ADVANCES IN SPECTRAL LEARNING WITH APPLICATIONS TO TEXT ANALYSIS AND BRAIN IMAGING.

Paramveer Singh Dhillon

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

____

Lyle H. Ungar
Professor
Computer and Information Science

Co-Supervisor of Dissertation

____

James C. Gee
Associate Professor
Computer and Information Science

Graduate Group Chairperson

____

Lyle H. Ungar, Professor, Computer and Information Science

Dissertation Committee:

Mark Liberman, Professor, Linguistics, University of Pennsylvania.

Mitch Marcus, Professor, Computer and Information Science, University of Pennsylvania.

Chris Callison-Burch, Assistant Professor, Computer and Information Science, University of Pennsylvania.

Tom M. Mitchell, Professor, Computer Science, Carnegie Mellon University.

ADVANCES IN SPECTRAL LEARNING WITH APPLICATIONS TO TEXT ANALYSIS AND BRAIN IMAGING.

Acknowledgements

First and foremost I would like to thank my advisors Prof. Lyle Ungar and Prof. Jim Gee. None of this would have been possible without their generous support and rigorous mentoring throughout my Ph.D. I couldn't have wished for better mentors! I also got a chance to work closely with Prof. Dean Foster and Dr. Brian Avants. They all taught me how important it is for a researcher to work independently as well as work as part of a larger team of collaborators. Moreover, in today's age of information and big data, no one single discipline has all the tools to solve the pressing problems, so I learned how invaluable it is to have a plethora of interdisciplinary problem solving skills. In addition, I also learned that good research is not just about solving a problem and possessing the necessary arsenal of tools to do so, but also finding the right problems to solve in the first place. I think my biggest take-away from the Ph.D has been the ability to recognize interesting research problems independently and having the interdisciplinary set of tools from Computer Science, Applied Mathematics, Statistics and Linguistics to solve them. And, I got immense support from Lyle, Jim, Dean and Brian in helping me attain these skills.

A big thanks goes to the remaining member of my thesis committee, Profs. Mark Liberman, Mitch Marcus, Chris Callison-Burch and Tom Mitchell. This thesis would have been incomplete without their feedback which has helped improve the thesis in multiple ways.

During my Ph.D I was also fortunate to collaborate with Sham Kakade and Prof. Michael Collins. I am always amazed by the amount that I learned from them. Besides this, I spent the summer of 2011 interning in the Machine Learning group at Yahoo! Research Santa Clara, where I had two amazing mentors in Sathiya Keerthi and Olivier Chapelle.

I also got a chance to interact with many other faculty members at Penn and elsewhere through courses, research or mentoring. They gave me an outside perspective on my research and valuable advise on research and otherwise. I really cherished my discussions with them. Here is the incomplete list (not in any order). Profs. Aravind Joshi, Kostas Daniilidis, Ben Taskar, Michael Kearns, Robert Stine, Shane Jensen, Shawndra Hill, Pete Fader, Eric Bradlow, Christophe Van den Bulte, Koby Crammer and Ani Nenkova.

A special thanks goes to Mike Felker, Charity Payne and Cheryl Hickey, who saved me on multiple occasions and ensured that the administrative matters ran smoothly.

During my stay at Penn and Philadelphia I had an amazing set of friends who provided for intellectual discussions and fun on topics ranging from science fiction to 70's classic rock to start-ups to European club soccer to human evolution! I will really miss you guys! Here is the incomplete list (not in any order). Bilwaj Gaonkar, Madhur Behl, Jordan Rodu, Yichao Lu, Ben Kandel, Annie Louis, Jessy Li, Ellie Pavlick, Wei Xu, Andrew King, Sanjian Chen, Jenny Gillenwater, Aris Sotiras, Harini Eavini, Erdem Varol, Harsha Bhattapady, Phil Cook, John Wu, Alex Liu, Allison Pouch, Dania Daye, Caroline Brun, Rachel Yotter, Ted Cary, Elizabeth Slamka, Katerina Fragiadaki, Gang Song, Girish Kulkarni, Rajan Arora, Saurabh Garg, Constantine Lignos, Peter-Michael Osera, Bradford Tuckfield, Joy Lu, Kory Johnson, Peichao Peng, Harjot Gill, Tanveer Gill, Neville Ryant, Andy Schwartz, Mickey Brautbar and David Weiss.

I would also fondly remember the time I spent at La Colombe, LaVa Cafe, Joe's Rittenhouse and Elixr Cafe. A significant portion of my thesis was conceived and written working from these places.

Last, but not the least, I would like to thank my family back home in Chandigarh, India who have supported me in every endeavor in life. None of this would have been possible without the great upbringing that they gave me and the qualities of hard-work, discipline, perseverance, punctuality and kindness that they instilled in me. Just like any other kid, I resented them while growing up, as its a hard path to follow, but now I realize their value.

ABSTRACT

ADVANCES IN SPECTRAL LEARNING WITH APPLICATIONS TO TEXT ANALYSIS AND BRAIN IMAGING.

Paramveer Singh Dhillon

Lyle H. Ungar

James C. Gee

Spectral learning algorithms are becoming increasingly popular in data-rich domains, driven in part by recent advances in large scale randomized SVD, and in spectral estimation of Hidden Markov Models. Extensions of these methods lead to statistical estimation algorithms which are not only fast, scalable, and useful on real data sets, but are also provably correct. Following this line of research, we make two contributions. First, we propose a set of spectral algorithms for text analysis and natural language processing. In particular, we propose fast and scalable spectral algorithms for learning word embeddings – low dimensional real vectors (called *Eigenwords*) that capture the "meaning" of words from their context. Second, we show how similar spectral methods can be applied to analyzing brain images.

State-of-the-art approaches to learning word embeddings are slow to train or lack theoretical grounding; We propose three spectral algorithms that overcome these limitations. All three algorithms harness the multi-view nature of text data i.e. the left and right context of each word, and share three characteristics:

1. They are fast to train and are scalable.

2. They have strong theoretical properties.

3. They can induce context-specific embeddings i.e. different embedding for "river bank" or "Bank of America".

They also have lower sample complexity and hence higher statistical power for rare words. We provide theory which establishes relationships between these algorithms and optimality criteria for the estimates they provide. We also perform thorough qualitative and quantitative evaluation of *Eigenwords* and demonstrate their superior performance over state-of-the-art approaches.

Next, we turn to the task of using spectral learning methods for brain imaging data.

Methods like Sparse Principal Component Analysis (SPCA), Non-negative Matrix Factorization (NMF) and Independent Component Analysis (ICA) have been used to obtain state-of-the-art accuracies in a variety of problems in machine learning. However, their usage in brain imaging, though increasing, is limited by the fact that they are used as out-of-the-box techniques and are seldom tailored to the domain specific constraints and knowledge pertaining to medical imaging, which leads to difficulties in interpretation of results.

In order to address the above shortcomings, we propose *Eigenanatomy* (EANAT), a general framework for sparse matrix factorization. Its goal is to statistically learn the boundaries of and connections between brain regions by weighing both the data and prior neuroanatomical knowledge.

Although EANAT incorporates some neuroanatomical prior knowledge in the form of connectedness and smoothness constraints, it can still be difficult for clinicians to interpret the results in specific domains where network-specific hypotheses exist. We thus extend EANAT and present a novel framework for prior-constrained sparse decomposition of matrices derived from brain imaging data, called Prior Based Eigenanatomy (p-Eigen). We formulate our solution in terms of a prior-constrained ℓ_1 penalized (sparse) principal component analysis. Experimental evaluation confirms that p-Eigen extracts biologically-relevant, patient-specific functional parcels and that it significantly aids classification of Mild Cognitive Impairment when compared to state-of-the-art competing approaches.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Spectral learning methods employ matrix factorization techniques to factor a matrix into low dimensional components typically weighted by its eigenvalues. Since the eigenvalues represent the “spectrum” of the matrix, these methods based on eigen-decomposition of the underlying matrix constitute spectral learning.

Spectral Learning algorithms are becoming increasingly popular for analyzing data sets that are large either in terms of number of observations (such as text analysis) or features (such as medical images), or both, driven in part by recent advances in large scale randomized Singular Value Decomposition (SVD) (Halko et al. 2011), and in spectral estimation of Hidden Markov Models (Hsu et al. 2009). Extensions of these methods lead to statistical estimation algorithms which are not only fast, scalable, and useful on real data sets, but are also provably correct.

This thesis advances the use of spectral algorithms by designing new algorithms for two application areas. First, we propose new spectral algorithms for Text Analysis/Natural Language Processing. Particularly, we propose fast and scalable spectral algorithms for learning word embeddings – low dimensional real vectors (called *eigenwords*) that capture the “meaning” of words from their context. Second, we show how similar spectral methods can be applied to analyzing brain images.

1.1 Spectral Word Embeddings

In recent years there has been immense interest in learning embeddings for words from large amounts of raw text. A word embedding is a mathematical object associated with each word. Ideally these embeddings should capture a rich variety of information about that word, including topic, part of speech, and word features such as animacy, sentiment, gender, or whether the numbers are years or small numbers. They are typically learned in a totally unsupervised manner by exploiting the co-occurrence structure of words in unlabeled text.

The most obvious embedding of a word is a vector the size of vocabulary of the corpora ($\sim 100k$) with only one entry (corresponding to the index of the word) set to one, hence also known as "one-hot" embedding. One-hot embeddings, being high dimensional suffer from the curse of dimensionality and are also inefficient for storage.

So, it is imperative to learn embeddings over a smaller 'k' dimensional (~ 50) vocabulary to rid of the aforementioned problems. They provide an efficient and condensed representation of words and are an easy way to improve the performance of a supervised Natural Language Processing system by providing an additional set of features to "plug" into the supervised classifier.

The NLP systems use labeled data to learn a model, and there is not a lot of labeled text available for these tasks (For English there is still a decent amount of labeled text but very little for other languages.). So, the word embeddings, having been learned from large amounts of raw data provide a highly discriminative set of features which enable the supervised learner to perform better.

Word embeddings have proven useful and have given state-of-the-art performance on many natural language processing tasks e.g. syntactic parsing (Täckström et al. 2012; Parikh et al. 2014), POS Tagging (Dhillon et al. 2012b; Huang et al.

2013), dependency parsing (Bansal et al. 2014; Koo et al. 2008; Dhillon et al. 2012a), sentiment analysis (Dhillon et al. 2012b), chunking (Turian et al. 2010; Dhillon et al. 2011), Named Entity Recognition (NER) (Turian et al. 2010; Dhillon et al. 2011), word analogies (Mikolov et al. 2013a,b) and word similarity (Huang et al. 2012) to name a few.

Word embeddings can be broadly classified into two types 1). Clustering based discrete embeddings or 2). Real valued dense embeddings. Figure 1.1 shows the two types of embeddings.

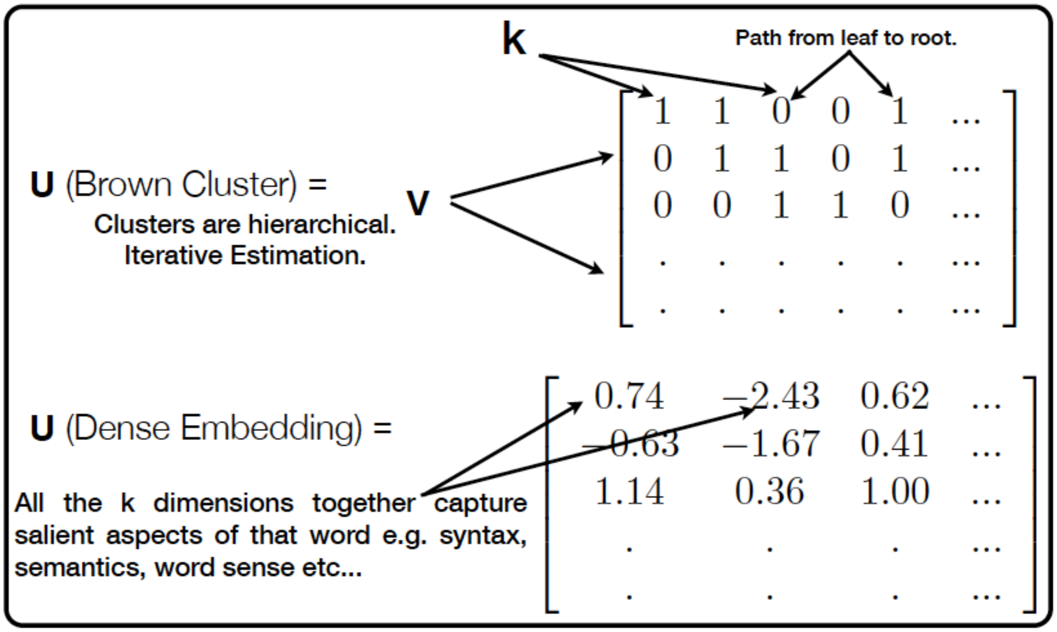


Figure 1.1: Clustering based vs Dense Embeddings

The state-of-the-art approaches (Collobert and Weston 2008; Mnih and Hinton 2007; Mikolov et al. 2013a,b) to learning word embeddings are slow to train and lack theoretical grounding. In addition, they only provide context oblivious embeddings i.e. same embedding for a given word (say) “bank” irrespective of the fact, if it is “river bank” or “Bank of America”, which can be sub-optimal in some domains e.g. Named Entity Recognition (NER) or Word Sense Disambiguation (WSD).

We propose three algorithms for learning word embeddings which address these issues. All the algorithms, 1) One Step CCA (OSCCA), 2) Two Step CCA (TSCCA) and 3) Low Rank Multi-View Learning (LR-MVL) have a Canonical Correlation Analysis (CCA) style eigen-decomposition at their core (which projects the high dimensional words to k (~ 50) dimensions (see Figure 1.2)), but at the same time are significantly novel.

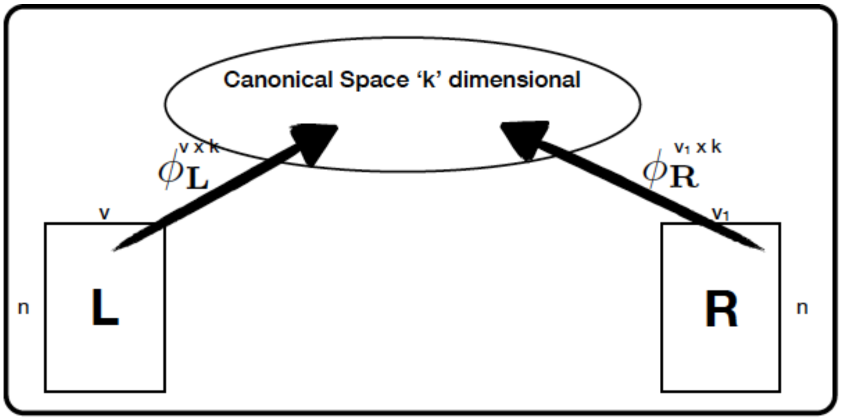

Canonical Space 'k' dimensional
$\phi_{\mathbf{L}}^{v \times k}$
$\phi_{\mathbf{R}}^{v_1 \times k}$
v
v_1
n
L
R
n

Figure 1.2: CCA Based dimensionality reduction.

They harness the multi-view nature of text data i.e. the left and right context of each word and share three characteristics:

1. They are fast to train and are scalable.

2. They have strong theoretical properties.

3. They can induce context-specific embeddings i.e. different embedding for "river bank" or "Bank of America".

In addition to this, TSCCA and LR-MVL also have lower sample complexity and hence higher statistical power for rare words. This thesis provides theory which establishes relationships between these algorithms and optimality criteria for the estimates they provide under the assumption that the data are generated by an HMM (standard assumption used in NLP).

Center Word	One Hot	OSCCA	TSCCA
communal (1166)	historical, moral, visual, modern, bureaucratic	eternal, outdoor, civic, spiritual, virtual	collective, cultural, civic, racial, Victorian
rabbi (1196)	physician, nurse, boy, waiter, bishop	priest, preacher, lawmaker, dentist, waiter	priest, bishop, preacher, Jew, pastor
cabbage (1196)	shrimp, potatoes, rice, asparagus, garden	yogurt, peppers, lettuce, peas, broccoli	broccoli, squash, mushrooms, lettuce, carrots
Cubs (1068)	Beatles, Angels, Communists, museum, Indians	Dodgers, Yankees, Redskins, Giants, Sox	Giants, Yankees, Dodgers, Redskins, Rangers

Table 1.1: (Rare Words) Nearest Neighbors of One Hot, OSCCA and TSCCA word embeddings. Counts of words in corpora in brackets.

As a sample, Table 1.1 shows the set of nearest neighbors found by the *eigenwords* (Since our word embeddings employ eigen-decomposition, we call them *eigenwords*) and the original high dimensional (one-hot) representation and it clearly shows the superiority of eigenwords.

This thesis also performs a thorough qualitative and quantitative evaluation of eigenwords. First, we show that when plotted they capture subtle syntactic and semantic aspects of the word with "similar" words being closer in this syntactic-semantic space. Next, we show that these word embeddings when included as additional features in supervised NLP systems, achieves state-of-the-art performance on tasks including Named Entity Recognition (NER), chunking, sentiment analysis, part of speech (POS) tagging, word similarity computation and syntactic and semantic word analogy tasks.

1.2 Spectral Learning for Brain Image Analysis

Next, we turn to the task of using Spectral Learning methods to harness the power of brain imaging data. This setting is complementary to the Text/ NLP setting as over here we have far fewer observations than features (i.e small n, large p), so dimensionality reduction becomes all the more important.

Dimensionality reduction methods like Sparse Principal Component Analysis (SPCA), Non-negative Matrix Factorization (NMF) and Independent Component Analysis (ICA) have been used to obtain state-of-the-art accuracies in a variety of problems in Machine Learning. However, their usage in brain imaging, though increasing, is limited by the fact that they are used as out-of-the-box techniques and are seldom tailored to the domain specific constraints/knowledge pertaining to medical imaging. For instance, uninformed, generic matrix decomposition methods, e.g. standard principal component analysis (PCA) or ICA, may be difficult to interpret because the solutions will produce vectors that are everywhere non-zero, i.e. involve the whole brain rather than its parts.

In order to address the above shortcoming, we propose Eigenanatomy (EANAT), a general framework for sparse matrix factorization that is closely related to SPCA, NMF, and ICA. The goal of EANAT is to statistically learn the boundaries of and connections between the brain regions by weighing both data and prior neuroanatomical guidance. Recent work points to the fact that exploiting problem-specific information can improve parts-based embeddings (Guan et al. 2011; Cai et al. 2010; Hosoda et al. 2009). EANAT component images, on the other hand, enable prior knowledge to enhance solution stability and are tied to a set of neuroanatomical coordinates that are *connected*, *smooth* and may also be defined by *non-negative* weights.

Although EANAT incorporates some neuroanatomical prior knowledge in the form of connectedness and smoothness constraints which aids clinical interpretability, it might still be difficult for clinicians to interpret the results in a specific domain where network-specific hypotheses exist. For example, someone studying fronto-temporal dementia would expect some or most of the signal to lie in frontal cortex, however if the voxels in frontal cortex don't explain the variance in data, these approaches won't highlight them.

So, going one step further, we present a novel framework for prior-constrained sparse decomposition of matrices derived from brain imaging data, called Prior Based Eigenanatomy (p-Eigen). p-Eigen stands in stark contrast with both the totally data-driven and totally prior driven approaches as it borrows strength from both these paradigms and leads to statistically refined definitions of ROIs based on information from data and hence provides a trade-off between the two approaches. In particular, p-Eigen seeks to identify a data-driven matrix decomposition like PCA/ICA, but at the same time it constrains the individual components by spatial anatomical priors (probabilistic regions of interest (ROIs)). We formulate our novel solution in terms of a prior-constrained ℓ_1 penalized (sparse) principal component analysis (SPCA). p-Eigen is shown in Fig. 1.3.

We use p-Eigen to address the problem of generating subject-specific functional connectivity networks. Using p-Eigen enables modeling of the inter-subject variability in the functional parcel boundaries and allows us to construct subject-specific functional networks with reduced sensitivity to ROI placement. We show that while still maintaining correspondence across subjects, p-Eigen extracts biologically-relevant and patient-specific functional parcels that facilitate hypothesis-driven network analysis. We show that using connectivity graphs derived from p-Eigen refined ROIs significantly aid classification of Mild Cognitive Impairment (MCI) as well as the prediction of scores in a Delayed Recall memory task when compared to graph metrics derived from state-of-the-art competing approaches. In a second set of experiments, we also show that the using the p-Eigen refined ROIs in structural cortical thickness images also aids classification of Mild Cognitive Impairment (MCI).

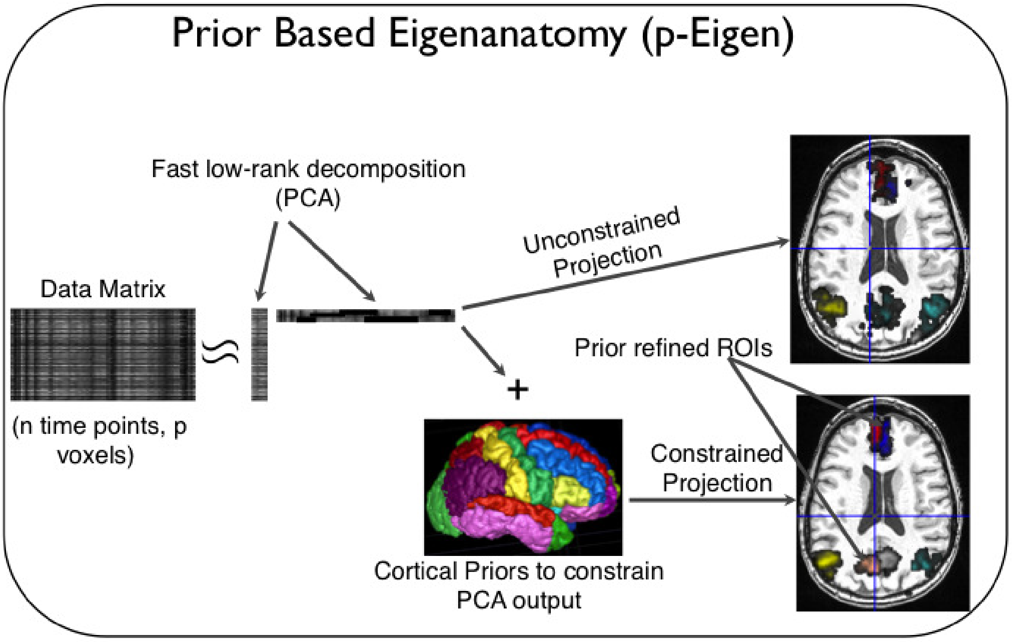


Figure 1.3: Prior Based Eigenanatomy (p-Eigen). An initial data matrix is decomposed into its eigenvectors, with each eigenvector being constrained by a corresponding cortical prior.

1.3 Summary & Broader Contribution

The two domains of Text/Natural Language Proecessing and brain imaging posit different challenges. Text applications have large n and large p, and brain imaging applications have small n, large p, where n is a token for text applications and a subject for brain imaging, and p is a word's context in text and a brain voxel in brain imaging. Broadly, this thesis shows that relatively simple linear models based on eigen-decomposition can help us attain state-of-the-art accuracies on these two domains and we do not require more complex non-linear models e.g. the ones based on Deep Neural Nets.

1.4 Organization of Thesis

This thesis is organized as follows. Chapters 2, 3 and 4 describe our three learning algorithms for *eigenwords* and show their empirical performance on multiple text analysis tasks. Chapters 5 describes our Eigenanatomy (EANAT) framework and uses the eigenvectors derived from it to classify control subjects vs Parkinson's patients for a clinically relevant population. Chapter 6 describes the Prior Based Eigenanatomy (p-Eigen) framework. Chapters 7 and Chapter 8 use it to to construct refined parcellations and subject specific functional connectivity networks from BOLD fMRI images and structural T-1 images respectively. Chapter 9 concludes the thesis and also provides avenues for future research.

Chapter 2

Eigenwords: CCA-Based Vector Space Models of Text

In[1] recent years there has been immense interest in learning embeddings for words from large amounts of raw text. Word embeddings map each word in text to a 'k' dimensional (~ 50) real valued vector. They are typically learned in a totally unsupervised manner by exploiting the co-occurrence structure of words in unlabeled text. Ideally these embeddings should capture a rich variety of information about that word, including topic, part of speech, word features such as animacy, sentiment, gender, whether the numbers are years or small numbers, and the direction of sentiment (happy vs. sad).

Their importance has been amplified by the fact that over the past decade there has been increased interest in using unlabeled data to supplement the labeled data in semi-supervised learning settings. This is mainly to overcome the inherent data sparsity and get improved generalization accuracies in high dimensional domains like NLP. Approaches like (Ando and Zhang 2005; Suzuki and Isozaki 2008) have

[1] This chapter is based on work in (Dhillon et al. 2011),(Dhillon et al. 2012b) and (Dhillon et al. 2014 (Under Review)

been empirically very successful and have achieved excellent accuracies on a variety of NLP tasks. However, it is often difficult to adapt these approaches to use in conjunction with an existing supervised NLP system as these approaches enforce a particular choice of model.

So, an increasingly popular alternative is to learn representational embeddings for words from a large collection of unlabeled data (typically using a generative model), and to use these embeddings to augment the feature set of a supervised learner, thereby improving the performance of a state-of-the-art NLP system e.g. a sentiment analyzer, parser, part of speech tagger etc.

Word embeddings have proven useful and have given state-of-the-art performance on many natural language processing tasks e.g. syntactic parsing (Täckström et al. 2012; Parikh et al. 2014), POS Tagging (Dhillon et al. 2012b; Huang et al. 2013), dependency parsing (Bansal et al. 2014; Koo et al. 2008; Dhillon et al. 2012a), sentiment analysis (Dhillon et al. 2012b), chunking (Turian et al. 2010; Dhillon et al. 2011), Named Entity Recognition (NER) (Turian et al. 2010; Dhillon et al. 2011), word analogies (Mikolov et al. 2013a,b) and word similarity (Huang et al. 2012) to name a few.

These NLP systems use labeled data to learn a model, and there is limited labeled text available for these tasks (For English there is still a reasonable amount of labeled text but much less for other languages.). So, the word embeddings, having been learned from large amounts of raw data provide a highly discriminative set of features which enable the supervised learner to perform better.

As mentioned earlier, embedding methods produce features in low dimensional spaces or over a small vocabulary size, unlike the traditional approach of working in the original high dimensional vocabulary space with only one dimension “on” at a given time.

Broadly speaking, the embedding methods fall into two categories (as also shown in Figure 2.1):

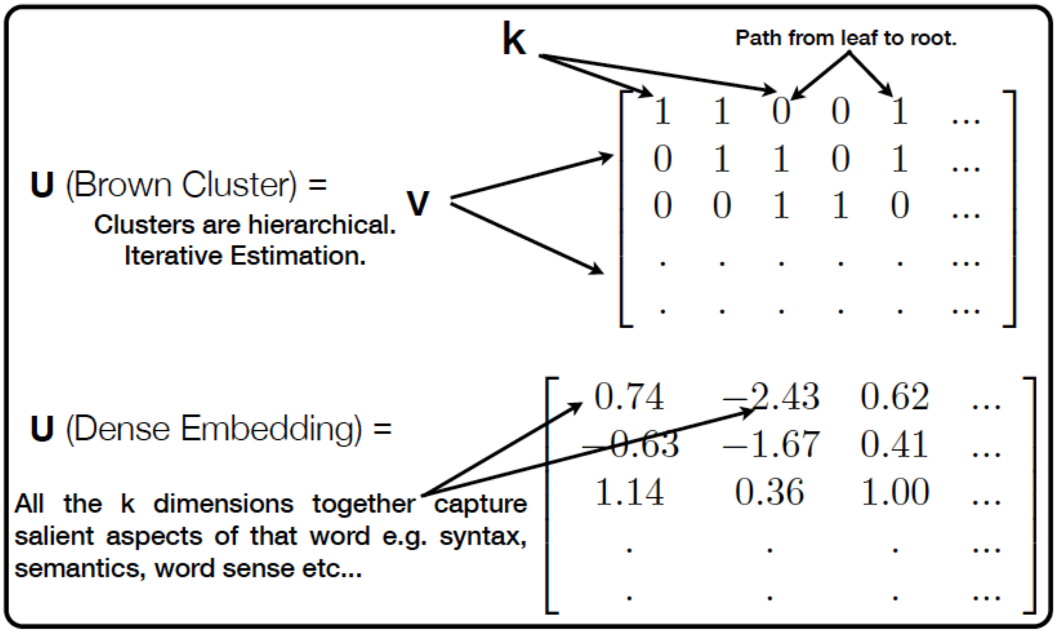


Figure 2.1: Clustering based vs Dense Embeddings

1. *Clustering based word embeddings*: Clustering methods, often hierarchical, are used to group distributionally similar words based on their contexts. The two dominant approaches are Brown Clustering (Brown et al. 1992a) and (Pereira et al. 1993a). As recently shown, HMMs can also be used to induce a multinomial distribution over possible clusters (Huang and Yates 2009).

2. *Dense embeddings*: These embeddings are dense, low dimensional and real-valued. Each dimension of these embeddings captures latent information about a combination of syntactic and semantic word properties. They can either be induced using neural networks like C&W embeddings (Collobert and Weston 2008), *Hierarchical log-linear* (HLBL) embeddings (Mnih and Hinton 2007), word2vec embeddings (Mikolov et al. 2013a,b) or by eigen-decomposition of the word co-occurrence matrix, e.g. *Latent Semantic Analysis/Latent Semantic*

Indexing (LSA/LSI) (Dumais et al. 1988).

The most classic and successful algorithm for learning word embeddings is Latent Semantic Analysis (LSA) (Landauer et al. 2008) which works by performing SVD on word by document matrix.

Unfortunately, the state-of-the-art embedding methods suffer from a number of shortcomings: 1). They are slow to train (especially, the traditional Deep Learning based approaches (Collobert and Weston 2008; Mnih and Hinton 2007). Though, recently, (Mikolov et al. 2013a,b) have proposed neural network based embeddings which avoid using the hidden layers which are typical in Deep Learning. This, coupled with good engineering allows their embeddings to be trained in minutes. 2). They are sensitive to the scaling of the embeddings (especially ℓ_2 based approaches like LSA/PCA). 3). They learn a single embedding for a given word type; i.e. all the occurrences of the word "*bank*" will have the same embedding, irrespective of whether the context of the word suggests it means "*a financial institution*" or "*a river bank*". Recently, (Huang et al. 2012) have proposed context specific word embeddings, but their Deep Learning based approach is slow and can not scale to large vocabularies.

In this chapter we provide spectral algorithms (based on eigen-decomposition) for learning word embeddings, as they have been shown to be fast and scalable for learning from large amounts of unlabeled data (Turney and Pantel 2010), have a strong theoretical grounding, and are guaranteed to converge to globally optimal solutions (Hsu et al. 2009). Particularly, we are interested in Canonical Correlation Analysis (CCA) (Hotelling 1935) based methods as:

- *First*, unlike PCA or LSA based methods they are scale invariant.

- *Second*, unlike LSA they can capture multi-view information. In text applications the left and right contexts of the words provide a natural split into two views which is totally ignored by LSA as it throws the entire context into a bag of words while constructing the term-document matrix.

We propose a variety of dense embeddings; they learn real-valued word embeddings by performing Canonical Correlation Analysis (CCA) (Hotelling 1935) between the past and future views of the data (Imagine the present view as the current token, then the previous and the next tokens are the past and future views) . All our embeddings have a number of common characteristics and address the shortcomings of the current state-of-the-art embeddings. In particular, they are:

1. Fast, scalable and scale invariant.

2. Provide better sample complexity for rare words.

3. Can induce context-specific embeddings i.e. different embeddings for "*bank*" based on whether it means "*a financial institution*" or "*a river bank*".

4. Strong theoretical foundations.

The remainder of the chapter is organized as follows. In the next section we give a brief overview of CCA, which forms the core of our method. The following section describes our proposed algorithms.

2.1 Brief Review: Canonical Correlation Analysis (CCA)

CCA (Hotelling 1935) is the analog to Principal Component Analysis (PCA) for pairs of matrices. PCA computes the directions of maximum covariance between elements in a single matrix, whereas CCA computes the directions of maximal correlation between a pair of matrices. Like PCA, CCA can be cast as an eigenvalue problem on a covariance matrix, but can also be interpreted as deriving from a generative mixture model (Bach and Jordan 2005). See (Hardoon et al. 2004) for a review of CCA with applications to machine learning.

More specifically, given a set of n paired observation vectors $\{(l_1, r_1), ..., (l_n, r_n)\}$– in our case the two matrices are the left (**L**) and right (**R**) context matrices of a word–we would like to simultaneously find the directions $\boldsymbol{\Phi_l}$ and $\boldsymbol{\Phi_r}$ that maximize the correlation of the projections of **L** onto $\boldsymbol{\Phi_l}$ with the projections of **R** onto $\boldsymbol{\Phi_r}$. This is expressed as

$$\max_{\mathbf{\Phi}_l, \mathbf{\Phi}_r} \frac{\mathbb{E}[\langle \mathbf{L}, \mathbf{\Phi}_l \rangle \langle \mathbf{R}, \mathbf{\Phi}_r \rangle]}{\sqrt{\mathbb{E}[\langle \mathbf{L}, \mathbf{\Phi}_l \rangle^2] \mathbb{E}[\langle \mathbf{R}, \mathbf{\Phi}_r \rangle^2]}} \tag{2.1}$$

where $\mathbb{E}$ denotes the empirical expectation. We use the notation $\mathbf{C_{lr}}$ ($\mathbf{C_{ll}}$) to denote the cross (auto) covariance matrices between **L** and **R** (i.e. **L'R** and **L'L** respectively.).

The left and right canonical correlates are the solutions (eigenvectors) $\langle \mathbf{\Phi}_l, \mathbf{\Phi}_r \rangle$ of the following equations:

$$\begin{aligned} \mathbf{C_{ll}}^{-1} \mathbf{C_{lr}} \mathbf{C_{rr}}^{-1} \mathbf{C_{rl}} \mathbf{\Phi}_l &= \lambda \mathbf{\Phi}_l \\ \mathbf{C_{rr}}^{-1} \mathbf{C_{rl}} \mathbf{C_{ll}}^{-1} \mathbf{C_{lr}} \mathbf{\Phi}_r &= \lambda \mathbf{\Phi}_r \end{aligned} \tag{2.2}$$

We keep the k left and right singular vectors ($\mathbf{\Phi_l}$ and $\mathbf{\Phi_r}$) corresponding to the $\mathbf{k}$ largest singular values. These computations can be performed easily using $eig()$ function in MATLAB or R.

The basic intuition behind CCA is shown in Figure 2.2.

