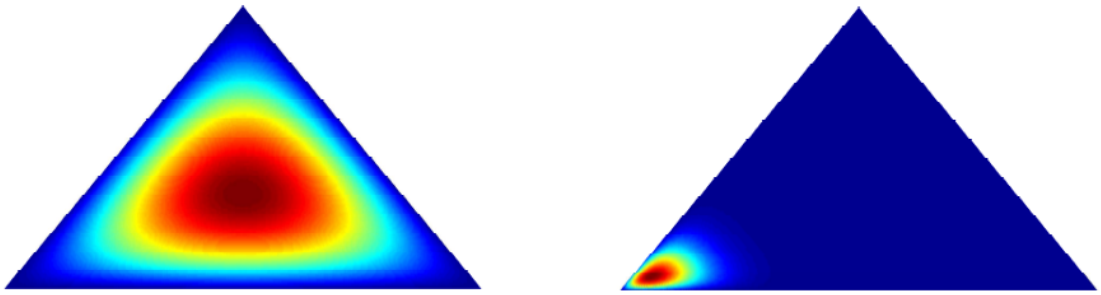


Figure 4.2: Dirichlet Distribution for $k = 3$ and (a) $\alpha = (2, 2, 2)$ (b) $\alpha = (20, 2, 2)$

and the product $p(x|\theta) \times p(\theta)$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is a conjugate prior. The conjugate priors of some distributions which are relevant to this work is given in the Table 4.1.

We also come across the term hyperparameters, which are PDF parameters that have their own prior distributions and can be estimated using Markov chain Monte Carlo (MCMC) methods [20]. Additional information on the Dirichlet-Multinomial and Normal-Wishart conjugate prior are provided in Appendices C, D respectively.

4.3 Dirichlet Process and Distribution

A DP is described as a distribution over probability measures G, $G(\theta) \geq 0$ and $\int G(\theta)d\theta = 1$, in other words it is a distribution over distributions [21]. If for any partition $(T_1, ..., T_k)$ it holds:

$$(G(T_1), ..., G(T_k)) \sim Dir(\alpha H(T_1), ..., \alpha H(T_k)) \tag{4.4}$$

then G is sampled from a Dirichlet process

It is shown as $G \sim DP(\alpha, H)$ where α is the concentration parameter and H is the base distribution [22].

Table 4.1: Table of Conjugate Distributions

Likelihood	**Model parameters**	**Conjugate prior distribution**	**Prior hyperparameters**	**Posterior hyperparameters**	**Interpretation of hyperparameters**	**Posterior predictive**
Categorical (Discrete)	p (probability vector), k (number of categories, i.e. size of p)	Dirichlet	α	$\alpha + (c_1, ..., c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i	$p(\tilde{x} = i) = \frac{\alpha'_i}{\sum_i \alpha'_i} = \frac{\alpha_i + c_i}{\sum_i \alpha'_i + n}$
Multinomial (Discrete)	p (probability vector), k (number of categories, i.e. size of p)	Dirichlet	α	$\alpha + \sum_{i=1}^{n} x_i$	$\alpha_i - 1$ occurrences of category i	DirMult($\tilde{x}$ $\|\alpha'$)(Dirichlet-Multinomial)
Multivariate normal (continuous)	μ (mean vector) and λ (precision matrix)	Normal-Wishart	$\mu_0, \kappa_0, \nu_0,$ $\mathbf{V}$	*	**	$t_{\nu'_0 - p + 1}$ $(\tilde{x}\|\mu'_0,$ $\frac{\kappa'_0 + 1}{\kappa'_0(\nu'_0 - p + 1)}$ $\mathbf{V}'^{-1})$

In Table 4.1, a star implies that

* mean was estimated from κ_0 observations with sample mean μ_0; covariance matrix was estimated from ν_0 observations with sample mean μ_0 and with sum of pairwise deviation products $\mathbf{V}^{-1}$

**

$$\frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}, \kappa_0 + n, \nu_0 + n,$$
$$(\mathbf{V}^{-1} + \mathbf{C} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T)^{-1}$$
$\bar{x}$ is the sample mean
$$\mathbf{C} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

The Dirichlet distribution is defined as:

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k - 1},\ \alpha_0 = \sum_{k=1}^{K}\alpha_k$$
$$0 \leq \mu_k \leq 1,\ \sum_{k=1}^{K}\mu_k = 1$$

It is the conjugate prior for the multinomial distribution. The parameters can be interpreted as the effective number of observations for every state. The parameter α_0 controls the strength of the distribution and α_k control the location of the peaks.

Every sample from a Dirichlet distribution is a vector of K positive values that sum up to 1, which means that the sample itself is a finite distribution. Accordingly, a sample from a Dirichlet process is an infinite discrete distribution.

4.4 Construction of the Dirichlet Process in this Problem

We have our feature vector X as defined in the beginning of this chapter, which we are going to feed as an in input to DP-GMM, and find the number of clusters

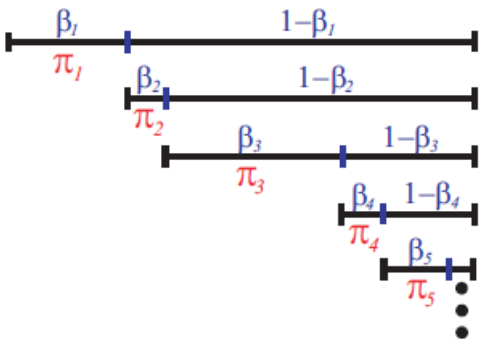


Figure 4.3: Stick Breaking Construction

based on these features. Let us assume that it is mixture of multivariate Gaussians with unknown $\theta_k = \{\mu_k, \Sigma_k\}$. So we sample θ_k from its conjugate prior which is a Normal-Wishart distribution $H(\lambda)$ with parameter λ as shown in Table 4.1.

