Discovering Alzheimer's Disease Subtypes by Voxel-based Morphometry and Topic Modeling

by

Xiuming Zhang

Submitted to the Department of Electrical & Computer Engineering
in partial fulfillment of the requirements for the degree of

Bachelor of Engineering (Electrical Engineering)

at the

NATIONAL UNIVERSITY OF SINGAPORE

April 2015

Author .....
Department of Electrical & Computer Engineering
April 6, 2015

Advised by .....
B. T. Thomas Yeo
Assistant Professor of Electrical & Computer Engineering
Thesis Supervisor

Examined by .....
Shuicheng Yan
Associate Professor of Electrical & Computer Engineering

Discovering Alzheimer's Disease Subtypes by Voxel-based Morphometry and Topic Modeling

by

Xiuming Zhang

Submitted to the Department of Electrical & Computer Engineering
on April 6, 2015, in partial fulfillment of the
requirements for the degree of
Bachelor of Engineering (Electrical Engineering)

Abstract

Alzheimer's disease (AD) is the most common type of dementia that usually attacks the elderly population. Great heterogeneity has been noticed within AD, suggesting the possible existence of different subtypes. As such, researchers are actively exploring AD subtypes to enable the disease treatment.

Two atypical variants have been manually defined by pathological autopsy, but this post-mortem classification may not be as helpful in treatment and suffer from human bias. On the other hand, neuroimaging-based classification, despite its advantage of being in-vivo, is difficult due to the overwhelming amount of high-dimensional data.

In this thesis, we address this challenge by machine learning of the neuroimaging data. Specifically, we first quantify brain atrophy with voxel-based morphometry and then model each AD patient as a mixture of AD subtypes and each subtype as a mixture of atrophied voxels under the framework of topic modeling.

By doing so, we are able to learn, in an unsupervised manner, three subtypes—hippocampal, cerebellum, and cortical—that show great disparity in memory and executive function both at the baseline and during the disease progression. Furthermore, our model, when applied to mild cognitive impairment (MCI) subjects, provides predictive information about possible future conversion into AD.

Hopefully, our neuroimaging-based classification could facilitate AD understandings as well as the development of subtype-specific treatments.

Thesis Supervisor: B. T. Thomas Yeo
Title: Assistant Professor of Electrical & Computer Engineering

Acknowledgments

I would like to express the utmost gratitude and appreciation to my thesis advisor, Assistant Prof. Thomas Yeo, who has been an invaluable wellspring of knowledge, in both research and life, throughout my senior year at the University. This thesis would never be completed without his continuous guidance and enlightening insights. It is especially his rigorous scientific approach that greatly motivated me to keep seeking improvements. I am also sincerely grateful for his generous support for my graduate school application, which opened up the exciting opportunity for me to achieve more in scientific research.

I would also like to extend my gratitude to our collaborator, Assistant Prof. Mert Sabuncu at Massachusetts General Hospital Martinos Center for Biomedical Imaging, for his significant input and positive encouragement during our every Skype meeting.

Special thanks go to my thesis examiner, Associate Prof. Shuicheng Yan, for his constructive comments and warm approval during the continuous assessment.

Data used in preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, I would like to thank the ADNI investigators (full list available at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.) for generously sharing the data.

Last but not least, I would like to thank my family and friends for their continuous support throughout my undergraduate study.

Contents

List of Figures

List of Tables

Abbreviations

AD Alzheimer's disease

ADNI Alzheimer's Disease Neuroimaging Initiative

Ce Cerebellum subtype

CN Cognitive normal

Co Cortical subtype

EF Executive function score

EM Expectation-maximization

GLM General linear model

GM Gray matter

H Hippocampal subtype

ICV Intracranial volume

LDA Latent Dirichlet allocation

MCI Mild cognitive impairment

MEM Memory score

MMSE Mini-mental state examination

MR Magnetic resonance

MRI Magnetic resonance imaging

S Subcortical subtype

VBM Voxel-based morphometry

Chapter 1

Introduction

Alzheimer's disease (AD) is the most common type of dementia that usually attacks the elderly population. Its symptoms include short-term memory loss in the early stage, confusion, irritability, aggression, and long-term memory loss in the later stage [1]. Eventually, the patients would have to completely rely on care givers in daily activities, which places huge pressure on both their families and the governments.

So far, none of the treatments for AD has been proven effective. Yet, scientists have noticed and verified the great heterogeneity within AD in clinical, imaging, and pathological aspects. In the hope that understanding the AD subtypes might shed light on personalized treatments, researchers are seeking to discover the atypical AD variants.

[2] is one such endeavor where Murray et al. pathologically define two AD subtypes—hippocampal-sparing (HpSp) and limbic-predominant (Lp)—besides the typical AD. Yet, the classification is post-mortem and highly supervised. Noh and colleagues hierachical-cluster AD into three subtypes based on cortical thickness [3], but it only utilizes the cortical thickness information and fails to account

for a possible mixture of subtypes. Hence, it still remains as an open issue how to properly model AD subtypes and thereby discover them.

In this thesis, we model AD subtypes with latent Dirichlet allocation (LDA) [4], a popular topic model originally proposed to discover latent topics underlying a corpus. Specifically, we quantify brain atrophy in structural magnetic resonance (MR) images with voxel-based morphometry (VBM) [5] and then process them to resemble text documents for LDA. The discovered subtypes are then examined in both their baseline characteristics and longitudinal progressions.

Our three contributions can be summarized as follows.

- Complementing the traditional classification methods (such as [2]) that largely rely on domain knowledge, our method is *unsupervised* and thus could potentially provide a more *holistic* picture, which may eventually lead to counterintuitive but interesting findings.

- Our method studies in-vivo MR images instead of performing post-mortem autopsy, therefore enabling *early classification*, which is a vital prerequisite for patients to receive personalized treatments.

- Given the complexity of neurodegenerative diseases such as AD, it seems impractical to classify a patient as one deterministic subtype. By considering each subject as a *mixture of subtypes*, we open up opportunities for neuroscientists to explain some of the composite effects, which are otherwise unaccountable.

Hopefully, this work could provide a new insight into AD subtype classification and thus facilitate the development of mature early-stage subtype classification. If AD subtype composition could be confirmed in an early stage, subtype-specific treatments could then be developed to retard, stop, or even reverse the disease

progression.

The remainder of the thesis is organized as follows. Chapter 2 provides the fundamentals needed to understand our methods. Chapter 3 summarizes the related work. Chapter 4 presents our model and methods. The discovered subtypes are shown and validated in Chapter 5. Finally, Section 6 concludes the thesis and points out future work.

Chapter 2

Background

2.1 Alzheimer's Disease

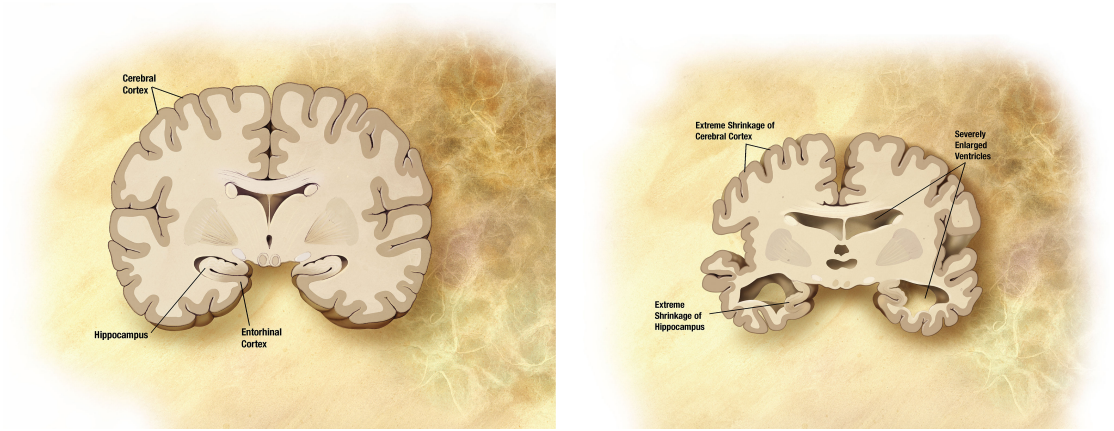
This section first illustrates the motivations of this project by providing an overview of the disease. Next, the pathological aspect of the disease is briefly discussed. Finally, we introduce several cognitive tests that we will utilize to analyze our subtypes in Chapter 5.

2.1.1 Overview

Alzheimer's disease (AD) is the most common type of dementia that usually attacks the elderly population. Its symptoms include short-term memory loss in the early stage, confusion, irritability, aggression, and long-term memory loss in the later stage [1]. **Mild cognitive impairment (MCI)** is a prodromal stage of AD. AD starts in the entorhina cortex, a region near the hippocampus [6]. Upon onset, the neurons start becoming enervated and losing the communication ability. Soon, AD spreads to the hippocampus (the component in charge of learning and converting short-term memories into long-term ones), causing new memory difficult to form. As AD progresses, the patients' language and problem solving abilities deteriorate, and they start losing the ability to control their emotions and

make sense of things. In the late stage, the patients can no longer sustain their old memories and have to completely depend on others for care [6].

The AD patients' brains are abundant in two abnormal structures—amyloid plaques and neurofibrillary tangles—that are made of misfolded proteins [2]. From the neuroimaging perspective, AD causes the hippocampus, cerebral cortex to shrink, and ventricles to expand, all of which are observable in the structural magnetic resonance imaging (MRI) images [6]. See Figure 2-1.



(a) Normal brain [7]. (b) AD brain [8].

Figure 2-1: Comparison between a normal brain and an AD brain.

As of September 2014, there have been over 1400 clinical trials studying possible AD treatments [9], but none of them has been proven effective. As [10] surveys, the heterogeneity of AD in clinical, imaging, and pathological aspects has been noticed and verified by many researchers. With the hope that understanding the AD subtypes might shed light on personalized treatments, researchers are seeking to discover the atypical AD variants. One such endeavor is [2], where Murray et al. suggest the existence of two AD subtypes besides the typical AD, hippocampal-sparing (HpSp) and limbic-predominant (Lp).

2.1.2 Pathological Biomarkers

AD biomarkers are biochemical indicators used to predict, diagnose the disease, and quantitatively measure its progression. Two widely accepted biomarkers for AD are amyloid-β (Aβ) that forms plaques and tau protein that forms tangles, both of which reside in the cerebrospinal fluid (CSF). More generally speaking, AD biomarkers fall into two categories: one assessing Aβ deposition and the other measuring neurodegeneration, defined as progressive loss of neurons and their functions. For instance, CSF Aβ 1-42 belongs to the former category, whereas CSF total tau (t-tau), phosphorylated tau (p-tau), and atrophy on MRI belong to the latter [11]. Unlike CSF Aβ 1-42, whose level drops in AD due to the plaque formation [12], CSF t-tau and p-tau increase in AD [13].

Besides CSF Aβ 1-42, t-tau, and p-tau, researchers also start studying the role of other proteins, such as CSF clusterin, in the neurodenegeration process. Morris et al. reveal a strong interaction between CSF Aβ 1-42 and clusterin on the entorhinal cortex atrophy rate but not the hippocampal atrophy rate [14]. In addition, the interaction between CSF Aβ 1-42 and p-tau 181p is also found significant [14]. As a genetic determinant of AD risk, apolipoprotein E's (ApoE) different alleles affect AD differently: individuals carrying ApoE ε4 are at an increased risk of AD than those carrying ApoE ε3, whereas ApoE ε2 helps reduce the risk [15].

2.1.3 Cognitive Tests

Yet, these biomarkers alone sometimes may not suffice to diagnose the disease or fully describe its progression. Hence, cognitive tests, such as **mini-mental state exam (MMSE)**, are usually used as a supplement [16]. MMSE is a 30-point

questionnaire[1] that measures cognitive impairment by testing attention, calculation, recall, language, ability to follow simple commands, and orientation. In MMSE, a higher score indicates better cognition, and the longitudinal scores reflect the trajectory of the cognitive change.

Besides MMSE, there are other standard tests such as Alzheimer's disease assessment scale-cognitive subscale (ADAS-Cog) and clinical dementia rating (CDR), both of which assign a higher score to greater cognition dysfunction (opposite to MMSE). To discern possible divergence in memory and executive function declines, researchers have derived from the available neuropsychological battery two separate composite scores, one for **executive functioning (EF)** [17] and the other for **memory (MEM)** [18]. Same as MMSE, a higher value in both MEM and EF stands for a healthier condition.