There is an equivalent formulation of CCA which allows us to compute the solution via SVD of $\mathbf{C_{ll}}^{-1/2} \mathbf{C_{lr}} \mathbf{C_{rr}}^{-1/2}$. (See the appendix for proof).

$$\mathbf{C_{ll}}^{-1/2} \mathbf{C_{lr}} \mathbf{C_{rr}}^{-1/2} \equiv \mathbf{\Phi}_l \Lambda \mathbf{\Phi}_r \tag{2.3}$$

where $\langle \mathbf{\Phi}_l, \mathbf{\Phi}_r \rangle$ are the left and right singular vectors and Λ is the diagonal matrix

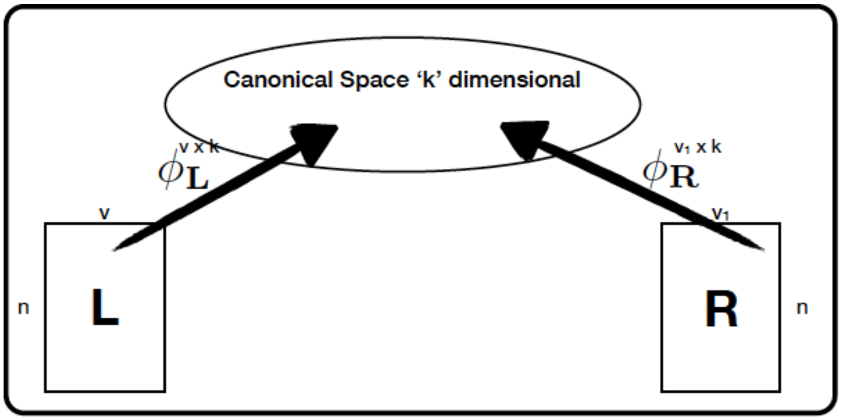


Figure 2.2: CCA Based dimensionality reduction.

of singular values. Finally, the CCA projections are gotten by "de-whitening" [2] as $\mathbf{\Phi}_{lproj} = \mathbf{C}_{ll}^{-1/2}\mathbf{\Phi}_l$ and $\mathbf{\Phi}_{rproj} = \mathbf{C}_{rr}^{-1/2}\mathbf{\Phi}_r$.

For most of the embeddings proposed in this chapter, the SVD formulation is preferred since it avoids fewer multiplications of large sparse matrices which is an expensive operation.

2.1.1 Suitability of CCA for Learning Word Embeddings

Recently, (Foster et al. 2008) showed that CCA can exploit multi-view nature of the data and provide sufficient conditions for CCA to achieve dimensionality reduction without losing predictive power. They assume that the data was generated by the model shown in Figure 2.3. The two assumptions that they make are that 1) Each of the two views are independent conditional on a k-dimensional hidden state (H) and that 2) The two views provide a redundant estimate of the hidden state (H).

These two assumptions are generalization of the assumptions made by co-training (Blum and Mitchell 1998) (Figure 2.4), as co-training conditions on the observed labels (Y) and not on a more flexible representation i.e. a hidden state (H).

[2] One way to think about CCA is as "whitening" the covariance matrix. Whitening converts covariances into correlations.

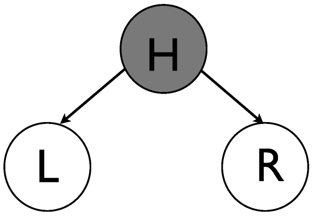


Figure 2.3: Multi-View Assumption. Grey color indicates that state is hidden.

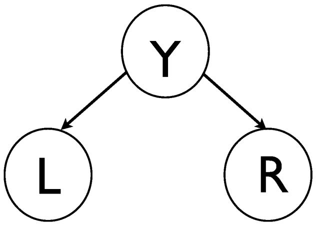


Figure 2.4: Co-training Assumption.

In text and Natural Language Processing (NLP) applications, its typical to assume a Hidden Markov Model (HMM) as the data generating model (Jurafsky and Martin 2000). Its easy to see that a Hidden Markov Model (HMM) satisfies the multi-view assumption. Hence, the left and right context of a given word provides two natural views and one could use CCA to estimate the hidden state (H).

Furthermore, as mentioned earlier, CCA is scale invariant and provides a natural scaling (inverse or square root of the inverse of the auto-covariance matrix, depending on whether we use Eigen-decomposition or SVD formation) for the observations. If we further use the SVD formulation, then it also allows us to harness the recent advances in large scale randomized SVD (Halko et al. 2011), which allows the embeddings learning algorithms to be fast and scalable.

The invariance of CCA to linear data transformations allows proofs that keeping the dominant singular vectors (those with largest singular values) will faithfully capture any state information (Kakade and Foster 2007). Also, CCA extends more naturally than LSA to sequences of words.[3] Remember that LSA uses “bags of words”, which are good for capturing topic information, but fail for problems like part of speech (POS) tagging which need sequence information.

Finally, as we show in the next chapter the CCA formulation can be naturally extended to a two step procedure that, while equivalent in the limit of infinite data, gives higher accuracies for finite corpora and provides better sample complexity.

So, in summary we estimate a hidden state associated with words by computing the dominant canonical correlations between target words and the words in their immediate context. The main computation, finding the singular value decomposition of a scaled version of the co-occurrence matrix of counts of words with their contexts, can be done highly efficiently. Use of CCA also allows us to prove theorems about the optimality of our reconstruction of the state.

In the next section we show how to efficiently compute a vector that characterizes each word type by using the left singular values of the above CCA to map from the word space (size v) to the state space (size k). We call this mapping the *eigenword dictionary* for words, as it associates with every word a vector that captures that word’s syntactic and semantic attributes. As will be made clear below, the *e*igenword dictionary is arbitrary up to a rotation, but captures the information needed for any linear model to predict properties of the words such as part of speech or word sense.

2.2 Problem Formulation

Our goal is to estimate a vector for each word *type* that captures the distributional properties of that word in the form of a low dimensional representation of the correlation between that word and the words in its immediate context.

More formally, assume a document (in practice a concatenation of a large number

[3] It is important to note that it is possible to come up with PCA variants which take sequence information into account.

of documents) consisting of n tokens $\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_n}\}$, each drawn from a vocabulary of v words. Define the left and right contexts of each token $\mathbf{w_i}$ as the h words to the left or right of that token. The context sits in a very high dimensional space, since for a vocabulary of size v, each of the $2h$ words in the combined context requires an indicator function of dimension v. The tokens themselves sit in a v dimensional space of words which we want to project down to a k dimensional state space. We call the mapping from word types to their latent vectors the *eigenword dictionary.*

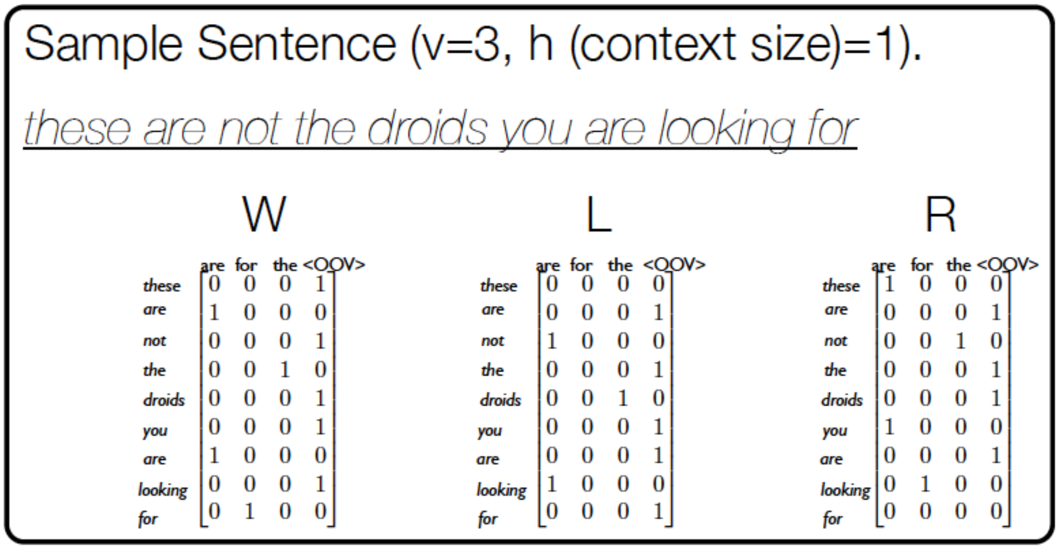
For a set of documents containing n tokens, define $\mathbf{L_{n\times vh}}$ and $\mathbf{R_{n\times vh}}$ as the matrices specifying the left and right contexts of the tokens, and $\mathbf{W_{n\times v}}$ as the matrix of the tokens themselves. In $\mathbf{W}$, we represent the presence of the j^{th} word type in the i^{th} position in a document by setting matrix element $\mathbf{w_{ij}} = \mathbf{1}$. $\mathbf{L}$ and $\mathbf{R}$ are similar, but have columns for each word in each position in the context. (For example, in the sentence "I ate green apples yesterday.", for a context of size $h = 2$, the left context of "green" would be "I ate" and the right context would be "apples yesterday" and the third row of $\mathbf{W}$ would have a "1" in the column corresponding to the word "green".) Figure 2.5 shows the $\mathbf{W}$, $\mathbf{L}$ and $\mathbf{R}$ matrices for a sample sentence.

Define the complete context matrix $\mathbf{C}$ as the concatenation $[\mathbf{L}\ \mathbf{R}]$. Thus, for a trigram representation with vocabulary size v words, history size $h = 1$, $\mathbf{C}$ has $2v$ columns – one for each possible word to the left of the target word and one for each possible word to the right of the target word.

$\mathbf{A_{wc}} = \mathbf{W}^\top\mathbf{C}$ then contains the counts of how often each word $\mathbf{w}$ occurs in each context $\mathbf{c}$, the matrix $\mathbf{A_{cc}} = \mathbf{C}^\top\mathbf{C}$ gives the covariance of the contexts, and $\mathbf{A_{ww}} = \mathbf{W}^\top\mathbf{W}$, the word covariance matrix, is a diagonal matrix with the counts of each word on the diagonal.[4]

We want to find a vector representation of each of the v word types such that

[4] We will pretend that the means are all in fact zero and refer to these $\mathbf{A_{cc}}$ etc. as covariance matrices, when in fact they are actually second moment matrices.



Sample Sentence (v=3, h (context size)=1).

these are not the droids you are looking for

$$\mathbf{W} = \begin{array}{c|cccc} & are & for & the & \text{<OOV>} \\ \hline these & 0 & 0 & 0 & 1 \\ are & 1 & 0 & 0 & 0 \\ not & 0 & 0 & 0 & 1 \\ the & 0 & 0 & 1 & 0 \\ droids & 0 & 0 & 0 & 1 \\ you & 0 & 0 & 0 & 1 \\ are & 1 & 0 & 0 & 0 \\ looking & 0 & 0 & 0 & 1 \\ for & 0 & 1 & 0 & 0 \end{array}$$

$$\mathbf{L} = \begin{array}{c|cccc} & are & for & the & \text{<OOV>} \\ \hline these & 0 & 0 & 0 & 0 \\ are & 0 & 0 & 0 & 1 \\ not & 1 & 0 & 0 & 0 \\ the & 0 & 0 & 0 & 1 \\ droids & 0 & 0 & 1 & 0 \\ you & 0 & 0 & 0 & 1 \\ are & 0 & 0 & 0 & 1 \\ looking & 1 & 0 & 0 & 0 \\ for & 0 & 0 & 0 & 1 \end{array}$$

$$\mathbf{R} = \begin{array}{c|cccc} & are & for & the & \text{<OOV>} \\ \hline these & 1 & 0 & 0 & 0 \\ are & 0 & 0 & 0 & 1 \\ not & 0 & 0 & 1 & 0 \\ the & 0 & 0 & 0 & 1 \\ droids & 0 & 0 & 0 & 1 \\ you & 1 & 0 & 0 & 0 \\ are & 0 & 0 & 0 & 1 \\ looking & 0 & 1 & 0 & 0 \\ for & 0 & 0 & 0 & 0 \end{array}$$

Figure 2.5: Various matrices for a sample sentence.

words that are distributionally similar (ones that have similar contexts) have similar state vectors. We will do this using Canonical Correlation Analysis (CCA) (Hotelling 1935; Hardoon and Shawe-Taylor 2008), by taking the CCA between the combined left and right contexts $\mathbf{C} = [\mathbf{L}\ \mathbf{R}]$ and their associated tokens, $\mathbf{W}$.

2.3 One Step CCA (OSCCA)

Using the above, we can define a "One step CCA" (OSCCA), procedure to estimate the *eigenword dictionary* as follows:

$$\mathbf{CCA}(\mathbf{W}, \mathbf{C}) \rightarrow (\boldsymbol{\Phi}_{\mathbf{W}}, \boldsymbol{\Phi}_{\mathbf{C}}) \tag{2.4}$$

where the $v \times k$ matrix $\boldsymbol{\Phi}_{\mathbf{W}}$ contains the *e*igenword dictionary that characterizes each of the v words in the vocabulary using a k dimensional vector. More generally, the "state" vectors $\mathbf{S}$ for the n tokens can be estimated either from the context as $\mathbf{C}\boldsymbol{\Phi}_{\mathbf{C}}$ or (trivially) from the tokens themselves as $\mathbf{W}\boldsymbol{\Phi}_{\mathbf{W}}$. Its important to note that both these estimation procedures give a redundant estimate of the same hidden "state."

The left canonical correlates found by OSCCA give an optimal approximation to the state of each word, where "optimal" means that it gives the linear model of a given size, k that is best able to estimate labels that depend linearly on state, subject to only using the word and not its context. The right canonical correlates similarly give optimal state estimates given the context.

OSCCA, as defined in Equations 2.4 thus gives an efficient way to calculate the eigenword dictionary $\mathbf{\Phi_W}$ for a set of v words given the context and associated word matrices from a corpus.

2.3.1 Theoretical properties

We now discuss how well the hidden state can be estimated from the target word. (A similar result can be derived for estimating hidden state from the context.) The state estimated is arbitrary up to any linear transformation, so all our comments address our ability to use the state to estimate some label which depends linearly on the state.

Keeping the dominant singular vectors in $\mathbf{\Phi}_W$ and $\mathbf{\Phi}_C$ provides two different bases for estimated state. Each is optimal in its own way, as explained below.

The following Theorem 1 shows that the left canonical correlates give an optimal approximation to the state of each word (in the sense of being able to estimate an emission or label Y for each state), subject to only using the word and not its context.

Theorem 1. *Let* $(\mathbf{W}_t, \mathbf{C}_t, Y_t)$ *for* $t = 1 \ldots n$ *be* n *tuples of random variables drawn i.i.d. from some distribution (pdf or pmf)* $\mathbb{D}(w_t, c_t, y_t)$*. We call the pair* $(Y_1 \ldots Y_n, \beta)$ *a* linear context problem *if*

1. *Y_t is a linear function of the context (i.e. $Y_t = \alpha^\top \mathbf{C}_t$)*

2. *$\beta^\top \mathbf{W}_t$ is the best linear estimator of Y_t given $\mathbf{W}_t$, namely β minimizes $\mathbb{E}\sum_t (Y_t - \beta^\top \mathbf{W}_t)^2$ and*

3. *$Var(Y_t) \leq 1$.*

Define ${\mathbf{\Phi_l}}^i$ to be the i'th left singular vector for the SVD of Eq. 2.3 with $\mathbf{C}_{rr} = \mathbb{E}(\mathbf{C}^\top \mathbf{C})$, $\mathbf{C}_{lr} = \mathbb{E}(\mathbf{W}^\top \mathbf{C})$, and $\mathbf{C}_{ll} = \mathbb{E}(\mathbf{W}^\top \mathbf{W})$ where $(\mathbf{W}, \mathbf{C})$ are drawn from the marginal distribution $\mathbb{D}(w, c)$. Then, for all $\epsilon > 0$ there exists a k such that for any linear context problem $(Y_1 \ldots Y_n, \beta)$, there exists a $\gamma \in \mathbb{R}^k$ such that $\widehat{Y}_t = \sum_{i=1}^k \gamma_i \phi_{it}$ is a good approximation to Y_t in the sense that $\mathbb{E}(\widehat{Y}_t - \beta^\top \mathbf{W})^2 \leq \epsilon$.

Please see the Appendix for the proof.

To understand the above theorem, note that we would have liked to have a linear regression predicting some label Y from the original data $\mathbf{W}$. However, the original data is very high dimensional. Instead, we can first use CCA to map high dimensional vectors $\mathbf{W}$ to lower dimensional vectors $\mathbf{\Phi_W}$, from which Y can be predicted. For example with a few labeled examples of the form $(\mathbf{W}, Y)$, we can recover the γ_i parameters using linear regression. The $\mathbf{\Phi_W}$ subspace is guaranteed to hold a good approximation. A special case of interest occurs when estimating a label Z $(= \alpha_\top \mathbf{C})$ plus zero mean noise. In this case, one can pick $Y = \mathbb{E}(Z)$ and proceed as above. This effectively extends the theorem to the case where the mapping from $\mathbf{C}$ to Y is random, not deterministic.

Note that if we had used covariance rather than correlation as done by LSA/PCA then in the worst case, the key singular vectors for predicting state could be those with arbitrarily small singular values. This corresponds to the fact that for principle component regression (PCR), there is no guarantee that the largest principle components will prove predictive of an associated label.

One can think of Theorem 1 as implicitly estimating a k-dimensional hidden state from the observed $\mathbf{W}$. This hidden state can be used to estimate Y. Note that for Theorem 1, the state estimate is "trivial" in the sense that because it comes from the words, not the context, every occurrence of each word must give the same state estimate. This is attractive in that it associates a latent vector with every word type, but limiting in that it does not allow for any word ambiguity. The right canonical vectors allow one to estimate state from the context of a word, giving different state estimates for the same word in different contexts, as is needed for word sense disambiguation. We relegate that discussion to the next chapter, when we discuss induction of context-specific word embeddings. For now, we focus on the simpler use of left canonical covariates to map each word type to a k dimensional vector.

2.4 Discussion

We have argued that for many problems, CCA can give better feature vectors for words than PCA. CCA, in our application, finds the components of maximum correlation between the context words, taking into account their location in the context, with the word of interest, unlike PCA on n-grams, which treats all the words in the n-gram equivalently.

PCA and CCA share deep similarities, not just in both being spectral methods. If the word co-occurrence matrix for PCA is normalized by scaling each word by dividing by the square of its overall frequency, then in the special case of bigrams, PCA and CCA become identical. In this special case, the context and the target word covariances $\mathbf{C}'\mathbf{C}$ and $\mathbf{W}'\mathbf{W}$ become (after normalization) the identity matrix. Since CCA scales by the inverse of these covariance matrices, if they are identity matrices, PCA and CCA will have identical singular vectors.

In the more common case of a larger context, PCA will devote more degrees of freedom to finding the structure within the context, while CCA will focus on finding the correlation between context and target word. If we could afford to compute and use larger state spaces, this would not be too serious, but because we are working with large corpora and large vocabularies, even a five-fold reduction in the number of components that is kept matters.

More broadly, we have argued that a single vector for each word can capture a wide range of attributes of that word including, part of speech, animacy, sex, edibility, etc. One could instead have clustered words based on distributional similarity using, e.g., (Pereira et al. 1993b) or (Brown et al. 1992b), but one would need some complex multi-faceted hierarchical clustering scheme to come close to capturing the different dimensions represented in the attribute vectors. For example, should "he" and "she" be in the same or different clusters? The words are very similar on many dimensions, but opposed on at least one. Using vector models also has advantages over categories in allowing word meaning to sit on a continuum, rather than being binned into discrete categories. There is substantial evidence from human studies that word meanings are often interpreted on a graded scale (Erk and McCarthy 2009), rather than categorically.

Words typically follow a Zipfian distribution, i.e. most words are rare, so getting better estimates of their state from just a few samples can help us get predictive performance when the eigenwords are used as features in a supervised learning task. So, in the next chapter, we propose two eigenword learning algorithms which have better sample complexity for rare words.

Chapter 3

Efficient computation of eigenwords with better sample complexity

OSCCA[1] is optimal only in the limit of infinite data. In practice, data is, of course, always limited. In languages, lack of data comes about in two ways. Some languages are resource poor; one just does not have that many tokens of them (especially languages that lack a significant written literature). Even for most modern languages, many of the individual words in them are quite rare. Due to the Zipfian distribution of words, many words do not show up very often. A typical year's worth of Wall Street Journal text only has "lasagna" or "backpack" a handful of times and "ziti" at most once or twice. To overcome these issues we propose a two-step procedure which gives rise to two algorithms, Two Step CCA (TSCCA) and Low-Rank Multi-View Learning (LR-MVL) that have better sample complexity for rare words.

[1] This chapter is based on work in (Dhillon et al. 2011),(Dhillon et al. 2012b) and (Dhillon et al. 2014 (Under Review).

3.1 Benefits of CCA for rare words

As mentioned earlier, the standard way of coding a word's representation is using a "one-hot" embedding i.e. a sparse vector which is 1 (or "on") at only one position and zero everywhere else. For instance, if we are considering a context window of 1 on either side of a word and a vocabulary of 15,000 words then the context matrix will be of size 15001×30002. The rows represent 15000 words and one generic "out of vocabulary" word, the columns represent 15001 words to the left and to the right of that word.

These simple embeddings are not ideal for multiple reasons. First, they are computationally expensive and tedious to store and operate on; one has to ensure that one only performs sparsity preserving operations as it might become infeasible to store the entire dense matrix in the memory. Second, these representations suffer from the curse of dimensionality, which is a well known phenomenon in Machine Learning which states that in high dimensions all the vectors look similar. Though, these high-dimensional vectors might still capture meaningful information for the more frequent words they perform poorly for rare words.

The OSCCA embeddings proposed in the last chapter, address both these issues. They learn low dimensional (typically 50-100) vectors for each word which makes it feasible to store the dense matrix efficiently in memory and to perform algebraic operations on it. Second, we get rid of the curse of dimensionality and can actually find meaningful neighbors of words, rather than all the words looking like their neighbors. This also implies that the vectors capture meaningful syntactic and semantic information about that word and hence can be used as out-of-the-box features in supervised learning tasks. The OSCCA embeddings or any dense embedding, for that matter, can also be seen as smoothing the aggregate context information of that word and in that process capturing relevant (syntactic and semantic) information about that word.

One potential shortcoming of OSCCA embeddings, as highlighted above, is that the vectors it learns for rare words can be poor owing to the paucity of data. The TSCCA and LR-MVL embeddings that we describe in the next section learn better embeddings for rarer words.

3.1.1 Illustration

We use an American newswire corpus (Reuters) containing $\sim$ 290 million tokens of text to illustrate our points; in particular that the "one-hot" representations suffer due to curse of dimensionality and that OSCCA and TSCCA smooth the representations giving meaningful word vectors. In particular we use vectors learned by "one-hot", OSCCA and TSCCA and plot the nearest neighbors for some of the frequent and rare words. We learned 100 dimensional vectors for OSCCA and TSCCA, using a context size of 2 words to the left and right and vocabulary of 15,000 words.

The results for randomly selected frequent and rare words are shown in Tables 3.1 and 3.2. As can be seen, for frequent words, all three approaches give reasonable neighbors which capture both the syntax and the semantics. For the rarer words TSCCA neighbors capture the most relevant information for that word. This can be attributed to its borrowing strength from more frequent distributionally similar words. Even OSCCA performs decently for the rare words, giving mostly relevant neighbors. However, the "one-hot" embeddings perform poorly and mostly return gibberish, even though the words we chose for demonstration occurred $\sim 1K$ times in the corpora and thus were not extremely rare.

Center Word	One Hot	OSCCA	TSCCA
Republican (169363)	Democratic, GOP, conservative, liberal, party	Democratic, GOP, Conservative, conservative, centrist	Democratic, GOP, incumbent, Senate, centrist
California (96166)	Florida, Virginia, Maryland, Texas, Michigan	Virginia, Maryland, Texas, Louisiana, Florida	Virginia, Maryland, Oregon, Texas, Connecticut
security (95138)	military, intelligence, safety, government, defense	intelligence, counterterrorism, military, enforcement, humanitarian	safety, operational, intelligence, health, technical
news (147332)	media, press, television, business, newspaper	media, television, press, radio, newspaper	television, broadcast, TV, radio, telephone

Table 3.1: (Frequent Words) Nearest Neighbors of One Hot, OSCCA and TSCCA word embeddings. Counts of words in corpora in brackets.

Center Word	One Hot	OSCCA	TSCCA
communal (1166)	historical, moral, visual, modern, bureaucratic	eternal, outdoor, civic, spiritual, virtual	collective, cultural, civic, racial, Victorian
rabbi (1196)	physician, nurse, boy, waiter, bishop	priest, preacher, lawmaker, dentist, waiter	priest, bishop, preacher, Jew, pastor
cabbage (1196)	shrimp, potatoes, rice, asparagus, garden	yogurt, peppers, lettuce, peas, broccoli	broccoli, squash, mushrooms, lettuce, carrots
Cubs (1068)	Beatles, Angels, Communists, museum, Indians	Dodgers, Yankees, Redskins, Giants, Sox	Giants, Yankees, Dodgers, Redskins, Rangers

Table 3.2: (Rare Words) Nearest Neighbors of One Hot, OSCCA and TSCCA word embeddings. Counts of words in corpora in brackets.

Algorithm 1 Two step CCA

```
Input: $\mathbf{L}, \mathbf{W}, \mathbf{R}$
$\mathbf{CCA}(\mathbf{L}, \mathbf{R}) \rightarrow (\mathbf{\Phi_L}, \mathbf{\Phi_R})$
$\mathbf{S} = [\mathbf{L\Phi_L} \;\; \mathbf{R\Phi_R}]$
$\mathbf{CCA}(\mathbf{S}, \mathbf{W}) \rightarrow (\mathbf{\Phi_S}, \mathbf{\Phi_W})$
Output: $\mathbf{\Phi_W}$, the eigenword dictionary
```

3.2 Two Step CCA (TSCCA) for estimating Eigenword dictionary.

We now introduce our two step procedure TSCCA of computing an *eigenword dictionary* and show theoretically that it gives better estimates than the OSCCA method described in the last section.

In the two-step method, instead of taking the CCA between the combined context [**L R**] and the words **W**, we first take the CCA between the left and right contexts and use the result of that CCA to estimate the state **S** of all the tokens in the corpus from their contexts. Note that we get partially redundant state estimates from the left context and from the right context; these are concatenated to make combined state estimate. This will contain some redundant information, but will not lose any of the differences in information from the left and right sides. We then take the CCA between **S** and the words **W** to get our final *eigenword dictionary*. This is summarized in Algorithm 1. The first step, the CCA between **L** and **R**, must produce at least as many canonical components as the second step, which produces the final output.

The two step method requires fewer tokens of data to get the same accuracy in estimating the *eigenword dictionary* because its final step estimates fewer parameters $O(vk)$ than the OSCCA does $O(v^2)$.

Before stating the theorem, we first explain this intuitively. Predicting each word as a function of all other word combinations that can occur in the context is far sparser than predicting low dimensional state from context, and then predicting word from state. Thus, for relatively infrequent words, OSCCA should have significantly lower accuracy than the two step version. Phrased differently, mapping from context to state and then from state to word (TSCCA) gives a more parsimonious model than mapping directly from context to word (OSCCA).

The relative ability of OSCCA to estimate hidden state compared to that of TSCCA can be summarized as follows:

Theorem 2. *Given a matrix of words,* $\mathbf{W}$ *and their associated left and right contexts,* $\mathbf{L}$ *and* $\mathbf{R}$ *with vocabulary size* v*, context size* h*, and corpus of* n *tokens. The ratio of the dimension of the hidden state that needs to be estimated by TSCCA in order to recover with high probability the information in the true state to the corresponding dimension needed for OSCCA is* $\frac{h+k}{hv}$.

Please see the appendix for a proof of the above theorem.

Since the corpora we care about usually have $vh \gg h+k$, the TSCCA procedure will in expectation correctly estimate hidden state with a much smaller number of components k than the one step procedure. Or, equivalently, for an estimated hidden state of given size k, TSCCA will correctly estimate more of the hidden state components.

As mentioned earlier, words have a Zipfian distribution so most words are rare. For such rare words, if one computes a CCA between them and their contexts, one will have very few observations, and hence will get a low quality estimate of their eigenword vector. If, on the other hand, one first estimates a state vector for the rare words, and then does a CCA between this state vector and the context, the rare words can be thought of as borrowing strength from more common distributionally similar words. For example, “umbrage” (56,020) vs. “annoyance” (777,061) or “unmeritorious” (9,947) vs. “undeserving” (85,325). The numbers in parentheses are the number of occurrences of these words in the Google n-gram collection used in some of our experiments.

3.3 Low Rank Multi-View Learning (LR-MVL)

The context around a word, consisting of the h words to the right and left of it, sits in a high dimensional space, since for a vocabulary of size v, each of the h words in the context requires an indicator function of dimension v. So, we propose an algorithm Low Rank Multi-View Learning (LR-MVL), where we work in the k dimensional space to begin with.

The key move in LR-MVL is to project the hv-dimensional $\mathbf{L}$ and $\mathbf{R}$ matrices down to a k dimensional state space before performing the first CCA. This is where it differs from TSCCA. Thus, all eigenvector computations are done in a space that is v/k times smaller than the original space. Since a typical vocabulary contains at least $100,000$ words, and we use state spaces of order $k \approx 100$ dimensions, this gives a 1,000-fold reduction in the size of calculations that are needed.

LR-MVL iteratively updates the real-valued state of a token $\mathbf{Z_t}$, till convergence. Since, the state is always real-valued, this also allows us to replace the projected left and right contexts with exponential smooths (weighted average of the previous (or next) token's state i.e. $\mathbf{Z_{t-1}}$ (or $\mathbf{Z_{t+1}}$) and previous (or next) token's smoothed state i.e. $\mathbf{S_{t-1}}$ (or $\mathbf{S_{t+1}}$).), of them at a few different time scales. One could use a mixture of both very short and very long contexts which capture short and long range dependencies as required by NLP problems as NER, Chunking, WSD etc. Since exponential smooths are linear, we preserve the linearity of our method.

We now describe the LR-MVL algorithms.

3.3.0.1 The LR-MVL Algorithms

Based on our theory (described in next subsection), various algorithms are possible for LR-MVL. We provide two algorithms, Algorithms 2, 3 (without and with exponential smooths).

Algorithm 2 LR-MVL (I) Algorithm - Learning from Large amounts of Unlabeled Data (no exponential smooths).

```
Input: Token sequence $\mathbf{W}_{n \times v}$, state space size $k$.
Initialize the eigenfeature dictionary $\hat{\mathbf{A}}_{v \times k}$ to random values $\mathcal{N}(0, 1)$.
repeat
    Project the left and right context matrices $\mathbf{L}_{n \times vh}$ and $\mathbf{R}_{n \times vh}$ down to 'k' dimensions
    and compute CCA between them. $[\mathbf{\Phi_L}, \mathbf{\Phi_R}]$=CCA($\mathbf{L\hat{A}_h}$, $\mathbf{R\hat{A}_h}$). //$\mathbf{A}_h$ is the stacked
    version of $\hat{\mathbf{A}}$ matrix as many times as the context length 'h'.
    Normalize $\Phi_L^{(k)}$ and $\Phi_R^{(k)}$. //Divide each row by the maximum absolute value in that
    row (Scales between -1 and +1).
    Compute a second CCA between the estimated state and the word itself
    $[\mathbf{\Phi}_W, \mathbf{\Phi}_C]$=CCA($\mathbf{W}$, $[\mathbf{L\hat{A}_h\Phi_L^{(k)}}$, $\mathbf{R\hat{A}_h\Phi_R^{(k)}}]$).
    $\hat{\mathbf{A}} = \mathbf{\Phi}_W^{(k)}$
    Compute the change in $\hat{\mathbf{A}}$ from the previous iteration
until $|\Delta\hat{\mathbf{A}}| < \epsilon$
Output: $\mathbf{\Phi}_L$, $\mathbf{\Phi}_R$, $\hat{\mathbf{A}}$ .
```

A few iterations (~ 10) of the above algorithms are sufficient to converge to the solution. (Since the optimizations are convex, there is a single solution, so there is no issue of local minima.) If the assumptions (1, 1A, 2 and 3) (in appendix) are satisfied, our methods converge equally rapidly to the true canonical variates.

3.3.0.2 Theoretical Properties of LR-MVL

We now present the theory behind the LR-MVL algorithms; particularly we show that the reduced rank matrix $\mathbf{A}$ allows a significant data reduction while preserving the information in our data and the estimated state does the best possible job of capturing any label information that can be inferred by a linear model.

The key difference from TSCCA is that we can initialize the state of each word randomly and work in a low (k) dimensional space from the beginning, iteratively

Algorithm 3 LR-MVL (II) Algorithm - Learning from Large amounts of Unlabeled Data (with exponential smooths).

```
**Input:** Token sequence $\mathbf{W}_{n \times v}$, state space size $k$, smoothing rates $\alpha^j$
Initialize the eigenfeature dictionary $\hat{\mathbf{A}}$ to random values $\mathcal{N}(0,1)$.
**repeat**
    Set the state $Z_t$ $(1 < t \le n)$ of each token $w_t$ to the eigenword vector of the corresponding word.
        $Z_t = (\hat{\mathbf{A}}_w : w = w_t)$
    Smooth the state estimates before and after each token to get a pair of views for each smoothing rate $\alpha^j$.
        $S_t^{(l,j)} = (1 - \alpha^j) S_{t-1}^{(l,j)} + \alpha^j Z_{t-1}$ // left view $\mathbf{L}$
        $S_t^{(r,j)} = (1 - \alpha^j) S_{t+1}^{(r,j)} + \alpha^j Z_{t+1}$ // right view $\mathbf{R}$.
    where the $t^{th}$ rows of $\mathbf{L}$ and $\mathbf{R}$ are, respectively, concatenations of the smooths $S_t^{(l,j)}$ and $S_t^{(r,j)}$ for each of the $\alpha^{(j)}$s.
    Find the left and right canonical correlates, which are the eigenvectors $\mathbf{\Phi}_l$ and $\mathbf{\Phi}_r$ of
        $(\mathbf{L'L})^{-\mathbf{1}}\mathbf{L'R}(\mathbf{R'R})^{-\mathbf{1}}\mathbf{R'L}\mathbf{\Phi}_l = \lambda\mathbf{\Phi}_l$.
        $(\mathbf{R'R})^{-\mathbf{1}}\mathbf{R'L}(\mathbf{L'L})^{-\mathbf{1}}\mathbf{L'R}\mathbf{\Phi}_r = \lambda\mathbf{\Phi}_r$.
    Project the left and right views on to the space spanned by the top $k$ left and right CCAs respectively
        $\mathbf{X_l} = \boldsymbol{L}\mathbf{\Phi}_l^{(k)}$ and $\mathbf{X_r} = \boldsymbol{R}\mathbf{\Phi}_r^{(k/2)}$
    where $\mathbf{\Phi}_l^{(k)}$, $\mathbf{\Phi}_r^{(k)}$ are matrices composed of the singular vectors of $\mathbf{\Phi}_l$, $\mathbf{\Phi}_r$ with the $k$ largest magnitude singular values. Estimate the state for each word $w_t$ as the union of the left and right estimates: $\mathbf{Z} = [\mathbf{X_l}, \ \mathbf{X_r}]$
    Compute a second CCA between the estimated state and the word itself $[\mathbf{\Phi}_W, \mathbf{\Phi}_Z]$=CCA($\mathbf{W}$, $\mathbf{Z}$).
    $\hat{\mathbf{A}}_w = \mathbf{\Phi}_W$.
    Normalize $\hat{\mathbf{A}}_\mathbf{w}$. //Divide each row by the maximum absolute value in that row (Scales between -1 and +1).
    Compute the change in $\mathbf{A}$ from the previous iteration
**until** $|\Delta\hat{\mathbf{A}}| < \epsilon$
**Output:** $\mathbf{\Phi}_l^k$, $\mathbf{\Phi}_r^k$, $\hat{\mathbf{A}}$ .
```

refining the state until convergence and still we can recover the eigenword dictionary ($\mathbf{A}$).