Next we have to construct the weight parameter π_k from the Dirichlet distribution. The weight vector as defined before belongs to a discrete multinomial discrete distribution, so we sample it from its conjugate pair which is the Dirichlet distribution $Dir(\alpha)$ as shown in Table 4.1. The DP can be constructed using the "Stick Breaking" analogy [23]. Let us imagine a stick of length 1, we select a random number β between 0 and 1 from a Beta-distribution (univariate marginal and conditional distributions being beta). Then we break the $\pi = \beta$ length of the stick, save it and repeat this infinite times as shown in Figure. 4.3. But practically in this case we set a truncation limit to M based on the truncation error [14] [24] which is give by,

$$4N \exp(\frac{-(M-1)}{\alpha}) \tag{4.5}$$

So we have,

$$\beta_k \sim \text{Beta}(1, \alpha)$$
$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l) = \beta_k(1 - \sum_{l=1}^{k-1}\pi_l)$$

Now we derive $\bar{\theta_i}$ from π_k and θ_k, given as

$$G(\bar{\theta_i}) = \sum_{k=1}^{\infty} \pi_k \delta(\bar{\theta_i} - \theta_k) \tag{4.6}$$

The size of each successive break is representative of $\pi_k = p(\bar{\theta_i} - \theta_k)$

The weight vector is constructed using the Chinese Restaurant analogy [25], which states that every time a new $\bar{\theta_i}$ comes in, its probability to enter the weights, which is more filled is more that the weights which are less [26] [27] [28] [29]. It can shown that the probability for a new $\bar{\theta_i}$ is

$$p(\bar{\theta}_{N+1} = \theta | \bar{\theta}_{1:N}, \alpha, H) = \frac{1}{\alpha + N}\left(\alpha H(\theta) + \sum_{k=1}^{K} N_k \delta(\bar{\theta}_k, \theta)\right) \tag{4.7}$$

4.5 Blocked Gibbs Sampler

Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution (i.e., from the joint probability distribution of two or more random variables), when direct sampling is difficult. This sequence can be used to approximate the joint distribution (e.g., to generate a histogram of the distribution); to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables); or to compute an integral (such as the expected value of one of the variables).

The conjugate prior relationship is used extensively to simplify calculations for posterior distributions estimated using the blocked Gibbs sampler algorithm. Considering the mixture model discussed before and using the same notation the blocked Gibbs sampler, at the ith iteration in the Markov chain estimates [24] [30] :

$$\theta_k^j \sim p(\theta_k|c^{j-1}, x_n), k = 1, ..., M$$
$$c_n^j \sim p(c_n|\theta_i^j, \pi^{j-1}, x_n), n = 1, ..., N$$
$$\pi_k^j \sim p(\pi_k|c^j), k = 1, ..., M$$

These can be expressed in terms of conjugate prior relationships:

$$p(\theta_k|c, x_n) \propto H(\theta_k) \prod_{n:c_n=k} p(x_n|\theta), k = 1, ..., M$$
$$p(c_n|\theta_i, w, x_n) = \sum_{k=1}^{M} (\pi_k p(x_n|\theta_k))\delta(c_n - k), n = 1, ..., N$$
$$p(\pi_k|c) = \beta_k \prod_{j=1}^{k} -1(1 - \beta_j), k = 1, ..., M$$

where β_k is defines as

$$\beta_k = Beta(1 + N_k^*, \alpha + \sum_{k'=k+1}^{M} N_k^*)$$

and $n : c_n = k$ denotes the indices in c such that $c_n = k$ and N_k^* is the number of elements in c that are equal to k. The DP-GMM algorithm with blocked Gibbs Sampling is shown in Appendix B

The clustering results from DP-GMM are shown in the next chapter.

Chapter 5

SIMULATION AND RESULTS

We use the local field potential (LFP) signals collected from twelve Parkinson's disease (PD) patients to demonstrate our clustering methods. The signal segments associated with different behavioral tasks were labeled by physicians during data collection. The behavioral tasks are: simple motor task ($m = 1$), language with motor task ($m = 3$), and language without motor task ($m = 4$). The language tasks ($m = 2$) combines tasks 3 and 4. The signal sampling rate is 4 kHz, and for different behavioral tasks, the number of data segments varied from 80 to 109. For DP-GMM clustering, first MPD features were extracted from signals from each behavioral class. The MPD was run for 15 iterations for each signal set which gave a 300×4 feature matrix. The best clustering results for two classes were obtained using the amplitude and time-shift MPD parameters as the feature set, $F_{i,p}^m = [\alpha_{i,p}^m \ \tau_{i,p}^m]^T$. Let D be the dimension on the feature vector. The parameters used in the DP-GMM were set to: innovation parameter $\alpha = 0.6$, truncation error $err = 0.01$, truncation size for DP $M = \text{round}(1 - \alpha * \log(\text{err}/4/N))$; 2000 burn-in and 1000 sampling iterations for the Gibbs sampler. Parameters for the base (Normal-Wishart) distribution were, $\mu_0 = [0\,0]$, $\tau_0 = \frac{1}{1000}$, W is an identity matrix of size D and $df = D + 1$. Figures 5.1, 5.3 and 5.5 show the contour plots for two classes clustering. Figures 5.2, 5.4 and 5.6 show the weight distribution for these two classes clustering.

For clustering of three classes (Simple motor, $m = 1$, Language with motor, $m = 3$ and Language without motor, $m = 4$), the feature set consisted of the MPD time shift and scale change parameters $F_{i,p}^m = [\tau_{i,p}^m \ \ \kappa_{i,p}^m]^T$. The DP-GMM parameters were chosen to be the same as in the previous cases.