2.2 Neuroimaging

Neuroimaging refers to an array of imaging techniques that allow neurologists to image the brain structures for analysis purpose in a non-invasive or low-invasiveness way. Given that the advent of neuroimaging techniques enables in-vivo brain examination, two Nobel prizes have been awarded to the neuroimaging pioneers—one to Allan Cormack and Godfrey Hounsfield for the invention of computerized axial tomography (CT) and the other to Peter Mansfield and Paul Lauterbur for developing magnetic resonance imaging (MRI).

Equally important as the imaging technology is the subsequent image analysis. Hence in this section, we first introduce very briefly the structural MR image—our data format—and then explain the technique that we use to compute the

[1] A sample is available at http://www.health.gov.bc.ca/pharmacare/adti/clinician/pdf/ADTI%20SMMSE-GDS%20Reference%20Card.pdf.

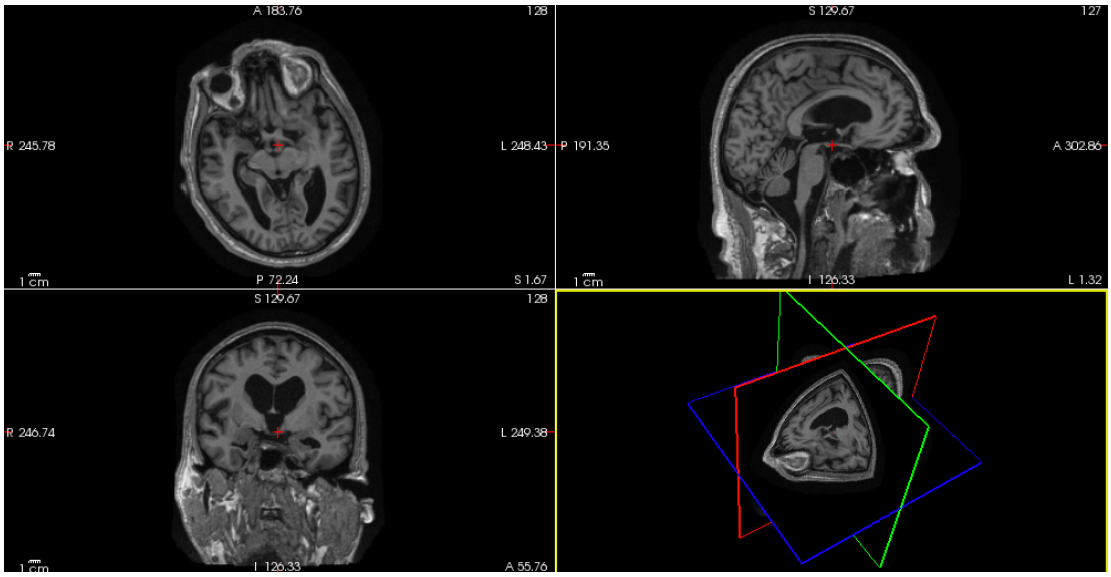


Figure 2-2: An unprocessed structural MR image viewed in FreeSurfer[2].

voxel-wise atrophy.

2.2.1 Structural Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) produces high-quality three-dimensional (3D) images of brain structures by magnetic field and radio wave. Two common types are structural MRI for static examinations and functional MRI (fMRI) for temporal brain activity recording. Unless otherwise specified, the subjects studied in this thesis are all structural MR images.

Similar to pixels constituting 2D images, voxels compose 3D MR images. As seen in Figure 2-2, besides gray matter (GM)—the matter of interest in this study, there are also skulls, brain stems, cerebrospinal fluid, white matter, and etc. present in an unprocessed MR image. Hence, one needs to perform necessary preprocessing (to be discussed in 4.3), such as image segmentation, prior to analyzing the GM.

[2] http://surfer.nmr.mgh.harvard.edu/

2.2.2 Voxel-based Morphometry

Voxel-based morphometry (VBM) [5] is a statistical approach that allows one to compare voxel-wise local GM concentrations between two subject groups. VBM differs from some other global-shape techniques that VBM focuses on the *local* differences with the macroscopic shape differences already discounted. Briefly speaking, VBM discounts these global differences by spatially normalizing the images to a common brain template.

In this thesis, FSL-VBM pipeline[3], as a part of FSL [19], is adopted, and its steps can be summarized as follows.

1. **Brain Extraction & GM Segmentation.** First, non-brain structures, such as skulls and eyes, are stripped by BET [20]. Then, all the brain-extracted images are segmented into GM, white matter, and cerebrospinal fluid by FAST4 [21].

2. **Template Creation.** This step constructs a study-specific GM template from the subjects. First, GM images are affinely registered (using FLIRT [22][23]) to the GM MNI152 template, then concatenated, and averaged. The averaged image is then flipped about the y-aixs and re-averaged with its unflipped copy, producing the first-pass affine template. The whole process is then repeated, but this time using non-linear registration (using FNIRT [24][25]) to the first-pass template (instead of the MNI152 template) to obtain the final non-linear study-specific template.

3. **Modulation & Smoothing.** All the GM images can now be non-linearly registered to the study-specific template, after which each voxel is modulated by the Jacobian of the warp field so as to preserve the absolute amount rather than relative concentration of the GM. The modulated GM images are then

[3] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fslvbm

concatenated into a 4D image and smoothed by a range of Gaussian kernels with full width at half maximum (FWHM) usually chosen around 10 mm (8 mm in [26] and 10 mm in [27][28]). This makes each voxel the average GM concentration from the voxels around, making the subsequent voxel-wise analysis similar to a region of interest approach. In addition, smoothing also renders, by the central limit theorem, the data more normally distributed, improving the validity of the subsequent parametric statistical tests.

4. **Statistical Analysis.** Statistical analysis relies on the general linear model (GLM, to be discussed in the next section) to identify the regions that are significantly different in GM amount between groups [5]. For each voxel, a GLM is fitted, which in turn provides us with the p-value against the null hypothesis $\mathcal{H}_0$: the two groups do not have statistically different GM amounts for this particular voxel.

2.3 Machine Learning

This section first discusses general linear model (GLM) and its usefulness in testing variable dependency, group comparison, and partialing out effects by nuisance variables. Next, it introduces latent Dirichlet allocation (LDA) by explaining its graphical model, generative process, and the variational expectation-maximization (EM) algorithm for posterior inference and parameter estimation.

2.3.1 General Linear Model

In this subsection, scalars are expressed in the regular lowercase, such as y; column vectors are in the bold lowercase, e.g., $\mathbf{x}$; matrices are written in the bold uppercase, like $\mathbf{X}$.

General linear model (GLM[4]) predicts a response y as a linear[5] combination of the explanatory variables (or conditions) $\mathbf{x} = (1, x_1, x_2, \dots,)^T$ with a residual $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots)^T$ are the coefficients. More generally, when multiple sets of responses and conditions are observed, the model becomes, in matrix form,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix} = \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \end{pmatrix},$$

where $\mathbf{y} = (y_1, y_2, \ldots)^T$ is the column vector of observations, each row of the **design matrix X** specifies a set of conditions $\mathbf{x}^T$, and $\boldsymbol{\epsilon}$ is the column vector of independent and identically distributed (i.i.d.) residuals. From a probabilistic perspective, GLM becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix} = \mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \operatorname{diag}(\sigma^2)\right) = \mathcal{N}\left(\begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \end{pmatrix} \boldsymbol{\beta}, \begin{pmatrix} \sigma^2 & 0 & \ldots \\ 0 & \sigma^2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}\right),$$

and our goal is to find the "best" estimate of $\boldsymbol{\beta}$ given $\mathbf{y}$ and $\mathbf{X}$, for which we use the maximum likelihood estimation (MLE) method

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log p\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2\right)$$

[4]Some literature abbreviates generalized linear model also as GLM. Yet, they shall not be confused. In fact, general linear model can be viewed as a special case of generalized linear model with identity link and responses normally distributed. "Identity link" simply means that the linear combination of the explanatory variables is already the response mean (i.e., without any transformation).

[5]Note that the "linear" here refers to the linearity of the coefficients $\boldsymbol{\beta}$. Hence, for example, $y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^4 + \epsilon$ is also a GLM.

$$
\begin{aligned}
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2\right) \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2 \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2,
\end{aligned}
$$

which means that estimating $\boldsymbol{\beta}$ by MLE is actually equivalent to by minimizing $\sum_{n=1}^{N}(y_n - \mathbf{x}_n^T\boldsymbol{\beta})^2$, the sum of square errors. As a result, this method is also known as **least squares**. Then, $\boldsymbol{\beta}$ can be easily solved to be $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, provided that the design matrix $\mathbf{X}$ is full-rank.

Testing Variable Dependency

GLM is a rather useful framework that, for example, allows us to test by how much one variable is linearly dependent on the other. Suppose we want to test whether the response y is dependent on one explanatory variable x given 200 (x, y) pairs. In this case, $\mathbf{X}$ is 200×2, and $\boldsymbol{\beta}$ contains only a intercept β_0 and a gradient β_1. We further define a **contrast** $\mathbf{c} = (0, 1)^T$ that allows us to only consider β_1 if we compute $\gamma = \mathbf{c}^T\boldsymbol{\beta}$. Thus, our goal becomes to test how likely γ is 0 given the observed data.

Given our model assumptions, γ actually follows a Student's t-distribution centered around zero under the null hypothesis $\mathcal{H}_0$: y is independent of x. Thus, one can compute the probability of our γ estimate assuming it is drawn from that t-distribution due to $\mathcal{H}_0$. This probability, often referred to as p-value, indicates how confident we are in rejecting $\mathcal{H}_0$ at some significance level (usually 1% or 5%). For instance, if $p < 0.01$, we know there is only a $< 1\%$ chance for us to draw such a γ assuming $\mathcal{H}_0$ is true. Hence, we could reject $\mathcal{H}_0$, concluding that y is dependent on x.

Comparing Two Groups Adjusted for Covariates

Recall that in the final step of VBM, we utilize GLM to compare two groups' mean values at a particular voxel. This is just one easy extension from the previous use. Now that we have two groups, we need to split $\mathbf{X}$'s all-one column into two binary complementary ones so as to account for two different intercepts[6]. In our VBM case, also present in $\mathbf{X}$ are three covariate columns for age, gender, and intracranial volume (ICV). After fitting the GLM, we set our contrast $\mathbf{c} = (1, -1, 0, \ldots)^T$ or $\mathbf{c} = (-1, 1, 0, \ldots)^T$ (assuming the first two are the two membership columns) to determine whether the two groups' averages are statistically different while controlling for age, gender, and ICV.

Partialing Out Nuisance Variables

Besides controlling for the nuisance variables during group comparison, GLM is also capable of doing so even when there is no comparison. This happens when we wish to preprocess the raw data (e.g., hippocampus volume) so that they are free of nuisance effects (such as those by age, gender, and ICV).

For this purpose, we first center the nuisance variables by demeaning each of them and then construct the design matrix $\mathbf{X}$. After fitting the model, we take the sum of only the offset and residuals as the processed data. This is equivalent to not centering the variables first and then using each variable's mean value to predict the response $\mathbf{y}$. Intuitively, we are replacing the raw data with the data predicted by a GLM that assumes the same age, gender, and ICV for everyone. By doing so, the effects by the nuisance variables are removed (partialed out), whereas the effects by the variable of interest, in this case the disease, are preserved.

[6] If the inference of interest is whether the gradients are different between the two groups, then other covariate columns need also to be bisplit to account for different gradients.

2.3.2 Latent Dirichlet Allocation

In this subsection, we introduce the overview and intuition of the latent Dirichlet allocation (LDA) model, then formulate the model with its graphical model representation and generative process, and finally outline the variational inference for LDA.

Overview

Latent Dirichlet allocation (LDA) [4] is a Bayesian generative model originally proposed to discover latent topics underlying a corpus. It considers a document as a mixture of topics and a topic as a mixture of words (see Figure 2-3). Therefore, it is *hierarchical* and *mixed-membership*. The learning of LDA is *unsupervised*; during learning, it encourages sparse word and topic distributions, which eventually leads to co-occurring words getting clustered together.