As earlier, let $\mathbf{L}$ be an $n \times hv$ matrix giving the words in the left context of each of the n tokens, where the context is of length h, $\mathbf{R}$ be the corresponding $n \times hv$ matrix for the right context, and $\mathbf{W}$ be an $n \times v$ matrix of indicator functions for the words themselves.

Lemma 1. *Define* $\mathbf{A}$ *by the following limit of the right singular vectors:*

$$CCA_k(\mathbf{W}, [\mathbf{L}, \mathbf{R}])_{left} \approx \mathbf{A}.$$

Under assumptions 2, 3 and 1A, (in appendix) such that if $CCA_k(\mathbf{L}, \mathbf{R}) \equiv [\mathbf{\Phi}_L, \mathbf{\Phi}_R]$ *then*

$$CCA_k(\mathbf{W}, [\mathbf{L}\mathbf{\Phi}_L, \mathbf{R}\mathbf{\Phi}_R])_{left} \approx \mathbf{A}.$$

Please see the appendix for the proof.

Lemma 3 shows that instead of finding the CCA between the full context and the words, we can take the CCA between the Left and Right contexts, estimate a k dimensional state from them, and take the CCA of that state with the words and get the same result. Lemma 3 is similar to Theorem 2, except that it does not provide ratios of the estimated state sizes.

Let $\tilde{\mathbf{A}}_h$ denote a matrix formed by stacking h copies of $\mathbf{A}$ on top of each other. Right multiplying $\mathbf{L}$ or $\mathbf{R}$ by $\tilde{\mathbf{A}}_h$ projects each of the words in that context into the k-dimensional reduced rank space.

The following theorem addresses the core of the LR-MVL algorithm, showing that there is an $\mathbf{A}$ which gives the desired dimensionality reduction. Specifically, it shows that the previous lemma also holds in the reduced rank space.

Theorem 3. *Under assumptions 1, 1A and 2 (in appendix) there exists a unique matrix* $\mathbf{A}$ *such that if*

$$CCA_k(\mathbf{L}\tilde{\mathbf{A}}_\mathbf{h}, \mathbf{R}\tilde{\mathbf{A}}_\mathbf{h}) \equiv [\tilde{\mathbf{\Phi}}_L, \tilde{\mathbf{\Phi}}_R]$$

then

$$CCA_k(\mathbf{W}, [\mathbf{L}\tilde{\mathbf{A}}_\mathbf{h}\tilde{\mathbf{\Phi}}_\mathbf{L}, \mathbf{R}\tilde{\mathbf{A}}_\mathbf{h}\tilde{\mathbf{\Phi}}_\mathbf{R}])_{left} \approx \mathbf{A}$$

where $\tilde{\mathbf{A}}_h$ *is the stacked form of* $\mathbf{A}$*.*

See the appendix for the Proof [2].

Because of the Zipfian distribution of words, many words are rare or even unique. So, just as in the case of TSCCA, CCA between the rare words and context will not be informative, whereas finding the CCA between the projections of left and right contexts gives a good state vector estimate even for unique words. One can then fruitfully find the CCA between the contexts and the estimated state vector for their associated words.

3.4 Generating Context Specific Embeddings

Once we have estimated the CCA model using any of our algorithms (i.e. OSCCA, TSCCA, LR-MVL), it can be used to generate context specific embeddings for the tokens from training, development and test sets (as described in Algorithm 4). These embeddings could be further supplemented with other baseline features and used in a supervised learner to predict the label of the token.

[2] Our matrix $\mathbf{A}$ corresponds to the matrix $\hat{U}$ used by (Hsu et al. 2009; Siddiqi et al. 2010). They showed that U is sufficient to compute the probability of a sequence of words generated by an HMM; although we do not show it here, our A provides a more statistically efficient estimate of U than their $\hat{U}$, and hence can also be used to estimate the sequence probabilities.

Algorithm 4 Inducing Context Specific Embeddings for Train/Dev/Test Data

```
Input: Model ($\mathbf{\Phi}_l^k$, $\mathbf{\Phi}_r^k$, $\mathbf{A}$) output from above algorithm and Token sequences $\mathbf{W^{train}}$, ($\mathbf{W^{dev}}$, $\mathbf{W^{test}}$)
Project the left and right views $L$ and $R$ onto the space spanned by the top $k$ left and right CCAs respectively. If algorithm is Algorithm 3, then, smooth $\mathbf{L}$ and $\mathbf{R}$ first.
    $\mathbf{X_l} = \boldsymbol{L}\mathbf{\Phi}_l^k$ and $\mathbf{X_r} = \boldsymbol{R}\mathbf{\Phi}_r^k$
    and the words onto the eigenfeature dictionary    $\mathbf{X_w} = \boldsymbol{W}^{train}\boldsymbol{A}$
Form the final embedding matrix $\mathbf{X_{train:embed}}$ by concatenating these three estimates of state
    $\mathbf{X_{train:embed}} = [\mathbf{X_l} \ , \mathbf{X_w} \ , \mathbf{X_r}]$
Output: The embedding matrices $\mathbf{X_{train:embed}}$, ($\mathbf{X_{dev:embed}}$, $\mathbf{X_{test:embed}}$) with context-specific representations for the tokens.
```

Note that we can get context "oblivious" embeddings i.e. one embedding per word type, just by using the eigenfeature dictionary ($\mathbf{A_{v \times k}}$).

3.5 Efficient Estimation

As mentioned earlier, CCA can be done by taking the singular value decomposition of a matrix. For small matrices, this can be done using standard functions in e.g. MATLAB, but for very large matrices (e.g. for vocabularies of tens or hundreds of thousands of words), it is important to take advantage of the recent advances in SVD algorithms. For our experiments we use the method of (Halko et al. 2011), which uses random projections to compute SVD of large matrices.

The key idea is to find a lower dimensional basis for $\mathbf{A}$, and to then compute the singular vectors in that lower dimensional basis. The initial basis is generated randomly, and taken to be slightly larger than the eventual basis. If $\mathbf{A}$ is $v \times hv$, and we seek a state of dimension k, we start with a $hv \times (k + l)$ matrix $\mathbf{\Omega}$ of random numbers, where l is number of "extra" basis vectors between 0 and k. We then

Algorithm 5 Randomized singular value decomposition

```
1: **Input:** matrix $\mathbf{A}$ of size $v \times hv$, the desired hidden state dimension $k$, and the number of "extra" singular vectors, $l$
2: Generate a $hv \times (k + l)$ random matrix $\mathbf{\Omega}$
3: **for** i =1:5 **do**
4:    $\mathbf{M} = \mathbf{A\Omega}$.
5:    $[\mathbf{Q}, \mathbf{R}]$=QR($\mathbf{M}$) //Find $v \times (k + l)$ orthogonal matrix $\mathbf{Q}$.
6:    $\mathbf{B} = \mathbf{Q}^\top \mathbf{A}$
7: **end for**
8: Find the SVD of $\mathbf{B}$. $[\hat{\mathbf{U}}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{V}}^\top]$ =SVD($\mathbf{B}$), and keep the $k$ components of $\hat{\mathbf{U}}$ with the largest singular values.
9: $\tilde{\mathbf{A}} = \mathbf{Q}\hat{\mathbf{U}}$. //Compute the rank-k projection.
10: **Output:** The rank-k approximation $\tilde{\mathbf{A}}$. (Similar procedure can be repeated to get the right singular values and the corresponding projections.)
```

project $\mathbf{A}$ onto this matrix and take the SVD decomposition of the resulting matrix ($\mathbf{A} \approx \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}^\top}$).

Since $\mathbf{A\Omega}$ is $v \times (k+l)$, this is much cheaper than working on the original matrix $\mathbf{A}$. We keep the largest k components of $\mathbf{U}$ and of $\mathbf{V}$, which form a left and a right basis for $\mathbf{A}$ respectively.

This procedure is repeated for a few (~ 5) iterations. The algorithm is summarized in Algorithm 5.

(Halko et al. 2011) prove a number of nice properties of the above algorithm. In particular, they guarantee that the algorithm, even without the extra iterations in steps 3 and 6 produces an approximation whose error is bounded by a small polynomial factor times the size of the largest singular value whose singular vectors are *not* part of the approximation, σ_{k+1}. They also show that using a small number of "extra" singular vectors (l) results in a substantial tightening of the bound, and that the extra iterations, which correspond to power iteration, drive the error bound exponentially quickly to one times the largest non-included singular value, σ_{k+1}

3.6 Conclusion

We proposed two new and improved spectral method, two-step alternative (TSCCA) and (Low Rank Multi-View Learning) LR-MVL which provide more accurate state estimates (lower sample complexity) for small corpora than standard OSCCA.

LR-MVL can also be viewed as a type of co-training (Blum and Mitchell 1998): The state of each token w_t is similar to that of the tokens both before and after it, and it is also similar to the states of the other occurrences of the same word elsewhere in the document. LR-MVL takes advantage of these two different types of similarity by alternately estimating word state using CCA on the states of the words before and after each target token and using the average over the states (second CCA) associated with all other occurrences of that word.

Chapter 4

Evaluating Eigenwords

In this chapter[1] we provide qualitative and quantitative evaluation of the various eigenword algorithms. The state estimates for words capture a wide range of information about them that can be used to predict part of speech, linguistic features, and meaning. Before presenting a more quantitative evaluation of predictive accuracy, we present some qualitative results showing how word states, when projected in appropriate directions usefully characterize the words.

We compare our approach against a variety of state-of-the-art word embeddings:

1. Turian Embeddings (C&W and HLBL) (Turian et al. 2010).

2. SENNA Embeddings (Collobert et al. 2011).

3. word2vec Embeddings (Mikolov et al. 2013a,b).

We also compare against simple PCA/LSA embeddings and other model based approaches wherever applicable.

We downloaded the Turian embeddings (C&W and HLBL), from `http://metaoptimize.com/projects/wordreprs` and use the best 'k' reported in the paper (Turian et al.

[1] This chapter is based on work in (Dhillon et al. 2011),(Dhillon et al. 2012b) and (Dhillon et al. 2014 (Under Review).

2010) i.e. k=200 and 100 respectively. SENNA embeddings were downloaded from `http://ronan.collobert.com/senna/`. word2vec code was downloaded from `https://code.google.com/p/word2vec/`. Since they made the code available we could train them on the exact same corpora, had the exact same context window and vocabulary size as the eigenword embeddings. The PCA baseline used is similar to the one that has recently been proposed by (Lamar et al. 2010) except that here we are interested in supervised accuracy and not the unsupervised accuracy as in that paper.

In the results presented below (qualitative and quantitative), we trained all the algorithms (including eigenwords) on Reuters RCV1 corpus (Rose et al. 2002) for uniformity of comparison[2]. Case was left intact and we did not do any other "cleaning" of data. Tokenization was performed using NLTK tokenizer (Bird and Loper 2004). RCV1 corpus contains Reuters newswire from Aug '96 to Aug '97 and containing about 215 million tokens after tokenization.

Unless otherwise stated, we consider a fixed window of two words (h=2) on either side of a given word and a vocabulary of 100,000 most frequent words for all the algorithms[3], in order to ensure fairness of comparison.

Eigenword algorithms are robust to the dimensionality of hidden space (k), so we did not tune it and fixed it at 200. For other algorithms, we report results using their best hidden space dimensionality.

Our theory and CCA in general (Bach and Jordan 2005) rely on normality assumptions, however the words follow Zipfian (heavy tailed) distribution. So, we took the square root of the word counts in the context matrices before running OSCCA, TSCCA and LR-MVL(I). This squishes the word distributions and makes them look more normal. We ran LR-MVL(I) and LR-MVL(II) for 10 iterations and

[2] word2vec, PCA and Turian (C&W and HLBL) embeddings are all trained on Reuters RCV1, but SENNA embeddings (training code not available) were trained on a larger Wikipedia corpus.

[3] Turian (C&W and HLBL), SENNA embeddings had much bigger vocabulary sizes of 268,000 and 130,000, though they also use a window of 2 as context.

only used one exponential smooth of 0.5 for LR-MVL(II).

4.1 Qualitative Evaluation of OSCCA

To illustrate the sorts of information captured in our state vectors, we present a set of figures constructed by projecting selected small sets of words onto the space spanned by the second and third largest principal components of their eigenword dictionary values, which are simply the left canonical correlates calculated from Equation 2.4. (The first principle component generally just separates the selected words from other words, and so is less interesting here.)

Figure 4.1 shows plots for three different sets of words. The left column uses the eigenword dictionary learned using OSCCA (the other eigenword algorithms gave similar results), while the right column uses the corresponding latent vectors derived using PCA on the same data. In all cases, the 200-dimensional vectors have been projected onto two dimensions (using a second PCA) so that they can be visualized.

The PCA algorithm differs from CCA based (eigenword) algorithms in that it does not whiten the matrices via (${\mathbf{C_{ll}}}^{-1/2}$ and ${\mathbf{C_{rr}}}^{-1/2}$) before performing SVD. If one considers a word and its two grams to the left and right as a document, then its equivalent to the Latent Semantic Analysis (LSA) algorithm.

The results for various (handpicked) semantic categories are shown in Figure 4.1 and 4.2

The top row shows a small set of randomly selected nouns and verbs. Note that for CCA, nouns are on the left, while verbs are on the right. Words that are of similar or opposite meaning (e.g. “agree” and “disagree”) are distributionally similar, and hence close. The corresponding plot for PCA shows some structure, but does not give such a clean separation. This is not surprising; predicting the part of speech of words depends on the exact order of the words in their context (as we

Center Word	**OSCCA NN**	**PCA NN**
market	markets, trade, currency, sector, activity.	dollar, economy, government, sector, industry.
company	firm, group, giant, operator, maker.	government, group, dollar, following, firm.
Ltd	Limited, Bhd, Plc, Co, Inc.	Corp, Plc, Inc, name, system.
President	Governor, secretary, Chairman, leader, Director.	Commerce, General, fuel, corn, crude.
Nomura	Daiwa, UBS, HSBC, NatWest, BZW.	Chrysler, Sun, Delta, Bre-X, Renault.
jump	drop, fall, rise, decline, climb.	surge, stakes, slowdown, participation, investing.
rupee	peso, zloty, crown, pound, franc.	crown, CAC-40, FTSE, Nikkei, 30-year.

Table 4.1: Nearest Neighbors of OSCCA and PCA word embeddings.

capture in CCA); a PCA-style bag-of-words can't capture part of speech well.

The bottom row in Figure 4.1 shows names of numbers or the numerals representing numbers and years. Numbers that are close to each other in value tend to be close in the plot, thus suggesting that state captures not just classifications, but also more continuous hidden features.

The plots in Figure 4.2 show a similar trend i.e., eigenword embeddings are able to provide a clear separation between different syntactic/semantic categories and capture a rich set of features characterizing the words, whereas PCA mostly just squishes them together.

Table 4.1 shows the five nearest neighbors for a few representative words using OSCCA and PCA. As can be seen, the OSCCA based nearest neighbors capture subtle semantic and syntactic cues e.g Japanese investment bank (Nomura) having another Japanese investment bank (Daiwa) as the nearest neighbor, whereas the PCA nearest neighbors are more noisy and capture mostly syntactic aspects of the word.

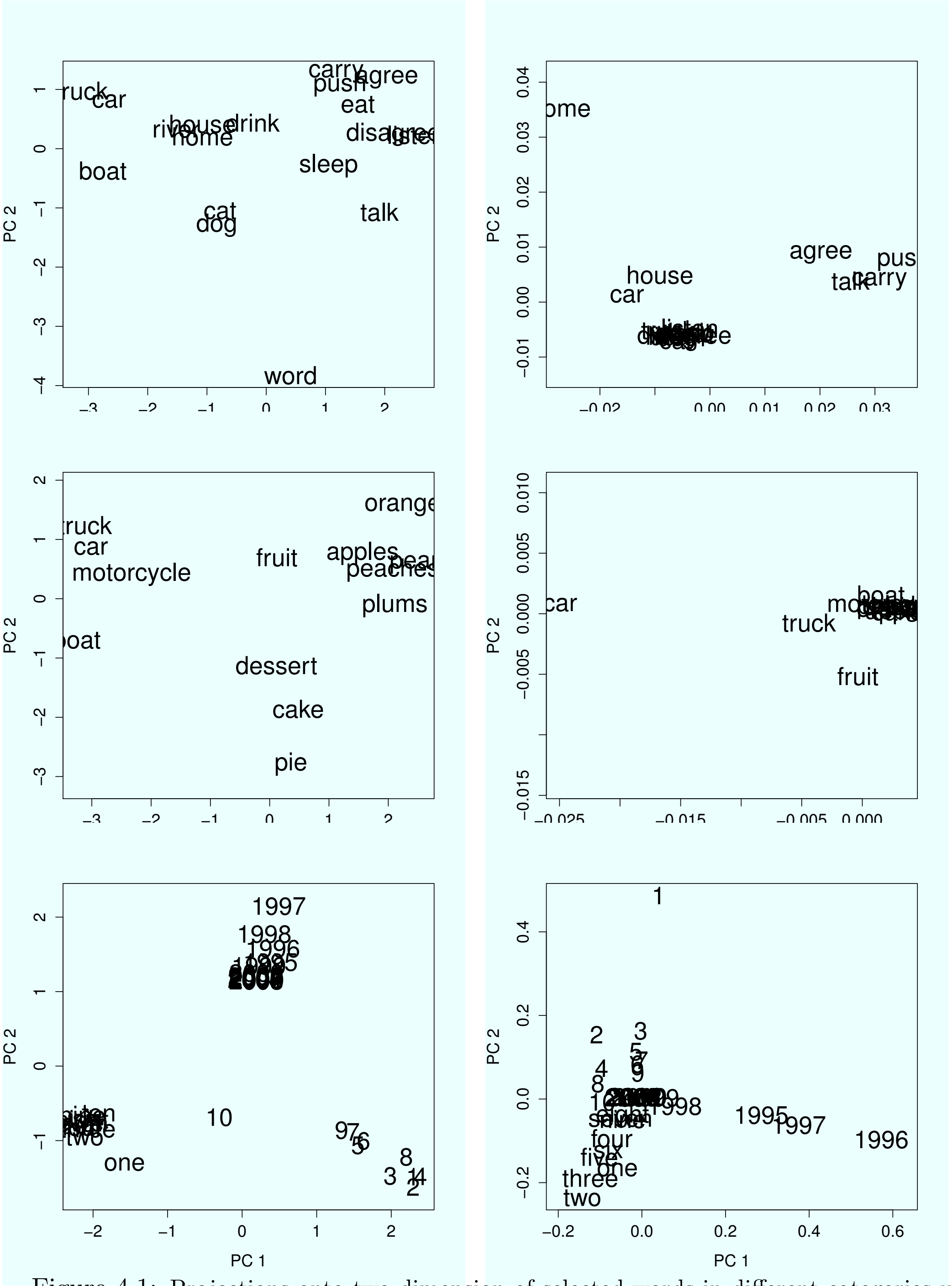


Figure 4.1: Projections onto two dimension of selected words in different categories using both OSCCA (left) and PCA (Right). Top to bottom: 1). (Nouns vs Verbs): house, home, dog, truck, boat, word, river, cat, car, sleep, eat, push, drink, listen, carry, talk, disagree, agree. 2). (Eateries vs vehicles): apples, pears, plums, oranges, peaches, fruit, cake, pie, dessert, truck, boat, car, motorcycle. 3). (Numerals vs letter numbers vs years): one, two, three, four, five, six, seven, eight, nine, ten, 1, 2,. . ., 10, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009.

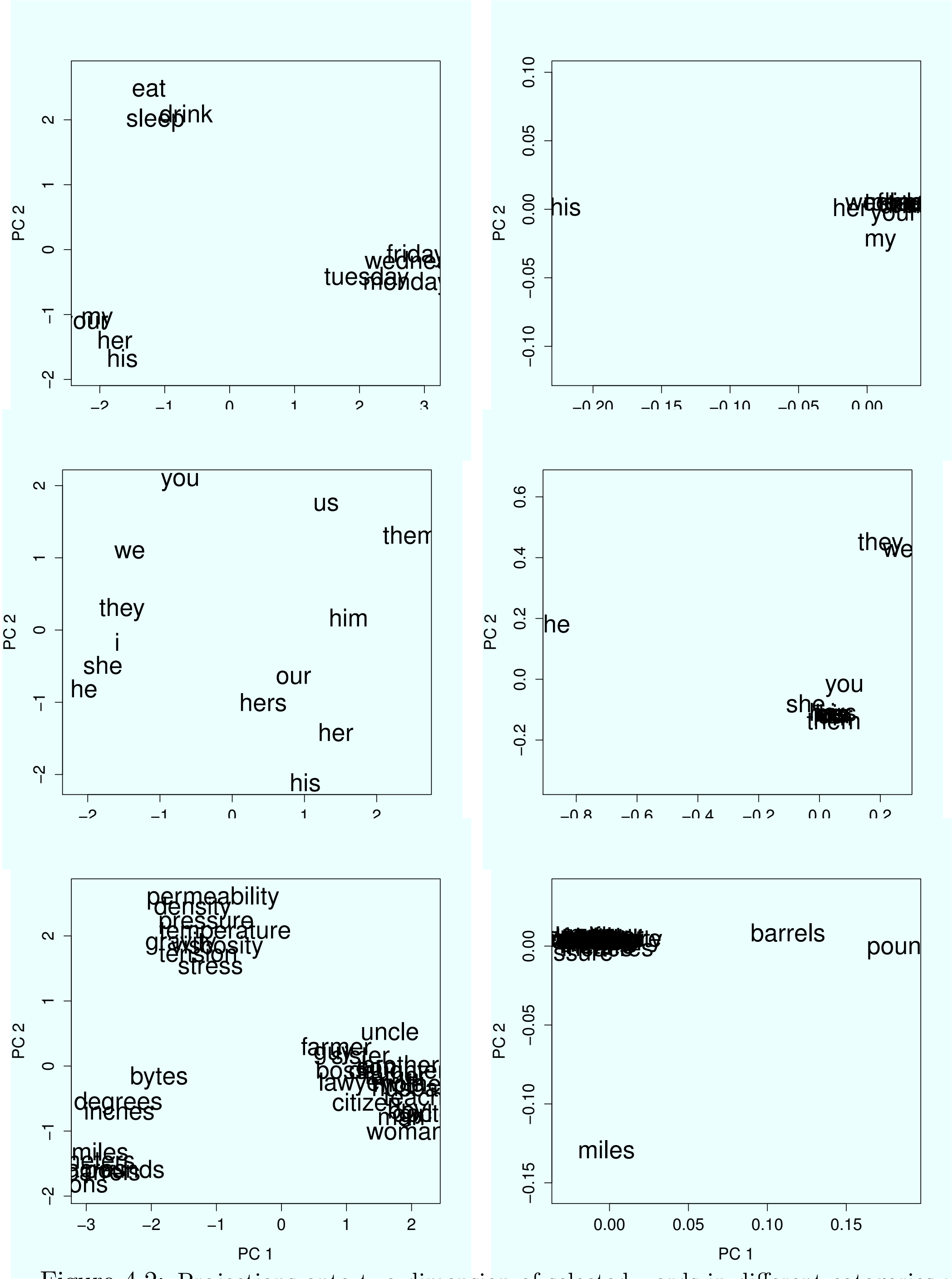


Figure 4.2: Projections onto two dimension of selected words in different categories using both OSCCA (left) and PCA (Right). Top to bottom: 1). (Weekdays vs verbs vs pronouns): monday, tuesday, wednesday, sunday, friday, eat, drink, sleep, his, her, my, your. 2). (Different kinds of pronouns): i, you, he, she, they, we, us, them, him, her, our, his, hers. 3). (Nouns vs Adjectives vs Units of measurement): man, woman boy, girl, lawyer, doctor, guy, farmer, teacher, citizen, mother, wife, father, son, husband, brother, daughter, sister, boss, uncle, pressure, temperature, permeability, density, stress, viscosity, gravity, tension, miles, pounds, degrees, inches, barrels, tons, acres, meters, bytes.

4.2 Quantitative Evaluation

This section describes the performance (accuracy and richness of representation) of various eigenword algorithms. We evaluate the quality of the *eigenword dictionary* by using it in a supervised learning setting to predict a wide variety of labels that can be attached to words.

We first compare OSCCA against TSCCA, LR-MVL(I) and LR-MVL II) embeddings on a set of Part of Speech (POS) tagging problems for different languages, looking at how the predictive accuracy scales with corpus size for predictions on a fixed vocabulary. These results use small corpora and highlight that TSCCA, LR-MVL(I) and LR-MVL(II) perform better for rarer words.

Next, we perform experiments for a variety of NLP tasks including, Word Similarity, Sentiment Classification, Named Entity Recognition (NER), chunking and Google semantic and syntactic analogy tasks to demonstrate the richness of the state learned by eigenwords and that they perform comparably or better than other state-of-the-art approaches. For these tasks, we report results using the best eigenwords for compactness, though all the four algorithms gave similar performances.

4.2.1 Part of Speech (POS) Tagging

We compare the performance of various eigenword algorithms on the task of POS tagging in four different languages. Note that this experiment is performed only to show the improved performance of TSCCA, LR-MVL (I) and LR-MVL (II) for rarer words compared to OSCCA and not to show the superior performance of eigenwords compared to other state-of-the-art algorithms.

Table 1 provides statistics of all the corpora used, namely: the Wall Street Journal portion of the Penn treebank (Marcus et al. 1993) (we consider the 17 tags of (PTB 17) (Smith and Eisner 2005)), the Bosque subset of the Portuguese Flo-

Language	Number of POS tags	Number of tokens
English	17	100311
Danish	25	100238
Bulgarian	12	100489
Portuguese	22	100367

Table 4.2: Description of the POS tagging datasets

resta Sinta(c)tica Treebank (Afonso et al. 2002), the Bulgarian BulTreeBank (Simov et al. 2002) (with only the 12 coarse tags), and the Danish Dependency Treebank (DDT) (Kromann 2003).

Note that some corpora like English have $\sim$ 1 million tokens whereas Danish only has $\sim$ $100k$ tokens. To address this data imbalance we kept only the first $\sim$ $100k$ tokens of the larger corpora so as to perform a uniform evaluation across all corpora.

Theorem 2 implies that the difference between OSCCA and TSCCA/LR-MVL(I)/LR-MVL(II) should be more pronounced at smaller sample sizes, where the errors are higher and that they should have similar predictive power asymptotically when we learn them using large amounts of data. So, we evaluate the performance of the methods on varying data sizes ranging from $5k$ to the entire $100k$ tokens. As mentioned earlier, we take a context size of $h = 2$ for eigenwords i.e. a word to the left and a word to the right; for PCA this reduces to a bag of 5-grams. It is important to note that for POS tagging usually a small context (in our case $h = 2$) is sufficient to get state-of-the-art performance as can be substantiated by trigram POS taggers e.g. (Merialdo 1994), so we need not consider longer contexts.

As mentioned earlier, for the unlabeled learning part i.e. learning using eigenwords/PCA we are interested in seeing the eigenword dictionary estimates for the word types (for a fixed vocabulary) get better with more data. So, when varying the unlabeled data from $5k$ to $100k$ we made sure that they had the exact same vocabulary and that the performance improvement is not coming from word types not present in the $5k$ tokens but present in the total $100k$.

To evaluate the predictive accuracy of the descriptors learned using different amounts of unlabeled data, we learn a multi-class logistic regression to predict the POS tag of each type. We trained using 80% of the word types chosen randomly and then tested on the remaining 20% types and this procedure was repeated 10 times. Its important to note that our train and test sets do not contain any of the same word types.[4]

The accuracy of using OSCCA, TSCCA, LR-MVL(I), LR-MVL(II) and PCA features in a supervised learner are shown in Figure 4.3 for the task of POS tagging. As can be seen from the results, eigenword embeddings are significantly better (5% significance level in a paired t-test) than the PCA-based supervised learner. Among the eigenwords, TSCCA, LR-MVL(I) and LR-MVL(II) are significantly better than OSCCA for small amounts of data, and (as predicted by theory) the two become comparable in accuracy as the amount of unlabeled data used to learn the CCAs becomes large.

4.2.2 Word Similarity Task (WordSim-353)

A standard dataset for evaluating vector-space models is the WordSim-353 dataset (Finkelstein et al. 2001) which consists of 353 pairs of nouns. Each pair is presented without context and associated with 13 to 16 human judgments on similarity and relatedness on a scale from 0 to 10. For example, (professor, student) received an average score of 6.81, while (professor, cucumber) received an average score of 0.31.

For this task, its interesting to see how well the cosine similarity between the word embeddings correlates with the human judgement of similarity between the same two words. The results in Table 4.3 show the Spearman's correlation between

[4] We are doing non-disambiguating POS tagging i.e. each word type has a single POS tag, so if the same word type occurred in both the training and testing data, a learning algorithm that just memorized the training set would perform reasonably well.

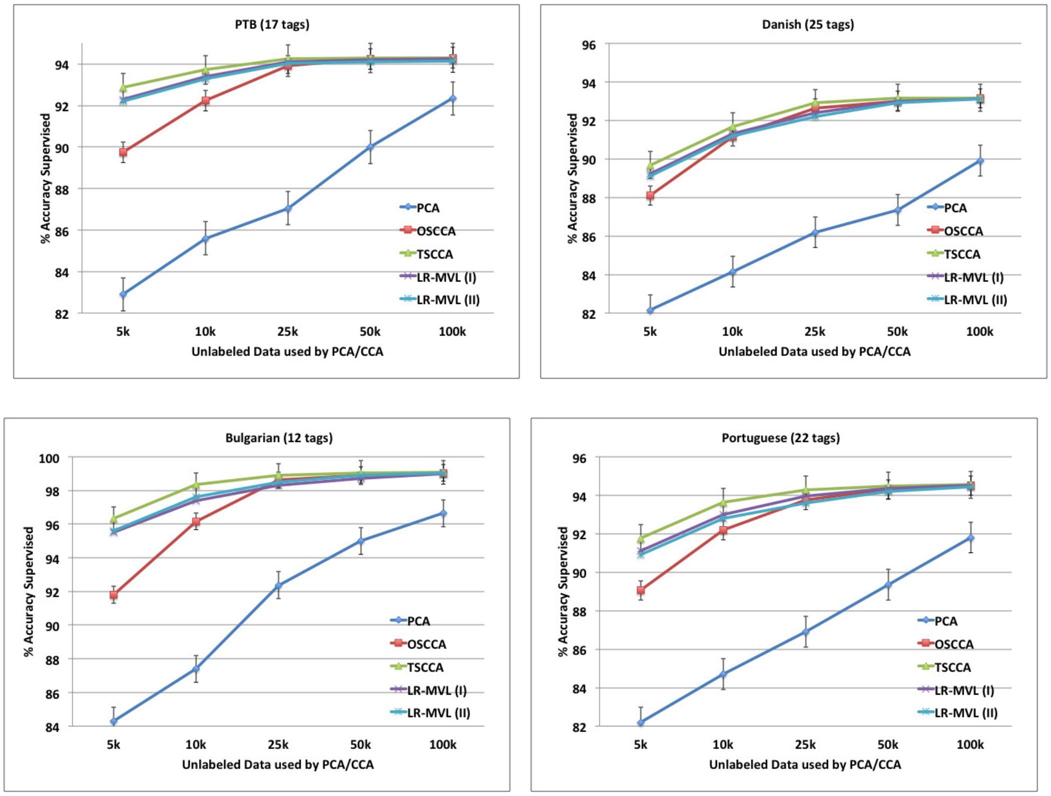


Figure 4.3: Plots showing accuracy as a function of number of tokens used to train the PCA/eigenwords for various languages. **Note:** The results are averaged over 10 random, 80 : 20 splits of word types.

the cosine similarity of the respective word embeddings and the human judgements.

As can be seen, eigenwords are statistically significantly (computed using resampled bootstrap) better than all embeddings except SENNA.

4.2.3 Sentiment Classification

It is often useful to group words into semantic classes such as colors or numbers, professionals or disciplines, happy or sad words, words of encouragement or discouragement, etc.

Many people have collected sets of words that indicate positive or negative sentiment. More generally, substantial effort has gone into creating hand-curated words

Model	$\rho\times$ 100
PCA	30.25
Turian (C&W)	28.08
Turian (HLBL)	35.24
SENNA	44.32
word2vec (SK)	42.73
word2vec (CB)	42.97
eigenwords (best)	**44.86**

Table 4.3: Table showing the Spearman correlation between the word embeddings based similarity and human judgement based similarity. Note that the numbers for word2vec are different from the ones reported elsewhere, which is due to the fact that we considered a 100,000 vocabulary and a context window of 2 just like eigenwords, in order to make a fair comparison.

that can be used to capture a variety of opinions about different products, papers, or people. For example (Teufel 2010) contains dozens of carefully constructed lists of words that she uses to categorize what authors say about other scientific papers. Her categories include "problem nouns" (caveat, challenge, complication, contradiction, ...), "comparison nouns" (accuracy, baseline, comparison, evaluation, ...), "work nouns" (account, analysis, approach, ...) as well as more standard sets of positive, negative, and comparative adjectives.

In the example below, we use words from a set of five dimensions that have been identified in positive psychology under the acronym PERMA (Seligman 2011):

- *Positive emotion* (aglow, awesome, bliss, ...),
- *Engagement* (absorbed, attentive, busy, ...),
- *Relationships* (admiring, agreeable, ...),
- *Meaning* (aspire, belong, ...)
- *Achievement* (accomplish, achieve, attain, ...).

For each of these five categories, we have both positive words – ones that connote, for example, *achievement*, and negative words, for example, *un-achievement* (amateurish, blundering, bungling, ...). We would hope (and we show below that this is in fact true), that we can use eigenwords not only to distinguish between different

Word sets	**Number of observations**	
	Class I	Class II
Positive emotion or not	81	162
Meaningful life or not	246	46
Achievement or not	159	70
Engagement or not	208	93
Relationship or not	236	204

Table 4.4: Description of the datasets used. All the data was collected from the PERMA lexicon.