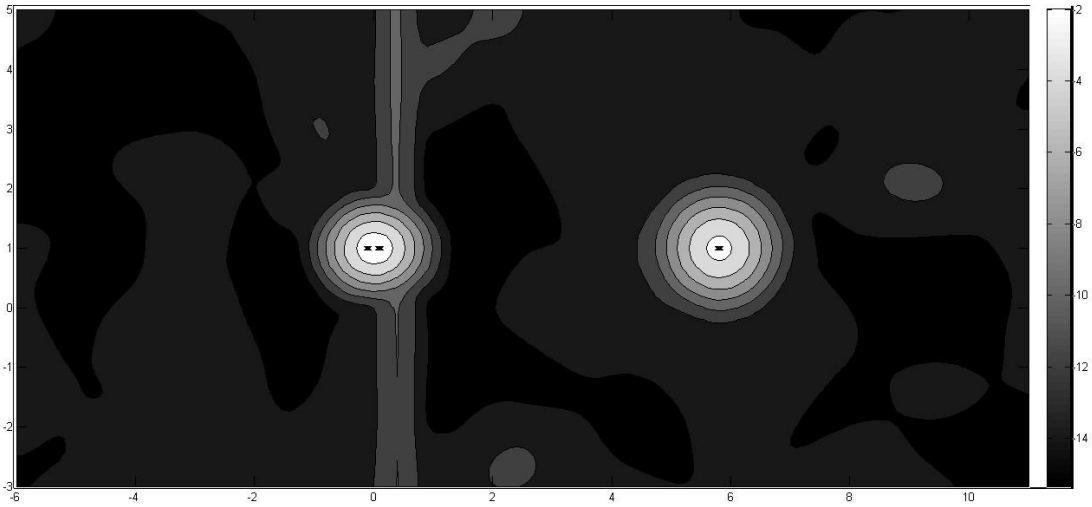


Figure 5.1: Simple Motor ($m = 1$) vs Language with Motor ($m = 3$) Contour Plot

Task m	1	3
1	**0.92**	0.08
3	0.22	**0.78**

Table 5.1: Simple Motor ($m = 1$), Language with Motor ($m = 3$)

Since equal weights of features from each class were taken,the weight distribution should show equal proportions, 0.5 in case of clustering 2 classes, and 0.33 in case of 3 classes.

Tables 5.1, 5.2, 5.3 and 5.4 show the Confusion Matrices that summarize the identification results:

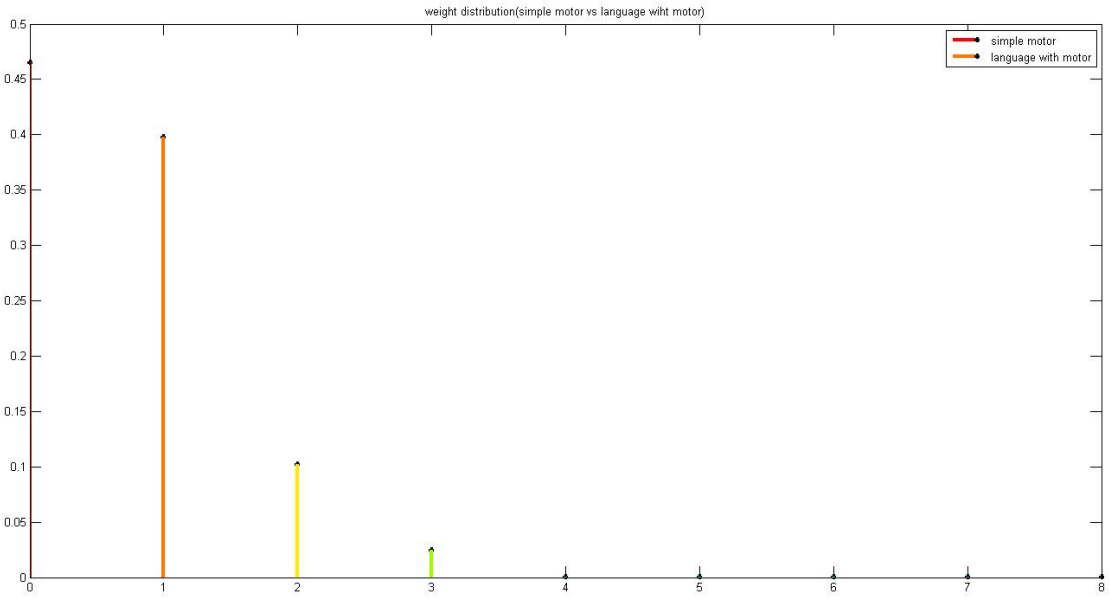


Figure 5.2: Simple Motor ($m = 1$) vs Language with Motor ($m = 3$) Weight Distributions

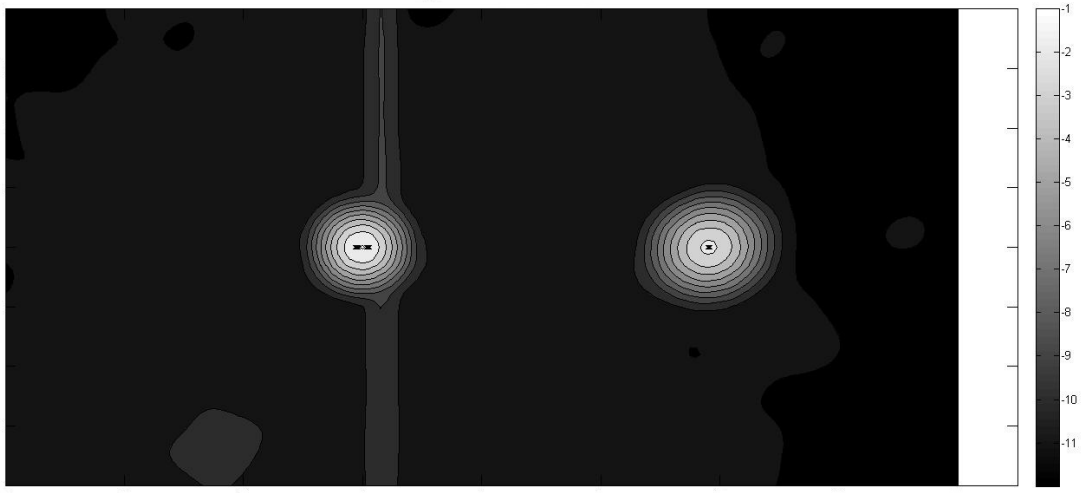


Figure 5.3: Simple Motor($m = 1$) vs Language Without Motor ($m = 4$) Contour Plot