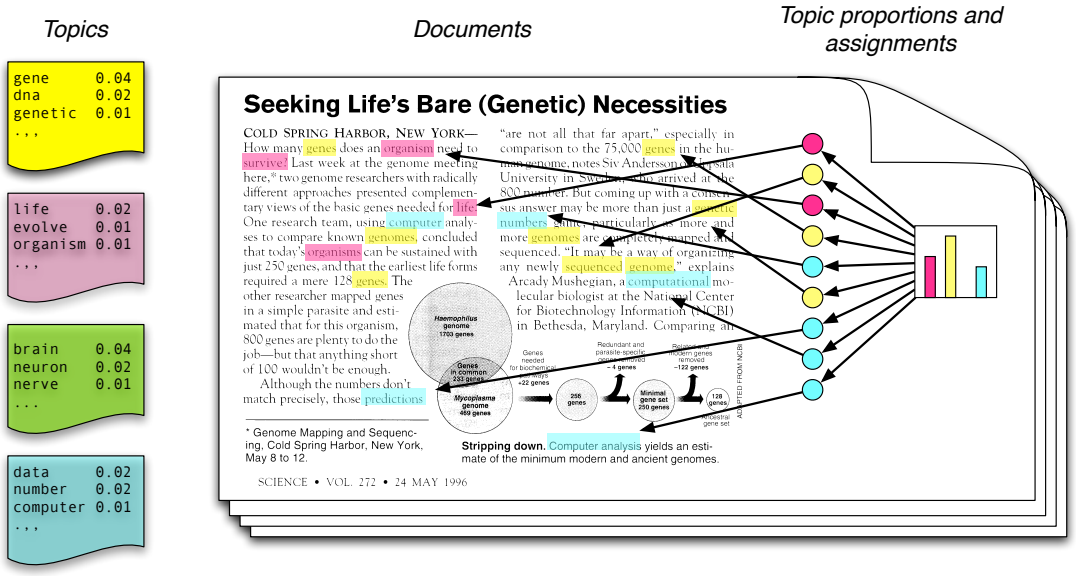


Figure 2-3: A document is a mixture of topics, each of which is in turn a mixture of words [29].

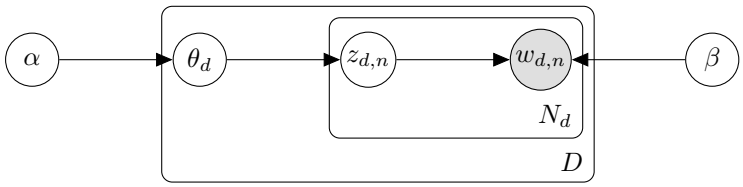


Figure 2-4: LDA graphical model in the plate notation, where circles, shaded circles, plates, and arrows stand for latent random variables or parameters, observed random variables or parameters, replicates, and statistical dependencies, respectively.

Graphical Model & Generative Process

Consider a V-term dictionary and a K-topic corpus that contains D documents, each N_d words long. The dth document's nth word $w_{d,n}$ (observable) is jointly dependent on (1) a topic $z_{d,n}$ drawn from a multinomial distribution θ_d, which itself is sampled once per document from $Dir(\alpha)$, an exchangeable Dirichlet distribution with a scalar parameter α, and (2) a topic-vocabulary probability matrix β, whose element $\beta_{k,v}$ is the probability of kth topic recruiting vth term in the vocabulary dictionary[7].

LDA's generative process for an N-word document is stated as follows.

1. Choose the topic mixture $\theta \sim Dir(\alpha)$.
2. For each of the N words, independently
   (a) choose a topic $z_n \sim Mult(\theta)$;
   (b) choose a word $w_n \sim p\left(w_n \mid z_n, \beta\right)$.

Posterior Inference & Parameter Estimation

See Appendix A for the full derivations of the variational inference and model parameter estimation. Only the procedural outline is given in this section.

[7]Hence, every row of β sums to 1.

The inference task is, for each document, finding the posterior distribution of the hidden variables (the topic mixture θ and word-level topics $\mathbf{z}$) given the document (observed words $\mathbf{w}$), i.e.,

$$p\left(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta\right) = \frac{p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)}{p\left(\mathbf{w} \mid \alpha, \beta\right)}.$$

Note that the document subscript is dropped for simplicity, since we are considering only one document. Parametrized by α and β, the numerator is given by

$$\begin{aligned}
p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right) &= p\left(\theta \mid \alpha\right) p\left(\mathbf{z}, \mathbf{w} \mid \theta, \beta\right) = p\left(\theta \mid \alpha\right) \prod_{n=1}^{N} p\left(z_n, w_n \mid \theta, \beta\right) \\
&= p\left(\theta \mid \alpha\right) \prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n, \beta\right) = p\left(\theta \mid \alpha\right) \prod_{n=1}^{N} \theta_{z_n} \beta_{z_n, w_n} \\
&= p\left(\theta \mid \alpha\right) \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \left(\theta_k \beta_{k,v}\right)^{\mathbb{1}_w(n,v) \mathbb{1}_z(n,k)} \right) \\
&= \left(\frac{\Gamma\left(K\alpha\right)}{\Gamma^K\left(\alpha\right)} \prod_{k=1}^{K} \theta_k^{\alpha-1} \right) \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \left(\theta_k \beta_{k,v}\right)^{\mathbb{1}_w(n,v) \mathbb{1}_z(n,k)} \right),
\end{aligned}$$

where $\mathbb{1}_w(n, v)$ and $\mathbb{1}_z(n, k)$ are indicator functions such that

$$\mathbb{1}_w(n, v) = \begin{cases} 1, w_n \text{ is the } v\text{th dictionary term;} \\ 0, \text{otherwise.} \end{cases} \qquad \mathbb{1}_z(n, k) = \begin{cases} 1, z_n \text{ is the } k\text{th topic;} \\ 0, \text{otherwise.} \end{cases}$$

The denominator, also referred to as evidence, is obtained by marginalizing out θ and $\mathbf{z}$ in $p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)$.

$$\begin{aligned}
p\left(\mathbf{w} \mid \alpha, \beta\right) &= \int \sum_{\mathbf{z}} p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right) \mathrm{d}\theta \\
&= \int \sum_{\mathbf{z}} \left(\frac{\Gamma\left(K\alpha\right)}{\Gamma^K\left(\alpha\right)} \prod_{k=1}^{K} \theta_k^{\alpha-1} \right) \prod_{n=1}^{N} \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \left(\theta_k \beta_{k,v}\right)^{\mathbb{1}_w(n,v) \mathbb{1}_z(n,k)} \right) \mathrm{d}\theta
\end{aligned}$$

$$= \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \int \left(\prod_{k=1}^{K} \theta_k^{\alpha-1} \right) \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_k \beta_{k,v})^{\mathbb{1}_w(n,v)} \right) \mathrm{d}\theta,$$

a function that is intractable because of the coupling between θ and β. As a result, we resort to variational (instead of exact) inference techniques, specifically the **variational expectation-maximization (EM) algorithm**, to approximate the true posterior distribution $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)$.

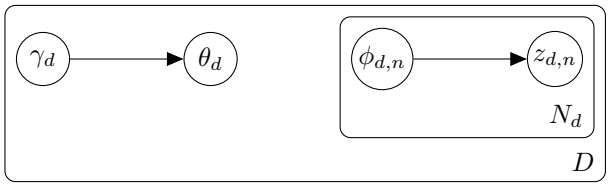


Figure 2-5: Variational distribution used to approximate the actual LDA posterior.

By removing the problematic edges in the original graphical model, we construct a simpler variational distribution $q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$, where the Dirichlet parameter γ and the multinomial parameter ϕ are the free variational parameters (see Figure 2-5). We then optimize γ and ϕ such that the constructed distribution is close in Kullback-Leibler (KL) divergence to the true posterior. That is,

$$(\gamma^*, \phi^*) = \underset{(\gamma,\phi)}{\operatorname{argmin}} D_{KL}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).$$

This optimization problem can be proven equivalent to maximizing the log likelihood's lower bound $\mathcal{L}(\gamma, \phi \mid \alpha, \beta)$ w.r.t. γ and ϕ. That is,

$$\begin{aligned}(\gamma^*, \phi^*) &= \underset{(\gamma,\phi)}{\operatorname{argmin}} D_{KL}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)) = \underset{(\gamma,\phi)}{\operatorname{argmax}} \mathcal{L}(\gamma, \phi \mid \alpha, \beta) \\ &= \underset{(\gamma,\phi)}{\operatorname{argmax}} E_q\{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} - E_q\{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\}.\end{aligned}$$

Solving the optimization problem yields the following update equations for the

variational parameters γ and ϕ

$$\begin{cases} \phi_{n,k} \propto \beta_{k,v} \exp\left(E_q\left\{\log\left(\theta_k\right) \mid \gamma\right\}\right) \\ \gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{n,k}. \end{cases}$$

Interestingly, the two update equations have intuitive interpretations: the Dirichlet parameter γ_k is updated by adding the expected occurrences of the kth topic among the observed N words; the update of the multinomial parameter $\phi_{n,k}$ resembles $p(z_n \mid w_n) \propto p(w_n \mid z_n)p(z_n)$.

After tightening the lower bound by adjusting γ and ϕ (E-step of the variational EM algorithm), we now fix γ and ϕ (i.e., the variational distributions) and maximize the lower bound w.r.t. the model parameters α and β (M-step of the variational EM algorithm).

For more details on inference and estimation, see Appendix A. The variational EM algorithm for LDA is summarized in pseudocode as follows.

Algorithm 1 Variational EM Algorithm for LDA

INPUT: Topic number K; D documents
OUTPUT: Approximated posterior $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)$; model parameters α and β
ALGORITHM:

▷ Initializing model parameters
$\alpha := 50/K$
$\beta_{k,v} := 1, \forall k \leq K, \forall v \leq V$

repeat

▷ E-step: tightening the lower bound of the log likelihood
for each document d **do**
▷ Initializing variational parameters
$\phi^0_{d,n,k} := 1/K, \forall n \leq N_d, \forall k \leq K$
$\gamma^0_{d,k} := \alpha + N_d/K, \forall k \leq K$
repeat
for each word n **do**
for each topic k **do**
$\phi^{t+1}_{d,n,k} := \beta_{k,w_n} \exp\left(\Psi(\gamma^t_{d,k})\right)$
normalize $\sum_{k=1}^{K} \phi^{t+1}_{d,n,k} = 1$
for each topic k **do**
$\gamma^{t+1}_{d,k} := \alpha + \sum_{n=1}^{N_d} \phi^{t+1}_{d,n,k}$
until convergence

▷ M-step: maximizing the constructed lower bound
for each topic k **do**
for each dictionary term v **do**
for each document d **do**
for each word n **do**
$\beta_{k,v} := \beta_{k,v} + \phi_{d,n,k} \mathbb{1}_w(d, n, v)$
normalize $\sum_{v=1}^{V} \beta_{k,v} = 1$
update α via the linear-time Newton-Raphson algorithm

until lower bound converged

Chapter 3

Related Work

3.1 Alzheimer's Disease Subtypes

Survey paper [10] reviews the long-standing theory of Alzheimer's disease (AD) heterogeneity and summarizes six potential AD subtypes from various literature. As the prototype, typical AD is late-onset with amnestic impairment in association with hippocampal and temporal-parietal atrophy. Also late-onset is the temporal (pure amnestic) variant that shows a slower rate of cognitive decline clinically and, pathologically, has the plaques and neurofibrillary tangles limited to the limbic regions with little or no spread to the neocortical areas. Left (language) variant is an early-onset AD variant typified by non-fluent speech with agrammatism and phonemic paraphasia. It differs from the language symptom of the late-stage typical AD, which is generally fluent in nature. Another similar language-related variant is logopenic progressive aphasia whose speech has the grammar and articulation preserved but suffers from impaired repetition. Frontal (executive) variant is a rare early-onset variant with prominent apathy, loss of empathy, and socially inappropriate behaviors. Finally, the right (visuoperceptive) variant patients experience difficulty with visually guided tasks and possess subtle greater-than-left right temporal and parietal atrophy.

In post-mortem study [2], Murray et al. categorize AD into three pathological subtypes—hippocampal-sparing (HpSp), limbic-predominant (LP), and typical AD. Compared with typical AD, HpSp has more neurofibrillary tangles in the cortex and fewer in the hippocampus, whereas the opposite pattern is observed in LP. The categorization is done by a quartile classification of the ratio between average hippocampal to cortical neurofibrillary tangle count (by domain knowledge, hence supervised). The median counts of neurofibrillary tangles are then used to further offset any over- or underestimation.

Continuing from [2], Whitwell and colleagues confirm in [30] that MRI atrophy patterns indeed differ across these pathologically defined AD subtypes. More specifically, they compute by VBM the head size-corrected gray matter (GM) volumes for hippocampus and three association cortices (lateral frontal, temporal, and parietal). It is found that the ratio of hippocampal to cortical volumes produces the most significant differences and hence the best discrimination between groups. This study thus proves the correlation between neurofibrillary tangles (in pathology) and volume loss (in neuroimaging). Note, however, [30] does not define the subtypes with MRI, but rather presents the neuroimaging manifestation of the pathological subtypes in [2].