PERMA categories, but also to address the harder task of distinguishing between positive and negative terms in the same category. (The latter task is harder because words that are opposites, such as "large" and "small," often are distributionally similar.)

The description of the PERMA datasets is given in Table 4.4 and Table 4.2.3 shows results for the five PERMA categories. As earlier, we used logistic regression for the supervised binary classification.

As can be seen from the plots, the eigenwords perform significantly (5% significance level in a paired t-test) better than all other embeddings in 3/5 cases and for the remaining 2 cases they perform significantly better than all embeddings except word2vec.

	eigenwords (best) $(\mu \pm \sigma)$	PCA $(\mu \pm \sigma)$	Turian (C&W) $(\mu \pm \sigma)$	Turian (HLBL) $(\mu \pm \sigma)$	SENNA $(\mu \pm \sigma)$	word2vec (SK) $(\mu \pm \sigma)$	word2vec (CB) $(\mu \pm \sigma)$
Positive	25.3± 5.8	33.14 ± 5.6	32.6 ± 5.6	29.7 ± 6.2	29.9 ± 5.1	27.7± 7.3	28.1 ± 6.7
Engagement	**15.3 ± 5.0**	29.6 ± 5.7	26.3 ± 6.2	23.7 ± 5.9	20.9 ±5.1	20.9 ± 5.3	20.5 ± 5.6
Relationship	12.6 ± 3.7	46.3 ± 4.9	36.1 ± 4.3	28.3 ± 4.3	18.9 ± 3.4	15.4 ± 4.3	16.5 ± 3.8
Meaningful	**8.5 ± 3.2**	15.4 ± 3.9	15.9 ± 4.0	16.2 ± 4.3	14.6 ± 3.5	11.6 ± 3.9	14.5 ± 4.4
Achievement	**16.0 ± 5.7**	31.3 ± 5.6	30.3 ± 6.1	23.1 ± 5.6	20.4 ± 4.9	24.3 ±7.4	29.0 ± 6.1

Table 4.5: Binary Classification % test errors ($\sum_{i=1}^{n} \frac{\mathbb{I}[y_i \neq \hat{y}_i]}{n}$) averaged over 100 random 80/20 train/test splits for sentiment classification. Bold (3/5 cases) indicates the cases where eigenwords are significantly better (5% level in a paired t-test) compared to all other embeddings. In the remaining 2/5 cases eigenwords are significantly better than all embeddings except word2vec.

4.2.4 Named Entity Recognition (NER) & Chunking

In this section we present the experimental results of eigenwords on Named Entity Recognition (NER) and chunking. For the previous evaluation tasks we were performing classification of individual words in isolation, however NER and chunking tasks involve assigning tasks to running text. This allows us to induce context specific embeddings i.e. a different embedding for a word based on its context.

4.2.4.1 Datasets and Experimental Setup

For the NER experiments we used the data from CoNLL 2003 shared task and for chunking experiments we used the CoNLL 2000 shared task data[5] with standard training, development and testing set splits. The CoNLL '03 and the CoNLL '00 datasets had $\sim 204K/51K/46K$ and $\sim 212K/-/47K$ tokens respectively for Train/Dev./Test sets.

Named Entity Recognition (NER): We use the same set of baseline features as used by (Zhang and Johnson 2003; Turian et al. 2010) in their experiments. The detailed list of features is as below:

- Current Word w_i; Its type information: all-capitalized, is-capitalized, all-digits and so on; Prefixes and suffixes of w_i
- Word tokens in window of 2 around the current word i.e. $d = (w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$; and capitalization pattern in the window.
- Previous two predictions y_{i-1} and y_{i-2} and conjunction of d and y_{i-1}
- Embedding features (eigenwords, C&W, HLBL, Brown etc.) in a window of 2 around the current word (if applicable).

[5] More details about the data and competition are available at `http://www.cnts.ua.ac.be/conll2003/ner/` and `http://www.cnts.ua.ac.be/conll2000/chunking/`

Following (Ratinov and Roth 2009) we use regularized averaged perceptron model with above set of baseline features for the NER task. We also used their BILOU text chunk representation and fast greedy inference as it was shown to give superior performance.

We also augment the above set of baseline features with gazetteers, as is standard practice in NER experiments.

Chunking: For our chunking experiments we use a similar base set of features as above:

- Current Word w_i and word tokens in window of 2 around the current word i.e. $d = (w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$;
- POS tags t_i in a window of 2 around the current word.
- Word conjunction features $w_i \cap w_{i+1}$, $i \in \{-1, 0\}$ and Tag conjunction features $t_i \cap t_{i+1}$, $i \in \{-2, -1, 0, 1\}$ and $t_i \cap t_{i+1} \cap t_{i+2}$, $i \in \{-2, -1, 0\}$.
- Embedding features in a window of 2 around the current word, including the current word (when applicable).

Since the CoNLL 00 chunking data does not have a development set, we randomly sampled 1000 sentences from the training data (8936 sentences) for development. So, we trained our chunking models on 7936 training sentences and evaluated their F1 score on the 1000 development sentences and used a CRF[6] as the supervised classifier. We tuned the magnitude of the ℓ_2 regularization penalty in CRF on the development set. The regularization penalty that gave best performance on development set was 2. Finally, we trained the CRF on the entire ("original") training data i.e. 8936 sentences.

[6] `http://www.chokkan.org/software/crfsuite/`

4.2.4.2 Results

The results for NER and chunking are shown in Tables 4.6 and 4.7, respectively, which show that eigenwords perform significantly better than state-of-the-art competing methods on both NER and chunking tasks.

		F1-Score	
Embedding/Model		Dev. Set	Test Set
Baseline	No Gazetteers	90.03	84.39
Brown 1000 clusters		92.32	88.52
Turian (C&W)		92.46	87.46
Turian (HLBL)		92.00	88.13
SENNA		-	88.67
word2vec (SK)		92.54	89.40
word2vec (CB)		92.08	89.20
eigenwords (Best)		93.19	**89.99**
Brown, 1000 clusters	With Gazetteers	93.25	89.41
Turian (C&W)		92.98	88.88
Turian (HLBL)		92.91	89.35
SENNA		-	89.59
word2vec (SK)		92.99	89.69
word2vec (CB)		92.93	89.89
eigenwords (Best)		93.97	**90.59**

Table 4.6: NER Results. **Note:** F1-score= Harmonic Mean of Precision and Recall. Note that the numbers reported for eigenwords here are different than those in (Dhillon et al. 2011) as we use a different vocabulary size and different dimensionality than there.

Embedding/Model	Test Set F1-Score
Baseline	93.79
Brown 3200 Clusters	94.11
Turian (HLBL)	94.00
Turian (C&W)	94.10
SENNA	93.94
word2vec (SK)	94.02
word2vec (CB)	94.16
eigenwords (best)	**94.23**

Table 4.7: Chunking Results. Note that the numbers reported for eigenwords here are different than those in (Dhillon et al. 2011) as we use a different vocabulary size and different dimensionality than there.

4.2.4.3 Modeling Context: Context Sensitive Embeddings

Modeling the context in embeddings gives decent improvements in accuracies on both NER and chunking problems. For the case of NER, the polysemous words were mostly like *Chicago, Wales, Oakland etc.*, which could either be a *location* or *organization* (Sports teams, Banks etc.), so when we don't use the gazetteer features, (which are known lists of cities, persons, organizations etc.) we got higher increase in F-score by modeling context, compared to the case when we already had gazetteer features which captured most of the information about polysemous words for NER dataset and modeling the context didn't help as much. The polysemous words for the chunking dataset included words such as *spot (VP/NP), never (VP/ADVP), more (NP/VP/ADVP/ADJP) etc.* and in this case embeddings with context helped significantly, giving $3.1 - 6.5\%$ relative improvement in accuracy over context oblivious embeddings.

4.2.5 Google Semantic and Syntactic Relations Task

(Mikolov et al. 2013a,b) present new syntactic and semantic relation datasets composed of analogous word pairs. The syntactic relations dataset contains word pairs that are different syntactic forms of a given word e.g. write : writes :: eat : eats There are nine such different kinds of relations: adjective-adverb, opposites, comparative, superlative, present participle, nation-nationality, past tense, plural nouns and plural verbs

The semantic relations dataset contains pairs of tuples of word relations that follow a common semantic relation e.g. in Athens : Greece :: Canberra : Australia, where the two given pairs of words follow the country-capital relation. There are three other such kinds of relations: country-currency, man-woman, city-in-state and overall 8869 such pairs of words. The task here is to find a word d that best fits the following relationship: a : b :: c : d given a, b and c. They use the vector offset method, which assumes that the words can be represented as vectors in vector space and computes the offset vector: $y_d = e_a - e_b + e_c$ where e_a, e_b and e_c are the vector embeddings for the words a, b and c. Then, the best estimate of d is the word in the entire vocabulary whose embedding has the highest cosine similarity with y_d. Note that this is a hard problem as it is a v class problem, where v is the vocabulary size.

Table 4.8 shows the performance of various embeddings for semantic and syntactic relation tasks. Here, as earlier, we trained eigenwords on a Reuters RCV1 with a window size of 2, however as can be seen it performed significantly better compared to all the embeddings except word2vec. We conjectured that it could be due to the fact that we were taking too small a context window which mostly captures syntactic information, which was sufficient for the earlier tasks. So, we experimented with a window size of 10 with the hope that a broader context window should be able to capture semantic and topic information. For this configuration, the eigenwords' performance was comparable to word2vec and as we had intuited most of the improvement in performance took place on the semantic relation task.

Embedding/Model	Semantic Relation	Syntactic Relation	Total Accuracy
Turian (C&W)	1.41	2.20	1.84
Turian (HLBL)	3.33	13.21	8.80
SENNA	9.33	12.35	10.98
eigenwords (Window size= 2)	12.21	29.40	21.70
word2vec (Window size= 10) (SK)	33.91	32.81	33.30
word2vec (Window size= 10) (CB)	31.05	**36.21**	**33.90**
eigenwords (Window size= 10) (Best)	**34.79**	31.01	32.70

Table 4.8: Accuracies for Semantic, Syntactic Relation Tasks and total accuracies.

4.2.6 Discussion

In this chapter performed a thorough qualitative and quantitative evaluation of eigenwords on a variety of natural language processing tasks. As we saw, eigenwords capture syntactic and semantic information about the words and give superior performance compared to state-of-the-art embeddings. They perform significantly better than all the embeddings except word2vec on all the supervised learning tasks. They perform comparably to or sometimes better than word2vec, though there is no clear pattern of superiority. Thus, word2vec is a viable alternative to the eigenwords embeddings as they capture similar syntactic and semantic information. However, eigenwords have better sample complexity for rare words and can perform better on resource poor languages for which relatively little unlabeled data is available. The spectral methods used to compute eigenwords also have a clearer theoretical foundation than the word2vec algorithm.

Chapter 5

Eigenanatomy: A Framework for Sparse Component Analysis for Brain Imaging

High[1] dimensional datasets are frequently collected in medicine and biology. Magnetic resonance imaging (MRI), gene expression and genotype all contain thousands to millions of measurements per individual. The individual discrete measurements comprising these modalities are related to each other through the lens of both the quantitative technology and the underlying biology. Therefore, the resulting datasets often exhibit strong covariation and suffer from the curse of dimensionality. When this type of data is addressed with univariate statistics, studies may be underpowered or fail to capture the intrinsically multivariate nature of the underlying biological signal.

Data-driven dimensionality reduction and feature selection techniques provide a potentially optimal strategy for analyzing "big data" in biological domains. Dimensionality reduction methods find (weighted) combinations of the univariate mea-

[1] This chapter is based on work in (Dhillon et al. 2014 (Under Review).

surements such that the original data is well-described by a relatively small set of summary measurements. Given the presence of collinearity, dimensionality reduction methods may improve statistical power by collecting related measurements together. This also facilitates data inspection which is challenging when using univariate approaches to data with several thousand or more variables.

There are three commonly used methods for data driven dimensionality reduction:

- **Principal Component Analysis (PCA)** (Jolliffe 2005): It is one of the most widely used dimensionality reduction algorithms. PCA, however, has the disadvantage that the low-dimensional components consist of contributions from every component of the high-dimensional space, which makes interpretation of the low-dimensional space difficult. Many techniques have been proposed to deal with this issue. One common method is Sparse PCA (SPCA) (Jolliffe and Uddin 2000; Shen and Huang 2008b; Guan and Dy 2009). Sparse PCA incorporates penalties on the matrix decomposition to encourage each component to consist of contributions from only a few components from the higher dimensional space. Another related technique is sparse coding (Lee et al. 2006; Mairal et al. 2010) (or dictionary learning) (Varoquaux et al. 2011; Abraham et al. 2013), which is motivated more from a neurological perspective and way the human cortex processes information (Olshausen and Field 2004).

- **Non-negative Matrix Factorization (NMF)** (Paatero and Tapper 1994): It constrains the components to be positive i.e. each learned component of the dictionary is a positive sum of positive "parts" rather than a sparse sum of positive or negative parts. The main motivation behind NMF is to come up with "parts-based representations" which transforms unstructured data into more interpretable pieces (Lee and Seung 1999a; Hoyer and Dayan 2004; Lee and Seung 1999b; Berry et al. 2007). Alternatively, one could drop the positivity constraint and impose the sparsity constraint. Another variant involves imposing sparsity in addition to positivity which is called non-negative sparse coding (Hoyer and Dayan 2004).

- **Independent Component Analysis (ICA)** (Hyvärinen and Oja 2000): It is motivated by the "blind source separation" problem: Given a data matrix that contains information from a variety of sources, how can we uncover the original sources? It optimizes statistical independence among the basis vectors. Note that PCA, owing to its gaussianity assumption identifies the true sources only upto a rotation, so we need a non-gaussianity assumption (prior/regularization) on the sources to recover the true sources. Because of the different motivation of ICA, and slightly different notation it makes comparisons between the various methods difficult.

All these methods fall into the more general class of sparse matrix factorization framework– each making a different set of assumptions. They can be viewed through a common lens as we show below.

These methods have been used to obtain state-of-the-art accuracies in a variety of problems in Machine Learning. However, their usage in brain imaging, though increasing, is limited by the fact that they are used as out-of-the-box techniques and are seldom tailored to the domain specific constraints/knowledge pertaining to medical imaging. For instance, uninformed, generic matrix decomposition methods, e.g. standard principal component analysis (PCA) or ICA, may be difficult to interpret because the solutions will produce vectors that are everywhere non-zero, i.e. involve the whole brain rather than its parts.

This limits their popularity especially among clinicians who are often as interested in clinical interpretability as predictive accuracy. Sparse methods have sought to resolve this issue (Hoyer and Dayan 2004; Witten and Tibshirani 2010; Friedman et al. 2010; Cherkassky and Ma 2009; Friedman et al. 2008a). However, these recent sparse multivariate methods are anatomically uninformed and which may lead to unstable results (Xu et al. 2012). In this chapter, we propose to bridge this gap by providing matrix decomposition techniques regularized by neuroanatomically inspired smoothness and connectedness terms.

In order to address the above shortcomings, in this chapter, we propose Eigenanatomy (EANAT). Eigenanatomy (EANAT) is a general framework for sparse matrix factorization that is closely related to SPCA, NMF, and a version of ICA [2]. The goal of EANAT is to statistically learn the boundaries of and connections between brain regions by weighing both data and prior neuroanatomical guidance. Recent work points to the fact that exploiting problem-specific information can improve parts-based representations (Guan et al. 2011; Cai et al. 2010; Hosoda et al. 2009). EANAT component images, on the other hand, enable prior knowledge to enhance solution stability and are tied to a set of neuroanatomical coordinates that are *connected*, *smooth* and may also be defined by *non-negative* weights.

The specific constraints implemented within each of these methods alter the patterns that are extracted. Each algorithm has a different history, employs different theoretical arguments and is favored in a different community. The optimization methods are also heterogeneous, making it challenging to know, when comparing different implementations of the algorithms, whether the theoretical or practical differences are the root of performance variation. To address this concern, we implement EANAT decompositions with respect to a consistent data-term and then enforce smoothness and sparsity constraints inspired by SPCA, ICA and NMF. This allows us to compare positively constrained decompositions (as with NMF) to ICA and SPCA-like decompositions.

[2] Log hyperbolic cosine sparsity penalty.

Our main contributions are:

1. From a theoretical standpoint, we provide a unified objective function for sparse matrix factorization inspired by SPCA, NMF and ICA which allows us to incorporate different aspects of each kind of decomposition into a single framework.

2. From a practical standpoint, we provide uniform optimization algorithms, that allow for easy comparison of the decompositions without the confounds of different implementation strategies.

3. Customization of EANAT for neuroimaging by using domain specific sparsity and smoothness constraints which aid interpretability of results and give superior predictive performance.

4. A thorough evaluation involving comparison with standard approaches; and finally the public availability of our toolkit.

The remaining chapter is organized as follows; we first describe the various dimensionality reduction methods and provide a unified objective for them. Then we describe our EANAT theoretical framework and detail how the framework may be extended to implement neuro or brain imaging specific decompositions. Finally, we outline our optimization algorithm and provide experimental evaluation.

5.1 Eigenanatomy (EANAT)

The class of methods encompassing NMF, ICA, SPCA and singular value decomposition (Sill et al. 2011b; Lee et al. 2010; Yeung et al. 2002) form the basis for the approach proposed here. So, firstly, we describe SPCA, ICA and NMF through a common lens and provide a unified objective for them. Finally, we show how one can define hybrid matrix decomposition methods by combining different approaches.

5.1.1 Notation

Define a $n \times p$ (rows by columns) matrix $\mathbf{X}$, where n is the number of observations (subjects) and p is the number of total voxels. For instance, $\mathbf{X}$ can be the cortical thickess measurements or the fMRI time series for all the n subjects. We seek a sparse decomposition of the $\mathbf{X}$ matrix into two matrices $\mathbf{U}$ and $\mathbf{V}$, (such that $\mathbf{X} \approx \mathbf{U}\mathbf{V}^\top$). The $\mathbf{V}$ matrix, which is of size $p \times k$ is typically called the *factor loading matrix* and each of its columns is a basis vector for approximating the $\mathbf{X}$ matrix or sometimes it is also called a *dictionary*, in the sparse coding literature where each of its columns is an *atom*. The $\mathbf{U}$ matrix of size $n \times k$ is called the *coefficient matrix* or the *latent vectors*.

$$\arg\min_{U,V} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_F^2 \tag{5.1}$$

Now, there are three keys aspects of the above objective that a researcher might care about and might want to tailor to the requirements of a specific domain.

1. Sparsity: Should the $\mathbf{U}$ and/or $\mathbf{V}$ matrices be sparse?

2. Orthogonality: Should each of $\mathbf{U}$ and $\mathbf{V}$ matrices have orthogonal columns i.e. $\forall \;\; l1 \leq i < k \;\; u_i \perp u_j$ and $v_i \perp v_j$? In addition to this, should the columns of $\mathbf{U}$ be orthogonal to the columns of $\mathbf{V}$.

3. Positivity: Should the entries of $\mathbf{U}$ and/or $\mathbf{V}$ matrices be only positive?

Broadly, this gives rise to four different algorithms

5.1.2 Principal Component Analysis (PCA)

In standard PCA, $\mathbf{U}$, and $\mathbf{V}$ are orthogonal and are related by $\mathbf{U} = \mathbf{X}\mathbf{V}$. The matrix $\mathbf{V}$ is estimated by performing singular value decomposition (SVD) on the correlation matrix $\mathbf{X}^\top \mathbf{X}$ and contains the 'k' eigenvectors with the largest eigenvalues as its columns.

A potential problem with standard PCA is that the low-dimensional components it finds consist of contributions from each of the p components of the high-dimensional space, which can be undesirable in a domain like brain imaging.

5.1.3 Sparse Principal Component Analysis (SPCA)

Sparse PCA (SPCA) augments the objective function presented in Equation 5.1 by putting an ℓ_1 sparsity constraint on the columns of $\mathbf{V}$. Its worth noting that unlike standard PCA, SPCA does not normally include an explicit orthogonality constraint. The reconstruction error term makes the orthogonality less important (Le et al. 2011). If we were to enforce sparsity on the coefficients matrix $\mathbf{U}$ instead of the loadings matrix $\mathbf{V}$, then we get the *sparse coding* (Lee et al. 2006; Mairal et al. 2010) or *sparse dictionary learning* (Varoquaux et al. 2011; Abraham et al. 2013) objective.

$$\underset{U,V}{\arg\min} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_2^2 + \lambda \sum_i \|\mathbf{V}_i\|_1 \tag{5.2}$$

5.1.4 Non-negative Matrix Factorization

NMF requires both $\mathbf{U}$ and $\mathbf{V}$ to be non-negative. There are no orthogonality constraints as earlier.

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_2^2 \\ \text{subject to} \quad & \mathbf{U}, \mathbf{V} \succeq 0, \end{aligned} \tag{5.3}$$

where $\succeq$ indicates element-wise inequality. This constraint fits naturally to problems in which the input data is non-negative, as is the case with text mining where each document can be seen as composed of words (Berry et al. 2007) or music analysis (Févotte et al. 2009). If we further add a sparsity penalty on the coefficients matrix $\mathbf{U}$, then we get non-negative sparse coding (Hoyer and Dayan 2004).

5.1.5 Independent Component Analysis (ICA)

The motivation for ICA comes from the blind source separation problem where "blind" means we know nothing about the source of signals, and we have to recover all the true sources generating the data.

ICA seeks to impose statistical "independence" on the *sources*. It is related to sparse coding, in that if we replace the ℓ_1 sparsity penalty in sparse coding with a non-gaussianity promoting penalty, we get ICA. The notation of ICA is also a bit different and in literature $\mathbf{A}$ and $\mathbf{s}$ are used instead of $\mathbf{U}$ and $\mathbf{V}$ respectively. However, to be consistent, we will stick to our notation of $\mathbf{U}$ and $\mathbf{V}$.

$$\underset{U,V}{\arg\min} \quad \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_2^2 + \lambda p(\mathbf{U}), \tag{5.4}$$

where $p(\mathbf{U})$ penalizes the non-independence of the columns of $\mathbf{U}$. Although there are many ways to optimize for statistical independence (or "non-Gaussianity"), e.g. skewness, kurtosis; a common practical way of enforcing the independence is to use the log hyperbolic cosine penalty. The log hyperbolic cosine is a close approximation to the ℓ_1 norm, Equation 5.4 is closely approximated by

$$\underset{U,V}{\arg\min} \quad \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_2^2 + \lambda \sum_i \|\mathbf{U}_i\|_1. \tag{5.5}$$

The optimization of the general ICA objective is also different from SPCA, sparse coding, NMF all of which can be optimized by gradient descent, arnoldi iteration or SVD. The most standard algorithm for ICA optimization is an approximate Newton method, called FastICA (Hyvarinen 1999).

It is worth noting that though we have cast SPCA, sparse coding, NMF and ICA in the same objective but still there are differences between the two in addition to what is described above.

Firstly, the connection is tied to the assumption of using a log hyperbolic cosine penalty; if we were to use some other penalty then there might not be an obvious similarity between the methods, which makes sense as SPCA considers only upto second order moments (gaussianity assumption) whereas ICA optimizes fourth order moments.

Secondly, since SPCA tries to find directions capturing decreasing variance (second order moment); there is a natural ordering to the components of SPCA. However, the basis vectors found by standard ICA are not ranked in any order. So, above, we assume that the components of ICA are also ranked according to decreasing variance.

5.2 EANAT Objective

EANAT objective uses a hybrid decomposition schemes which borrow ideas from SPCA and ICA and further augments it with neuroanatomically specific penalty terms. EANAT seeks to represent each component image with a set of neuroanatomical coordinates that are *connected*, *smooth* and are defined by *non-negative* weights. Although this latter constraint can be relaxed, non-negativity improves our ability to interpret data by preventing weights from being both positive and negative within the same eigenanatomy component. Also, non-negativity means that the projections of eigenanatomy into subject space are simple weighted averages of the input data (e.g. cortical thickness values) for each subject.

$$\arg\min_{U,V} \|\mathbf{X} - \mathbf{U}\mathbf{V}^\top\|_2^2 + \lambda_1 S(\mathbf{U}) + \lambda_2 S(\mathbf{V}), \tag{5.6}$$

where λ_1 controls the contribution of the ICA (sparse coding) sparsity component of the penalty and λ_2 controls the sparsity of the SPCA component of the penalty. $S(\cdot)$, the sparsity penalty, is defined as follows,

$$S(\mathbf{V}) = \sum_{i=1}^{k} \|G \times v_i\|_1, \tag{5.7}$$

$$\forall_{i \neq j} \langle \mathbf{v}_j, \mathbf{v}_i \rangle = 0 \ , \ \mathbf{v}_i \succeq 0 \ , \ \|S(\mathbf{v})\|_1 = \gamma$$

Where, γ is a user defined sparsity parameter which controls the number of non-zero entries in the solution. G is a kernel matrix which enforces smoothness and connectedness among the different EANAT components along the anatomical manifold and is similar in spirit to the wavelet or discrete cosine basis transform (Becker et al. 2011). When the G operator is equal to I (identity matrix), it reduces to a simple ℓ_1 penalty.

Its worth noting a further few points about the objective:

- We enforce orthogonality between the various components. In other words, $\forall$ $1 \leq i, j \leq p$, $u_i \perp u_j$ and $v_i \perp v_j$ but unlike standard PCA, $u_i \neq v_i \cdot x$.
- Non-negativity of the components means that the projections of eigenanatomy into subject space are simply weighted averages of the input data (e.g. cortical thickness values) for each subject. Although this constraint can be relaxed, non-negativity improves our ability to interpret data by preventing weights from being both positive and negative within the same eigenanatomy component. As such, one may compute effect sizes and interpret statistics directly, for example, "reductions in posterior cingulate cortical thickness reduce performance on memory-related psychometrics."

To the best of our knowledge, directly exploring the interaction between sparseness, orthogonality and non-negativity for automated parcellation of the brain is novel. This penalty set gives us anatomically reasonable results as we show in the experiments section.

5.2.1 Optimization

There are a variety of ways that one could optimize the above objective. (Mairal et al. 2010) formulate a convex alternative for the above objective which uses an elastic net type penalty on $\mathbf{V}$. However, we propose an alternating optimization approach, also called an analysis-synthesis loop (Murphy 2012). As a broader template, we optimize $\mathbf{U}$ keeping $\mathbf{V}$ fixed. Next, we deflate the $\mathbf{X}$ matrix using the optimized $\mathbf{U}$s, and then optimize $\mathbf{V}$ with $\mathbf{U}$ fixed. This alternating procedure is repeated till convergence. Each of our sparse optimization for $\mathbf{U}$ and $\mathbf{V}$ is performed via iterative soft-thresholding on the conjugate gradient of the Rayleigh Quotient.

Iterative soft-thresholding (soft$(a, \delta) \triangleq$ sign(a)$(\|a\| - \delta)_+$ with x_+ =max(x,0)) falls in the class of proximal gradient methods and has been shown to have better convergence (Bredies and Lorenz 2008) and scalability properties compared to other sparse optimization algorithms e.g. Least Angle Regression (LARS) (Yang et al. 2010). Furthermore, deflation has been shown to give better sparse PCA solutions (Mackey 2008); so adding a deflation step between the alternating optimizations helps us get better solutions.

5.2.2 Implementation Details

We know that the best rank 'k' reconstruction of a matrix i.e. $\text{argmin}_{\hat{\mathbf{X}}} \|\mathbf{X} - \hat{\mathbf{X}}\|_{\mathbf{F}}^{\mathbf{2}}$, is provided by its first 'k' eigenvectors (Eckart and Young 1936) i.e. $\hat{\mathbf{X}} = \sum_{i=1}^{k} d_k u_k v_k^\top$.

Hence, the best rank-1 approximation of $\mathbf{X}$, i.e. the $n \times 1$ and $p \times 1$ vectors $\tilde{u}$, $\tilde{v}$ such that,

$$\tilde{u}^*, \tilde{v}^* = min_{\tilde{u},\tilde{v}} \|\mathbf{X} - \tilde{u}\tilde{v}^\top\|_F^2 \tag{5.8}$$

is given by the SVD solution– $\tilde{u} = u_1$ and $\tilde{v} = d_1 v_1$, where u_1, v_1 and d_1 are the first left and right eigenvectors and the eigenvalue, respectively, of the $\mathbf{X}$ matrix.

Proceeding this way, $d_2 u_2 v_2^\top$ provide the best rank-1 approximation of the "deflated" matrix $\mathbf{X} - d_1 u_1 v_1^\top$ and so on.

As pointed by (Shen and Huang 2008a), with $\tilde{v}$ fixed, the above optimization over $\tilde{u}$ is equivalent to a least squares regression of $\mathbf{X}$ on $\tilde{v}$. However, in our case, we have sparsity on $\tilde{u}$ also, so it becomes a sparse optimization problem (Equation 6.4).

Similarly, with $\tilde{u}$ fixed, the optimization over $\tilde{v}$ is also a sparse optimization problem (Equation 5.10). As mentioned in the last section, we solve both these by iterative soft thresholding on the conjugate gradient of Rayleigh Quotient.

As described earlier, our implementation alternates between optimization of Equations 6.4, 5.10 (shown below for iteration number 'm') till convergence .

$$\mathbf{U^*}_m = \underset{U, \|U\|=1, u_i^\top u_j = 0, i \neq j}{\operatorname{argmin}} (\mathbf{X} - \mathbf{U}\mathbf{V}_{m-1}^\top)^2 + \lambda_1 \|G\mathbf{U}\|_1 \tag{5.9}$$

$$\{v_i^*\}_m = \underset{v_i, \|v_i\|=1, v_i^\top v_j = 0, i \neq j}{\operatorname{argmin}} (\mathbf{X}_{\setminus i} - \mathbf{U}_m v_i^\top)^2 + \lambda \|G v_i\|_1 \tag{5.10}$$

where $\mathbf{X}_{\setminus i} \triangleq \mathbf{X} - \sum_{j=1, j\neq i}^{k} \tilde{u}_j \tilde{v}_j^\top$ is the "deflated" $\mathbf{X}$ matrix.

The sparseness, smoothness and non-negativity are enforced as discussed in the previous section.

The details of our algorithm can be found in Algorithms 6, 7.

5.3 Experiments

In this section we benchmark the performance of Eigenanatomy (EANAT) on Parkinson's Progressive Markers Initiative (PPMI) dataset.

The data consists of T1 images of 613 individuals with 3 diagnostic statuses: Control, (Scans without evidence for dopaminergic deficit) SWEDD or (Parkinson's

Algorithm 6 Eigenanatomy: EANAT (Main Algorithm)

```
Input:  $\mathbf{X}$, $\lambda$
Standardize the data matrix $\mathbf{X}$: Mean center it and scale to unit variance
Initialize the eigenvectors randomly $\mathbf{V} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{1})$
$\mathbf{U} \leftarrow \texttt{SCGP}(\mathbf{X}, \mathbf{V}, \lambda_1, 1:k)$
while $\Delta\|\mathbf{V}\| \leq \epsilon$ do
    for i=1 to k do
        $\mathbf{X}_{\setminus\mathbf{i}} \leftarrow \mathbf{X} - \sum_{j=1,i\neq j}^{k} u_j v_j^\top$ //Deflate $\mathbf{X}$
        $v_i \leftarrow \texttt{SCGP}(\mathbf{X}_{\setminus i}, \mathbf{V}, \lambda, i)$
        $\mathbf{U} \leftarrow \texttt{SCGP}(\mathbf{X}, \mathbf{V}, \lambda_1, 1:k)$ //Where $\mathbf{V}$ is the matrix with only its $i^{th}$ column
        updated. Other columns are the same as previous iteration
    end for
end while
Output: $\mathbf{V}$
```

Algorithm 7 EANAT (Sub Algorithm): Sparse Conjugate Gradient Projection (SCGP) for finding v_i

```
Input: $\mathbf{X}$, $\mathbf{V}$, $\lambda$,i
$k \leftarrow 1$
$c^k \leftarrow \mathbf{V}_{:,i}$ //$\mathbf{V}_{:,i}$ contains the $i^{th}$ eigenvector.
$c^k \leftarrow \texttt{S}(c^k,\gamma)$ //Soft-Max Thresholding.
$g^{k-1} \leftarrow 1$ //Initialize Gradient.
while $\Delta c \leq \epsilon$ do
    $g^k \leftarrow (\mathbf{X}^\top \mathbf{X}) c^k$ //Gradient of Rayleigh quotient.
    $\gamma \leftarrow \frac{g^k \cdot g^k}{g^{k-1} . g^{k-1}}$ //Conjugate Gradient.
    $d^k \leftarrow g^k + \gamma \cdot d^{k-1}$
    $c^{k+1} \leftarrow c^k + d^k$
    $c^{k+1} \leftarrow \texttt{Orthogonalize}(c^{k+1}, \mathbf{V}_{:,\setminus\mathbf{i}})$ //Orthogonalize w.r.t. all the other $k-1$
    eigenvectors.
    $c^{k+1} \leftarrow \texttt{S}(c^{k+1},\gamma)$
    $c^{k+1} \leftarrow \frac{c^{k+1}}{\|c^{k+1}\|}$ //Normalize
    $k \leftarrow k + 1$
end while
Output: $c^{k+1}$
```

disease) PD. SWEDD term can refer to any subject that looks as if they have PD but where subsequent functional imaging assessments do not confirm this. SWEDD phenotypes therefore vary in much the same way as PD phenotypes do. Other details about image acquisition and diagnosis can be found here (`http://www.ppmi-info.org/access-data-specimens/`).

The basic statistics for the cohort are given in Table 5.1.

Characteristic	Entire Cohort (613) ($\mu \pm \sigma$)	Only Controls (175) ($\mu \pm \sigma$)	Only PD (379) ($\mu \pm \sigma$)	Only SWEDD (59) ($\mu \pm \sigma$)
Age	61.06 ± 10.30	59.99 ± 11.36	61.54 ± 9.76	61.09 ± 10.34
Gender (Female)	220	64	135	21
Weights (kg)	81.20 ± 16.37	79.79 ± 15.88	81.40 ± 16.86	84.16 ± 14.26

Table 5.1: Basic statistics for the cohort.

The data was preprocessed using ANTs (Avants et al. 2009) and the pipeline described in (Tustison et al. 2014) was used to generate cortical thickness and Jacobian measurement images. Since the cerebellum is thought to be involved in Parkinson's disease (Wu and Hallett 2013), the Jacobian image should capture discriminative information which should aid in Parkinson's classification.

5.3.1 Evaluation

We use the demographics along with the top 10 EANAT eigenvectors derived from cortical thickness and Jacobian measurement images to learn a multiclass logistic regression classifier to classify subjects into controls, PD and SWEDD.

We randomly divided the dataset into training/testing (80/20) instances and repeated this procedure 100 times. The results (average classification accuracy on the test set) are shown in Table 5.2.