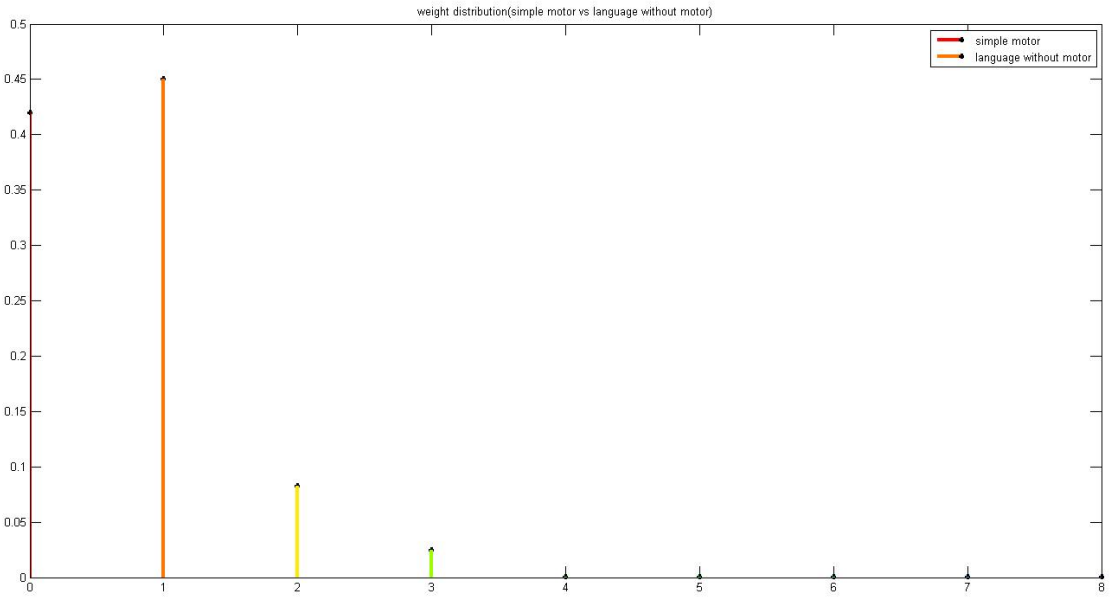


Figure 5.4: Simple Motor ($m = 1$) vs Language without Motor ($m = 4$) Weight Distributions

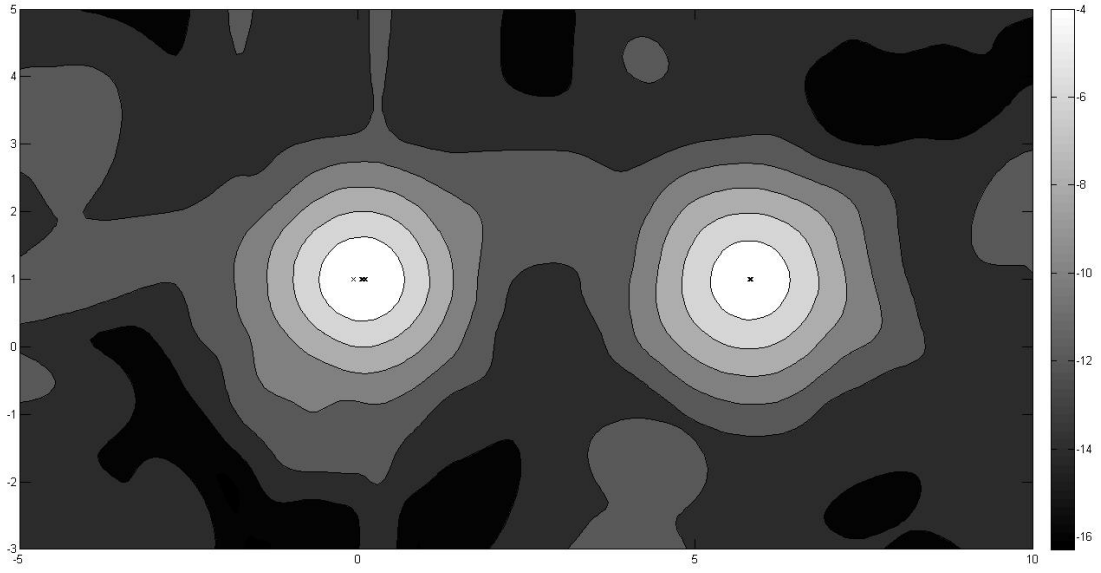


Figure 5.5: Language with Motor ($m = 3$) vs Language without Motor ($m = 4$) Contour Plot

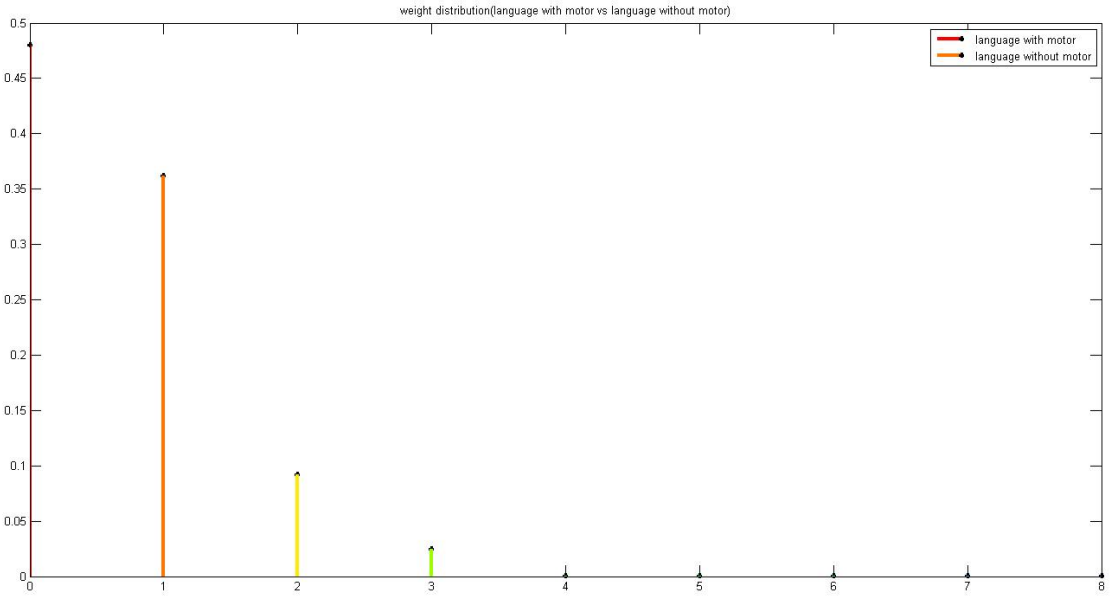


Figure 5.6: Language with Motor ($m = 3$) vs Language without Motor ($m = 4$) Weight Distributions

Task m	1	4
1	**0.84**	0.16
4	0.10	**0.9**

Table 5.2: Simple Motor ($m = 1$), Language Without Motor ($m = 4$)

Task m	3	4
3	**0.96**	0.04
4	0.28	**0.72**

Table 5.3: Language with Motor ($m = 3$), Language without Motor ($m = 4$)