Noh et al. make the classification unsupervised by performing Ward's hierarchical clustering on the cortical thickness data [3] and discover three subtypes—medial temporal, parietal dominant, and diffuse atrophy. Although the classification is unsupervised, it does not account for mixed membership—a subject may be a mixture of several subtypes. Furthermore, it considers only the cortical region, but atrophy in the subcortical region, such as cerebellum, does provide some subtype discrimination as shown in Chapter 5.

3.2 Topic Modeling in Neuroscience

So far, there are very few works that tackle neuroscience problems by topic modeling, and they are all very recent (in late 2014). One such study is [31], where Koch et al. discover six latent sleep states by LDA.

As an extension of LDA, the author-topic model [32] is proposed to include author information as an additional layer on top of the original model. Utilizing this author-topic model, Yeo et al. explore the flexibility and specialization of the human association cortex by viewing experiments as documents, tasks as authors, cognitive components as topics, voxels as vocabulary, and activation foci as words [33].

Chapter 4

Methods

4.1 Our Model

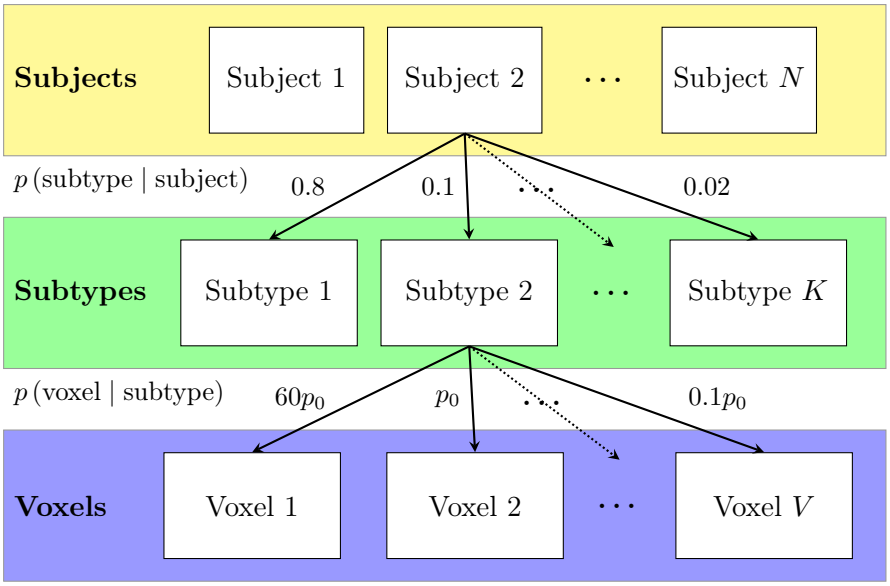


Figure 4-1: The relationships among subjects, AD subtypes, and implicated voxels. Since it is the relative value that matters in p (voxel | subtype), p_0 is adopted as the "base probability" for clarity.

Similar to the latent Dirichlet allocation (LDA) model, our model is also structured in a hierarchical manner as in Figure 4-1. More specifically, instead of hard-assigning a subject to one AD subtype, we consider each subject is a mixture of

subtypes (i.e., mixed membership), each of which is in turn a mixture of atrophied voxels. For example in Figure 4-1, Subject 2's top two prominent subtypes are Subtypes 1 and 2; Subtype 2 most likely implicates Voxel 1. Hence, in our problem, to be estimated are each subject's subtype mixture $p\,(\text{subtype} \mid \text{subject})$ and each subtype's atrophy pattern $p\,(\text{voxel} \mid \text{subtype})$.

The reasonableness of our model is two-fold. First, the mixed membership accounts for a subject possessing several subtypes simultaneously and therefore allows us to consider mixed effects. Second, in our model, one subtype implicates many voxels, and conversely, several subtypes may be responsible for a voxel's atrophy—this nicely captures the actual scenario in the disease.

4.2 Data Acquisition

Data used in this thesis were obtained from the **Alzheimer's Disease Neuroimaging Initiative (ADNI)**[1]. The ADNI was launched in 2003 as a database for researchers to study whether neuroimaging techniques, biomarkers, and neuropsychological assessments can be fused to provide better AD understandings. The subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 subjects, but ADNI has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 55 to 90-year-old participants consisting of cognitively normal (CN) individuals, people with mild cognitive impairment (MCI), and AD subjects. The follow-up examination dates of each subject are specified in the ADNI-1, ADNI-2, and ADNI-GO protocols. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2.

The AD subtypes are learned from 188 AD subjects at their respective base-

[1] http://adni.loni.usc.edu

lines in the ADNI-1 cohort. Their longitudinal follow-up data, whenever available, are also analyzed for validation purpose as to be shown in Chapter 5. Note that although only the 188 AD subjects are used for machine learning, yet the whole cohort, additionally including 228 CN subjects and 394 MCI subjects, is used in data preprocessing, such as VBM (to create our VBM template so as to avoid template bias) and linear regression.

4.3 Data Preprocessing

We preprocess the raw MR images so that they resemble text documents, which in turn serve as LDA's input.

4.3.1 Voxel-based Morphometry

As outline in Section 2.2.2, the FSL-VBM pipeline[2] is adopted. Hence, we first reorder the voxels and adjust the header information so that the images conform to FSL's convention. After this, the MR images are ready for VBM.

In brief, the MR images are first brain-extracted [20] and gray matter (GM)-segmented [21] before being affinely registered [22][23] to the MNI152 standard space. The resulting images are averaged, flipped around the y-axis, and then re-averaged to create a left-right symmetric, first-pass GM template. The same procedure is then repeated but this time with non-linear registration [24][25], producing the final customized GM template. Next, all native GM images are non-linearly registered to this final template and modulated by the Jacobians of the warp field to correct for local expansion (or contraction) due to the non-linear component of the spatial transformation. Finally, the modulated GM images are smoothed with an isotropic Gaussian kernel with its full width at half maximum

[2] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fslvbm

(FWHM) being 10 mm. Up to this point, the lower a voxel's intensity is, the more atrophied that voxel is.

In fact, the whole VBM procedure includes another final and probably the most important step, voxel-wise statistical analysis, that displays which voxels are significantly different across groups. This step is achieved by the second use of GLM as illustrated in Section 2.3.1. Yet, this step is only used for VBM sanity check in this project and thus not shown here.

4.3.2 Partialing Out the Effects by Nuisance Variables

Before the linear regression, we take $\log_{10}(\cdot)$ of the modulated GM density from the previous step and obtain a stretched range, wherein a more negative value means severer atrophy.

As detailed in Section 2.3.1, we can utilize a general linear model (GLM) to partial out the effects by the nuisance variables. In this project, since we are only concerned with the disease-related atrophy, age, gender, and intracranial volume (ICV) are the nuisance variables. Intuitively, aging partially accounts for brain atrophy; men tend to have larger brain volume than women, which may confuse women's brains as more atrophied when registered to the same template; the same logic applies to ICV.

Implementation-wise, for our 810 subjects, we construct our GLM as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{810} \end{pmatrix} = \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} 1 & a_1 & g_1 & v_1 \\ 1 & a_2 & g_2 & v_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{810} & g_{810} & v_{810} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{810} \end{pmatrix},$$

where y is the logarithm of GM density, a is the demeaned age, g is the demeaned gender, and v is the demeaned ICV. We then find the least squares solution of $\hat{\boldsymbol{\beta}}$ (and the corresponding $\hat{\boldsymbol{\epsilon}}$), with which we partial out age, gender, and ICV

$$\begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_{810} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_{810} \end{pmatrix}.$$

4.3.3 z-normalization, Thresholding, & Quantization

For each voxel, we perform z-normalization across all the subjects (including CN and MCI subjects) and threshold any value above zero as zero. By far, each voxel either carries a zero value (considered as no atrophy) or a negative value. Then, we multiply the values by -10 so that now a larger (positive) number presents severer atrophy. Finally, we make them integer counts by taking the floor function.

Thus far, each subject's every voxel carries a nonnegative integer, which we name as "atrophy count". For instance, a voxel's atrophy count of 10 means that that voxel is 1 standard deviation below (more atrophied than) that voxel's cohort mean (all in the log-space).

4.4 Learning the Subtypes

Now that each voxel has an atrophy count, we could translate LDA to our problem in the following fashion (also see Table 4.1). Consider a cohort of D brains, each of which is composed of N_d atrophied voxels. In the cohort lies a total of K hidden atrophy topics, each formed by possibly overlapping subsets of the V MNI152[3] voxels. Under this mapping, the number of times that a word appears

[3]MNI152 is a standard space defined by the Montreal Neurological Institute.

(i.e., word count) corresponds to the atrophy level of a voxel.

Table 4.1: Translating the original LDA model to our problem.

From	corpus	dictionary words	topic	document	word count
To	cohort	MNI152 voxels	subtype	brain	atrophy level

Now, the dth brain's nth atrophied voxel $w_{d,n}$ (observable) is jointly dependent on (1) an atrophy topic $z_{d,n}$ drawn from a multinomial distribution θ_d, which itself is sampled once per brain from $Dir(\alpha)$, an exchangeable Dirichlet distribution with a scalar parameter α, and (2) a subtype-voxel probability matrix β, whose element $\beta_{k,v}$ is the probability of kth subtype implicating vth voxel of the MNI152 voxels[4].

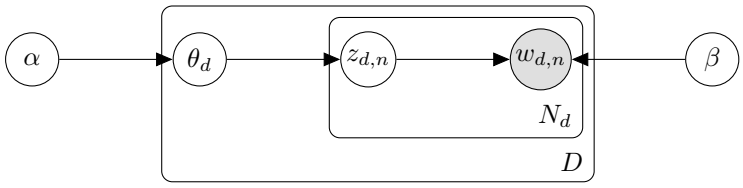


Figure 4-2: The LDA model revisited.

Similar to the original LDA model, the generative process producing N atrophied voxels given α and β can be summarized as below.

1. Choose the atrophy topic mixture $\theta \sim Dir(\alpha)$.
2. For each of the N atrophied voxels, independently
   (a) choose an atrophy topic $z_n \sim Mult(\theta)$;
   (b) choose an atrophied voxel $w_n \sim p(w_n \mid z_n, \beta)$.

Using the variational expectation-maximization (EM) algorithm detailed in Appendix A, we can infer the posteriors and estimate the model parameters from

[4]Hence, every row of β sums to 1.

the 188 AD subjects' baseline VBM output. By iterating between the E-step and M-step until convergence, we obtain the approximate posterior and the model parameters. More specifically, we compute a subject's ith subtype probability p_i as

$$p_i = \frac{\gamma_i}{\sum_{k=1}^{K} \gamma_k},$$

the ith element of the expectation of $Dir(\gamma)$. To obtain the probabilities of this subtype implicating different voxels, i.e., $p(\text{voxel} \mid \text{subtype})$, we simply take the i-th row of the estimated β matrix.

Since EM algorithm is susceptible to local optima, we randomly initialize 20 runs, among which we choose the one with maximum likelihood as our final estimation. Our inspection is that all the runs appear to be almost the same.

We pick the run with maximum likelihood as our final estimation among the 20 random initializations. As a result, we know the subtype composition for each subject. Furthermore, given a subtype, we can statistically pinpoint which voxels are likely to be implicated.

Chapter 5

Results & Discussions

5.1 Two Subtypes: Subcortical & Cortical

Setting the desired number of subtypes $K = 2$, we obtain two subtypes, which we name as subcortical (S) and cortical (Co) subtypes. As shown in Figure 5-1, S mainly implicates the subcortical structures, such as hippocampus and cerebellum, whereas Co largely affects the cortical region.