The main findings are that cortical thickness and Jacobian measurements help on top of just demographics based covariates though not significantly. However, using both the measurements i.e. cortical thickness and Jacobian significantly (5% level in a paired t-test) helps classification of Controls vs SWEDD vs PD.

Covariates Used	% Classification Error ($\mu \pm \sigma$)
D	38.9 ± 2.8
D+J	37.5 ± 2.4
D+T	38.0 ± 2.1
D+T+J	36.7 ± 1.9

Table 5.2: Test Set % Classification Errors (Averaged over 100 random splits). Note D= Demographics (Age, Weight, Gender). T= (Top 10) Eigenvectors derived from Cortical thickness images. J= (Top 10) Eigenvectors derived from Jacobian images.

Chapter 6

Prior Based Eigenanatomy (p-Eigen): Incorporating Prior Knowledge into Eigenanatomy

As[1] described in the last chapter, there have been significant breakthroughs in medical imaging machinery over the past few decades. This has led to an increase in the amount and diversity of data being available e.g. structural and functional modalities, neuro-cognitive batteries, genetics, and environmental measurements etc. This has lead to a substantial interest in using sophisticated statistical methods to analyze and explore this data. Methods like Principal Component Analysis (PCA), Independent Component Analysis (ICA), Canonical Correlation Analysis (CCA) and their robust and sparse variants (Witten et al. 2009; Avants et al. 2010) have been the workhorse of brain and neuro-imaging fields as they provide key insights into the data in a totally data driven way. In the previous chapter, we proposed a unified framework, Eigenanatomy (EANAT), which learns the boundaries of and connections between the brain regions by weighing both data and prior neuroanatomical

[1] This chapter is based on work in (Dhillon et al. 2014).

guidance.

However, one potential pitfall of these approaches, which has made clinicians and other researchers cautious with their use is the lack of control over the areas they highlight. Since, they work in a totally data driven way, the end clinician/researcher has little control over the areas of brain they chose. Clinicians usually have some form of prior knowledge as to which areas may contain the signal they are looking for, for example, someone studying fronto-temporal dementia would expect some or most of the signal to lie in frontal cortex, however if the voxels in frontal cortex don't explain the variance in data these approaches won't highlight them.

So, this has led the clinicians and researchers to work with totally prior driven approaches e.g. Region of Interest (ROI) analysis (Poldrack 2007), where they chose a pre-defined region manually or based on past study (for instance *Brodmann Areas*) and study it exclusively. The assumption is that the ROI is correct and contains all relevant signal. However, important signal may have slightly different boundaries than the scientist's conception. The data representation (or spatially varying noise) may also lead to strong or weak signal within different parts of the ROI. Such dataset specific information is not taken into account by a traditional ROI. When effects are localized to the selected region, and that region is well-defined, an ROI analysis may provide the most sensitive testing method. However, some conditions involve a network of regions that may not be fully identified. In such cases—in addition to the general case of an exploratory analysis in a small dataset—dimensionality reduction may provide advantages over an ROI analysis or mass univariate voxel-based morphometry (Mechelli et al. 2005).

Our approach provides a principled way of incorporating priors in an otherwise totally data driven approach based on Sparse Principal Component Analysis (SPCA) (Zou et al. 2006; Witten et al. 2009; d'Aspremont et al. 2007; Shen and Huang 2008a).

p-Eigen allows an initial binary or probabilistic ROI to adapt to the underlying subject specific covariation within the data. At the same time, p-Eigen maintains proximity to (and the locality of) the original region and thus retains the advantages of the standard seed based approach. p-Eigen also maintains non-negativity in the estimated anatomically-constrained eigenvector, thereby keeping ROI interpretability. This allows us to modify the definitions of labels to capture the variation in dataset while still staying close to the initial ROI definitions. p-Eigen therefore produces labelings with "soft" weighted averages and as we show in the experimental sections (in the next two Chapters), are more sensitive to the underlying brain data than a standard ROI.

Given an ROI set, p-Eigen has only one key parameter to tune- the weight of the prior term guiding the decomposition. Therefore, our optimization objective provides a tradeoff between 1). staying close to the initial ROI definitions and 2). allowing data to lead the exploratory analysis by explaining variance through PCA. A good way to think about this is as ROI definitions forcing us to be *conservative* and staying close to the initial brain parcellation; on the other hand the SPCA component gives us *liberty* to be either more exploratory or more focused on the content of the given dataset. The tradeoff between the two competing paradigms is defined by user tunable (prior strength) parameter, which is chosen via cross validation.

Our proposed approach is shown in Fig. 6.1.

In the remaining chapter we provide the details of Prior Based Eigenanatomy (p-Eigen) and provide an algorithm.

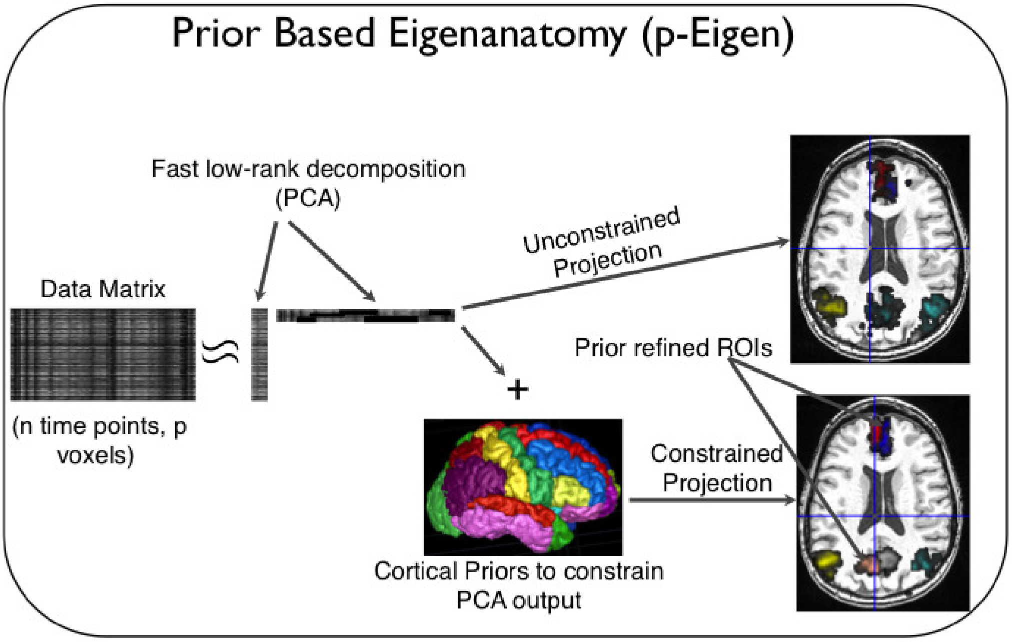
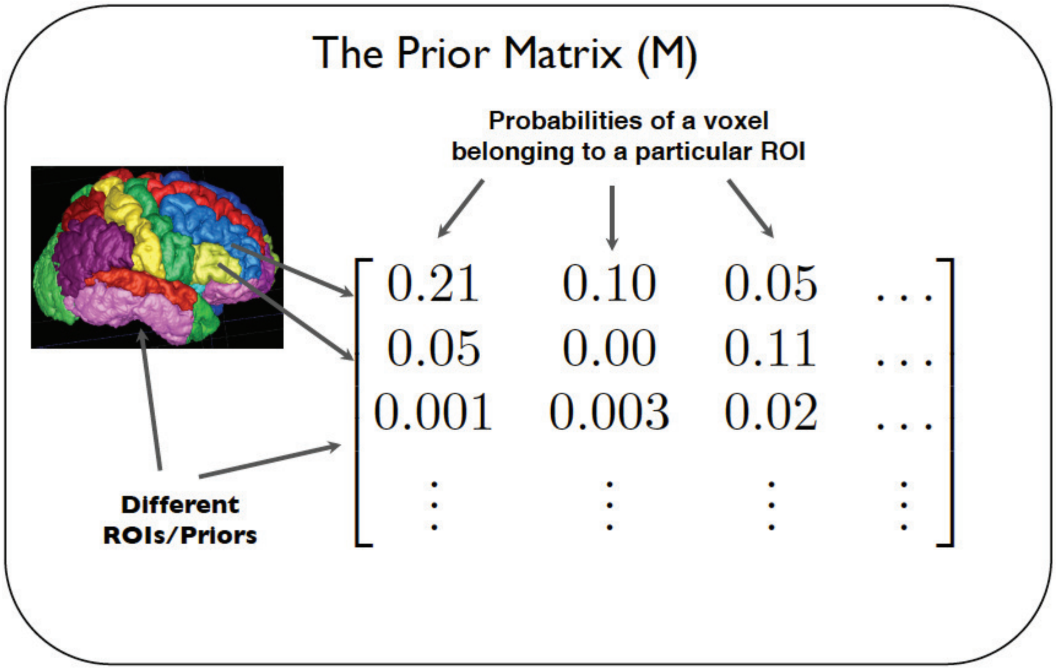

Prior Based Eigenanatomy (p-Eigen)
Fast low-rank decomposition (PCA)
Data Matrix
≅
(n time points, p voxels)
Unconstrained Projection
Prior refined ROIs
+
Constrained Projection
Cortical Priors to constrain PCA output

Figure 6.1: Prior Based Eigenanatomy (p-Eigen). An initial data matrix is decomposed into its eigenvectors, with each eigenvector being constrained by a corresponding cortical prior.

6.1 Prior Based Eigenanatomy: p-Eigen

p-Eigen is based on the methods of sparse principal components analysis (SPCA) (Zou et al. 2006; Witten et al. 2009; d'Aspremont et al. 2007; Shen and Huang 2008a) and singular value decomposition (Sill et al. 2011a).

Define a $n \times p$ (rows by columns) matrix $\{\mathbf{X}\}$ where n is the number of subjects, p is the number of total voxels where each $\mathbf{X}$ matrix could derive from the subjects' T1-structural imaging data or from each subject's BOLD fMRI image (in which case 'n' will be the number of timepoints in that subject's timeseries image.). Also, assume that we have a prior matrix $\mathbf{M}$ (Fig. 6.2) each of whose $\mathbf{k}$ rows corresponds to a separate prior and each of whose **p** columns contains the probability of a particular voxel belonging to that prior.



The Prior Matrix (M)

Probabilities of a voxel belonging to a particular ROI

$$\begin{bmatrix} 0.21 & 0.10 & 0.05 & \dots \\ 0.05 & 0.00 & 0.11 & \dots \\ 0.001 & 0.003 & 0.02 & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Different ROIs/Priors

Figure 6.2: Prior Matrix (M). Each row corresponds to a different ROI (Total k of them) and each column corresponds to a different voxel in the brain (Total p of them).

We seek a sparse decomposition of the $\mathbf{X}$ matrix constrained by the anatomical priors $\mathbf{M}$ which should give us a $n \times k$ matrix where each of the k eigenvectors explains the variance in the corresponding anatomical region specified by the prior.

Our objective is described by Equation 6.1.

$$v_i^* = \underset{v_i, \|v_i\|=1, v_i^\top v_j=0, i\neq j, v_i \succeq 0}{\operatorname{argmax}} \quad {v_i}^\top \left(\mathbf{C} + \theta \cdot m_i^\top m_i\right) v_i - \lambda_i \|v_i\|_1^+ \qquad (6.1)$$

where $\mathbf{m}_i$ is the i^{th} prior and is itself a vector of size $(1 \times p)$. $\mathbf{C}$ is the covariance matrix $\mathbf{X}^\top \mathbf{X}$. θ is a user tunable parameter which controls the tradeoff between the influence of data and the prior and should be typically tuned on a held out validation set in the absence of other knowledge. Smaller values of θ suggest that we trust data more and as θ is increased the eigenvectors v are increasingly influenced by the prior. The $\|v_i\|_1^+$ term ensures sparsity and non-negativity in the eigenvectors; in addition to this we enforce unit norm and orthogonality constraints on sparse eigenvectors.

Connection to Sparse PCA: Our objective (Equation 6.1) is intimately connected to the variance maximization formulation of sparse principal components analysis (SPCA) (Zou et al. 2006; Witten et al. 2009) :

$$v_i^* = \underset{v_i, \|v_i\|=1, v_i^\top v_j = 0, i \neq j}{\operatorname{argmax}} \quad {v_i}^\top \mathbf{C} v_i - \lambda \|v_i\|_1 \tag{6.2}$$

where terms have the same meaning as in Equation 6.1.

As can be seen, our objective entails that instead of finding the eigenvectors of the data covariance matrix as done by SPCA, we find the eigenvectors of the transformed data covariance matrix obtained by "regularizing" it by the prior information. An important consequence of this is that we are not confining our data driven priors to lie in the original ROIs but rather we are encouraging them to find ways to explain data variance in this new "prior regularized" space.

One could optimize the p-Eigen objective in Equation 6.1 using an iterative approach like power iteration. However, in our experience we found the standard power iteration to be unstable and got more efficient and stable solutions using an optimization approach which performs iterative soft-thresholding on the conjugate gradient of the Rayleigh Quotient. In addition, we *deflate* our data matrix $\mathbf{X}$ (factoring out the effect of other eigenvectors) between computations of different eigenvectors, which lead to better solutions (Mackey 2008). The resulting method is related to the Non-linear Iterative Partial Least Squares (NIPALS) algorithm (Wold et al. 1987) for large scale PCA which also combines deflation with estimation of the principal eigenvector.

6.2 An Algorithm for p-Eigen

In this section we provide an alternating optimization approach, also called an analysis-synthesis loop (Murphy 2012) for solving the p-Eigen optimization problem. As was briefly mentioned at the end of the last section, our optimization performs iterative soft-thresholding on the conjugate gradient of the Rayleigh Quotient and further relies on deflation to get better quality solutions.

Iterative soft-thresholding (soft$(a, \delta) \triangleq$ sign(a)$(\|a\| - \delta)_+$ with x_+ =max(x,0)) falls in the class of proximal gradient methods and has been shown to have better convergence (Bredies and Lorenz 2008) and scalability properties compared to other sparse optimization algorithms e.g. Least Angle Regression (LARS) (Yang et al. 2010). Furthermore, deflation has been shown to give better sparse PCA solutions (Mackey 2008); so we added a deflation step between the alternating optimizations.

The deflation based optimization of Equation 6.1 entails performing an additional ordinary least squares regression (OLS) step and can be motivated as follows.

We know that the best rank 'k' reconstruction of a matrix i.e. argmin$_{\hat{\mathbf{X}}} \|\mathbf{X} - \hat{\mathbf{X}}\|^{\mathbf{2}}$, is provided by its first 'k' eigenvectors (Eckart and Young 1936) i.e. $\hat{\mathbf{X}} = \sum_{i=1}^{k} d_k u_k v_k^\top$.

Hence, the best rank-1 approximation of $\mathbf{X}$, i.e. the $n \times 1$ and $p \times 1$ vectors $\tilde{u}$, $\tilde{v}$ such that,

$$\tilde{u}^*, \tilde{v}^* = min_{\tilde{u},\tilde{v}} \|\mathbf{X} - \tilde{u}\tilde{v}^\top\|^2 \tag{6.3}$$

is given by the SVD solution– $\tilde{u} = u_1$ and $\tilde{v} = d_1 v_1$, where u_1, v_1 and d_1 are the first left and right eigenvectors and the eigenvalue, respectively, of the $\mathbf{X}$ matrix.

Proceeding this way, $d_2 u_2 v_2^\top$ provide the best rank-1 approximation of the "deflated" matrix $\mathbf{X} - d_1 u_1 v_1^\top$ and so on.

As pointed by (Shen and Huang 2008a), with $\tilde{v}$ fixed, the above optimization over $\tilde{u}$ is equivalent to a least squares regression of $\mathbf{X}$ on $\tilde{v}$.

Similarly, with $\tilde{u}$ fixed, the optimization over $\tilde{v}$ is a sparse optimization problem. As mentioned in the last section, we solve this by iterative soft thresholding on the conjugate gradient of Rayleigh Quotient.

So, our implementation alternates between the optimization of Equations 6.4, 6.5 (shown below for iteration number 's') till convergence .

$$\mathbf{U^*}_s = \underset{U,\|U\|=1,u_i^\top u_j=0,i\neq j}{\operatorname{argmin}} \|\mathbf{X}-\mathbf{U}\mathbf{V}_{s-1}^\top\|^2 \tag{6.4}$$

$$\{v_i^*\}_s = \underset{v_i,\|v_i\|=1,v_i^\top v_j=0,i\neq j,v_i\succeq 0}{\operatorname{argmax}} v_i^\top\left(\mathbf{C}_s^{\backslash i}+\theta\cdot m_i^\top m_i\right)v_i - \lambda_i\|v_i\|_1^+ \tag{6.5}$$

where symbols have the same meaning as in Equation 6.1. $\mathbf{C}_s^{\backslash i}$ is the covariance matrix created from the "deflated" $\mathbf{X}$ matrix. $\mathbf{C}_s^{\backslash i} \triangleq \mathbf{X}_{\backslash \mathbf{i}}^\top \mathbf{X}_{\backslash \mathbf{i}}$ where $\mathbf{X}_{\backslash i} \triangleq \mathbf{X} - \sum_{j=1,j\neq i}^k u_j^{(s)} v_j^{\top(s-1)}$.

The sparseness is enforced by a soft-thresholding algorithm as in (Zou et al. 2006; Witten et al. 2009). We denote this function as $S(v,\lambda)$ and choose λ in a data driven way as $\lambda_i = \sum_{j=1}^p \frac{m_{ij}}{p}$. In other words, we are constraining the sparsity of our eigenvectors to be equal to the weighted size of the corresponding prior ROI. Defining sparsity in this manner via neuro-anatomical priors has biological motivation, as the sizes of ROIs are approximately equal to the sizes of different areas of the brain that we are modeling.

In addition to the sparsity penalty, we also include an optional minimum cluster size threshold, as is commonly performed in Voxel Based Morphometry (VBM)-type analyses. We have found that including a minimum cluster threshold size generally improves robustness of results by getting rid of isolated voxels and also helps prevent overfitting. In the experiments presented in this chapter, we chose the minimum cluster threshold as 100 voxels.

Similar to (Zass and Shashua 2006; Hoyer 2002), non-negativity in the eigenvectors is enforced by repeated projection onto the feasible (non-negative) set. In between different iterations of our algorithm, the negative values in the eigenvectors are zeroed and the optimization is continued.

The details of our algorithm can be found in Algorithms 8, 9. Note that the OLS regression step is just required to compute $\mathbf{u}$ which is required for deflation of the data matrix.

Algorithm 8 Prior Based Eigenanatomy: p-Eigen (Main Algorithm)

```
Input: $\mathbf{X}$, $\mathbf{M}$, $\theta$
Standardize the data matrix $\mathbf{X}$: Mean center it and scale to unit variance
Initialize the eigenvectors $\mathbf{V} \leftarrow \mathbf{M}$ based on the corresponding priors // $v_i \leftarrow m_i$ where
  $m_i$ is the $i^{th}$ row of $\mathbf{M}$
$\mathbf{U} \leftarrow \mathtt{ReconOpt}(\mathbf{X}, \mathbf{V})$
repeat
  for i=1 to k do
    $\mathbf{X}_{\setminus \mathtt{i}} \leftarrow \mathbf{X} - \sum_{j=1, i \neq j}^{k} u_j v_j^{\top}$ // Deflate $\mathbf{X}$
    $v_i \leftarrow \mathtt{SPP}(\mathbf{X}_{\setminus i}, \mathbf{V}, m_i, \theta, i)$
    $\mathbf{U} \leftarrow \mathtt{ReconOpt}(\mathbf{X}, \mathbf{V})$ // Where $\mathbf{V}$ is the matrix with only its $i^{th}$ column updated.
      Other columns remain the same as the previous iteration
  end for
until $\Delta \|\mathbf{V}\| \leq \epsilon$
Output: $\mathbf{V}$
```

Algorithm 9 p-Eigen Sub-algorithm (ReconOpt)

```
Input: $\mathbf{X}$, $\mathbf{V}$ //Optimize Reconstruction error for finding $u_i$
$\mathbf{U} \leftarrow (\mathbf{X}\mathbf{X}^{\top})^{-1}\mathbf{X}\mathbf{V}$ //Performs Ordinary Least Squares Regression.
Output: $\mathbf{U}$
```

6.3 Conclusion

We proposed a novel approach, Prior Based Eigenanatomy (p-Eigen), for fMRI network analysis which integrates ideas from the matrix decomposition and the ROI paradigms. p-Eigen leads to statistically refined definitions of ROIs based on local covariance structure of the data matrix and provides a principled way of

Algorithm 10 p-Eigen Sub-algorithm Sparse Prior (Based) Projection for finding v_i (SPP)

```
**Input:** $\mathbf{X}$, $\mathbf{V}$, $m$, $\theta$, i
$k \leftarrow 1$
$c^k \leftarrow \mathbf{V}_{:,i}$ //$\mathbf{V}_{:,i}$ contains the $i^{th}$ eigenvector.
$c^k \leftarrow \mathtt{S}(c^k,\lambda_i)$ //Soft-Max Thresholding.
$g^{k-1} \leftarrow 1$ //Initialize Gradient.
**repeat**
    $g^k \leftarrow (\mathbf{X}^\top \mathbf{X} + \theta \cdot m^\top m)c^k$ //Gradient of Rayleigh quotient.
    $\gamma \leftarrow \frac{g^k \cdot g^k}{g^{k-1} \cdot g^{k-1}}$ //Conjugate Gradient.
    $d^k \leftarrow g^k + \gamma \cdot d^{k-1}$
    $c^{k+1} \leftarrow c^k + d^k$
    $c^{k+1} \leftarrow \mathtt{Orthogonalize}(c^{k+1}, \mathbf{V}_{:,\backslash\mathbf{i}})$ //Orthogonalize w.r.t. all the other $k-1$ eigenvectors.
    $c^{k+1} \leftarrow \mathtt{S}(c^{k+1},\lambda_i)$
    $c^{k+1} \leftarrow \frac{c^{k+1}}{\|c^{k+1}\|}$ //Normalize
    $k \leftarrow k + 1$
**until** $\Delta c \leq \epsilon$
**Output:** $c^{k+1}$
```

incorporating prior information in the form of probabilistic or binary ROIs while still allowing the data to softly modify the original ROI definitions. In the next two chapters, we show the empirical performance of p-Eigen on BOLD fMRI and T1 (cortical thickness) imaging data.

Chapter 7

Prior Based Eigenanatomy (p-Eigen) for Generating Subject Specific Functional Connectivity Networks.

In[1] this chapter we show the performance of p-Eigen on resting-state BOLD fMRI data. In particular, we use p-Eigen to perform a prior constrained sparse decomposition of each subject's time series image separately. The signal in the p-Eigen refined ROIs is then correlated to create subject specific functional connectivity networks.

Functional connectivity is defined as the temporal co-activation of neuronal activation patterns between anatomically separated regions of the brain (Aertsen et al. 1989) and is thought to be an indicator of functional communication between these different regions. Typically, functional connectivity studies measure the level of correlation between the time-series of the resting state BOLD signal of the different

[1] This chapter is based on work in (Dhillon et al. 2014).

brain regions (Biswal et al. 1997; Damoiseaux et al. 2006; Salvador et al. 2005). Studying the brain as an integrative network of functionally interacting brain regions can shed new light on large scale neuronal communication in the brain and how this communication is impaired in neurological diseases (Bullmore and Sporns 2009; Mohammadi et al. 2009; Seeley et al. 2009).

7.1 Related Work

There are two predominant approaches for the analysis of functional connectivity:

- **Seed (ROI) Based Approaches**: These are straightforward and operate in the traditional confirmatory network paradigm (Tukey 1977). They involve computing the correlation between the time series of a given (preselected) *seed* brain region (ROI) [2] against all the other brain regions, resulting in a set of functional connectivity maps of the given brain regions (Biswal et al. 1997; Cordes et al. 2000). These functional connectivity maps can then be used to construct *resting-state-networks* of functionally correlated regions in the brain (Fox et al. 2005). The *seed* region can either be selected based on prior clinical knowledge or it can be selected from the activation map of a separate task dependent fMRI scan.

- **Learning Based Approaches**: These approaches use statistical techniques to explore functional connectivity in the brain, obviating the need to define a *seed* region. Typical methods employed are Principal Component Analysis (PCA) (Friston 1998), Independent Component Analysis (ICA) or its variants e.g. Group ICA (Beckmann and Smith 2004; Beckmann et al. 2005; Damoiseaux et al. 2006; Varoquaux et al. 2010a; Petrella et al. 2011) or hierarchical

[2] One can compute these correlations either voxel wise or by averaging over the voxels in an entire ROI.

methods (Cordes et al. 2002; Salvador et al. 2005; Blumensath et al. 2013). These methods strive to find a set of orthogonal or independent signals in the time series that can explain the resting state activity patterns. ICA based methods are the popular methods in this setting as they can find a set of independent signals from whole brain voxel-wise data and also due to the public availability of tools like MELODIC in FSL (Jenkinson et al. 2012) for ICA and Group ICA of fMRI Toolbox (GIFT) (Calhoun et al. 2001). Subsequently, one can create brain connectivity networks from the outputs of these approaches by computing correlations between the different (independent/orthogonal) signals they find.

The brain networks found by the above approaches are represented as a set of vertices (brain regions) connected by edges which represent the strength of correlation between those two regions (He and Evans 2010; Stam et al. 2007). Various independent studies (surveyed here (van den Heuvel and Hulshoff Pol 2010)) have consistently found a set of eight functional connectivity networks in the brain. One can use a set of key properties of the network graph e.g. clustering coefficient, centrality and modularity to get further insights into the flow of neuronal signals within a network (He and Evans 2010; Stam et al. 2007).

The above mentioned approaches for analyzing functional connectivity and constructing brain networks suffer from a variety of problems. The Group ICA based approaches do a group decomposition of the time series' images of the entire cohort; they have an averaging effect and erode away any subject specific characteristics of the network. So, the Group ICA analysis is usually followed by a back reconstruction step to generate subject specific functional connectivity maps (Smith et al. 2011). However, it is unclear how to choose a statistically justified threshold to binarize these maps.

The seed based approaches also suffer from the problem of averaging the signal and may be sensitive to ROI placement (Zhang et al. 2012), co-registration errors and the specific ROI boundaries. These approaches assume that the signal lies totally within a predefined region. However, the important signal may have slightly different boundaries than the scientist's conception. The data representation (or spatially varying noise) may also lead to strong or weak signal within different parts of the ROI. Such dataset specific information is not taken into account by a traditional seed based approach. When effects are localized to the selected region, and that region is well-defined, a seed based analysis may provide the most sensitive testing method. However, some conditions involve a network of regions that may not be fully identified.

Furthermore, it has been shown that decreased/impaired functional connectivity in certain brain networks, for instance, the Default Mode Network (DMN) has association with neurodegenerative disorders e.g. Alzheimer's Disease (AD) (Greicius et al. 2004; Sheline et al. 2010), schizophrenia (Liu et al. 2008; Whitfield-Gabrieli et al. 2009), multiple sclerosis (MS) (Lowe et al. 2008), mild cognitive impairment (MCI) (Petrella et al. 2011; Agosta et al. 2012; Hedden et al. 2009; Bai et al. 2009). So, it has become even more imperative to improve statistical analysis methods to efficiently leverage the scarce patient BOLD fMRI data that is typically available.

We have drawn a clear contrast between our approach and the two related approaches namely seed based approaches (no influence of data) and Group ICA/PCA based approaches (only data driven). That said, there has also been substantial work on incorporating prior information across subjects to build subject specific functional networks as proposed by this paper.

Some early work that performed PCA on fMRI signal within ROIs (Nieto-Castanon et al. 2003) clearly foreshadowed p-Eigen. (Thirion et al. 2006) also proposed a spectral learning based technique for parcellation that delineates homogeneous and connected regions across subjects, providing subject specific functional networks.

The research that is perhaps closest to ours is (Ng et al. 2009a), (Deligianni et al. 2011) and (Blumensath et al. 2013). (Ng et al. 2009a) used group replicator dynamics (GRD) for finding sparse functional networks that are common across subjects but have subject specific weightings of the brain regions. (Deligianni et al. 2011) used brain anatomical connectivity to constrain the conditional independence structure of functional connectivity via a multivariate autoregressive model. (Blumensath et al. 2013) perform hierarchical parcellation of the brain with a further clustering of the parcels to derive spatially contiguous parcels. Closely related is the work (Ng et al. 2009b) which constrains the PCA output by employing neighborhood information to learn spatially contiguous clusters.

p-Eigen is complementary to these set of approaches and proposes a new formulation to derive subject specific functional parcels (and hence connectivity networks) and also the first one to use the networks to derive covariates for MCI and Delayed Recall prediction. Moreover, unlike any of the above approaches p-Eigen, helps us maintain a direct correspondence between the anatomy of the same regions across different subjects hence leading to better clinical interpretability.

7.2 Experiments

Our data consists of time series images of 59 individuals (28 females and 31 males) with 34 controls and 25 subjects diagnosed clinically with Mild Cognitive Impairment (MCI). Patients were diagnosed according to the criterion of Petersen (2004). The memory measure used to get the delayed recall score was the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Word List Memory (WLM) test (Morris et al. 1989). Each time series had a total of 120 points. The basic statistics for the cohort are given in Table 7.1.

Images were acquired on a 3T Siemens Trio scanner. The imaging protocol included the following sequences: 1 mm^3 T1-weighted structural MRI and 3×3×3 mm^3 resting-state BOLD fMRI covering the entire brain (TR/TE = 4000/30 ms; Matrix = 64×64; 40 axial slices).

Characteristic	Entire Cohort $(\mu \pm \sigma)$	Only Controls $(\mu \pm \sigma)$	Only Patients $(\mu \pm \sigma)$
Age	70.4 ± 8.5	69.9 ± 9.4	71.1 ± 7.0
Education	16.7 ± 2.7	16.4 ± 3.0	17.2 ± 2.3
Delayed Recall	6.2 ± 3.2	8.4 ± 1.6	3.1 ± 1.8
Total Hippocampal Vol.	4065.2 ± 766.8	4281.7 ± 642.0	3770.8 ± 835.0

Table 7.1: Basic statistics for the cohort.

7.2.1 Data Preprocessing

We used ANTs (Avants et al. 2009) to preprocess the data and used a subset of the AAL labels (80 cortical labels out of the total 116 labels) as our seed ROIs (Tzourio-Mazoyer et al. 2002). The list of 80 cortical labels is in Appendix A. Firstly, we registered the AAL labels which were in template space to subject T1 space and then subsequently registered them to the BOLD space. Once in BOLD space, we multiplied the labels with a gray matter probability mask to smooth the labels and convert them to probabilities as well as to reject labelings in non-cortical regions.

We embrace a minimalistic approach to resting-state fMRI processing that seeks to use widely accepted methods to factor out nuisance variables within subject-space (Glasser et al. 2013). The first step in our process involves motion correcting each time-slice of the BOLD image to the average BOLD image in order to capture motion parameters. The first five time slices are discarded before further processing. We then identify physiological noise with the CompCor algorithm (Behzadi et al. 2007). Motion and CompCor parameters are then residualized off of the time series matrix. Subsequently, we apply a band pass filter to the time-series data with lower and upper frequencies of 0.01 and 0.1 respectively. No spatial smoothing is performed and all computations are undertaken within the original subject's BOLD space.

After subject-specific preprocessing, we used p-Eigen to create the constrained eigenvectors $\{\mathbf{v_i}\}_{\mathbf{i=1}}^{\mathbf{k}}$. Once we had the eigenvectors, we used a total of eight labels for our study on the default mode and hippocampus network. The eight AAL ROIs that we used along with their modified counterparts after running p-Eigen are shown in Figures 7.3, 7.4, 7.5, 7.6, 7.7.

It is worth clarifying that since the algorithms for PCA and p-Eigen optimization are iterative in nature, they need an initialization of eigenvectors. In order to make a fair comparison we initialize the eigenvectors with the AAL ROIs both for PCA as well as p-Eigen. We did not have to initialize them this way for PCA, as it is oblivious to any prior information, and has no correspondence between the eigenvectors and the ROIs; however we did that just in order to make a fair comparison. In the case of PCA the eigenvectors drift away freely whereas for p-Eigen they remain close to the priors based on the strength of the prior weight.

Next, we computed the regularized partial correlations between the mean BOLD signal across all the 120 time series points, which were used to construct the functional connectivity graphs for all the three methods AAL, PCA and p-Eigen.

Regularized partial correlations have been shown to be a more robust measure of graph connectivity than simple correlations or partial correlations (Smith et al. 2011; Varoquaux et al. 2010b). We estimated them using Graphical Lasso, which also imposes sparseness on the estimated partial correlations, making the derived network more interpretable (Friedman et al. 2008b).

Note that we run p-Eigen and find partial correlations using all the 80 cortical ROIs in order to explain the covariation in the entire cortex. Finally, we derive covariates for MCI classification and delayed recall prediction from only the eight nodes in the default mode and hippocampus network post hoc due to their strong association with MCI as has been pointed out by (Petrella et al. 2011; Agosta et al. 2012; Hedden et al. 2009; Bai et al. 2009).

7.2.2 Choosing tunable parameters

We used the Graphical Lasso R package (Friedman et al. 2008b) for estimating regularized partial correlations and the GLMNET R package (Friedman et al. 2009) for Elastic Net classification.

The Graphical Lasso has a tunable parameter (hyperparameter) ρ which controls the amount of regularization (sparsity in the estimated regularized partial correlations), with $\rho = 0$ corresponding to no regularization. Further, the Elastic Net has two tunable parameter α and β. α controls the sharing of strength between ℓ_2 and ℓ_1 penalties with a smaller α corresponding to a bigger ℓ_2 regularization. β scales the strength of the penalty terms (ℓ_2 and ℓ_1) relative to the data term, just as in any penalized regression. In addition to this, p-Eigen has a tunable parameter θ which controls the effect of the priors, with higher values corresponding to larger effect of prior.

We tuned all these parameters in a totally data driven manner via a nested leave 5 out cross validation (CV), where we tried values of ρ, $\alpha \in$ [0,1] in steps of 0.02, $\beta \in$ [0,2] in steps of 0.05 and $\theta \in$ [0.05, 0.95] in steps of 0.1. In order to be totally objective and be completely fair to all the methods, we tuned these parameters separately for all the three methods AAL, PCA and p-Eigen for the tasks of MCI vs. Normal classification and prediction of Delayed Recall in a memory task. Finally, we choose the parameters corresponding to the minimum CV error.

The work-flow showing the details of our approach is shown in Figures 7.1 and 7.2.