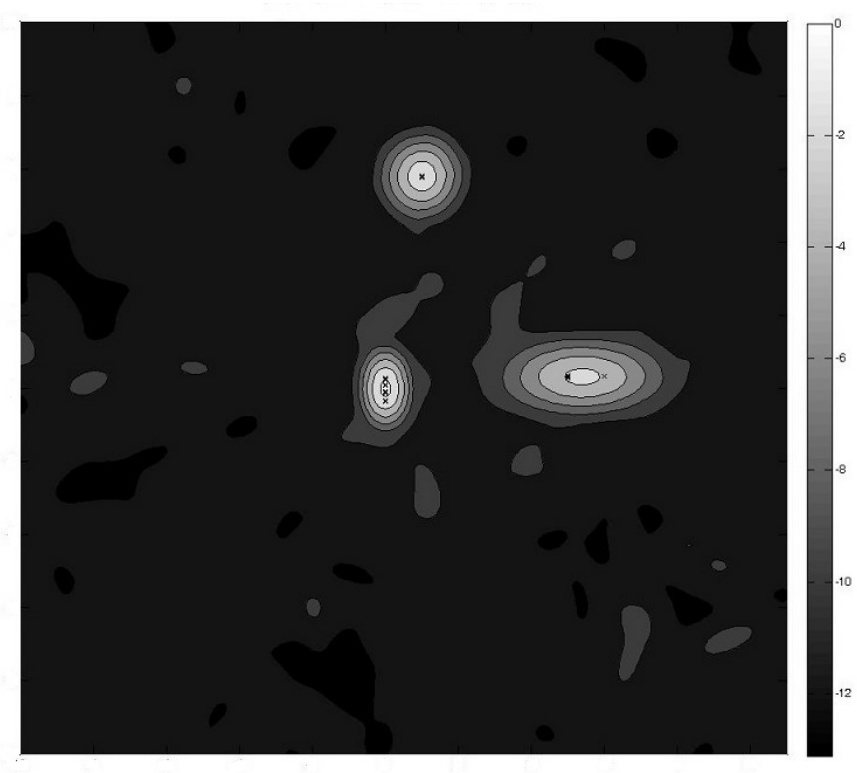


Figure 5.7: Simple Motor ($m = 1$) vs Language with Motor ($m = 3$) vs Language without Motor ($m = 4$) Contour Plot

Task m	1	3	4
1	**0.78**	0.11	0.11
3	0.035	**0.90**	0.035
4	0.095	0.095	**0.81**

Table 5.4: Simple motor ($m = 1$), Language with motor ($m = 3$) and Language without Motor ($m = 4$)

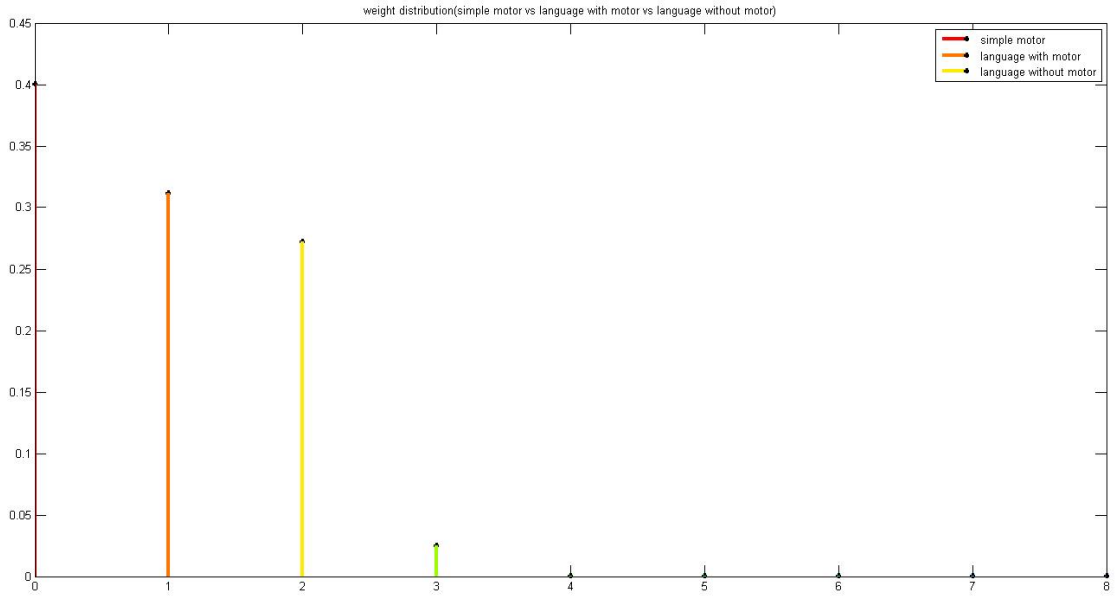


Figure 5.8: Simple Motor ($m = 1$) vs Language with Motor ($m = 3$) vs Language without Motor ($m = 4$) Weight Distributions

Chapter 6

CONCLUSION AND FUTURE WORK

We provided a description of the data collection experiments and a complete mathematical formulation for the overall proposed behavioral task identification system. We applied the matching pursuit decomposition (MPD) feature extraction method, and employed an unsupervised adaptive learning method to classify the different behavioral tasks.

From the results in Chapter 5, we have shown that it is possible to cluster the different behavior tasks performed by the Parkinson's disease (PD) patients. The confusion matrix shows the accuracy percentage of the classifications. For tasks simple motor and language with motor the accuracy was 92% and 78% respectively. For simple motor and language without motor it was 84% and 90% respectively. As for language with motor and language without motor 96% and 72% of the respective tasks were detected accurately.

In case of all the three classes, 78% of simple motor 90% of language with motor and 81% of language without motor were classified accurately. Comparing these results with that in [31] which was done using the same LFP data the hidden Markov model- support vector machine (HMM-SVM) gave an accuracy of 89% and 90% for classes simple motor and language with motor, 94% and 92% for classes simple motor and language without motor and 92% and 92% for language with motor and language without motor shown in Tables 6.1, 6.2 and 6.3.