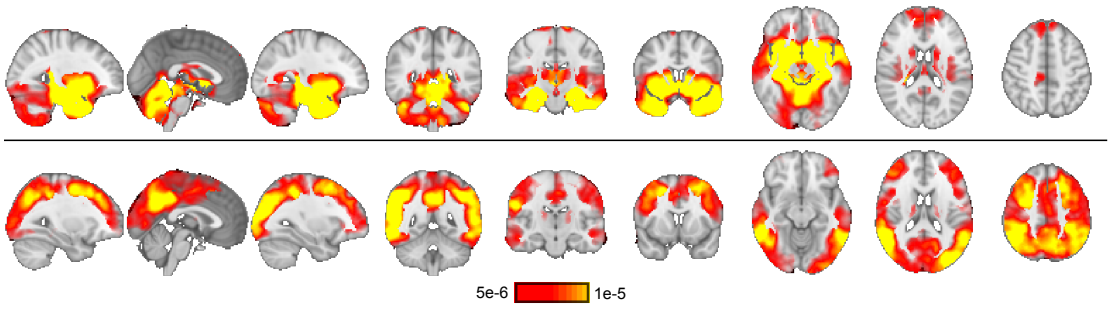


Figure 5-1: Atrophy patterns of the two subtypes discovered: subcortical (S, top) and cortical (Co, bottom). Each 3D subtype volume is presented as three sagittal, three coronal, and three axial slices (in order, from left to right), organized in one row. Heat map indicates which voxels are more likely to be atrophied given a certain subtype, i.e., $p(\text{voxel} \mid \text{subtype})$.

5.1.1 Age at Baseline

It has been reported in [2] that the age at onset is different for different subtypes: hippocampal-sparing (our Co) < typical < limbic-predominant (our S) with $p < 0.001$. We attempt to replicate the same trend with patients' baseline age. Although our baseline age is not exactly the age at onset, it still provides a good approximation, especially given the great difficulty of measuring the exact age at onset in practice. Since our model expresses subjects as subtype mixtures, instead of categorizing them deterministically, we consider the correlation between the baseline age and the probability of a subtype (say, Co), rather than manually split the subjects into subtype groups for group comparison.

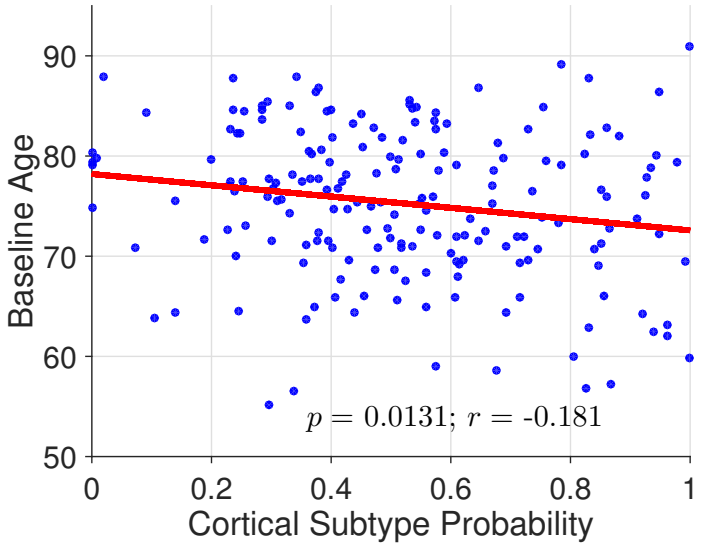


Figure 5-2: Correlation between age and AD subtype. The red line is the regression line by GLM; p is for the null hypothesis $\mathcal{H}_0$: subtype and age are uncorrelated, controlling for total atrophy; r is the partial correlation coefficient controlling for total atrophy.

Raw age and subtype probability are presented as blue datapoints in Figure 5-2; the p-value and partial correlation coefficient r are obtained while controlling for total atrophy, which is represented by the total log GM amount. As the figure indicates, the Co group is younger ($p < 0.05$) than the S group, consistent with [2].

5.1.2 Education

We also detect a difference in education between different subtypes. Using the same analysis as above, we test against the null hypothesis that education is the same between the two subtypes. As shown in Figure 5-3, we reject the null hypothesis at 5% and conclude that the Co group has a higher education than the S group. A possible explanation for the phenomenon is that the more education a man receives, the more often he repeatedly trains his hippocampus, the part in charge of converting short-term memories into long-term ones. As a result, his hippocampus gets increasingly resilient to AD atrophy, causing him to develop Co (instead of H) AD, if ever.

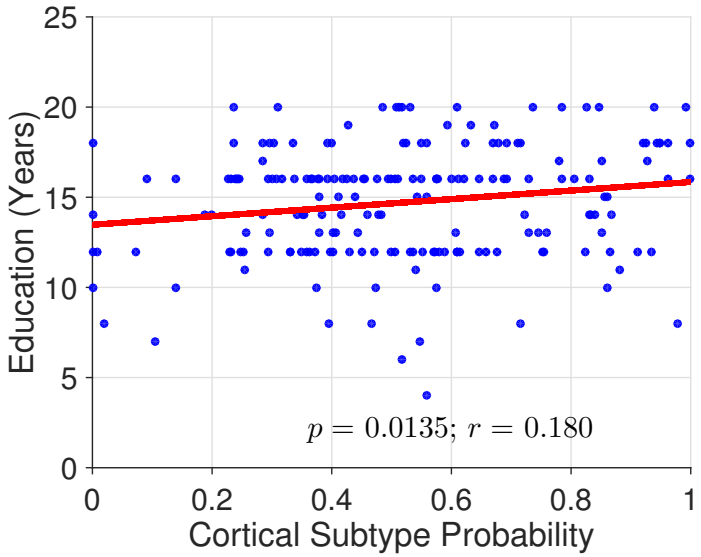
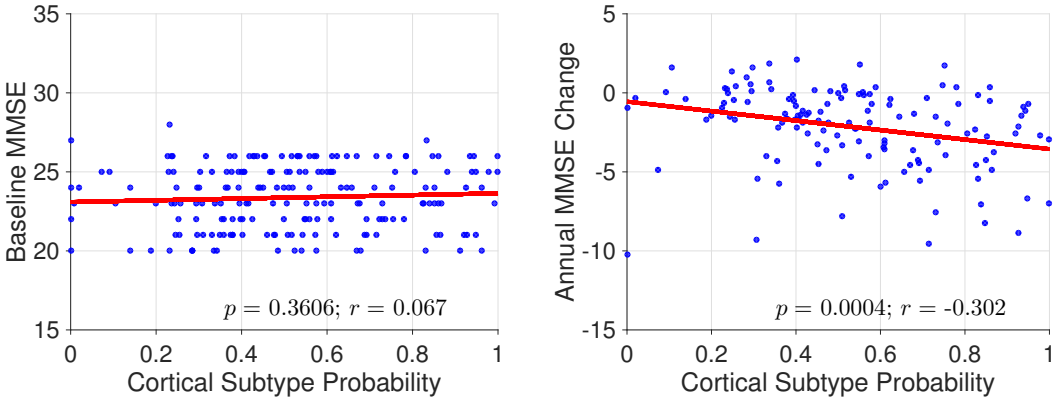


Figure 5-3: Correlation between education and AD subtype. The red line is the regression line by GLM; p is for the null hypothesis $\mathcal{H}_0$: subtype and education are uncorrelated, controlling for total atrophy; r is the partial correlation coefficient controlling for total atrophy.

5.1.3 Mini-mental State Exam

We now validate the two subtypes by proving their great disparity in the disease progression. Let us first describe the disease progression with mini-mental state

exam (MMSE), a widely-used 30-point questionnaire that measures cognitive impairment. We adopt the same analysis method as before. In Figure 5-4a and the subsequent figures, the "annual change" is computed as the slope of the fitting line for all (at least four) longitudinal datapoints.



(a) Baseline MMSE

(b) Longitudinal MMSE

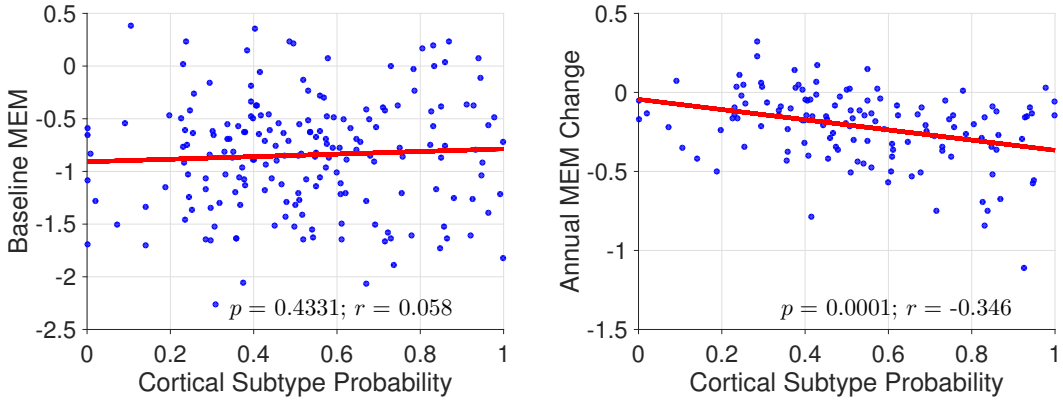
Figure 5-4: Correlations between MMSE and AD subtype. The red line is the regression line by GLM; p is for the null hypothesis $\mathcal{H}_0$: subtype and quantity of interest are uncorrelated, controlling for total atrophy; r is the partial correlation coefficient controlling for total atrophy.

As Figure 5-4a shows, there is no difference ($p = 0.36$) between S and Co groups in their baseline MMSE. However, in the longitudinal course, Co declines faster ($p < 0.001$) than S, as Figure 5-4b demonstrates. This suggests that Co is a more fast-deteriorating subtype than S.

5.1.4 Memory & Executive Function

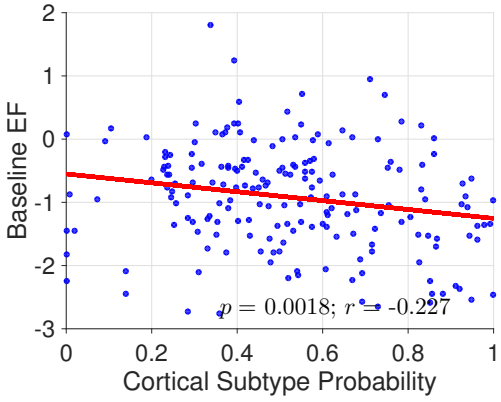
MMSE can be inherently noisy and unstable, possibly leading to a rise even when the subject remains at the same dementia level. Consequently, the computed annual change, i.e., the slope, may not be very accurate. As such, researchers have derived from MMSE and other tests a composite score for memory (MEM) [18] that is perceived to be more sensitive. In addition, they devise another composite score for executive functioning (EF) [17] to test particularly executive functioning.

We now analyze the differences between the subtypes in these two composite scores.

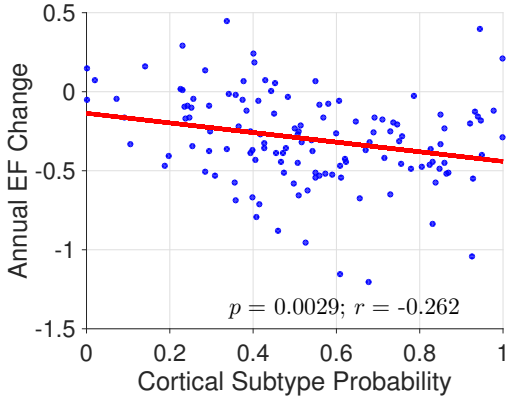


(a) Baseline MEM

(b) Longitudinal MEM



(c) Baseline EF



(d) Longitudinal EF

Figure 5-5: Correlations between memory, executive function scores and AD subtype. The red line is the regression line by GLM; p is for the null hypothesis $\mathcal{H}_0$: subtype and quantity of interest are uncorrelated, controlling for total atrophy; r is the partial correlation coefficient controlling for total atrophy.

We find that Co declines faster than S in both memory ($p = 0.0001$, Figure 5-5b) and executive function ($p = 0.0029$, Figure 5-5d); Co is worse in executive function at the start ($p = 0.0018$, Figure 5-5c). Yet, there is no significant difference in baseline memory between the two subtypes ($p = 0.43$, Figure 5-5a).

As Figure 5-5b shows, MEM indeed gives a stronger, presumably less noisy,

trend ($r = -0.35$) than MMSE. The same trend in EF annual change confirms that Co is a more fast-deteriorating subtype than S in both memory and executive function.

5.2 Three Subtypes: Hippocampal, Cerebellum, & Cortical

As we move from two subtypes to three subtypes, S further splits into two subtypes—hippocampal (H) and cerebellum (Ce), whereas Co remains (Figure 5-6). As their names suggest, H, Ce, and Co mostly implicates hippocampus, cerebellum, and cortical region, respectively.