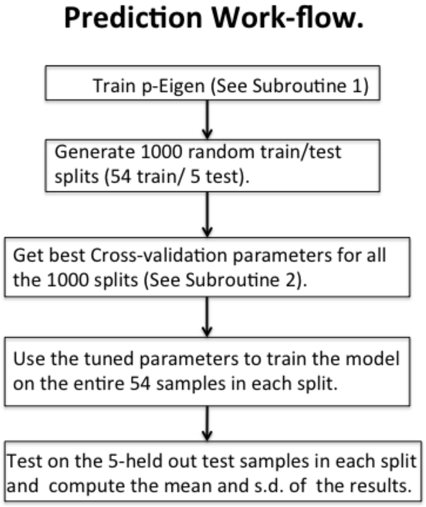

Prediction Work-flow.
Train p-Eigen (See Subroutine 1)
Generate 1000 random train/test splits (54 train/ 5 test).
Get best Cross-validation parameters for all the 1000 splits (See Subroutine 2).
Use the tuned parameters to train the model on the entire 54 samples in each split.
Test on the 5-held out test samples in each split and compute the mean and s.d. of the results.

Figure 7.1: Work-flow (Part 1) showing our prediction pipeline.

The results reported in the next subsection use the "best" values of these tunable parameters that we found in the cross validation (CV) and are reported in the Table 7.2.

	MCI vs Control			Delayed Recall (Patients)		
Parameter	p-Eigen	AAL	PCA	p-Eigen	AAL	PCA
ρ	0.2	0.26	0.66	0.74 (0.64)	0.88 (0.58)	0.88 (0.4)
α	0.94	0.44	0.08	0.02 (0.48)	0.02 (0.72)	0.02(0.36)
β	0.1	2	0.1	1.7 (0.4)	0.7 (1.7)	0.7 (0.9)
θ	0.80	N/A	N/A	0.80	N/A	N/A

Table 7.2: Best values (minimum CV error on most validation splits) of the tunable parameters chosen via CV. Parameters for Delayed Recall for Patients in parenthesis. Note that the implicit value of θ for AAL is 1 and for PCA is 0.

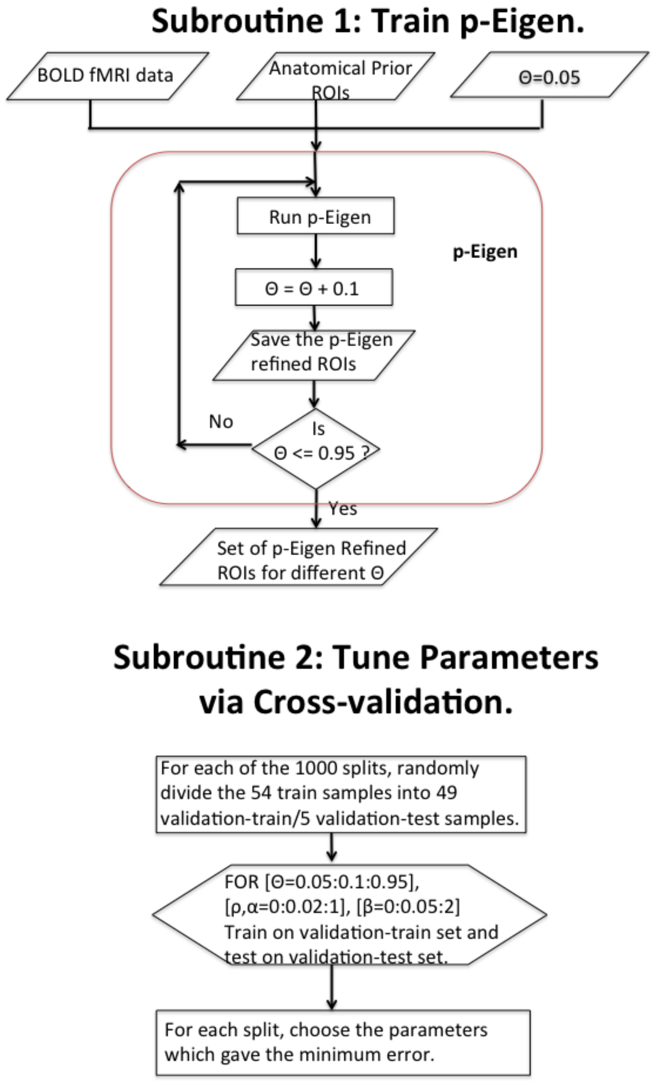

Subroutine 1: Train p-Eigen.
BOLD fMRI data
Anatomical Prior ROIs
Θ=0.05
p-Eigen
Run p-Eigen
Θ = Θ + 0.1
Save the p-Eigen refined ROIs
No
Is Θ <= 0.95 ?
Yes
Set of p-Eigen Refined ROIs for different Θ
Subroutine 2: Tune Parameters via Cross-validation.
For each of the 1000 splits, randomly divide the 54 train samples into 49 validation-train/5 validation-test samples.
FOR [Θ=0.05:0.1:0.95], [ρ,α=0:0.02:1], [β=0:0.05:2] Train on validation-train set and test on validation-test set.
For each split, choose the parameters which gave the minimum error.

Figure 7.2: Work-flow (Part 2) showing our prediction pipeline.

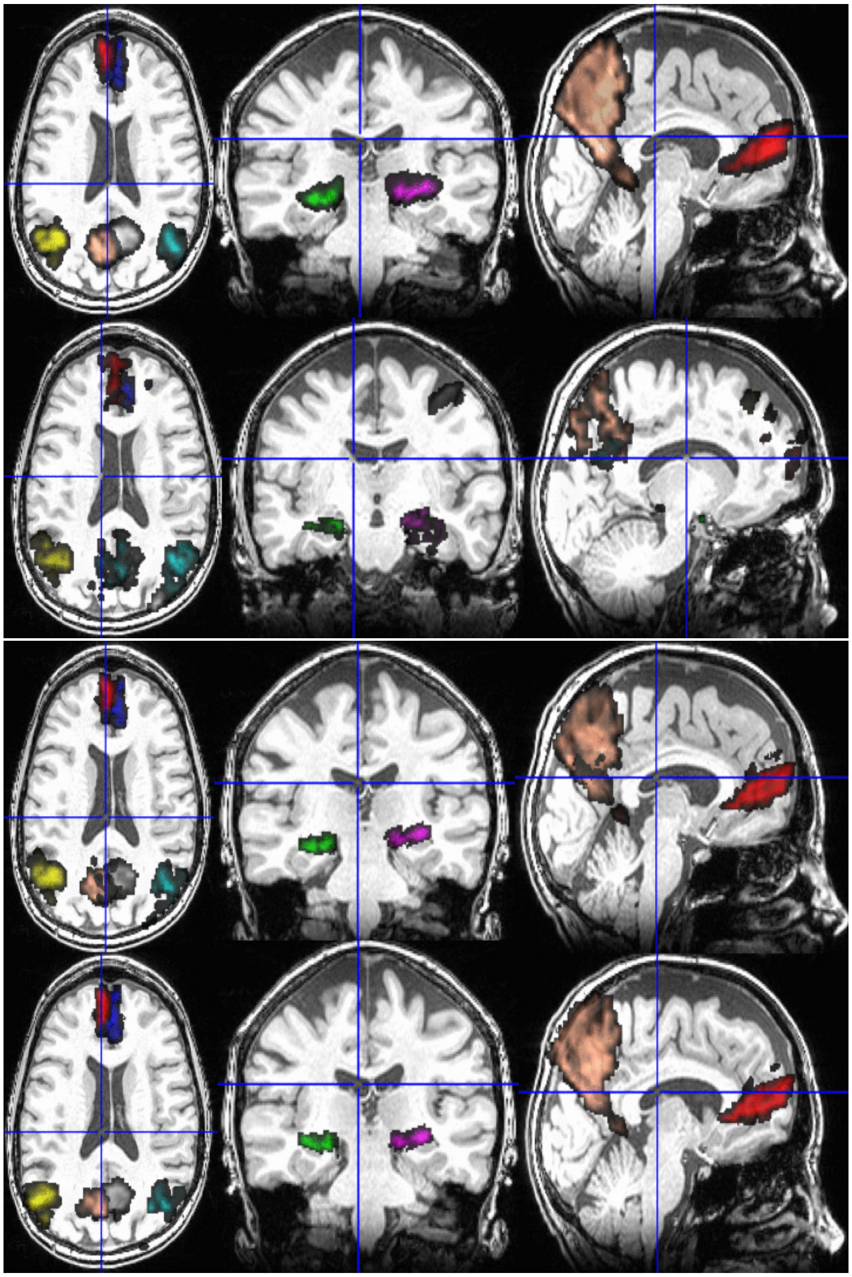


Figure 7.3: Row 1: The original 8 AAL ROIs; Row 2: PCA modified ROIs; Row 3: p-Eigen modified ROIs ($\theta = 0.5$); Row 4: p-Eigen modified ROIs ($\theta = 0.80$) for a randomly chosen subject. 6 of these 8 ROIs are from the default mode network (Precuneus (L/R), Angular Gyrus (L/R), Frontal Medial Orbital Lobe (L/R)) and the remaining two are from left and right parts of hippocampus.

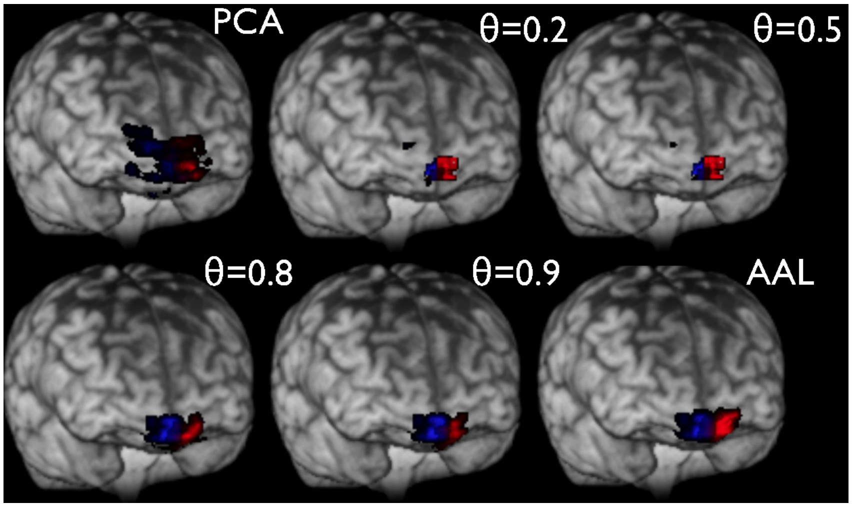


Figure 7.4: Modified Frontal Medial Orbital Lobe ROIs as a function of θ for a randomly chosen subject.

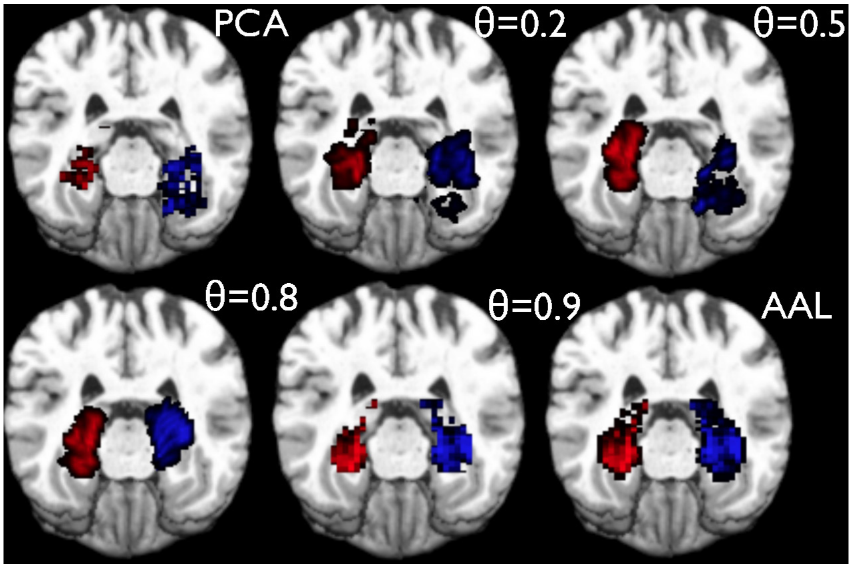


Figure 7.5: Modified Hippocampal ROIs as a function of θ for a randomly chosen subject.

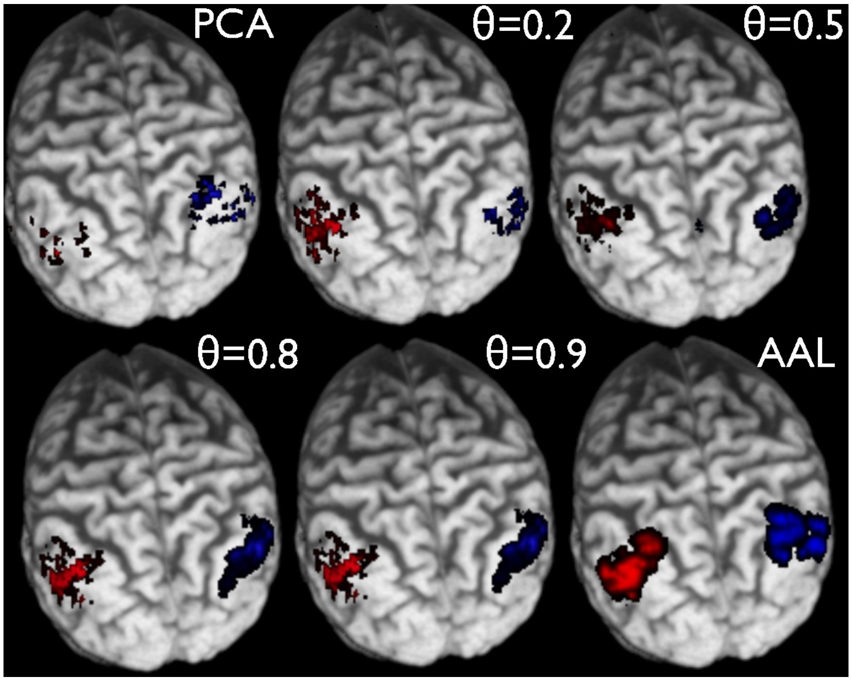


Figure 7.6: Modified Angular Gyrus ROIs as a function of θ for a randomly chosen subject.

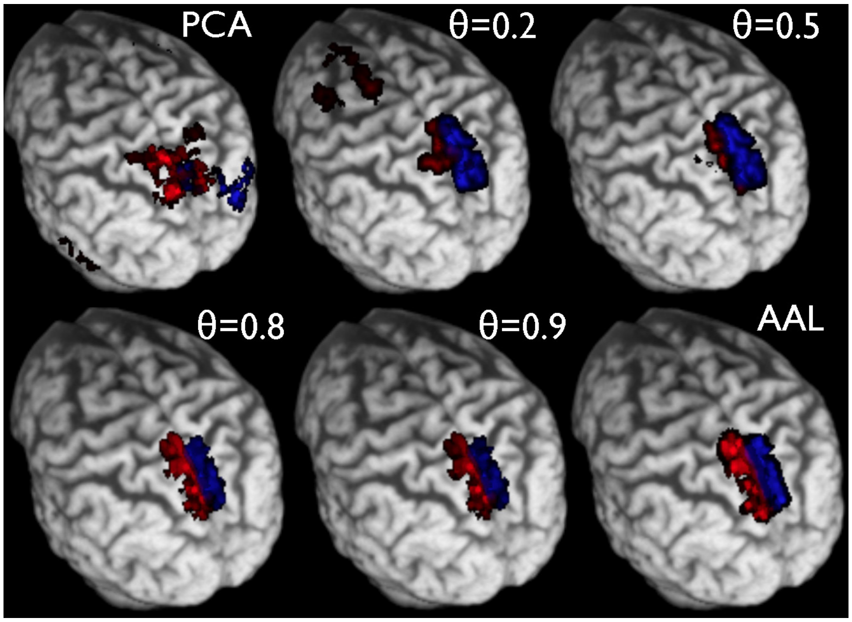


Figure 7.7: Modified Precuneus ROIs as a function of θ for a randomly chosen subject.

7.2.3 Results

We used our functional connectivity graph network information for classifying controls vs MCI. We also used the network information to predict the *Delayed Recall* for the entire cohort as well as for the patients.

In all the experiments reported below, we perform a leave-5-out cross validation (separate from the one used to tune the hyperparameters) i.e. training on 54 and testing on 5, and this procedure was repeated a 1000 times. We used an Elastic Net classifier whose parameters were tuned as described earlier.

The base features that we used in our classification tasks were, *age*, *education*, *gender* and *Hippocampal Volume* (Total left and right, as well as separately for left and right hemisphere) of the subject. As described earlier, in addition to this, we used the sparse inverse correlation matrix of the DMN network (which was estimated using Graphical Lasso) as features.

The results are summarized in the Tables 7.3 and 7.4.

Feature Set	MCI vs. Normal ($\mu \pm \sigma$)	Delayed Recall (All) ($\mu \pm \sigma$)	Delayed Recall (Patients) ($\mu \pm \sigma$)
Base Features	0.42 ± 0.08	2.68 ± 0.35	1.47 ± 0.30
+ p-Eigen	**0.24 ± 0.06**	**1.83 ± 0.29**	**0.97 ± 0.18**

Table 7.3: Results showing p-Eigen better than just using the Base Features. For MCI vs. Normal we report mean classification error (e.g. 0.24 $\rightsquigarrow$ 24%), whereas for Delayed Recall we report Mean Absolute Prediction Error ($\sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{n}$). The p-values from two sample t-tests for all columns are 2.2×10^{-16} .

#	FeatureSet	MCI vs. Normal ($\mu \pm \sigma$)	Delayed Recall (All) ($\mu \pm \sigma$)	Delayed Recall (Patients) ($\mu \pm \sigma$)
1.	AAL	0.36 ± 0.05	2.34 ± 0.21	1.41 ± 0.28
2.	PCA	0.34 ± 0.06	2.41 ± 0.35	1.35 ± 0.21
3.	p-Eigen	**0.24 ± 0.06**	**1.83 ± 0.29**	**0.97 ± 0.18**

Table 7.4: Results showing p-Eigen better than AAL and PCA ROI labels. For MCI vs. Normal we report mean classification error (e.g. 0.24 $\rightsquigarrow$ 24%), whereas for Delayed Recall we report Mean Absolute Prediction Error ($\sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{n}$). All the classifiers also used Base Features in addition to the graph measurement features from ROIs. The p-values from two sample t-test for p-Eigen vs AAL and PCA are 2.2×10^{-16}.

The obtained networks for the AAL, PCA and p-Eigen labels are shown in Figures 7.10, 7.12 and the corresponding heatmaps are show in Figures 7.11 and 7.13.

The Dice coefficient for p-Eigen as a function of the prior strength parameter θ is shown in Figure 7.8. A prior strength of zero corresponds to totally data driven (PCA) based decomposition and a prior strength of one corresponds to using only the prior. As expected, there is increasing overlap between the p-Eigen refined ROIs and the true AAL ROIs as the prior strength increases.

Our main findings are:

1. p-Eigen network measurements along with the base features are a significantly better predictor of both MCI status and Delayed Recall score compared to

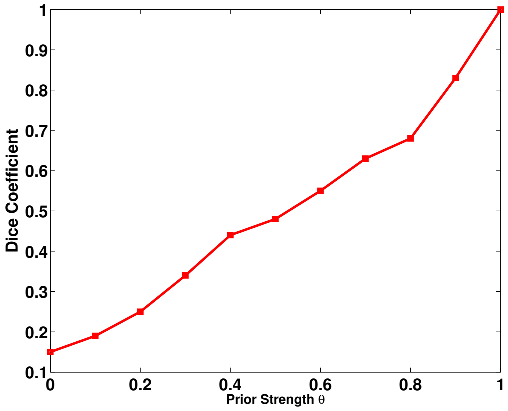


Figure 7.8: Dice coefficient for p-Eigen for the randomly chosen subject (same as Figure 7.3) with AAL ROIs as a function of the prior strength parameter θ. The plotted dice coefficient ($=2\frac{|A \cap B|}{|A|+|B|}$, where A is the p-Eigen modified ROI and B is the true AAL ROI) was computed as the average of dice coefficients over all the 8 DMN ROIs.

using the base features alone.

2. p-Eigen network measurements along with the base features are a significantly better predictor of both MCI status and Delayed Recall score compared to using AAL or PCA ROI graph measurements along with the base features.

7.2.4 Robustness of p-Eigen

In this section we first show some additional experiments which highlight the robustness of our approach. Particularly, we consider alternate definitions of initial ROIs and the use of other measures of correlation to construct networks.

Sensitivity to the Choice of ROIs.

To investigate the sensitivity of our results we considered an alternative ROI set constructed using Ward Clustering (Michel et al. 2012). We parcellated the average time-series image of the subjects using Ward Clustering (we used the implementation from NILEARN (`http://nilearn.github.io/`)) with 80 clusters and then used those as the prior ROIs for p-Eigen. The corresponding method is called p-Eigen (Ward).

In addition, we also compared p-Eigen against another baseline. We created subject specific parcellations by intersecting AAL ROI labels with the subject specific ROIs obtained by performing Ward Clustering on each subject's time series image separately and keeping only those clusters that were larger than 100 voxels. This provides a simple baseline method to construct subject specific parcellations and hence functional connectivity networks.

The results are shown in Table 7.5 and the corresponding intersected ROIs are shown in Figure 7.9. We can make two observations from the results. Firstly, p-Eigen with AAL ROIs is statistically significantly better than subject specific parcellations constructed by intersecting AAL labels with Ward clusters. This shows that subject specific functional parcellations constructed using p-Eigen contain more discriminative signal to aid MCI and delayed recall prediction.

Secondly, p-Eigen with AAL ROIs is significantly better (though marginally) than p-Eigen using Ward Clustered ROIs. We conjecture that it is due to the Ward Clustered ROIs being noisy as can be seen in Fig. 7.9. When we further smoothed them and used them to constrain p-Eigen (p-Eigen (Ward)), the difference was no longer significant.

So, p-Eigen is reasonably robust to the choice of ROIs used for priors, unless they are very noisy. In such cases, smoothing might improve performance.

Using Pearson's Correlation for Constructing Networks.

As mentioned earlier, regularized partial correlations have been shown to be a robust measure of graph connectivity. However, its estimation procedure via Graphical Lasso has a tunable parameter ρ, which needs to be chosen via cross validation. So, we were interested in knowing if we can obviate the need for it by constructing the functional connectivity network using Pearson's correlation which

#	FeatureSet	MCI vs. Normal ($\mu \pm \sigma$)	Delayed Recall (All) ($\mu \pm \sigma$)	Delayed Recall (Patients) ($\mu \pm \sigma$)
1.	Ward Clustering	0.33 ± 0.08	2.24 ± 0.23	1.21 ± 0.24
2.	p-Eigen (Ward)	0.28 ± 0.05	2.07 ± 0.26	1.10 ± 0.22
3.	p-Eigen (AAL)	**0.24 ± 0.06**	**1.83 ± 0.29**	**0.97 ± 0.18**
	p-value	(2. vs 3.) 0.045 (1. vs 3.) 2.7×10^{-6}	(2. vs 3.) 0.041 (1. vs 3.) 1.4×10^{-7}	(2. vs. 3) 0.066 (1. vs. 3) 3.4×10^{-5}

Table 7.5: Results comparing 1). p-Eigen with Ward Clustering based ROIs and 2). p-Eigen with AAL ROIs vs p-Eigen with Ward Clustering based ROIs. For MCI vs. Normal we report mean classification error (e.g. 0.24 ⇝ 24%), whereas for Delayed Recall we report Mean Absolute Prediction Error ($\sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{n}$). All the classifiers also used Base Features in addition to the graph measurement features from ROIs. Note: The reported p-values are from a two sample t-test.

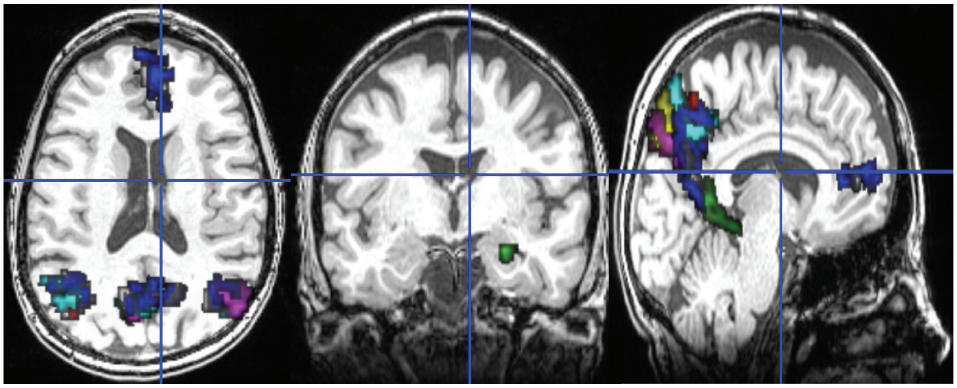


Figure 7.9: Figure showing only those Ward ROIs that had a non-zero intersection with any of the 8 AAL ROIs and were larger than 100 voxels. There were a total of 7 such ROIs.

has no tunable parameters.

The results are shown in Table 7.6. As can be seen p-Eigen with graph constructed using Graphical Lasso performs significantly better than p-Eigen with the graph constructed using Pearson's correlation. This can be due to the fact that since regularized partial correlations explain away the effect of all the other nodes in the network while computing correlation between a pair of nodes, they are more robust. Our finding is also in consonance with the finding by (Smith et al. 2011).

#	FeatureSet	MCI vs. Normal ($\mu \pm \sigma$)	Delayed Recall (All) ($\mu \pm \sigma$)	Delayed Recall (Patients) ($\mu \pm \sigma$)
1.	AAL (Pearson)	0.39 ± 0.05	2.51 ± 0.23	1.57 ± 0.28
2.	PCA (Pearson)	0.38 ± 0.06	2.59 ± 0.37	1.46 ± 0.21
3.	p-Eigen (Pearson)	0.31 ± 0.04	2.09 ± 0.31	1.11 ± 0.11
4.	p-Eigen (GLasso)	$\mathbf{0.24 \pm 0.06}$	$\mathbf{1.83 \pm 0.29}$	$\mathbf{0.97 \pm 0.18}$
	p-value	(3. vs 4.) 1.3×10^{-4}	(3. vs 4.) 2.1×10^{-3}	(3. vs. 4) 3.3×10^{-4}

Table 7.6: Results showing p-Eigen (AAL) with Graphical Lasso better than p-Eigen using Pearson's correlation. For MCI vs. Normal we report mean classification error (e.g. 0.24 ⇝ 24%), whereas for Delayed Recall we report Mean Absolute Prediction Error ($\sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{n}$). All the classifiers also used Base Features in addition to the graph measurement features from ROIs. All other p-values for each column were 2.2×10^{-16}.

7.3 Discussion

We proposed a new approach for deriving data-driven subject specific functional parcellations. The strong and robust empirical performance of p-Eigen makes it a viable alternative to the seed based or totally data driven approaches. p-Eigen also enhances the findings of several MCI clinical studies.

Multiple clinical studies have reported that there is change in connectivity in default mode network (DMN) for MCI patients as well as for the patients who progress onto clinical Alzheimer's Disease (AD) (Greicius et al. 2004; Sheline et al. 2010; Petrella et al. 2011; Agosta et al. 2012; Hedden et al. 2009; Bai et al. 2009; He and Evans 2010).

There are variations in ROI definitions across these studies and furthermore there are some variations on the ROIs thought to be the primary nodes. In this paper, we tried to choose the ROIs corresponding to one of the most widely accepted definitions of DMN. However, differences in definition of the primary nodes of the network maybe a source of some variation across the studies. Nonetheless, there are two consistent and general findings across most MCI studies.

- There is reduced mean connectivity across all the nodes in the DMN for the MCI patients compared to the healthy controls.
- There is reduced mean connectivity of the hippocampus with the other nodes of the DMN (Frontal Medial Orbital, Angular Gyrus, Precuneus in our case) for the MCI patients compared to the healthy controls.

The networks and heatmaps generated by our approach (Figures 7.10, 7.12, 7.11 and 7.13) for a randomly chosen patient and control illustrate these findings.

First, all the three approaches (AAL, PCA, p-Eigen) highlight that there is reduced mean DMN connectivity in the networks of MCI patients compared to the controls. However, this reduction (averaged over all the subjects) is only significant ($p \approx 0.03$ in Welch's t-test) for p-Eigen ROIs. The corresponding p-values for AAL and PCA are 0.09 and 0.11 respectively.

Second, the memory network is disrupted in patients; they have limited or no connectivity of hippocampus with the DMN. This contrast is more evident in the case of p-Eigen ROIs as the Hippocampus has strong connections to DMN on average for the controls but is not connected (or is weakly connected) to the nodes of the DMN for the patients. In the case of AAL, the Hippocampus also has strong connections to DMN for the controls but the change is less drastic when we compare to patients as some patients also have significant hippocampus connectivity to DMN. Lastly, in the case of PCA, there is no hippocampus connectivity with DMN in the case of patients, but the same is true for controls also, as the PCA networks have very few connections, in general.

A quantitative evaluation of mean connectivity of hippocampus with the nodes of the DMN, shows that the reduction in mean connectivity between controls and patients is highly significant ($p < 0.01$) for p-Eigen. For AAL and PCA it is again insignificant, however, with a trend towards statistical significance.

We conjecture that it is due to the ability of p-Eigen ROIs to highlight these differences (which have also been confirmed by multiple studies) between controls and patients, that it does a better job of MCI vs. Control classification and Delayed Recall prediction compared to AAL and PCA.

7.3.1 Limitations

In summary, we showed that p-Eigen better resolves subtle functional patterns that separate MCI from controls and that this is largely consistent with past research. It is also possible that changes in DMN may predict those who convert from MCI to AD and those who do not. While we cannot address this question, it is possible that our MCI data contains subjects of both types. Although this is a limitation in this study, it leads to the possibility that p-Eigen may be extracting signal that is relevant to separating those who do from those who do not progress to AD. This will be a topic of future research. A second limitation of our research is that we explored only a single parcellation scheme, i.e. the AAL. While other parcellation schemes e.g. (Klein and Tourville 2011; Yeo et al. 2011) may reveal different results,

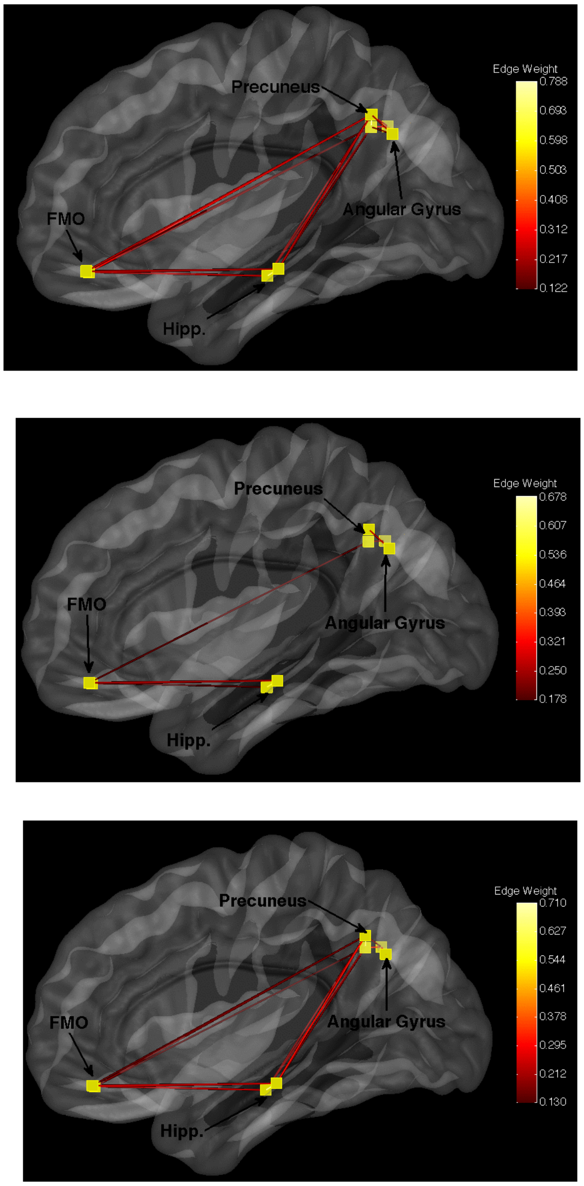


Figure 7.10: Default Mode Networks for a randomly chosen control. Top-to-Bottom AAL, PCA, pEigen. Key: FMO- Frontal Medial Orbital.

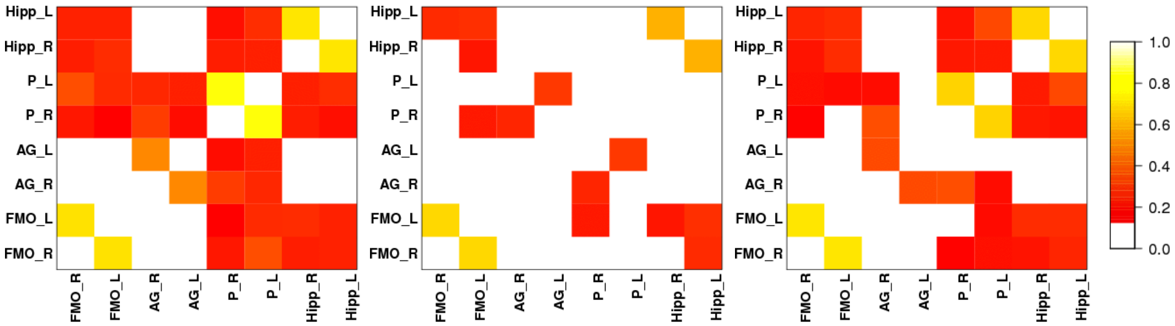


Figure 7.11: Heatmaps for brain networks connectivity for the same randomly chosen control as above. AAL, PCA, pEigen (L-to-R). Key: FMO- Frontal Medial Orbital, AG- Angular Gyrus, P- Precuneus, Hipp.- Hippocampus.

our focus on the relatively consistently defined DMN mitigates this possibility. We did show results which used data-driven Ward Clustering based parcellations and showed that p-Eigen is reasonably robust to the definition of ROIs. However, we think that this needs to be explored more as connectome analyses are known to be sensitive to functional homogeneity (Zuo et al. 2013). We also note that a full exploration of the parameter space of fMRI pre-processing decisions may alter the results reported here. For instance, though we perform motion correction but there might be some residual motion effects in the signal that could affect our results (Van Dijk et al. 2012; Power et al. 2012). We chose a minimal pre-processing pipeline with reasonable control for motion and other nuisance parameters widely recognized as problematic. The fact that we employ a prediction framework and consistent processing across all algorithms compared also mitigates this limitation, although it remains considerable. Finally, we note that we explored only one out of the many possible applications of p-Eigen. Additional work in structural imaging and structural-functional decomposition will be considered in the future.