This is the only work where the behavioral tasks performed by PD patients were classified using an unsupervised learning method which did not require any prior training and any prior knowledge of number of clusters. This work provides an im-

Task	simple motor	language with motor
Our method	92%	78%
DHMM	89%	90%
Hybrid DHMM-SVM	91%	90%

Table 6.1: Comparison of methods used to classify Simple motor ($m = 1$), Language with motor ($m = 3$)

Task	simple motor	language without motor
Our method	84%	90%
DHMM	94%	91%
Hybrid DHMM-SVM	91%	90%

Table 6.2: Comparison of methods used to classify Simple motor ($m = 1$), Language without motor ($m = 4$)

portant contribution to the construction of a closed-loop DBS system which can be the next revolutionary work for the treatment of PD patients.

6.1 Future Work

The immediate next step should be to extract relevant discriminatory information for clustering for other low/high level brain activities. That way we can have a general idea about the feature characteristics for different low/high level behavioral tasks. At the end of this we can come up with the algorithm that works best in classifying these

Task	language with motor	language without motor
Our method	96%	72%
DHMM	93%	91%
Hybrid DHMM-SVM	92%	92%

Table 6.3: Comparison of methods used to classify Language with motor ($m = 3$), Language without motor ($m = 4$)

tasks. In this work, three tasks were classified at the same time, so our future target should be to try our more than three tasks at a time. This will help us to construct the closed loop feedback DBS system which will adaptively try to adjust the DBS so that even during tremor and other motor symptoms the PD patients can perform simple tasks.

Also another area of research is the effect of DBS when on the PD patients. We have been working on this problem recently where the EEG data are collected from PD patients performing certain tasks with and without DBS. So our target is to investigate whether there is any change in the patients' behavior due to DBS. We can do that by seeing if there is any change in the feature vector and in the behavior task clustering before and after DBS.

Thirdly another area of research would be to track PD characteristics using dynamical system modeling, be it linear or nonlinear. It can be used to track tremors in Parkinson's disease. The test is to come up with a model using data to estimate the posterior PDF. This way we will be able to know the area in the brain which is responsible for the abnormal motor symptoms and how the neurons in those parts react during such symptoms. We can use tracking methods like Kalman filter or Particle filter to track the movement of the dipole based on the dynamic system model.

REFERENCES

[1] J. Vesper, S. Chabardes, V. Fraix, N. Sunde, and K. Ostergaard, "Dual channel deep brain stimulation system (kinetra) for parkinson's disease and essential tremor: a prospective multicentre open label clinical study," *Journal of Neuroly, Neurosurgy and Psychiatry,*, vol. 73, pp. 275–280, 2002.

[2] M. Clinic, "Diseases and conditions parkinson's disease." [Online]. Available: http://www.mayoclinic.org/diseases-conditions/parkinsons-disease/basics/definition/con-20028488

[3] A. O. Hebb, F. Darvas, and K. J. Miller, "Transient and state modulation of beta power in human subthalamic nucleus during speech production and finger movement," *Neuroscience*, vol. 202, 2012.

[4] D. Hirtz, D. J. Thurman, K. Gwinn-Hardy, M. Mohamed, A. R. Chaudhuri, and R. Zalutsky, "How common are the "common" neurologic disorders?" *Neurology*, vol. 6, pp. 326–327, 2007.

[5] A. O. Hebb, J. J. Zhang, M. H. Mahoor, C. Tsiokos, C. Matlack, H. J. Chizeck, and N. Pouratian, "Creating the feedback loop: Closed-loop neurostimulation," *Neurosurgery Clinics of North America*, vol. 25, no. 1, pp. 187–204, 2014.

[6] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, 2009.

[7] R. Yuvaraj, M. Murugappan, N. M. Ibrahim, K. Sundaraj, M. I. Omar, K. Mohamad, and R. Palaniappan, "Optimal set of EEG features for emotional state classification and trajectory visualization in parkinson's disease," *International Journal of Psychophysiology*, vol. 94, pp. 482–495, 2014.

[8] R. Yuvaraj, M. Murugappan, N. M. Ibrahim, and M. I. Omar, "On the analysis of EEG power, frequency and asymmetry in parkinsons disease during emotion processing," *Behavioral and Brain Functions*, 2014.

[9] S. Niketeghad, A.O.Hebb, J. Nedrud, S. J. Hanrahan, and M. H. Mahoor, "Single trial behavioral task classification using subthalamic nucleus local field potential signals." in *Engineering in Medicine and Biology Society (EMBC)*, 2014.

[10] A. O. Hebb, J. J. Zhang, M. H. Mahoor, C. Tsiokos, C. Matlack, H. J. Chizeck, and N. Pouratian, "Creating the feedback loop closed-loop neurostimulation," *Neurosurgery clinics of North America*, vol. 25, pp. 187–204, 2014.

[11] G. Buzsaki, C. Anastassiou, and C. Koch, "The origin of extracellular fields and currents–EEG, ECoG, LFP and spikes," *Nature Reviews Neuroscience*, vol. 13, no. 6, pp. 407–420, 2012.

[12] G. Buzsaki, "Large-scale recording of neuronal ensembles," *Nature Neuroscience*, vol. 7, pp. 446–451, 2004.

[13] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[14] D. Chakraborty, N. Kovvali, J. Wei, A. Papandreou-Suppappola, D. Cochran, and A. Chattopadhyay, "Damage classification structural health monitoring in bolted structures using time-frequency techniques," *Journal of Intelligent Material Systems and Structures*, vol. 20, pp. 1289–1305, 2009.

[15] K. Ni, Y. Qi, and L. Carin, "Multiaspect target detection via the in nite hidden markov model," *Journal of the Acoustical Society of America*, vol. 121, pp. 2731–2742, 2007.

[16] Y. Qi, J. W. Paisley, and L. Carin, "Music analysis using hidden markov mixture models," *IEEE Transactions on Signal Processing*, vol. 55, pp. 5209–5224, 2007.

[17] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. D. Jr., "Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model," *PLoS Computational Biology*, vol. 6, 2010.

[18] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

[19] D. Görür and C. E. Rasmussen, "Dirichlet process Gaussian mixture models: Choice of the base distribution," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 653–664, July 2010.

[20] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Harvard University and MIT Press, 1961.

[21] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, pp. 161–173, 2001.

[22] D. Young, "An overview of mixture models," *Statistics Surveys*, pp. 1–24, 2008.