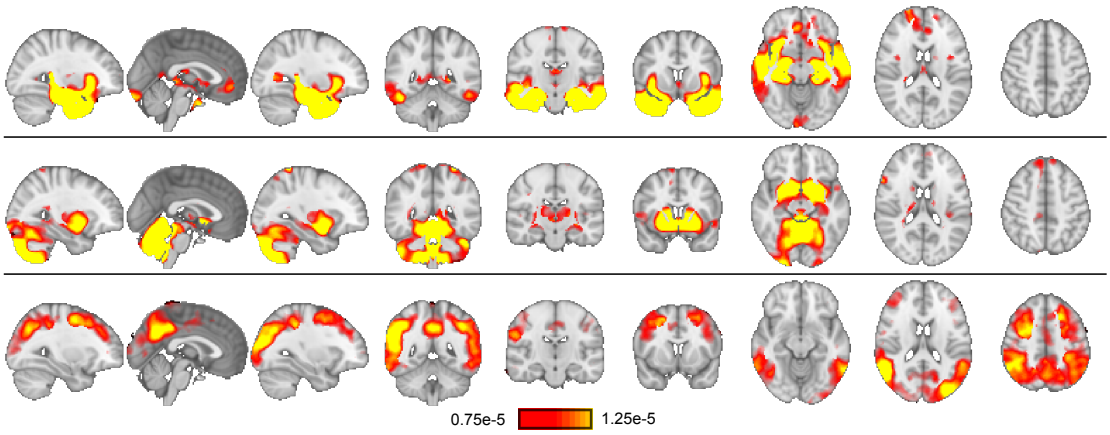


Figure 5-6: Atrophy patterns of the three subtypes discovered: hippocampal (H, top), cerebellum (Ce, middle), and cortical (Co, bottom). Each 3D subtype volume is presented as three sagittal, three coronal, and three axial slices (in order, from left to right), organized in one row. Heat map indicates which voxels are more likely to be atrophied given a certain subtype, i.e., $p(\text{voxel} \mid \text{subtype})$.

Since one more subtype may provide us with higher resolution (hence, better discrimination), we now re-examine age, education, MMSE, MEM, and EF. However, in three subtypes, we can no longer express the subtype composition with one axis as in two subtypes. Therefore, for each subtype, we create 1000 populations by bootstrapping our current 188 AD subjects and then, for each

population, compute its subtype-weighted average as its bootstrap statistic. So for each subtype, we have 1000 bootstrap statistics, which are the empirical "expected values of the quantity of interest" for this subtype. All of the bootstrap statistics are presented by subtype in the subsequent standard box plots.

Because we can generate an arbitrarily large number of populations by bootstrapping, it makes no sense to compute the p-value from these artificial populations, say, with analysis of variance (ANOVA). So all the p-values below are computed from GLM, for $\mathcal{H}_0$: $\beta_1 = \beta_2 = 0$ in $y = \beta_0 + \beta_1 \cdot p\,(\mathrm{H} \mid \text{subject}) + \beta_2 \cdot p\,(\mathrm{Ce} \mid \text{subject}) + \epsilon$. That is, we are testing whether the quantity of interest is independent of *all* the subtypes. Note that because $p\,(\text{subtype} \mid \text{subject})$ is a probability distribution, any subtype probability is linearly dependent on the other two, i.e., $p\,(\mathrm{H} \mid \text{subject}) + p\,(\mathrm{Ce} \mid \text{subject}) + p\,(\mathrm{Co} \mid \text{subject}) = 1$. Therefore, no matter which two subtype probabilities we put as explanatory variables in our GLM, the p-value will be the same.

5.2.1 Age at Baseline

In the case of age, having one more subtype does not seem to have provided powerful discrimination (see Figure 5-7). Although S further splits into H and Ce, H and Ce are statistically the same in age. Thus, our conclusion in age remains the same: the onset of Co AD tends to be earlier than other subtypes.

5.2.2 Education

Similar to age, one more subtype does not further distinguish the S group, because education is statistically the same between the H and Ce groups (see Figure 5-8). Same as before, the Co group receives longer education.

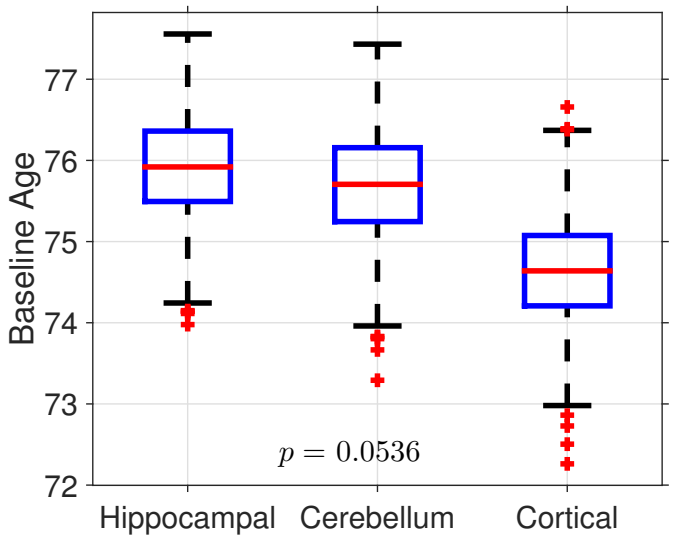


Figure 5-7: Expected age of each AD subtype shown in standard box plot, where the first Q_1 quartile and third quartile Q_3, median, $1.5\times(Q_3-Q_1)$, and outliers are indicated respectively by the blue rectangular, red horizontal bar, black whiskers, and red crosses.

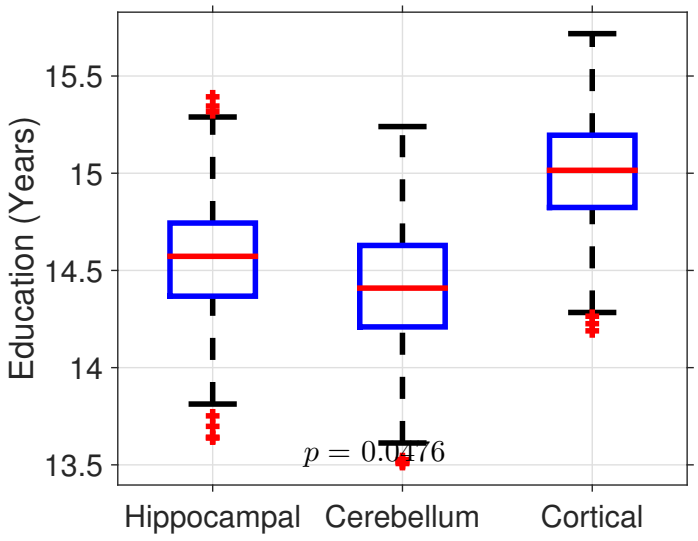
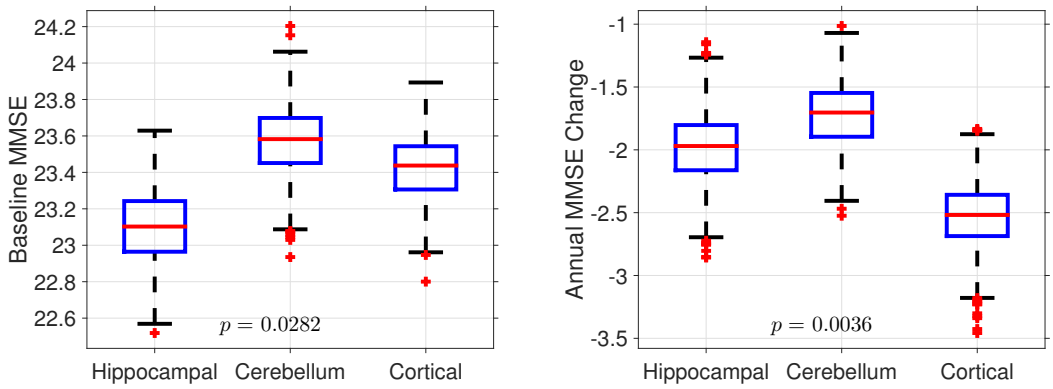


Figure 5-8: Expected education (years) of each AD subtype shown in standard box plot, where the first Q_1 quartile and third quartile Q_3, median, $1.5\times(Q_3-Q_1)$, and outliers are indicated respectively by the blue rectangular, red horizontal bar, black whiskers, and red crosses.

5.2.3 Mini-mental State Exam

Moving from two subtypes to three provides great separability in MMSE within the S group. When we only have two subtypes, we fail to detect any difference in baseline MMSE, despite that H and Ce in fact have significantly different baseline MMSE as shown in Figure 5-9a. When combined as S in the case of two subtypes, H and Ce simply balance each other out, which does not happen any more when we allow for three subtypes.



(a) Baseline MMSE

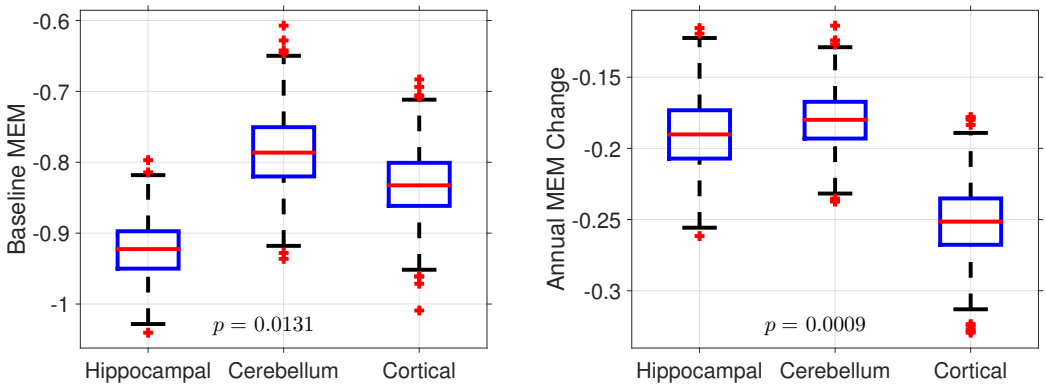
(b) Longitudinal MMSE

Figure 5-9: Expected baseline MMSE and annual decline of each AD subtype shown in standard box plot, where the first Q_1 quartile and third quartile Q_3, median, $1.5 \times (Q_3 - Q_1)$, and outliers are indicated respectively by the blue rectangular, red horizontal bar, black whiskers, and red crosses.

Figure 5-9a conveys that at baseline, the H group has an expected MMSE as low as 23.1, whereas the Ce group has a higher MMSE, around 23.6. In terms of MMSE decline rate, Co remains to be the fastest at around -2.5 points per year. Ce seems to be the mildest subtype that starts off better and declines slowly.

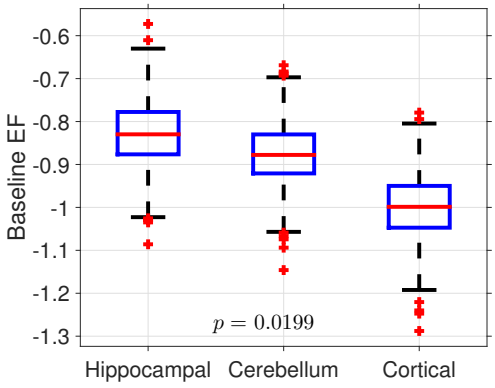
5.2.4 Memory & Executive Function

Similar separation is also observed in baseline MEM (Figure 5-10a), which again demonstrates the increased resolution by having one more subtype. Just as in two subtypes, MEM shares the same trend with MMSE, but provides stronger

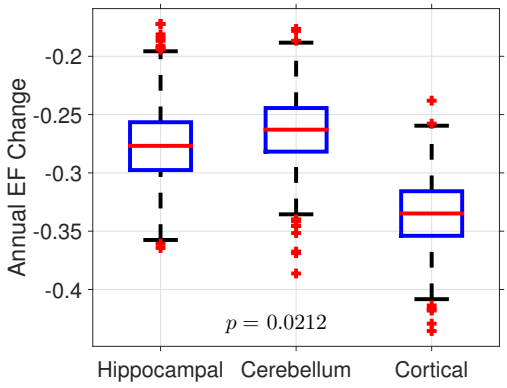


(a) Baseline MEM

(b) Longitudinal MEM



(c) Baseline EF



(d) Longitudinal EF

Figure 5-10: Expected baseline memory, executive function and their annual changes of each AD subtype shown in standard box plot, where the first Q_1 quartile and third quartile Q_3, median, $1.5 \times (Q_3 - Q_1)$, and outliers are indicated respectively by the blue rectangular, red horizontal bar, black whiskers, and red crosses.

separation. In both MMSE and MEM, H starts off the worst among the three subtypes, but Co declines faster in progression (Figure 5-10b). Co also declines fastest in executive function (Figure 5-10d). Furthermore, Co is already worse than Ce and H in executive function at the baseline (Figure 5-10c).