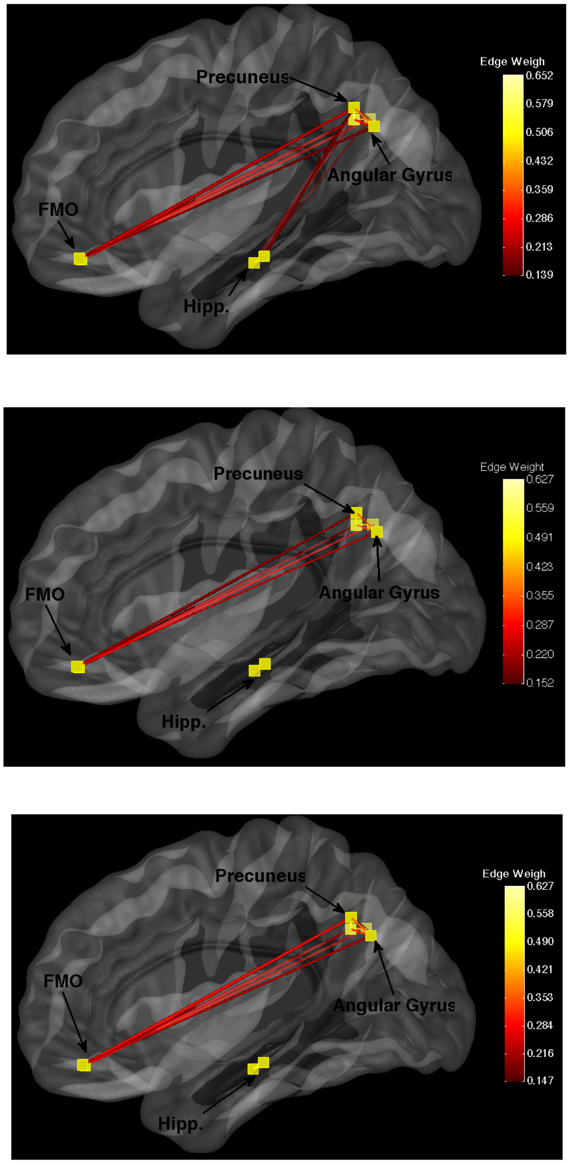


Figure 7.12: Default Mode Networks for a randomly chosen patient. Top-to-Bottom AAL, PCA, pEigen. Key: FMO- Frontal Medial Orbital.

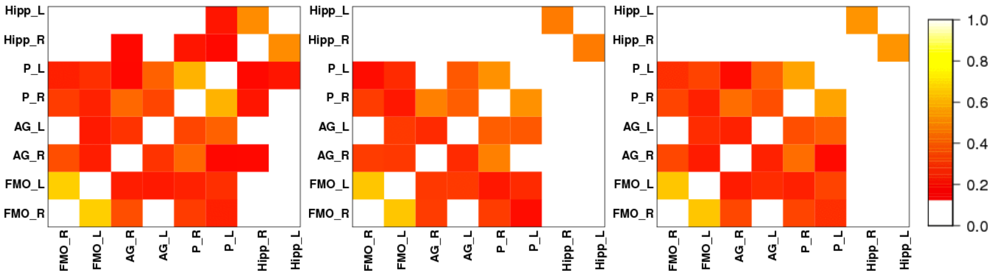


Figure 7.13: Heatmaps for brain networks connectivity for the same randomly chosen patient as above. AAL, PCA, pEigen (L-to-R). Key: FMO- Frontal Medial Orbital, AG- Angular Gyrus, P- Precuneus, Hipp.- Hippocampus.

7.4 Conclusion

The subject specific parcellations generated by p-Eigen were used to construct subject specific functional connectivity networks. These networks showed reduced sensitivity to ROI placement for a cohort of subjects which included people diagnosed with MCI. The network measures gathered from our refined ROIs significantly aid classification of early Mild Cognitive Impairment (MCI) as well as the prediction of Delayed Recall in a memory task when compared to metrics derived from standard registration-based ROI definitions, totally data driven methods, a model based on standard demographics plus hippocampal volume and state-of-the-art Ward Clustering parcellations. We showed that the use of p-Eigen enhances the detection of previously demonstrated findings in this population, namely that there is reduced mean connectivity and disruption of the connections of hippocampus with DMN for MCI patients.

Chapter 8

Prior Based Eigenanatomy (p-Eigen) for Generating Structural (T1) Imaging Brain Parcellations

In this chapter we show the performance of p-Eigen on structural (cortical thickness) imaging data. We use p-Eigen to create refined cortical labels which are then used as features in a classifier to classify Mild Cognitive Impairment (MCI).

8.1 Experimental Results

Our data consists of images from 222 individuals with equal number of males and females of whom 122 were diagnosed clinically with Mild Cognitive Impairment (MCI) and the remaining 100 were normal controls. The average age of the cohort was 71.33. All images were acquired with a Siemens Trio 3.0 Tesla MRI scanner. The analysis of T1 images was done using publicly available Advanced

Normalization Tools (ANTS, `http://www.picsl.upenn.edu/ANTS/`) and the associated pipelining framework PipeDream (`http://sourceforge.net/projects/neuropipedream/`) which mapped each subject to an existing, elderly/neurodegenerative population template, built from images acquired from the same scanner and imaging parameters.

We used two cortical label (probabilistic ROIs) definitions for our experiments 1). Non-Rigid Image Registration Evaluation Project (NIREP), (`http://www.nirep.org/`), 32 labels in total and 2). LONI Probabilistic Brain Atlas (LPBA40) (Shattuck et al. 2008), 55 labels in total.

We ran p-Eigen independently for each label set with varying values of prior strengths, $\theta = [0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0]$.

8.1.1 Qualitative Results

The figures below show slices of axial and sagittal section labelings for unconstrained PCA, true cortical labels (NIREP and LPBA40) and p-Eigen.

Note that in the unconstrained PCA, the same label is assigned to parts of the left frontal gyrus and the left superior temporal gyrus (shown by arrows in figure) since it has no notion of anatomy; on the other hand, p-Eigen, since it is anatomical prior driven does not join distant structures which aids interpretability.

Another interesting observation is that both p-Eigen and unconstrained PCA have assigned the same label to corresponding regions in the left and right hemispheres. This occurs less frequently in p-Eigen as we have constrained it via anatomical priors.

8.1.2 Classification of MCI vs Controls

We hypothesize that p-Eigen will improve over basic ROIs on (MCI vs Controls) classification results by allowing the summary measurements derived from the imag-

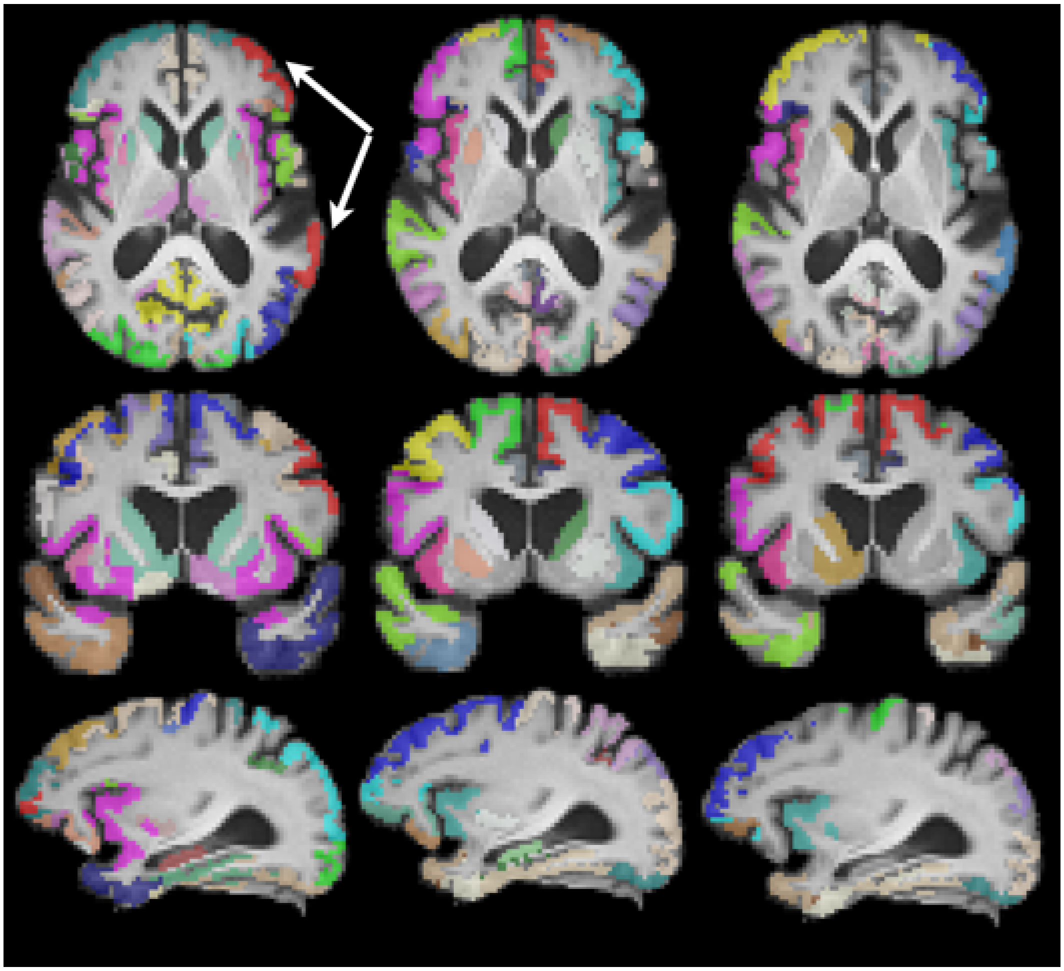


Figure 8.1: Results for LPBA40 labels. Left to right- Unconstrained PCA, prior LPBA labels, p-Eigen labels. The arrows show left frontal gyrus and the left superior temporal gyrus. The value of the prior strength θ for p-Eigen is 0.2, chosen on validation set. Note that not all the 55 anatomical priors can be seen in this slice.

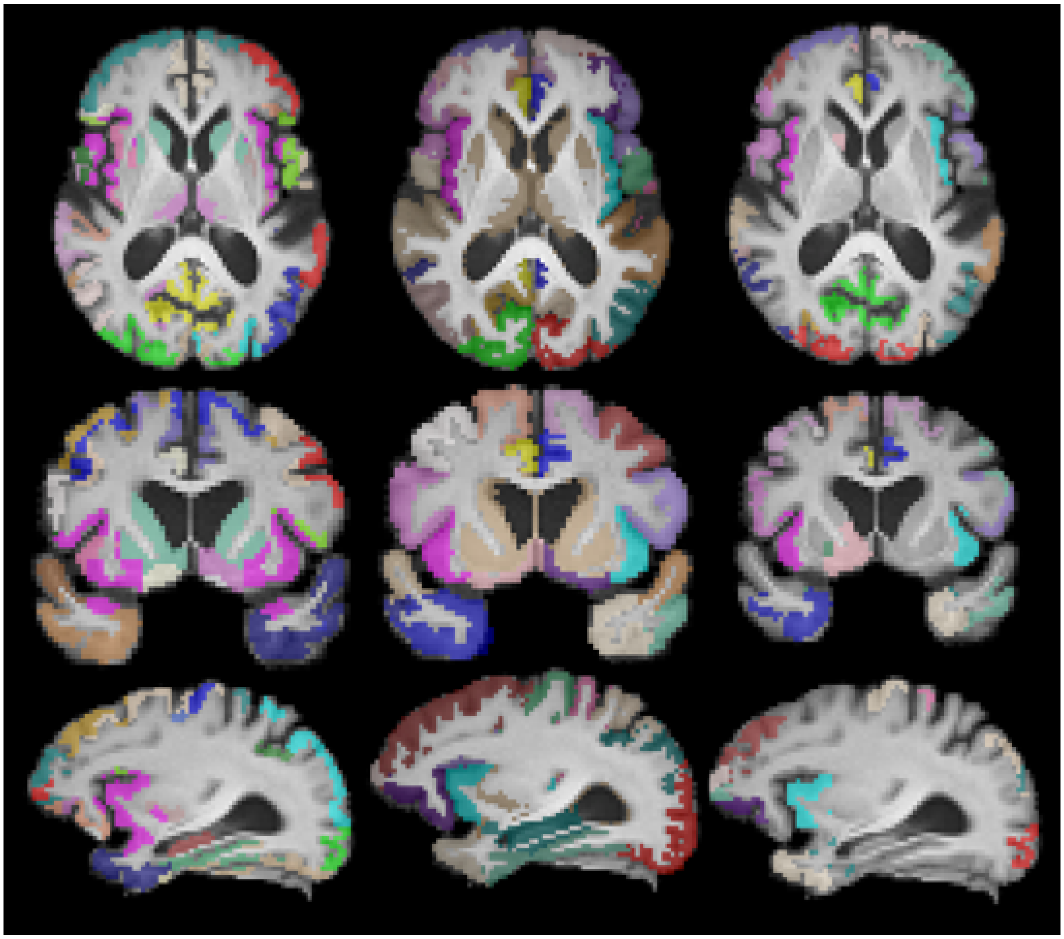


Figure 8.2: Results for NIREP labels. Left to right- Unconstrained PCA, prior NIREP labels, p-Eigen labels. The value of the prior strength θ for p-Eigen is 0.5, chosen on validation set. Note that not all the 32 anatomical priors can be seen in this slice.

ing data to adapt to the underlying signal. We also hypothesize that p-Eigen locality will improve over unconstrained PCA-based classification.

We use the projections (eigenvectors projected onto the data matrix) resulting from unconstrained PCA, p-Eigen and prior cortical labels to distinguish individuals with Mild Cognitive Impairment (MCI) from normal aging individuals (Zhou et al. 2011). Since it is a classification problem, we used logistic regression and randomly split the data into training (80%), validation (10%) and testing (10%). Firstly, we use cross validation to determine the best value of prior strength parameter (θ); we train on the training data and test on validation data with varying prior strengths as mentioned above and then choose the best value of prior strength as the one which gave the least classification error on the validation set. The best prior strength value was $\theta = 0.5$ for NIREP labels and $\theta = 0.2$ for LPBA40 labels.

Next, with these values of λ we trained p-Eigen on the entire training and validation set and tested on the test set. This whole procedure was repeated 10000 times and the classification accuracies are given in Table 8.1.

Algorithm ($\mu \pm \sigma$)	NIREP	LPBA40
Unconstrained PCA	$62.58\% \pm 4.5\%$	$62.58\% \pm 4.5\%$
ROI Cortical Labels	$66.05 \pm 3.9\%$	$65.95\% \pm 3.6\%$
p-Eigen	$\mathbf{67.96}\% \pm \mathbf{2.3}\%$	$\mathbf{67.15}\% \pm \mathbf{3.2}\%$

Table 8.1: Test Set classification accuracies averaged over 10000 runs (More details in text). p-Eigen is significantly ($p - val < 0.0001$ in paired t-test) better than Unconstrained PCA and ROI Cortical Labels

As can be seen from the table, the p-Eigen significantly (paired t-test) outperforms unconstrained PCA as well as the approach which just uses cortical ROI labels hence leading to a classifier which can better distinguish MCI patients from normal controls.

Chapter 9

Conclusion & Future Work

In this thesis we made multiple contributions. First, we proposed three algorithms for learning word embeddings (eigenwords) which are fast to train, have strong theoretical properties, can induce context specific embeddings and have better sample complexity for rare words. All the algorithms had a Canonical Correlation Analysis (CCA) style eigen-decomposition at their core. We performed a thorough evaluation of *eigenwords* learned using these algorithms, and showed that they were comparable to or better than other state-of-the-art algorithms when used as features in a set of NLP classification tasks. Eigenwords are able to capture nuanced syntactic and semantic information about the words. They also have a clearer theoretical foundation than the competing algorithms, which allows us to bound their error rate in recovering the true hidden state under linearity assumptions.

Second, we showed that in order to get good performance with spectral embeddings (or any embeddings which employ matrix factorization on word co-occurrence matrices) we need to transform the data, in particular, transform the word counts by taking their square-root. This makes the result look more Gaussian and hence provide better fits and better embeddings especially when using models of data generation which make Gaussianity assumptions e.g. CCA.

Third, we proposed Eigenanatomy (EANAT), a general framework for sparse matrix factorization for brain images. EANAT incorporates neuroanatomical prior knowledge in the form of connectedness and smoothness constraints. We further augmented it and proposed p-Eigen which incorporates even more domain knowledge and identifies a data-driven matrix decomposition constrained by probabilistic regions of interest (ROIs). We provided efficient optimization algorithms for EANAT and p-Eigen.

The parcellations generated by both EANAT and p-Eigen are able to extract highly discriminative information which helps us validate network-specific hypotheses and significantly improve classification of Mild Cognitive Impairment.

Fourth, we showed that linear models help us attain state-of-the-art performance on two domains– text and brain imaging and there is no need to move to more complex non-linear models, e.g. Deep Learning based models. In addition, spectral learning methods are highly scalable and parallelizable and can incorporate the latest advances in numerical linear algebra as black-box routines.

There are many open avenues for future research building on the above spectral methods.

1. Our word embeddings are based on modeling individual words based on their contexts; it will be interesting to induce embeddings for entire phrases or sentences. There are multiple possibilities here. One could directly model phrases by considering a phrase as a "unit" rather than a word, perhaps taking the context of a word or phrase from connected elements in a dependency or constituency parse tree. Another possibility is to learn embeddings for individual words but then combine them in some manner to get an embedding for a phrase or a sentence; some relevant work on this problem has been done by (Socher et al. 2012, 2013).

2. Closely related is the idea of semantic composition. Recent advances in spectral learning for tree structures e.g. (Dhillon et al. 2012a; Cohen et al. 2012) may be able to be extended to provide scalable principled alternative methods to the recursive neural networks of (Socher et al. 2012, 2013).

3. Also it will be fruitful to study embeddings where the contexts are left and right dependencies of a word rather than the neighboring words in the surface structure of the sentence. This might give more precise embeddings with smaller data sets.

4. We can also borrow ideas and approaches from the brain imaging work and incorporate more domain knowledge into learning of eigenwords. For example, one could envision using ontologies like WordNet (Fellbaum 1998) as priors in an otherwise data-driven embedding learning.

5. On the brain imaging front it will be desirable to model even more complicated priors e.g. arbitrary variance-covariance structures between the ROIs. This might require extending spectral methods to handle more more sophisticated Bayesian models with hierarchical priors.

6. Recent work modeling the brain activity associated with textual stimulus (Mitchell et al. 2008) could benefit from both parts of this thesis. One would ideally learn joint models of word or sentence embeddings along with functional time-series brain imaging data, akin to (Fyshe et al. 2014).

Appendix A

Appendix

A.1 Eigenwords

A.1.1 CCA by SVD

Proof of Eq. 2.3.

Assuming W is the $n \times v$ word matrix and C is the $n \times hv$ context matrix where n is the number of tokens in the corpus, h is the context size and v is the vocabulary size. Further $C_{wc} = W^\top C$, $C_{cc} = C^\top C$ and $C_{ww} = W^\top W$. The CCA objective is to find vectors Φ_w and Φ_c such that the linear combinations $s_w = \Phi_w^\top W$ and $s_{cc} = \Phi_c^\top C$ are maximally correlated.

$$\max_{\Phi_w, \Phi_c} \frac{\mathbb{E}[s_w^\top s_{cc}]}{\sqrt{\mathbb{E}[s_w^\top s_w]}\sqrt{\mathbb{E}[s_{cc}^\top s_{cc}]}} \tag{A.1}$$

i.e.

$$\max_{\Phi_w, \Phi_c} \frac{{\Phi_w}^\top C_{wc} \Phi_c}{\sqrt{{\Phi_w}^\top C_{ww} \Phi_w}\sqrt{{\Phi_c}^\top C_{cc} \Phi_c}} \tag{A.2}$$

This is equivalent to

$$\max_{\Phi_w, \Phi_c} {\Phi_w}^\top C_{wc} \Phi_c \tag{A.3}$$

subject to unit-norm constraints ${\Phi_w}^\top C_{ww}\Phi_w = I$ and ${\Phi_c}^\top C_{cc}\Phi_c = I$.

Then, performing full SVD on C_{ww} and C_{cc}, we get

$$\begin{aligned} C_{ww} &= V_w \Lambda_w V_w^\top \\ C_{cc} &= V_c \Lambda_c V_c^\top \end{aligned}$$

where $V_w^\top V_w = I_{v\times v}$ and $V_c^\top V_c = I_{hv\times hv}$.

Define change of basis as

$$\begin{aligned} u_w &= \Lambda_w^{-1/2} V_w^\top W \\ u_{cc} &= \Lambda_c^{-1/2} V_c^\top C \end{aligned}$$

Now, in this new transformed basis:

$\mathbb{E}[u_w^\top u_w] = \Lambda_w^{-1/2} V_w^\top W V_w^\top \Lambda_w V_w V_w \Lambda_w^{-1/2} = I_{v\times v}$ and similarly $\mathbb{E}[u_{cc}^\top u_{cc}] = I_{hv\times hv}$, as desired.

Transform the coefficients Φ_w and Φ_c, so that s_w and s_{cc} can be expressed as linear combination in the new basis:

$$\begin{aligned} s_w &= \Phi_w^\top W = g_{\Phi_w}^\top u_w \\ s_{cc} &= \Phi_c^\top C = g_{\Phi_c}^\top u_{cc} \end{aligned}$$

where $g_{\Phi_w} = \Lambda_w V_w \Phi_w$ and $g_{\Phi_c} = \Lambda_c V_c \Phi_c$.

So, the CCA optimization problem can be cast as the following maximization criteria

$$\max_{g_{\Phi_w}, g_{\Phi_c}} g_{\Phi_w}^\top D_{wc} g_{\Phi_c} \tag{A.4}$$

subject to unit-norm constraints $g_{\Phi_w}^\top g_{\Phi_w} = I$ and $g_{\Phi_c}^\top g_{\Phi_c} = I$.

where $D_{wc} = \Lambda_w^{-1/2} V_w^\top C_{wc} V_c \Lambda_c^{-1/2}$.

The solution to above is nothing but the SVD of D_{wc}.

Finally, we can construct the original coefficient matrices $\mathbf{\Phi}_w$ and $\mathbf{\Phi}_c$ as $\mathbf{\Phi}_w = V_w \Lambda_w^{-1/2} G_{\Phi_w}$ and $\mathbf{\Phi}_c = V_c \Lambda_c^{-1/2} G_{\Phi_c}$, where G_{Φ_w} and G_{Φ_c} are the matrices corresponding to the vectors g_{Φ_w} and g_{Φ_c} respectively.

Now, in our case $C_{ww} = W^\top W$ is the diagonal word occurrence matrix with the words counts in the corpus on the diagonal, so $\Lambda_w^{-1/2}$ is nothing but $C_{ww}^{-1/2}$ and $V_w = I$.

The context matrix $C_{cc} = C^\top C$, though is not diagonal but it can be approximated by its diagonal. One could also approximate it as a diagonal matrix plus its first order Taylor's expansion, but it would make the resulting matrix substantially more dense and hence the computations intense. In our experiments we found no improvement in prediction accuracy by adding the first order Taylor's term, so we approximate C_{cc} just by its diagonal.

A.1.2 Chapter 2: Theorem 1.

Proof:

With out loss of generality, we can assume that $\mathbf{W}$ and $\mathbf{C}$ have been transformed to their canonical correlations coordinate space. So $Var(\mathbf{W})$ is the identity and $Var(\mathbf{C})$ is the identity, and the $Cov(\mathbf{W}, \mathbf{C})$ is a diagonal with non-increasing values ρ_i on the diagonal (namely the correlations / singular values). We can write α and β in this coordinate system. By orthogonality we now have $\beta_i = \rho_i \alpha_i$. So, $\mathbb{E}(Y - \beta\mathbf{W})^2$ is simply $\sum(\alpha_i - \beta_i \rho_i)^2$. Which is $\sum \alpha_i^2 (1 - \rho_i^2)$. Our estimator will then have $\gamma_i = \beta_i$

for i smaller than k and $\gamma_i = 0$ otherwise. Hence $(\hat{Y} - \beta^\top \mathbf{W})^2 = \sum_{i=k+1}^{\infty} \beta_i^2$.

So if we pick k to include all terms which have $\rho_i \geq \sqrt{\epsilon}$ our error will be less than $\epsilon \sum_{i=k+1}^{\infty} \alpha_i^2 \ \leq \ \epsilon$.

q.e.d.

A.1.3 Chapter 3: Theorem 2.

Proof:

The key is that CCA can be understood using the same machinery as is used for analyzing linear regression. In this context we want to recover the word of length v given its context which can be expressed in terms of regression. For a more in-depth discussion of how CCA relates to regression, see Glahn (1968), for example. Thus, consider the case of predicting a vector $\mathbf{y}$ of length v (the word) from a vector $\mathbf{x}$ (the context, which is of dimension $2hv$ in the one step CCA case and dimension $2k$ in the two step CCA). We consider the linear model

$$y = \mathbf{x}\beta + \epsilon$$

Note that, we are predicting only one dimension of our v-dimensional vector $\mathbf{y}$ at a time.

We want to understand the variance of our prediction of a word given the context. As is typical in regression, we caluclate a standard error for each coefficient in our contexts, $\approx O(\frac{1}{\sqrt{n}})$. In the one step CCA, $\mathbf{X} = [\mathbf{L} \quad \mathbf{R}]$, and running a regression we will get a prediction error on order of $\frac{hv}{n}$, but since we have v such y's we get a total prediction error on the order of $\frac{hv^2}{n}$.

For the two-step case we take $\mathbf{X} = [\mathbf{L\Phi_L} \quad \mathbf{R\Phi_R}]$. As mentioned earlier, note that now we are working with about 2k predictors instead of 2hv predictors. If we knew the true $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$, and thus the true subspace covered by our predictors, the regression error would be on the order of $\frac{kv}{n}$ (again, since there are v entries in our vector $\mathbf{y}$). Instead, we have an estimation of $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$. If these were computed on infinite amounts of data (and hence we would be arbitrarily close to the true subspace)–we would be done. However since they come from a sample, we are using $\widehat{\mathbf{\Phi_L}}$ and $\widehat{\mathbf{\Phi_R}}$ which are approximation to the ideal $\mathbf{\Phi_L}$ and $\mathbf{\Phi_R}$. So our task is to understand the error introduced by this sample approximation of the true CCA. First, we develop some notation and concepts found in Stewart (1990).

Consider two subspaces $\mathcal{V}$ and $\hat{\mathcal{V}}$ and respective matrices containing an orthonormal basis for these subspaces $\mathbf{V}$ and $\hat{\mathbf{V}}$. Let $\gamma_1, \gamma_2, \ldots$ be the singular values of the matrix $\mathbf{V}^\top \hat{\mathbf{V}}$, then define

$$\theta_i = \cos^{-1} \gamma_i$$

and define the canonical angle matrix $\mathbf{\Theta} = \text{diag}(\theta_\mathbf{1}, \ldots, \theta_\mathbf{k})$.

These values of $\mathbf{\Theta}$ capture the effect of using estimated singular vectors, $\hat{\mathbf{V}}$ to form an underlying subspace, as compared to the true subspace formed by the true singular vectors $\mathbf{V}$ stemming from infinite data. The largest canonical angle captures the largest angle between any two vectors- one from the perturbed subspace and one from the true subspace. The second largest canonical angle captures the second largest angle between any two vectors given that they are orthogonal to the original two, and so on. In this proof we will only make use of the largest canonical angle to provide a loose upper bound on the error stemming from the imperfect estimation of the true subspace.

Now, consider a matrix $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ and take the thin singular value decomposition of $\mathbf{A}$ and $\hat{\mathbf{A}}$ (and here we take the liberty of applying diag in a block matrix sense)

$$\mathbf{A} = [\mathbf{U_1 U_2}]\text{diag}(\mathbf{\Lambda_1}, \mathbf{\Lambda_2})[\mathbf{V_1 V_2}]^\top$$
$$\hat{\mathbf{A}} = [\hat{\mathbf{U}}_\mathbf{1} \hat{\mathbf{U}}_\mathbf{2}]\text{diag}(\hat{\mathbf{\Lambda}}_\mathbf{1}, \hat{\mathbf{\Lambda}}_\mathbf{2})[\mathbf{V_1 V_2}]^\top$$

In our case we have that $\lambda_i = 0$ for all $\lambda_i \in \mathbf{\Lambda_2}$.

From Stewart and Sun (1990), we have that

$$\max\{||\sin \mathbf{\Theta}||_\mathbf{2}, ||\sin \mathbf{\Psi}||_\mathbf{2}\} \leq \mathbf{c}||\mathbf{E}||_\mathbf{2} \tag{A.5}$$

for some constant c where here $\mathbf{\Theta}$ is the matrix of canonical angles formed from the subspaces of $\mathbf{U}$ and $\hat{\mathbf{U}}$, and $\mathbf{\Psi}$ is the matrix of canonical angles formed between the subspaces of $\mathbf{V}$ and $\hat{\mathbf{V}}$. Note that since $\mathbf{\Theta}$ and $\mathbf{\Psi}$ are diagonal matrices the induced norms $|| \cdot ||_2$ recover the largest canonical angle of each subspace, and hence we can simultaneously derive an upper bound for the largest canonical angle of either subspace.

We have now developed the machinery we need to analyze the two step CCA.

Without loss of generality, assume that $\mathbf{L}^\top \mathbf{L} = \mathbf{R}^\top \mathbf{R} = \mathbf{I}$, then ultimately we are interested in projection onto the subspace spanned by $\mathbf{B} = [\mathbf{LU_1} \;\; \mathbf{RV_1}]$. Note that because of our assumption the projection onto $\mathbf{LU_1}$ is $\mathbf{LU_1 U_1^\top L^\top}$ and similarly for $\mathbf{RV_1}$. Furthermore, note from our assumptions that $\mathbf{LU_1}$ forms an orthonormal basis for the space spanned by $\mathbf{LU_1}$ (since

$$(\mathbf{LU_1})^\top (\mathbf{LU_1}) = \mathbf{U_1^\top L^\top L U_1} = \mathbf{I}$$

and similarly for $\mathbf{L}\hat{\mathbf{U}}_\mathbf{1}$, $\mathbf{RV_1}$, and $\mathbf{R}\hat{\mathbf{V}}_\mathbf{1}$).

Lastly, and critically, the singular values of $\mathbf{U_1^\top L^\top L}\hat{\mathbf{U}}_\mathbf{1}$ are identical to those of $\mathbf{U_1^\top}\hat{\mathbf{U}}_\mathbf{1}$ (similarly for $\mathbf{RV_1}$ etc.) and so from above we have that the matrix of canonical angles between the subspaces $\mathbf{LU_1}$ and $\mathbf{L\hat{U}_1}$ are identical to $\mathbf{\Theta}$, the matrix of canonical angles between $\mathbf{U_1}$ and $\mathbf{\hat{U}_1}$, and likewise the matrix of canonical angles between the subspaces $\mathbf{RV_1}$ and $\mathbf{R\hat{V}_1}$ are identical to $\mathbf{\Psi}$, the matrix of canonical angles between $\mathbf{V_1}$ and $\mathbf{\hat{V}_1}$, and thus the maximal angle enjoys the same bound derived above. If we can get a handle on the spectral norm of $\mathbf{E}$, which will come directly from random matrix theory, then we can bound the largest canonical angle of our two subspaces.

We know that $\mathbf{E}$ is a random matrix of iid Gaussian entries with variance $\frac{1}{n}$, and that the largest singular value of a matrix is the spectral norm of the matrix. From random matrix theory we know that the square of the spectral norm of $\mathbf{E}$ is $O(\frac{\sqrt{hv}}{\sqrt{n}})$, from say Rudelson and Vershynin (2010).

The strategy will be to divide the variance in the prediction of $\mathbf{y}$ into two separate parts. First the variance that comes from predicting using the incorrect subspace, and then the variance from regression (as stated above) if we had the correct subspace.

Let $\mathbf{\hat{X}} = [\mathbf{L\hat{\Phi}_L} \;\; \mathbf{R\hat{\Phi}_R}]$ (i.e. the incorrect subspace) and $\mathbf{X} = [\mathbf{L\Phi_L} \;\; \mathbf{R\Phi_R}]$ (the true version). To get a handle on predicting with the incorrect subspace (we will consider the subspaces $\mathbf{L\Phi_L}$ and $\mathbf{R\Phi_R}$ separately here, but note that from (A.5) the angles between the subspaces and their respective perturbed subspaces are bounded by a common bound) we note that, for the regression of $\mathbf{Y}$ on $\mathbf{X}$ we have

$$\beta|\mathbf{\hat{X}} = \frac{\mathrm{Cov}(\mathbf{Y}, \mathbf{\hat{X}})}{\mathrm{Var}(\mathbf{\hat{X}})}$$

and

$$\beta|\mathbf{X} = \frac{\mathrm{Cov}(\mathbf{Y}, \mathbf{X})}{\mathrm{Var}(\mathbf{X})}$$

and

$$\mathrm{Cov}(\mathbf{Y},\mathbf{X}) = \mathrm{Cov}(\mathbf{Y},\hat{\mathbf{X}})$$

so trivially

$$\begin{aligned}\beta|\hat{\mathbf{X}} &= \beta|\mathbf{X} * \frac{\mathrm{Var}(\mathbf{X})}{\mathrm{Var}\hat{\mathbf{X}}}\\ &= \beta|\mathbf{X} * \frac{\mathrm{Var}(\mathbf{X})}{\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}\end{aligned}$$

Let $\hat{y}$ be the the estimate of y from the true subspace, and $\hat{\hat{y}}$ be the estimate from the perturbed subspace. For the first part of our strategy, bounding the error that comes from predicting with the incorrect subspace, we want to bound $\mathbb{E}(\hat{y} - \hat{\hat{y}})^2$.

We have

$$\begin{aligned}\left[\hat{y} - \hat{\hat{y}}\right]^2 &= \left[\beta|\mathbf{X} * \mathbf{x} - \beta|\hat{\mathbf{X}} * \mathbf{x}\right]^2\\ &= \left[(\beta|\mathbf{X} - \beta|\hat{\mathbf{X}}) * \mathbf{x}\right]^2\\ &= \left[\left(\beta|\mathbf{X} - \beta|\mathbf{X}\frac{\mathrm{Var}(\mathbf{X})}{\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}\right) * \mathbf{x}\right]^2\\ &= \left[\beta|\mathbf{X}\left(\mathbf{1} - \frac{\mathrm{Var}(\mathbf{X})}{\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}\right) * \mathbf{x}\right]^2\\ &= \left[\beta|\mathbf{X} * \mathbf{x}\left(\frac{\mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}{\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}\right)\right]^2\\ &= \left[\hat{y} * \left(\frac{\mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}{\mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{X}-\hat{\mathbf{X}})}\right)\right]^2 \qquad \text{(A.6)}\end{aligned}$$

Because we are working with a ratio of variances instead of actual variances, then without loss of generality we can set $\mathrm{Var}(\hat{\mathbf{X}}) = 1$ for all predictors.

Now, we don't really care what the exact 'true' $\mathbf{X}$'s are (formed with the true singular vectors), because we only care about predicting y and not actually recovering the true β's associated with our SVD. This means we do not suffer from the usual constraints imposed on the erratic behavior of singular vectors. Usually one must handle this kind of error with respect to the entire subspace since singular vectors are highly unstable. In our case, however, we are free to compare to any 'true' vectors we like from the correct subspace, as long as they span the entire true subspace (and nothing more).