[23] J. Sethuraman, "A constructive de nition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[24] M. Jordan, "Bayesian nonparametric learning: Expressive priors for intelligent systems," *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pp. 167–186, 2010.

[25] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDPHMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, pp. 1020–1056, 2011.

[26] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[27] D. Blackwell and J. B. MacQueen, "Ferguson distributions via polya urn schemes," *The Annals of Statistics*, vol. 1, pp. 353–355, 1973.

[28] C. E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[29] M. West, P. Muller, and M. D. Escobar, *Hierarchical priors and mixture models, with applications in regression and density estimation*, P. R. Freeman and A. F. Smith, Eds. Aspects of Un- certainty, 1994.

[30] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Markov chain monte carlo in practice," *CRC press*, no. 2, 1996.

[31] N. Zaker, A. Dutta, A. O. Hebb, A. Maurer, N. Kovvali, J. J. Zhang, and A. P.-S. an S. Hanrahan, "Adaptive learning of behavioral tasks for patients with parkinsons disease using signals from deep brain stimulation," in *Asilomar Conference*, 2014.

Appendix

APPENDIX A

MATCHING PURSUIT DECOMPOSITION ALGORITHM

Algorithm 1 Matching Pursuit Decomposition

Input to algorithm:
Signal vector $\mathbf{y} = [y[0]\, y[1]\, \ldots,\, y[N-1]]^T$
Dictionary matrix $\mathbf{D} = [\mathbf{g}_1\, \mathbf{g}_2\, \ldots\, \mathbf{g}_\Gamma]^T$, with dictionary vector elements $\mathbf{g}_\gamma = [g_\gamma[0]\, g_\gamma[1]\, \ldots,\, g_\gamma[N-1]]^T$, $\gamma = 1, \ldots, \Gamma$; the γth vector element has feature vector $\mathbf{q}_\gamma = [\alpha_\gamma\, \tau_\gamma\, \nu_\gamma\, \kappa_\gamma]^T$
stopping iteration number I_{stop}
initialization:
initialize extracted signal matrix $\mathbf{s} = zeros(N, I_{\text{stop}})$ and extracted features matrix $\mathbf{F} = zeros(4, I_{\text{stop}})$
initialize: residue $\mathbf{r} = \mathbf{y}$,
for i=1:I_{stop} **do**
Compute the inner product of the residue with each dictionary element
$\alpha_\gamma = \langle \mathbf{r}, \mathbf{g}_\gamma \rangle = \sum_{n=0}^{N-1} y[n] g_\gamma^*[n]$
Find dictionary element vector $\mathbf{g}_p$ that yields the maximum inner product
$\mathbf{g}_i = \arg\max_{\gamma \in [1,\ldots,\Gamma]} \alpha_\gamma$
Update the residue vector: $\mathbf{r} = \mathbf{r} - \alpha_i \mathbf{g}_i$
Update extracted signal matrix and extracted feature matrix: $\mathbf{s} = \mathbf{s}(:, \mathbf{g}_i)$ $\mathbf{F} = \mathbf{F}(:, \mathbf{q}_i)$
set $i = i + 1$
end for
output:
Extracted signal matrix $\mathbf{s}$, extracted feature matrix $\mathbf{F}$ whose $[\mathbf{F}]_p = [\alpha_p\, \tau_p\, \nu_p\, \kappa_p]^T$ column corresponds to the features of the pth extracted signal, $p = 1, \ldots, I_{\text{stop}}$.

APPENDIX B

DIRICHLET PROCESS GAUSSIAN MIXTURE MODEL WITH BLOCKED GIBBS SAMPLING ALGORITHM

Algorithm 2 Blocked Gibbs sampling for DP-GMM using an D-dimensional dataset X

input: Dataset $X = \{x_1, ..., x_N\}$, DP innovation parameter α, Normal-Wishart hyperparameters $\mu_{\mathcal{N}}, \tau_{\mathcal{N}}, \xi_{\mathcal{W}}, \nu_{\mathcal{W}}$, DP truncation limit M

output: Samples $\{\mu_m^{(i)}, \Sigma_m^{(i)}, c^{(i)}, w^{(i)}\}_{i=1}^{L}$

for i=1:Gibbs iteration **do**

1. update for $\theta_m^{(i)} = \{\mu_m^{(i)}, \Sigma_m^{-1(i)}\} \sim p(\mu_m, \Sigma_m | c^{(i-1)}, X) m = 1, ..., M$

(a) Let $X_m = x_n : c_n^{(i-1)} = m$ and $N_m = |X_m|$, for $m = 1, ..., M$

(b) For all clusters, $m = 1, ..., M$, compute

$\mu_{x_m} = \frac{1}{N_m} \sum_n x_n$

$\Sigma_{x_m} = \frac{1}{N_m} \sum_n (x_n - \mu_{x_n})^2$

$\tilde{\mu}_{\mathcal{N},m} = \frac{\tau_{\mathcal{N}} \tilde{\mu}_{\mathcal{N}} + N_m \mu_{x_m}}{\tau_{\mathcal{N}} + N_m}$

$\tilde{\tau}_{\mathcal{N},m} = \tau_{\mathcal{N}} + N_m$

$\tilde{\nu}_{\mathcal{W},m} = \nu_{\mathcal{W}} + \Sigma_{x_m} + \frac{\tau_{\mathcal{N}} N_m}{\tau_{\mathcal{N}} + N_m} (m - \mu_{x_m})(m - \mu_{x_m})^T$

$\tilde{\xi}_{\mathcal{W},m} = \xi_{\mathcal{W}} + N_m$

(c) draw samples for $\Sigma_m^{-1(i)}$ from the Wishart distribution, $\mathcal{W}(\Sigma_m^{-1}, \tilde{\nu}_{\mathcal{W},m}, \tilde{\xi}_{\mathcal{W},m})$, for $m = 1, ..., M$

(d) draw sample for $\mu_m^{(i)}$ from the Normal distribution $\mathcal{N}(\mu_m; \tilde{\mu}_{\mathcal{N},m}, \frac{\Sigma_m^{(i)}}{\tau_{\mathcal{N},m}})$ for $m = 1, ..., M$