The decoupling between memory and executive function at the baseline is also interesting (Figure 5-10a, 5-10c): although H is worst in memory, yet it is better in executive function. This proves that our model is not merely splitting up groups

by disease severity, which, if true, will manifest itself with a subtype that is worst in both memory and executive function. Hence, these are real subtypes that may fall into different disease stages.

5.3 Predicting Conversion to Alzheimer's Disease

Because we include CN and MCI subjects in preprocessing, they also have atrophy counts just as the AD subjects do. We therefore can apply our learned model to infer the baseline-MCI subjects' posterior distributions. This amounts to examining whether their *pre*-AD subtype compositions can foretell anything. Since MCI is the prodromal stage of AD, a natural idea is to see if we can utilize MCI subjects' pre-AD subtype compositions inferred by our model to predict the possible future conversion into AD.

The tricky part is, however, that ascertaining whether an MCI subject converts or not is sometimes impossible given the data available. For instance, a subject may have already converted into AD, but has dropped out of the study, making his conversion record missing. Because of this limitedness, we need to define "convert" or "nonconvert" within a certain time window. In this thesis, we define a convert as a subject whose conversion (assigned to value 1) is seen within all the available data and a nonconvert (assigned to value 0) as a subject whose conversion is never seen within all the data, which itself has to span at least two years.

After inferring the 394 MCI subjects with our model, we compute each subtype's "expected conversion" by exactly the same bootstrap technique as in previous section. The results are shown in Figure 5-11. Note that because we have 198 converts and 131 nonconverts, the expected conversion values of *all* the subtypes will be skewed away from 0 towards 1. Therefore, it is their *relative* positions that carry useful information. As shown, MCI subjects with more H component

is more likely to convert into AD ($p = 0.0041$).

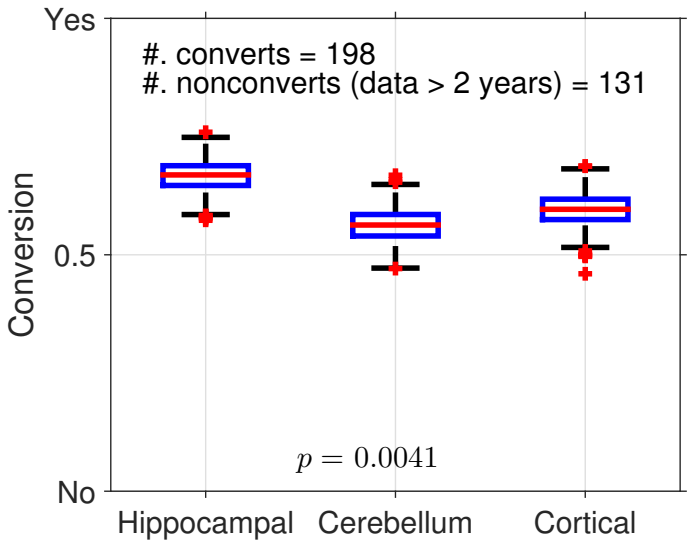


Figure 5-11: Expected conversion of each pre-AD subtype shown in standard box plot, where the first Q_1 quartile and third quartile Q_3, median, $1.5 \times (Q_3 - Q_1)$, and outliers are indicated respectively by the blue rectangular, red horizontal bar, black whiskers, and red crosses.

To conclude, among the three subtypes that we discover, Cortical Subtype (Co) receives the longest education and has the earliest age at onset. In terms of memory, Hippocampal Subtype (H) is the worst at baseline, but Cortical Subtype (Co) declines fastest during the disease progression. As for executive function, Cortical Subtype (Co) not only starts off worst, but also declines faster than Hippocampal Subtype (H) and Cerebellum Subtype (Ce). Overall, Cerebellum Subtype (Ce) seems to be a mild subtype that deteriorates slowly. Moreover, MCI of Hippocampal Subtype (H) is more likely to convert to AD than those with the other two subtypes.

Chapter 6

Conclusion & Future Work

In this thesis, we address the problem of unsupervisedly discovering Alzheimer's disease (AD) subtypes from a huge amount of high-dimensional magnetic resonance (MR) images. Specifically, we model each AD patient as mixture of AD subtypes, each of which is in turn a mixture of atrophied voxels. We first quantify voxel-wise atrophy with voxel-based morphometry (VBM) and then learn the hidden subtypes as well as subtype-voxel distributions by topic modeling. As a result, we discover three subtypes—hippocampal (H), cerebellum (Ce), and cortical (Co)—that, as their names suggest, largely implicate the hippocampus, cerebellum, and cortical region, respectively.

We then validate these subtypes by showing their great disparity in different aspects both at the baseline and during the disease progression. Specifically, Co receives the longest education and has the earliest age at onset. In the memory aspect, H starts off worst, but Co deteriorates fastest during the disease development. Co also declines fastest in executive function. Therefore, Co is a relatively acute subtype that develops fast. On the contrary, Ce seems to be a mild subtype that develops slowly.

Furthermore, we demonstrate our model's usefulness in predicting the disease conversion even when the patient is still in mild cognitive impairment (MCI). After applying our learned model to the MCI subjects, we find MCI patients with more H component is more likely to convert into AD.

Our future work includes

1. finding and justifying the most appropriate number of subtypes mathematically or computationally;
2. performing case studies that instantiate the power of mixed membership;
3. analyzing the baseline and longitudinal change simultaneously under a unified framework, possibly linear mixed-effects model;
4. modeling the functional magnetic resonance imaging (fMRI), which is perceived to be more responsive to early-stage AD development than structural MRI;
5. developing a new generative model whose observation is continuous, which frees us from the ad hoc preprocessing, such as quantization; and
6. modeling the disease development as a dynamic graphical model, which may lead to a trajectory-based classification of AD subtypes.

Appendix A

Variational Inference in Latent Dirichlet Allocation

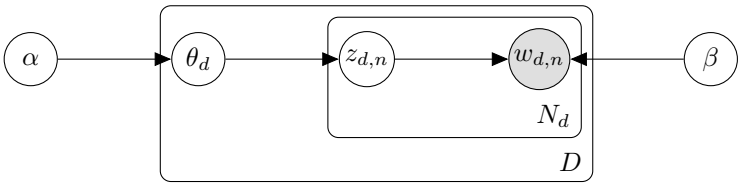


Figure A-1: LDA model revisited.

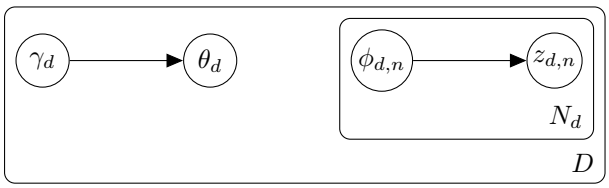


Figure A-2: Variational distribution revisited.

A.1 Constructing the Lower Bound

From Figure A-2, the variational distribution used to approximate the true posterior is factorizable as

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n).$$

The lower bound $\mathcal{L}(\gamma, \phi \mid \alpha, \beta)$ of the single-document[1] log likelihood $\log p(\mathbf{w} \mid \alpha, \beta)$ is computed using Jensen's inequality as follows

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} q(\theta, \mathbf{z} \mid \gamma, \phi) \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log E_q \left\{ \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \right\} \\
&\geq E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} - E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&\triangleq \mathcal{L}(\gamma, \phi \mid \alpha, \beta).
\end{aligned} \tag{A.1}$$

The difference between the log likelihood and its lower bound can be proven to be in fact the KL divergence between the variational distribution and the true posterior.

$$\begin{aligned}
&\log p(\mathbf{w} \mid \alpha, \beta) - \mathcal{L}(\gamma, \phi \mid \alpha, \beta) \\
&= E_q \{\log p(\mathbf{w} \mid \alpha, \beta)\} - E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} + E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&= E_q \left\{ \log \frac{p(\mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)} \right\} \\
&= E_q \left\{ \log \frac{q(\theta, \mathbf{z} \mid \gamma, \phi)}{p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)} \right\}
\end{aligned}$$

[1] This also explains why the document subscript is dropped for simplicity hereafter.

$$= D_{KL}(q\left(\theta, \mathbf{z} \mid \gamma, \phi\right) \parallel p\left(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta\right)).$$

Therefore, maximizing the lower bound is equivalent to minimizing the KL divergence $D_{KL}(q\left(\theta, \mathbf{z} \mid \gamma, \phi\right) \parallel p\left(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta\right))$. That is, the variational distribution automatically approaches to the real posterior as we maximize the lower bound.

A.2 Expanding the Lower Bound

To maximize the lower bound, we first need to spell out the lower bound $\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right)$ in terms of the model parameters (α, β) and the variational parameters (γ, ϕ). Continuing from (A.1), we have

$$\begin{aligned}
\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right) &= E_q\left\{\log p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)\right\} - E_q\left\{\log q\left(\theta, \mathbf{z} \mid \gamma, \phi\right)\right\} \\
&= E_q\left\{\log \frac{p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)}{q\left(\theta, \mathbf{z} \mid \gamma, \phi\right)}\right\} \\
&= E_q\left\{\log \frac{p\left(\theta \mid \alpha\right) p\left(\mathbf{z} \mid \theta\right) p\left(\mathbf{w} \mid \mathbf{z}, \beta\right)}{q\left(\theta \mid \gamma\right) q\left(\mathbf{z} \mid \phi\right)}\right\} \\
&= E_q\left\{\log p\left(\theta \mid \alpha\right)\right\} + E_q\left\{\log p\left(\mathbf{z} \mid \theta\right)\right\} + E_q\left\{\log p\left(\mathbf{w} \mid \mathbf{z}, \beta\right)\right\} \\
&\quad - E_q\left\{\log q\left(\theta \mid \gamma\right)\right\} - E_q\left\{\log q\left(\mathbf{z} \mid \phi\right)\right\}.
\end{aligned} \tag{A.2}$$

We now further expand each of the five terms in (A.2).

The first term is

$$\begin{aligned}
E_q\left\{\log p\left(\theta \mid \alpha\right)\right\} &= E_q\left\{\log \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right\} \\
&= E_q\left\{\log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) + \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_k - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right)\right\} \\
&= \sum_{k=1}^{K} (\alpha_k - 1) E_q\left\{\log \theta_k\right\} + \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right) \\
&= \sum_{k=1}^{K} (\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K} \gamma_i\right)\right) + \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right),
\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function. The final line is due to the following property of the Dirichlet distribution as a member of the exponential family. If $\theta \sim Dir(\alpha)$, then $E_{p(\theta|\alpha)}\{\log \theta_i\} = \Psi(\alpha_i) - \Psi(\sum_{i=1}^{K} \alpha_i)$.

The second term is

$$\begin{aligned}
E_q\{\log p(\mathbf{z} \mid \theta)\} &= E_q\left\{\log \prod_{n=1}^{N} p(z_n \mid \theta)\right\} \\
&= E_q\left\{\log \prod_{n=1}^{N}\prod_{k=1}^{K} \theta_k^{\mathbb{1}_z(n,k)}\right\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} E_q\{\mathbb{1}_z(n,k)\log\theta_k\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} E_q\{\mathbb{1}_z(n,k)\}\, E_q\{\log\theta_k\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{n,k}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right),
\end{aligned}$$

where $\phi_{n,k}$ is the probability of the nth word being produced by topic k, and $\mathbb{1}(\cdot)$ is the indicator function as defined in Section 2.3.2.

We expand **the third term** as

$$\begin{aligned}
E_q\{\log p(\mathbf{w} \mid \mathbf{z}, \beta)\} &= E_q\left\{\log \prod_{n=1}^{N} p(w_n \mid z_n, \beta)\right\} \\
&= E_q\left\{\log \prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{v=1}^{V} \beta_{k,v}^{\mathbb{1}_z(n,k)\mathbb{1}_w(n,v)}\right\} \\
&= E_q\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} \mathbb{1}_z(n,k)\mathbb{1}_w(n,v)\log\beta_{k,v}\right\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} E_q\{\mathbb{1}_z(n,k)\}\,\mathbb{1}_w(n,v)\log\beta_{k,v} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} \phi_{n,k}\mathbb{1}_w(n,v)\log\beta_{k,v}.
\end{aligned}$$

Very similar to the first term, **the fourth term** is expanded as

$$E_q\left\{\log q\left(\theta \mid \gamma\right)\right\} = \sum_{k=1}^{K}(\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) + \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\gamma_k\right).$$

Finally, **the fifth term** is expanded as

$$\begin{aligned}
E_q\left\{\log q\left(\mathbf{z} \mid \phi\right)\right\} &= E_q\left\{\log\prod_{n=1}^{N} q\left(z_n \mid \phi_n\right)\right\} \\
&= E_q\left\{\log\prod_{n=1}^{N}\prod_{k=1}^{K}\phi_{n,k}^{\mathbb{1}_z(n,k)}\right\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} E_q\left\{\mathbb{1}_z(n,k)\right\}\log\phi_{n,k} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{n,k}\log\phi_{n,k}.
\end{aligned}$$

Therefore, the fully expanded lower bound is

$$\begin{aligned}
\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right) = &\sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) + \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{n,k}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{n,k}\mathbb{1}_w(n,v)\log\beta_{k,v} \qquad \text{(A.3)} \\
&- \sum_{k=1}^{K}(\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) - \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{k=1}^{K}\log\Gamma\left(\gamma_k\right) \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{n,k}\log\phi_{n,k}.
\end{aligned}$$

A.3 Maximizing the Lower Bound

In this section, we maximize the lower bound w.r.t. the variational parameters ϕ and γ. Recall that as the maximization runs, the KL divergence between the

variational distribution and the true posterior drops (E-step of the variational EM algorithm).

A.3.1 Variational Multinomial

We first maximize Equation (A.3) w.r.t. $\phi_{n,k}$. Since $\sum_{k=1}^{K} \phi_{n,k} = 1$, this is a constrained optimization problem that can be solved by the Lagrange multiplier method. The Lagrangian w.r.t. $\phi_{n,k}$ is

$$\mathcal{L}_{[\phi_{n,k}]} = \phi_{n,k} \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) + \phi_{n,k} \log \beta_{k,v} - \phi_{n,k} \log \phi_{n,k} + \lambda_n \left(\sum_{i=1}^{K} \phi_{n,i} - 1 \right),$$

where λ_n is the Lagrange multiplier. Taking the derivative, we get

$$\frac{\partial}{\partial \phi_{n,k}} \mathcal{L}_{[\phi_{n,k}]} = \Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) + \log \beta_{k,v} - \log \phi_{n,k} - 1 + \lambda_n.$$

Setting this derivative to zero yields

$$\begin{aligned} \phi_{n,k} &= \beta_{k,v} \exp \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) + \lambda_n - 1 \right) \\ &\propto \beta_{k,v} \exp \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right). \end{aligned}$$

A.3.2 Variational Dirichlet

Now we maximize Equation (A.3) w.r.t. γ_k, the kth component of the Dirichlet parameter. Only the terms containing γ_k are retained.

$$\begin{aligned} \mathcal{L}_{[\gamma]} = &\sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) \\ &+ \sum_{n=1}^{N} \phi_{n,k} \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) \end{aligned}$$

$$- \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) - \log \Gamma \left(\sum_{i=1}^{K} \gamma_i \right) + \sum_{k=1}^{K} \log \Gamma (\gamma_k)$$

Taking the derivative w.r.t. γ_k, we have

$$\begin{aligned}
\frac{\partial}{\partial \gamma_k} \mathcal{L}_{[\gamma]} &= \left(\Psi'(\gamma_k) - \Psi' \left(\sum_{i=1}^{K} \gamma_i \right) \right) (\alpha_k - 1) \\
&+ \left(\Psi'(\gamma_k) - \Psi' \left(\sum_{i=1}^{K} \gamma_i \right) \right) \sum_{n=1}^{N} \phi_{n,k} \\
&- \left(\Psi'(\gamma_k) - \Psi' \left(\sum_{i=1}^{K} \gamma_i \right) \right) (\gamma_k - 1) - \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) \\
&- \frac{\Psi \left(\sum_{i=1}^{K} \gamma_i \right)}{\Gamma \left(\sum_{i=1}^{K} \gamma_i \right)} + \frac{\Psi(\gamma_k)}{\Gamma (\gamma_k)} \\
&= \left(\Psi'(\gamma_k) - \Psi' \left(\sum_{i=1}^{K} \gamma_i \right) \right) \left(\alpha_k + \sum_{n=1}^{N} \phi_{n,k} - \gamma_k \right) - \Psi(\gamma_k) + \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \\
&- \Psi \left(\sum_{i=1}^{K} \gamma_i \right) + \Psi(\gamma_k) \\
&= \left(\Psi'(\gamma_k) - \Psi' \left(\sum_{i=1}^{K} \gamma_i \right) \right) \left(\alpha_k + \sum_{n=1}^{N} \phi_{n,k} - \gamma_k \right).
\end{aligned}$$

Setting it to zero, we have

$$\gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{n,k}.$$

A.4 Estimating Model Parameters

The previous section is the E-step of the variational EM algorithm, where we tighten the lower bound w.r.t. the variational parameters; this section is the M-step, where we maximize the lower bound w.r.t. the model parameters α and β. Now we add back the document subscript to consider the whole corpus.

By the assumed exchangeability of the documents, the overall log likelihood of the corpus is just the sum of all the documents' log likelihoods, and the overall lower bound is just the sum of the individual lower bounds. Again, only the terms involving β are left in the overall lower bound. Adding the Lagrange multipliers, we obtain

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\log\beta_{k,v} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\beta_{k,v} - 1\right).$$

Taking the derivative w.r.t. $\beta_{k,v}$ and setting it to zero, we have

$$\begin{aligned}\frac{\partial}{\partial\beta_{k,v}}\mathcal{L}_{[\beta]} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\frac{1}{\beta_{k,v}} + \lambda_k = 0\\ \Rightarrow \beta_{k,v} = -\frac{1}{\lambda_k}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\\ \Rightarrow \beta_{k,v} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v).\end{aligned}$$

Similarly, for α, we have

$$\begin{aligned}\mathcal{L}_{[\alpha]} &= \sum_{d=1}^{D}\left(\sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right)\right) + \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)\right)\\ \frac{\partial}{\partial\alpha_k}\mathcal{L}_{[\alpha]} &= \sum_{d=1}^{D}\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right) + \Psi\left(\sum_{i=1}^{K}\alpha_i\right) - \Psi(\alpha_k)\right)\\ &= \sum_{d=1}^{D}\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right)\right) + D\left(\Psi\left(\sum_{i=1}^{K}\alpha_i\right) - \Psi(\alpha_k)\right).\end{aligned}$$

Since the derivative also depends on other $\alpha_{k'\neq k}$, we compute the Hessian

$$\frac{\partial^2}{\partial\alpha_k\partial\alpha_{k'}}\mathcal{L}_{[\alpha]} = D\Psi'\left(\sum_{i=1}^{K}\alpha_i\right) - D\delta(k-k')\Psi(\alpha_k),$$

and notice that its form allows for the linear-time Newton-Raphson algorithm.

Bibliography

[1] Alzheimer's Association. Alzheimer's disease and dementia. http://www.alz.org. Accessed: Sep. 20, 2014.

[2] Melissa E. Murray, Neill R. Graff-Radford, Owen A. Ross, Ronald C. Petersen, Ranjan Duara, and Dennis W. Dickson. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *The Lancet Neurology*, 10(9):785–796, 2011.

[3] Young Noh, Seun Jeon, Jong Min Lee, Sang Won Seo, Geon Ha Kim, Hanna Cho, Byoung Seok Ye, Cindy W Yoon, Hee Jin Kim, Juhee Chin, et al. Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. *Neurology*, 83(21):1936–1944, 2014.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] John Ashburner and Karl J. Friston. Voxel-based morphometry—the methods. *NeuroImage*, 11(6):805–821, 2000.

[6] US Department of Health and Human Services. Alzheimer's disease: unraveling the mystery. *NIH Publication*, 2008.

[7] File:Alzheimer's disease brain preclinical.jpg. http://commons.wikimedia.org/wiki/File:Alzheimer%27s_disease_brain_preclinical.jpg. Accessed: Dec. 21, 2014.

[8] File:Alzheimer's disease brain severe.jpg. http://commons.wikimedia.org/wiki/File:Alzheimer%27s_disease_brain_severe.jpg. Accessed: Dec. 21, 2014.

[9] US National Institutes of Health. Search of: alzheimer – List Results – ClinicalTrials.gov. http://www.clinicaltrials.gov/ct2/results?term=alzheimer. Accessed: Sep. 20, 2014.

[10] Benjamin Lam, Mario Masellis, Morris Freedman, Donald T. Stuss, and Sandra E. Black. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimer's Research & Therapy*, 5(1):1, 2013.

[11] Clifford R. Jack Jr. and David M. Holtzman. Biomarker modeling of Alzheimer's disease. *Neuron*, 80(6):1347–1358, 2013.

[12] Kaj Blennow, Henrik Zetterberg, and Anne M. Fagan. Fluid biomarkers in Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, page a006221, 2012.

[13] Kaj Blennow. Cerebrospinal fluid protein biomarkers for Alzheimer's disease. *NeuroRx*, 1(2):213–225, 2004.

[14] Rahul S. Desikan, Wesley K. Thompson, Dominic Holland, Christopher P. Hess, James B. Brewer, Henrik Zetterberg, Kaj Blennow, Ole A. Andreassen, Linda K. McEvoy, Bradley T. Hyman, et al. The role of clusterin in amyloid-β–associated neurodegeneration. *JAMA Neurology*, 71(2):180–187, 2014.

[15] Chia-Chan Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106–118, 2013.

[16] Tests for Alzheimer's Disease and Dementia. http://www.alz.org/alzheimers_disease_steps_to_diagnosis.asp. Accessed: Jan. 8, 2015.

[17] Laura E Gibbons, Adam C Carle, R Scott Mackin, Danielle Harvey, Shubhabrata Mukherjee, Philip Insel, S McKay Curtis, Dan Mungas, Paul K Crane, et al. A composite score for executive functioning, validated in the Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior*, 6(4):517–527, 2012.

[18] Paul K Crane, Adam Carle, Laura E Gibbons, Philip Insel, R Scott Mackin, Alden Gross, Richard N Jones, Shubhabrata Mukherjee, S McKay Curtis, Danielle Harvey, et al. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, 6(4):502–516, 2012.

[19] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E. J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23:S208–S219, 2004.

[20] Stephen M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[21] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, 20(1):45–57, 2001.

[22] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[23] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[24] Jesper L. R. Andersson, Mark Jenkinson, and Stephen Smith. Non-linear optimisation. *FMRIB Technical Report TR07JA1*, 2007.

[25] Jesper L. R. Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, a.k.a. spatial normalisation. *FMRIB Technical Report TR07JA2*, 2007.

[26] Simon Brunton, Cerisse Gunasinghe, Nigel Jones, Matthew J. Kempton, Eric Westman, and Andrew Simmons. A voxel-based morphometry comparison of the 3.0 T ADNI-1 and ADNI-2 volumetric MRI protocols. *International Journal of Geriatric Psychiatry*, 2014.

[27] Shannon L. Risacher, Li Shen, John D. West, Sungeun Kim, Brenna C. McDonald, Laurel A. Beckett, Danielle J. Harvey, Clifford R. Jack Jr., Michael W. Weiner, and Andrew J. Saykin. Longitudinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort. *Neurobiology of Aging*, 31(8):1401–1418, 2010.

[28] Shannon L. Risacher, Andrew J. Saykin, John D. West, Li Shen, Hiram A. Firpi, and Brenna C. McDonald. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Current Alzheimer Research*, 6(4):347, 2009.

[29] David M. Blei. Topic Models. *Prensentation Slides at Machine Learning Summer School, Cambridge, UK*, 2009.

[30] Jennifer L. Whitwell, Dennis W. Dickson, Melissa E. Murray, Stephen D. Weigand, Nirubol Tosakulwong, Matthew L. Senjem, David S. Knopman, Bradley F. Boeve, Joseph E. Parisi, and Ronald C. Petersen. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. *The Lancet Neurology*, 11(10):868–877, 2012.

[31] Henriette Koch, Julie A. E. Christensen, Rune Frandsen, Marielle Zoetmulder, Lars Arvastson, Soren R. Christensen, Poul Jennum, and Helge B. D. Sorensen. Automatic sleep classification using a data-driven topic model reveals latent sleep states. *Journal of Neuroscience Methods*, 235:130–137, 2014.

[32] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):4, 2010.

[33] B. T. Thomas Yeo, Fenna M. Krienen, Simon B. Eickhoff, Siti N. Yaakub, Peter T. Fox, Randy L. Buckner, Christopher L. Asplund, and Michael W. L. Chee. Functional specialization and flexibility in human association cortex. *Cerebral Cortex*, page 217, 2014.