2. Update for $c_n^{(i)} \sim p(c_n | \mu^{(i)}, \Sigma^{-1(i)}, w^{(i-1)}, X)$, $n = 1, ..., N$

(a) let $\pi_{m,n} = w_m^{(i-1)} \mathcal{N}(x_n; \mu_m^{(i)}, \Sigma_m^{(i)})$, $m = 1, ..., M$, $n = 1, ..., N$

(b) Normalize $\pi'_{m,n} = \frac{\pi_{m,n}}{\sum_{m=1}^{M} \pi_{m,n}}$, $m = 1, ..., M$, $n = 1, ..., N$

(c) draw samples for $c_n^{(i)} \sim \sum_{m=1}^{M} \pi'_{m,n} \delta(c_n, m)$, $n = 1, ..., n$

3. Update for $w_m^{(i)} \sim p(w_m, c^{(i)})$, $m = 1, ..., M$

(a) draw samples $\beta_j \sim Beta(1_{N_m}^*, \alpha + \sum_{m'=m+1}^{M} N_m'^*)$, where $N_m^* = |n : c_n^{(i)} = m|$, $m = 1, ..., M$

(b) finally evaluate $w_m^{(i)} = \beta_m \prod_{j=1}^{m-1} (1 - \beta_j)$, $m = 1, ..., M$

end for

APPENDIX C

DIRICHLET-CATEGORICAL CONJUGATE PRIOR

$$p_1, ..., p_K \sim Dir(\alpha_1, .., \alpha_K)$$

$$y \sim Cat(p_1, ..., p_K)$$

C.1 Posterior

$$\begin{aligned}
f(\theta|D) &\propto f(\theta, D) \\
&= f(p_1, ..., p_K|\alpha_1, ..., \alpha_K) \prod_{y_i \in D} f(y_i|p_1, ..., p_k) \\
&\propto \prod_{j=1}^{K} p_j^{\alpha_j - 1} \prod_{y_i \in D} \prod_{j=1}^{K} p_j^{I\{y_i = j\}} \\
&= \prod_{j=1}^{K} p_j^{\alpha_j - 1 + \sum_{y_i \in D} I\{y_i = j\}}
\end{aligned}$$

This density is exactly that of a Dirichlet distribution, except we have

$$\alpha'_j = \alpha_j + \sum_{y_i \in D} I\{y_i = j\}$$

That is, $f(\theta|D) = Dir(\alpha'_1, ..., \alpha'_K)$

$$
\begin{aligned}
f(y = x|D) &= \int f(y = x|\theta) f(\theta|D) d\theta \\
&= \int f(y = x|p_1, ..., p_K) f(p_1, ..., p_K|D) dS_k \\
&= \int p_x \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \prod_{j=1}^K p_j^{\alpha'_j - 1} dS_k \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \int \prod_{j=1}^K p_j^{I\{x=j\} + \alpha'_j - 1} dS_k \\
&= \frac{\Gamma(\sum_{j=1}^K \alpha'_j)}{\prod_{j=1}^K \Gamma(\alpha'_j)} \frac{\prod_{j=1}^K \Gamma(I\{x = j\} + \alpha'_j)}{\Gamma(1 + \sum_{j=1}^K) \alpha'_j} \\
&= \frac{\alpha'_x}{\sum_{j=1}^K \alpha'_j}
\end{aligned}
$$

where we used the fact that $\Gamma(n + 1) = \Gamma(n)$ to simplify the second to last line.

APPENDIX D

NORMAL-WISHART CONJUGATE PRIOR

The multivariate analog of Normal prior is the Normal-Wishart prior. Here we state the results. We assume X is a d-dimensional.

D.1 Likelihood

$$p(D|\mu,\Sigma) = (2\pi)^{nd/2}|\Sigma|^{n/2}\exp\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu)\right)$$

D.2 Prior

$$\begin{aligned}
p(\mu,\Sigma) &= NWi(\mu,\Sigma|\mu_0,\kappa,\nu,T) = \mathcal{N}(\mu|\mu_0,(\kappa\Sigma)^{-1})Wi_\nu(\Sigma|T) \\
&= \frac{1}{Z}\Sigma^{1/2}\exp(-\frac{\kappa}{2}(\mu-\mu_0)^T\Sigma(\mu-\mu_0))|\Sigma|^{(\kappa-d-1)/2}\exp(-1/2tr(T^{-1}\Sigma)) \\
Z &= (\frac{\kappa}{2\pi})^{d/2}|T|^{\kappa/2}2^{d\kappa/2}\Gamma d(\kappa/2)
\end{aligned}$$

Here T si the prior covariance. To see the connection to the scalar case, make the substitutions

$$\alpha_0 = \frac{\nu_o}{2}, \beta_0 = \frac{T_0}{2}$$

D.3 Posterior

$$
\begin{aligned}
p(\mu, \Sigma|D) &= \mathcal{N}(\mu|\mu_0, (\kappa\Sigma)^{-1}) Wi_\nu(\Sigma|T) \\
\mu_n &= \frac{\kappa\mu_0 + n\bar{x}}{\kappa + n} \\
T_n &= T + S + \frac{\kappa n}{\kappa + n}(\mu_0 - \bar{x})(\mu_0 - \bar{x})^T \\
S &= \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \\
\nu_n &= \nu + n \\
\kappa_n &= \kappa + n
\end{aligned}
$$

posterior marginals

$$
\begin{aligned}
p(\Sigma|D) &= Wi_{\nu_n}(T_n) \\
p(\mu|D) &= t_{\nu_n - d + 1}(\mu|\mu_n, \frac{T_n}{\kappa_n(\nu_n - d + 1)})
\end{aligned}
$$

The MAP estimates are given by

$$
\begin{aligned}
(\hat{\mu}, \hat{\Sigma}) &= argmax\ p(D|\mu, \Sigma) NWi(\mu, \Sigma) \\
\hat{mu} &= \sum_{i=1}^{n} x_i + \kappa_0\mu_0 N + \kappa_0 \\
\hat{\Sigma} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T + \kappa_0(\mu_0 - \bar{x})(\mu_0 - \bar{x})^T + T_0^{-1}}{N + \nu_0 - d}
\end{aligned}
$$

This reduces to the MLE if $\kappa_0 = 0, \nu_0 = d$ and $|T_0| = 0$

D.4 Posterior Predictive

$$
p(x|D) = t_{\nu_n - d + 1}(\mu_n, \frac{T_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)})
$$