We will define a theoretical set of predictors to compare with, then. We are doing this to obtain an upper bound for the total possible variance of $\mathrm{Var}(x - \hat{x})$ for any acceptable set of x's in the true underlying subspace (where we take acceptable to mean that the x's span the true subspace and nothing more).

We handle each subspace $\mathbf{L\hat{U}_1}$ and $\mathbf{R\hat{V}_1}$ separately. The construction is to take our first vector and 'choose' a vector from the true subspace that lies such that the angle between the two vectors is the maximal canonical angle between the true and perturbed subspaces.

We proceed to our second predictor and choose a vector from the true subspace such the second 'true' predictor is orthogonal to the first. Note that the angle between our second observed $\hat{x}$ and the second chosen x is at most the maximal canonical angle by assumption. Again, because we don't care about the β's associated with our true singular vectors, but only about prediction quality of our perturbed subspace, we need not be worried that our 'chosen' vectors might not be the true singular vectors. We continue in this manner until we have expired all of our predictors from both sets of spaces.

We know from above that the sine of the maximal angle of of both sets of subspaces is $O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)$ and so we have that the maximal variation

$$\frac{\mathrm{Var}(\mathbf{X} - \hat{\mathbf{X}})}{\mathrm{Var}(\hat{\mathbf{X}})} \sim O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)$$

and so from A.6 we have

$$\begin{aligned}\mathbb{E}(\hat{y}-\hat{\hat{y}})^2 &= \mathbb{E}\left[\hat{y} * O\left(\frac{\sqrt{hv}}{\sqrt{n}}\right)\right]^2 \\ &\approx O\left(\frac{hv}{n} * \frac{1}{v}\right) = O\left(\frac{h}{n}\right)\end{aligned}$$

We have v of these to predict, so we have a total error attributable to subspace estimation on the order of $\frac{hv}{n}$. Adding regression error as we did from above, which is on the order of $\frac{kv}{n}$ we get a total error of $\frac{(h+k)v}{n}$. We recall that the error from the one step CCA is on the order of $\frac{hv^2}{n}$ which yields an error ratio of $\frac{h+k}{hv}$.

q.e.d.

A.1.4 Chapter 3: Lemma 1 and Theorem 3.

Our goal is to find a $v \times k$ matrix $\mathbf{A}$ that maps each of the v words in the vocabulary to a k-dimensional state vector. We will show that the $\mathbf{A}$ we find preserves the information in our data and allows a significant data reduction.

Let $\mathbf{L}$ be an $n \times hv$ matrix giving the words in the left context of each of the n tokens, where the context is of length h, $\mathbf{R}$ be the corresponding $n \times hv$ matrix for the right context, and $\mathbf{W}$ be an $n \times v$ matrix of indicator functions for the words themselves.

We will use three assumptions at various points in our proof:

Assumption 1. $\mathbf{L}$*,* $\mathbf{W}$*, and* $\mathbf{R}$ *come from a rank* k *HMM i.e it has a rank* k *observation matrix and a rank* k *transition matrix both of which have the same domain.*

For example, if the dimension of the hidden state is k and the vocabulary size is v then the observation matrix, which is $k \times v$, has rank k. This rank condition is similar to the one used by Siddiqi et al. (2010).

Assumption 1A. *For the three views,* $\mathbf{L}$*,* $\mathbf{W}$ *and* $\mathbf{R}$ *assume that there exists a "hidden state H" of dimension $n \times k$, where each row H_i has the same non-singular variance-covariance matrix and such that $\mathbb{E}(L_i|H_i) = H_i\boldsymbol{\beta}_L^T$ and $\mathbb{E}(R_i|H_i) = H_i\boldsymbol{\beta}_R^T$ and $\mathbb{E}(W_i|H_i) = H_i\boldsymbol{\beta}_W^T$ where all $\boldsymbol{\beta}$'s are of rank k, where L_i, R_i and W_i are the rows of* $\mathbf{L}$*,* $\mathbf{R}$ *and* $\mathbf{W}$ *respectively.*

This assumption actually follows from the previous one.

Assumption 2. $\rho(\mathbf{L}, \mathbf{W})$*,* $\rho(\mathbf{L}, \mathbf{R})$ *and* $\rho(\mathbf{W}, \mathbf{R})$ *all have rank k, where* $\rho(\mathbf{X_1}, \mathbf{X_2})$ *is the expected correlation between* $\mathbf{X_1}$ *and* $\mathbf{X_2}$*.*

This is a rank condition similar to that in Hsu et al. (2009).

Assumption 3. $\rho([\mathbf{L}, \mathbf{R}], \mathbf{W})$ *has k distinct singular values.*

This assumption just makes the proof a little cleaner, since if there are repeated singular values, then the singular vectors are not unique. Without it, we would have to phrase results in terms of subspaces with identical singular values.

We also need to define the CCA function that computes the left and right singular vectors for a pair of matrices:

Definition 1 (CCA)**.** *Compute the CCA between two matrices* $\mathbf{X_1}$ *and* $\mathbf{X_2}$*. Let* $\mathbf{\Phi_{X_1}}$ *be a matrix containing the d largest singular vectors for* $\mathbf{X_1}$ *(sorted from the largest on down). Likewise for* $\mathbf{\Phi_{X_2}}$*. Define the function* $CCA_d(\mathbf{X_1}, \mathbf{X_2}) = [\mathbf{\Phi_{X_1}}, \mathbf{\Phi_{X_2}}]$*. When we want just one of these* $\mathbf{\Phi}$*'s, we will use* $CCA_d(\mathbf{X_1}, \mathbf{X_2})_{left} = \mathbf{\Phi_{X_1}}$ *for the left singular vectors and* $CCA_d(\mathbf{X_1}, \mathbf{X_2})_{right} = \mathbf{\Phi_{X_2}}$ *for the right singular vectors.*

Note that the resulting singular vectors, $[\Phi_{X_1}, \Phi_{X_2}]$ can be used to give two redundant estimates, $X_1\Phi_{X_1}$ and $X_2\Phi_{X_2}$ of the "hidden" state relating X_1 and X_2, if such a hidden state exists.

Definition 2. *Define the symbol "≈" to mean*

$$\mathbf{X_1} \approx \mathbf{X_2} \iff \lim_{n\to\infty} \mathbf{X_1} = \lim_{n\to\infty} \mathbf{X_2}$$

where n is the sample size.

Lemma 2. *Define A by the following limit of the right singular vectors:*

$$CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{right} \approx \mathbf{A}.$$

Under assumptions 2, 3 and 1A, such that if $CCA_k(\mathbf{L}, \mathbf{R}) \equiv [\mathbf{\Phi}_L, \mathbf{\Phi}_R]$ then we have

$$CCA_k([\mathbf{L\Phi}_L, \mathbf{R\Phi}_R], \mathbf{W})_{right} \approx \mathbf{A}.$$

This lemma shows that instead of finding the CCA between the full context and the words, we can take the CCA between the Left and Right contexts, estimate a k dimensional state from them, and take the CCA of that state with the words and get the same result.

Proof:

By assumption 1A, we see that:

$$\mathbb{E}(\mathbf{L}\boldsymbol{\beta}_{\mathbf{L}}|\mathbf{H}) = \mathbf{H}\boldsymbol{\beta}_L^T\boldsymbol{\beta}_L$$

and

$$\mathbb{E}(\mathbf{R}\boldsymbol{\beta}_{\mathbf{R}}|\mathbf{H}) = \mathbf{H}\boldsymbol{\beta}_R^T\boldsymbol{\beta}_R$$

Since, again by assumption 1A both of the $\boldsymbol{\beta}$ matrixes have full rank, $\boldsymbol{\beta}_L^T\boldsymbol{\beta}_L$ is a $k \times k$ matrix of rank k, and likewise for $\boldsymbol{\beta}_R^T\boldsymbol{\beta}_R$. So

$$\mathbb{E}(\boldsymbol{\beta}_R^T\mathbf{R}^T\mathbf{L}\boldsymbol{\beta}_L|\mathbf{H}) = \boldsymbol{\beta}_R^T\boldsymbol{\beta}_R\mathbf{H}^T\mathbf{H}\boldsymbol{\beta}_L\boldsymbol{\beta}_L^T$$

So,

$$\boldsymbol{\beta}_R^T \mathbb{E}(\mathbf{R}^T \mathbf{L}) \boldsymbol{\beta}_L = \boldsymbol{\beta}_R^T \boldsymbol{\beta}_R \mathbb{E}(\mathbf{H}^T \mathbf{H}) \boldsymbol{\beta}_L \boldsymbol{\beta}_L^T$$

since $\boldsymbol{\beta}_R^T \boldsymbol{\beta}_R$, $\mathbb{E}(\mathbf{H}^T \mathbf{H})$ and $\boldsymbol{\beta}_L^T \boldsymbol{\beta}_L$ are all $k \times k$ full rank matrices, $\boldsymbol{\beta}_R$ and $\boldsymbol{\beta}_L$ span the same subspace as the singular values of the CCA between $\mathbf{L}$ and $\mathbf{R}$ since by assumption 2 they have rank k also. Similar arguments hold when relating $\mathbf{L}$ with $\mathbf{W}$ and when relating $\mathbf{R}$ with $\mathbf{W}$. Thus if $CCA_k(\mathbf{L}, \mathbf{R}), \mathbf{W}) \equiv [\Phi_L, \Phi_R]$,

$$CCA_k(\mathbf{L}\Phi_L, \mathbf{R}\Phi_R)_{\text{right}} \approx CCA_k([\mathbf{L}\boldsymbol{\beta}_L, \mathbf{R}\boldsymbol{\beta}_R], \mathbf{W})_{\text{right}}$$

(where we have used assumption 3 to ensure that not only are the subspaces the same, but that the actual singular vectors are the same.)

Finally by 3 we know that the rank of $CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\text{right}}$ is k we see that

$$CCA_k([\mathbf{L}\boldsymbol{\beta}_L, \mathbf{R}\boldsymbol{\beta}_R], \mathbf{W})_{\text{right}} \approx CCA_k([\mathbf{L}, \mathbf{R}], \mathbf{W})_{\text{right}}.$$

Calling this common limit $\mathbf{A}$ yields our result.

q.e.d.

Let $\tilde{\mathbf{A}}_h$ denote a matrix formed by stacking h copies of $\mathbf{A}$ on top of each other. Right multiplying $\mathbf{L}$ or $\mathbf{R}$ by $\tilde{\mathbf{A}}_h$ projects each of the words in that context into the k-dimensional reduced rank space.

The following theorem addresses the core of a new LR-MVL(II) algorithm, showing that there is an $\mathbf{A}$ which gives the desired dimensionality reduction.

Theorem 4. *Under assumptions 1, 2 and 3 there exists a unique matrix A such that if $CCA_k(\mathbf{L}\tilde{\mathbf{A}}_{\mathbf{h}}, \mathbf{R}\tilde{\mathbf{A}}_{\mathbf{h}}) \equiv [\tilde{\boldsymbol{\Phi}}_L, \tilde{\boldsymbol{\Phi}}_R]$ then*

$$CCA_k([\mathbf{L}\tilde{\mathbf{A}}_{\mathbf{h}}\tilde{\boldsymbol{\Phi}}_{\mathbf{L}}, \mathbf{R}\tilde{\mathbf{A}}_{\mathbf{h}}\tilde{\boldsymbol{\Phi}}_{\mathbf{R}}], \mathbf{W})_{right} \approx \mathbf{A}$$

where $\tilde{\mathbf{A}}_h$ *is the stacked form of* $\mathbf{A}$.

Proof: We start by noting that assumption 1 implies assumption 1A. Thus, the previous lemma follows. So we know

$$CCA_k([\mathbf{L},\mathbf{R}],\mathbf{W})_{\text{right}} \approx CCA_k([\mathbf{L}\Phi_L,\mathbf{R}\Phi_R],\mathbf{W})_{\text{right}}$$

where, as usual, $CCA_k(\mathbf{L},\mathbf{R}) \equiv [\Phi_L,\Phi_R]$, which allows us to define $\mathbf{A}$. This $\mathbf{A}$ has the property that the rank of $CCA(\mathbf{WA},\mathbf{H})_{\text{left}}$ is the same as $CCA(\mathbf{W},\mathbf{H})_{\text{left}}$ where $\mathbf{H}$ is the hidden state process associated with our data. Hence anything which is not in the domain of $\mathbf{A}$ won't have any correlation with $\mathbf{H}$ and hence no correlation with other observed states. So $\mathbf{L}$ and $\mathbf{L}\tilde{\mathbf{A}}_h$ have the same "information." More precisely,

$$[\tilde{\mathbf{A}}_h\tilde{\Phi_L},\tilde{\mathbf{A}}_h\tilde{\Phi_R}] \approx CCA_k(\mathbf{L},\mathbf{R})$$

where $CCA_k(\mathbf{L}\tilde{\mathbf{A}}_h,\mathbf{R}\tilde{\mathbf{A}}_h) \equiv [\tilde{\Phi_L},\tilde{\Phi_R}]$ Putting this together with our first equation shows our desired result.

q.e.d.

A.2 p-Eigen

A.2.1 List of 80 AAL (Cortical) ROIs used

```
1,Precentral_R

2,Precentral_L

3,Frontal_Sup_R

4,Frontal_Sup_L

5,Frontal_Sup_Orb_R

6,Frontal_Sup_Orb_L

7,Frontal_Mid_R
```

```
8,Frontal_Mid_L
9,Frontal_Mid_Orb_R
10,Frontal_Mid_Orb_L
11,Frontal_Inf_Oper_R
12,Frontal_Inf_Oper_L
13,Frontal_Inf_Tri_R
14,Frontal_Inf_Tri_L
15,Frontal_Inf_Orb_R
16,Frontal_Inf_Orb_L
17,Rolandic_Oper_R
18,Rolandic_Oper_L
19,Supp_Motor_Area_R
20,Supp_Motor_Area_L
21,Olfactory_R
22,Olfactory_L
23,Frontal_Sup_Medial_R
24,Frontal_Sup_Medial_L
25,Frontal_Med_Orb_R
26,Frontal_Med_Orb_L
27,Rectus_R
28,Rectus_L
29,Insula_R
30,Insula_L
31,Cingulum_Ant_R
32,Cingulum_Ant_L
33,Cingulum_Mid_R
34,Cingulum_Mid_L
```

```
35,Cingulum_Post_R

36,Cingulum_Post_L

37,Hippocampus_R

38,Hippocampus_L

39,ParaHippocampal_R

40,ParaHippocampal_L

41,Amygdala_R

42,Amygdala_L

43,Calcarine_R

44,Calcarine_L

45,Cuneus_R

46,Cuneus_L

47,Lingual_R

48,Lingual_L

49,Occipital_Sup_R

50,Occipital_Sup_L

51,Occipital_Mid_R

52,Occipital_Mid_L

53,Occipital_Inf_R

54,Occipital_Inf_L

55,Fusiform_R

56,Fusiform_L

57,Postcentral_R

58,Postcentral_L

59,Parietal_Sup_R

60,Parietal_Sup_L

61,Parietal_Inf_R
```

```
62,Parietal_Inf_L

63,SupraMarginal_R

64,SupraMarginal_L

65,Angular_R

66,Angular_L

67,Precuneus_R

68,Precuneus_L

69,Heschl_R

70,Heschl_L

71,Temporal_Sup_R

72,Temporal_Sup_L

73,Temporal_Pole_Sup_R

74,Temporal_Pole_Sup_L

75,Temporal_Mid_R

76,Temporal_Mid_L

77,Temporal_Pole_Mid_R

78,Temporal_Pole_Mid_L

79,Temporal_Inf_R

80,Temporal_Inf_L
```

Bibliography

Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 607–615. Springer, 2013.

AM Aertsen, GL Gerstein, MK Habib, and G. Palm. Dynamics of neuronal firing correlation: modulation of “effective connectivity”. *Journal of Neurophysiology*, 61(5):900–917, 1989.

S. Afonso, E. Bick, R. Haber, and D. Santos. Floresta sinta(c)tica: a treebank for portuguese. In *In Proc. LREC*, pages 1698–1703, 2002.

Federica Agosta, Michela Pievani, Cristina Geroldi, Massimiliano Copetti, Giovanni B Frisoni, and Massimo Filippi. Resting state fmri in alzheimer’s disease: beyond the default mode network. *Neurobiology of aging*, 33(8):1564–1578, 2012.

R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ANTS). *Insight J*, 2009.

Brian B Avants, Philip A Cook, Lyle Ungar, James C Gee, and Murray Grossman. Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage*, 50(3):1004–1016, Apr 2010. doi: 10.1016/j.neuroimage.2010.01.041. URL `http://dx.doi.org/10.1016/j.neuroimage.2010.01.041`.

F.R. Bach and M.I. Jordan. A probabilistic interpretation of canonical correlation analysis. In *TR 688, University of California, Berkeley*, 2005.

Feng Bai, David R Watson, Hui Yu, Yongmei Shi, Yonggui Yuan, and Zhijun Zhang. Abnormal resting-state functional connectivity of posterior cingulate cortex in amnestic type mild cognitive impairment. *Brain research*, 1302:167–174, 2009.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. Tailoring continuous word representations for dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1): 1–39, 2011.

C.F. Beckmann, M. DeLuca, J.T. Devlin, and S.M. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1001–1013, 2005.

Christian F Beckmann and Stephen M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *Medical Imaging, IEEE Transactions on*, 23(2):137–152, 2004.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T. Liu. A component based noise correction method (CompCor) for bold and perfusion based fmri. *Neuroimage*, 37(1):90–101, Aug 2007. doi: 10.1016/j.neuroimage.2007.04.042. URL `http://dx.doi.org/10.1016/j.neuroimage.2007.04.042`.

Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

Steven Bird and Edward Loper. Nltk: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

B.B. Biswal, J.V. Kylen, and J.S. Hyde. Simultaneous assessment of flow and bold signals in resting-state functional connectivity maps. *NMR in Biomedicine*, 10 (45):165–170, 1997.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98*, pages 92–100, 1998.

Thomas Blumensath, Saad Jbabdi, Matthew F Glasser, David C Van Essen, Kamil Ugurbil, Timothy EJ Behrens, and Stephen M Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. *NeuroImage*, 2013.

Kristian Bredies and Dirk A Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5-6):813–837, 2008.

P. Brown, P.. deSouza, R. Mercer, V. Della Pietra, and J. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18:467–479, December 1992a. ISSN 0891-2017.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992b. URL `http://acl.ldc.upenn.edu/J/J92/J92-4003.pdf`.

E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell*, Dec 2010. doi: 10.1109/TPAMI.2010.231. URL `http://dx.doi.org/10.1109/TPAMI.2010.231`.

VD Calhoun, T. Adali, GD Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

Vladimir Cherkassky and Yunqian Ma. Another look at statistical learning theory and regularization. *Neural Netw*, 22(7):958–969, Sep 2009. doi: 10.1016/j.neunet.2009.04.005. URL `http://dx.doi.org/10.1016/j.neunet.2009.04.005`.

Shay B Cohen, Karl Stratos, Michael Collins, Dean P Foster, and Lyle Ungar. Spectral learning of latent-variable pcfgs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 223–231. Association for Computational Linguistics, 2012.

R. Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. ICML '08, pages 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

D. Cordes, V.M. Haughton, K. Arfanakis, G.J. Wendt, P.A. Turski, C.H. Moritz, M.A. Quigley, and M.E. Meyerand. Mapping functionally related regions of brain with functional connectivity mr imaging. *American Journal of Neuroradiology*, 21(9):1636–1644, 2000.

D. Cordes, V. Haughton, J.D. Carew, K. Arfanakis, and K. Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magnetic resonance imaging*, 20(4):305–317, 2002.

JS Damoiseaux, S. Rombouts, F. Barkhof, P. Scheltens, CJ Stam, S.M. Smith, and CF Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3): 434–448, 2007.

Fani Deligianni, Gael Varoquaux, Bertrand Thirion, Emma Robinson, David J Sharp, A David Edwards, and Daniel Rueckert. A probabilistic framework to infer brain functional connectivity from anatomical connections. In *Information Processing in Medical Imaging*, pages 296–307. Springer, 2011.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Multi-view learning of word embeddings via cca. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, 2011.

Paramveer S. Dhillon, Jordan Rodu, Michael Collins, Dean P. Foster, and Lyle H. Ungar. Spectral dependency parsing with latent variables. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL'12, 2012a.

Paramveer S. Dhillon, Jordan Rodu, Dean P. Foster, and Lyle H. Ungar. Two step cca: A new spectral method for estimating vector models of words. In *Proceedings of the 29th International Conference on Machine learning*, ICML'12, 2012b.

Paramveer S Dhillon, David A Wolk, Sandhitsu R Das, Lyle H Ungar, James C Gee, and Brian B Avants. Subject-specific functional parcellation via prior based eigenanatomy. *NeuroImage*, 2014.

Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Spectral word embeddings. In *Journal of Machine Learning Research (JMLR)*, 2014 (Under Review)a.

Paramveer S. Dhillon, Lyle H. Ungar, James C. Gee, and Brian B. Avants. Eigenanatomy: Sparse component analysis for medical imaging. In *NeuroImage*, 2014 (Under Review)b.

S. Dumais, G. Furnas, T. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *SIGCHI Conference on human factors in computing systems*, pages 281–285. ACM, 1988.

C. Eckart and D. Young. The approximation of one matrix by another of low rank. *Psychometrika*, page 211, 1936.

Katrin Erk and Diana McCarthy. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, 2009.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001.

Dean P Foster, Sham M Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. Technical report, Technical Report TR-2008-4, TTI-Chicago, 2008.

M.D. Fox, A.Z. Snyder, J.L. Vincent, M. Corbetta, D.C. Van Essen, and M.E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9673–9678, 2005.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, Jul 2008a. doi: 10.1093/biostatistics/kxm045. URL `http://dx.doi.org/10.1093/biostatistics/kxm045`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

K.J. Friston. The disconnection hypothesis. *Schizophrenia research*, 30(2):115–125, 1998.

Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. Interpretable semantic vectors from a joint model of brain-and text-based meaning. 2014.

Harry R. Glahn. Canonical Correlation and Its Relationship to Discriminant Analysis and Multiple Regression. *Journal of the Atmospheric Sciences*, 25(1):23–31, January 1968.

Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 2013.

M.D. Greicius, G. Srivastava, A.L. Reiss, and V. Menon. Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4637–4642, 2004.

Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans Image Process*, 20(7):2030–2048, Jul 2011. doi: 10.1109/TIP.2011.2105496. URL `http://dx.doi.org/10.1109/TIP.2011.2105496`.

Yue Guan and Jennifer Dy. Sparse probabilistic principal component analysis. In *Proceedings of AISTATS*, volume 5, pages 185–192, 2009.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev*, 2011.

D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, 2004.

David Hardoon and John Shawe-Taylor. Sparse cca for bilingual word generation. In *EURO Mini Conference, Continuous Optimization and Knowledge-Based Technologies*, 2008.

Y. He and A. Evans. Graph theoretical modeling of brain connectivity. *Current opinion in neurology*, 23(4):341–350, 2010.

Trey Hedden, Koene RA Van Dijk, J Alex Becker, Angel Mehta, Reisa A Sperling, Keith A Johnson, and Randy L Buckner. Disruption of functional connectivity in clinically normal older adults harboring amyloid burden. *The Journal of neuroscience*, 29(40):12686–12694, 2009.

Kenji Hosoda, Masataka Watanabe, Heiko Wersing, Edgar Krner, Hiroshi Tsujino, Hiroshi Tamura, and Ichiro Fujita. A model for learning topographically organized parts-based representations of objects in visual cortex: topographic nonnegative matrix factorization. *Neural Comput*, 21(9):2605–2633, Sep 2009. doi: 10.1162/neco.2009.03-08-722. URL `http://dx.doi.org/10.1162/neco.2009.03-08-722`.

H. Hotelling. Canonical correlation analysis (cca). *Journal of Educational Psychology*, 1935.

Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002.

Patrik O. Hoyer and Peter Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *COLT*, 2009.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

F. Huang and A. Yates. Distributional representations for handling sparsity in supervised sequence-labeling. ACL '09, pages 495–503, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9.

Fei Huang, Arun Ahuja, Doug Downey, Yi Yang, Yuhong Guo, and Alexander Yates. Learning representations for weakly supervised natural language processing tasks. *Computational Linguistics*, 2013.

A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. 10(3):626 –634, May 1999. ISSN 1045-9227. doi: 10.1109/72.761722.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012.

Ian Jolliffe. *Principal component analysis.* Wiley Online Library, 2005.

I.T. Jolliffe and M. Uddin. The simplified component technique: An alternative to rotated principal components. *Journal of Computational and Graphical Statistics*, 9(4):689–710, 2000.

Dan Jurafsky and James H Martin. *Speech & Language Processing.* Pearson Education India, 2000.

S M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2007.

Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171–171, 2011.

Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In *Proc. ACL*, 2008.

Matthias T. Kromann. The danish dependency treebank and the underlying linguistic theory. in second workshop on treebanks and linguistic theories (tlt). In *In Proc. LREC*, pages 217–220, 2003.

Michael Lamar, Yariv Maron, Mark Johnson, and Elie Bienenstock. Svd and clustering for unsupervised pos tagging. ACL Short '10, pages 215–219, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

TK Landauer, PW Foltz, and D Laham. An introduction to latent semantic analysis. In *Discourse processes*, 2008.

Quoc V. Le, Alex, re Karpenko, Jiquan Ngiam, and Andrew Y. Ng. ICA with reconstruction cost for efficient overcomplete feature learning. pages 1017–1025, 2011. URL `http://dblp.uni-trier.de/rec/bibtex/conf/nips/LeKNN11`.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999a. doi: 10.1038/44565. URL `http://dx.doi.org/10.1038/44565`.

D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999b. ISSN 0028-0836. doi: 10.1038/44565. URL `http://www.ncbi.nlm.nih.gov/pubmed/10548103`. PMID: 10548103.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

Mihee Lee, Haipeng Shen, Jianhua Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, Dec 2010. doi: 10.1111/j.1541-0420.2010.01392.x. URL `http://dx.doi.org/10.1111/j.1541-0420.2010.01392.x`.

Y. Liu, M. Liang, Y. Zhou, Y. He, Y. Hao, M. Song, C. Yu, H. Liu, Z. Liu, and T. Jiang. Disrupted small-world networks in schizophrenia. *Brain*, 131(4):945–961, 2008.

M.J. Lowe, E.B. Beall, K.E. Sakaie, K.A. Koenig, L. Stone, R.A. Marrie, and M.D. Phillips. Resting state sensorimotor functional connectivity in multiple sclerosis inversely correlates with transcallosal motor pathway transverse diffusivity. *Human brain mapping*, 29(7):818–827, 2008.

Lester Mackey. Deflation methods for sparse pca. *Neural Information Processing Systems (NIPS08)*, pages 1–8, 12/2008 2008.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19: 313–330, June 1993. ISSN 0891-2017.

Andrea Mechelli, Cathy J. Price, Karl J. Friston, and John Ashburner. Voxel-based morphometry of the human brain: Methods and applications. *Current Medical Imaging Reviews*, 1:105–113, 2005.

Bernard Merialdo. Tagging english text with a probabilistic model. *Comput. Linguist.*, 20, jun 1994.

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, Christine Keribin, and Bertrand Thirion. A supervised clustering approach for¡ i¿ fmri¡/i¿-based inference of brain states. *Pattern Recognition*, 45(6):2041–2049, 2012.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. ICML '07, pages 641–648, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: http://doi.acm.org/10.1145/1273496.1273577. URL `http://doi.acm.org/10.1145/1273496.1273577`.

B. Mohammadi, K. Kollewe, A. Samii, K. Krampfl, R. Dengler, T.F. Münte, et al. Changes of resting state brain networks in amyotrophic lateral sclerosis. *Experimental neurology*, 217(1):147, 2009.

JC Morris, A Heyman, RC Mohs, JP Hughes, et al. The consortium to establish a registry for alzheimer's disease (cerad): I. clinical and neuropsychological assessment of alzheimer's disease. *Neurology*, 1989.

Kevin P Murphy. *Machine learning: A Probabilistic Perspective.* MIT Press, Cambridge, MA, 2012.

Bernard Ng, Rafeef Abugharbieh, and Martin J McKeown. Discovering sparse functional brain networks using group replicator dynamics (grd). In *Information Processing in Medical Imaging*, pages 76–87. Springer, 2009a.

Bernard Ng, Rafeef Abugharbieh, and Martin J McKeown. Functional segmentation of fmri data using adaptive non-negative sparse pca (anspca). In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, pages 490–497. Springer, 2009b.

Alfonso Nieto-Castanon, Satrajit S. Ghosh, Jason A. Tourville, and Frank H. Guenther. Region of interest based analysis of functional imaging data. *Neuroimage*, 19(4):1303–1316, Aug 2003.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

Ankur P Parikh, Shay B Cohen, and Eric P Xing. Spectral unsupervised parsing with additive tree metrics. 2014.

F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *31st Annual Meeting of the ACL*, pages 183–190, 1993a.

Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993b.

Ronald C Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004.

JR Petrella, FC Sheldon, SE Prince, VD Calhoun, and PM Doraiswamy. Default mode network connectivity in stable vs progressive mild cognitive impairment. *Neurology*, 76(6):511–517, 2011.

Russell A. Poldrack. Region of interest analysis for fMRI. *Social cognitive and affective neuroscience*, 2(1):67–70, March 2007. ISSN 1749-5024. URL `http://dx.doi.org/10.1093/scan/nsm006`.

Jonathan D Power, Kelly A Barnes, Abraham Z Snyder, Bradley L Schlaggar, and Steven E Petersen. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. *Neuroimage*, 59(3):2142–2154, 2012.

L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CONLL*, pages 147–155, 2009.

Tony Rose, Mark Stevenson, and Miles Whitehead. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. In *LREC*, volume 2, pages 827–832, 2002.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values, 2010. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:1003.2990`.

R. Salvador, J. Suckling, M.R. Coleman, J.D. Pickard, D. Menon, and ED Bullmore. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral Cortex*, 15(9):1332–1342, 2005.

W.W. Seeley, R.K. Crawford, J. Zhou, B.L. Miller, and M.D. Greicius. Neurodegenerative diseases target large-scale human brain networks. *Neuron*, 62(1):42, 2009.

Martin Seligman. *Flourish: A Visionary New Understanding of Happiness and Well-being.* Free Press, 2011.

David W. Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L. Narr, Russell A. Poldrack, Robert M. Bilder, and Arthur W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064–1080, February 2008. ISSN 1053-8119.

Yvette I Sheline, John C Morris, Abraham Z Snyder, Joseph L Price, Zhizi Yan, Gina D'Angelo, Collin Liu, Sachin Dixit, Tammie Benzinger, Anne Fagan, et al. Apoe4 allele disrupts resting state fmri connectivity in the absence of amyloid plaques or decreased csf aβ42. *The Journal of Neuroscience*, 30(50):17035–17040, 2010.

Haipeng Shen and Jianhua Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *J Multivar Anal*, 99(6):1015–1034, July 2008a.

Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6): 1015–1034, 2008b.

S. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank hidden Markov models. In *AISTATS-2010*, 2010.

M. Sill, S. Kaiser, A. Benner, and A. Kopp-Schneider. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics*, 27 (15):2089–2097, 2011a.

Martin Sill, Sebastian Kaiser, Axel Benner, and Annette Kopp-Schneider. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics*, 27(15):2089–2097, Aug 2011b. doi: 10.1093/bioinformatics/btr322. URL `http://dx.doi.org/10.1093/bioinformatics/btr322`.

K. Simov, P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, E. Simov, and M. Kouylekov. Building a linguistically interpreted corpus of bulgarian: the bultreebank. In *In Proc. LREC*, 2002.

Noah A. Smith and Jason Eisner. Contrastive estimation: training log-linear models on unlabeled data. ACL '05, pages 354–362. Association for Computational Linguistics, 2005.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. Citeseer, 2013.

C.J. Stam, J.C. Reijneveld, et al. Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomed Phys*, 1(3), 2007.

G. W. Stewart. Perturbation theory for the singular value decomposition. In *IN SVD AND SIGNAL PROCESSING, II: ALGORITHMS, ANALYSIS AND APPLICATIONS*, pages 99–109. Elsevier, 1990.

G.W. Stewart and Jiguang Sun. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, 1990. ISBN 9780126702309. URL `http://books.google.com/books?id=l78PAQAAMAAJ`.

J. Suzuki and H. Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *In ACL*, 2008.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487. Association for Computational Linguistics, 2012.

Simone Teufel. *The structure of scientific articles*. CSLI Publications, 2010.

Bertrand Thirion, Guillaume Flandin, Philippe Pinel, Alexis Roche, Philippe Ciuciu, and Jean-Baptiste Poline. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fmri datasets. *Human brain mapping*, 27 (8):678–693, 2006.

John W Tukey. Exploratory data analysis. *Reading, MA*, 231, 1977.

J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. ACL '10, pages 384–394, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1858681.1858721`.

P.D. Turney and P. Pantel. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

Nicholas J Tustison, Philip A Cook, Arno Klein, Gang Song, Sandhitsu R Das, Jeffrey T Duda, Benjamin M Kandel, Niels van Strien, James R Stone, James C Gee, and Brian B. Avants. Large-scale evaluation of ants and freesurfer cortical thickness measurements. *NeuroImage*, 2014.

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

M.P. van den Heuvel and H.E. Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010.

Koene RA Van Dijk, Mert R Sabuncu, and Randy L Buckner. The influence of head motion on intrinsic functional connectivity mri. *Neuroimage*, 59(1):431–438, 2012.

G. Varoquaux, S. Sadaghiani, P. Pinel, A. Kleinschmidt, J. B. Poline, and B. Thirion. A group model for stable multi-subject ica on fmri datasets. *Neuroimage*, 51(1):288–299, May 2010a. doi: 10.1016/j.neuroimage.2010.02.010. URL `http://dx.doi.org/10.1016/j.neuroimage.2010.02.010`.

Gaël Varoquaux, Alexandre Gramfort, Jean Baptiste Poline, and Bertrand Thirion. Brain covariance selection: better individual functional connectivity models using population prior. *arXiv preprint arXiv:1008.5071*, 2010b.

Gael Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information processing in medical imaging*, pages 562–573. Springer, 2011.

S. Whitfield-Gabrieli, H.W. Thermenos, S. Milanovic, M.T. Tsuang, S.V. Faraone, R.W. McCarley, M.E. Shenton, A.I. Green, A. Nieto-Castanon, P. LaViolette, et al. Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proceedings of the National Academy of Sciences*, 106(4):1279–1284, 2009.

Daniela M. Witten and Robert Tibshirani. A framework for feature selection in clustering. *J Am Stat Assoc*, 105(490):713–726, Jun 2010. doi: 10.1198/jasa.2010.tm09415. URL `http://dx.doi.org/10.1198/jasa.2010.tm09415`.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, Jul 2009. URL `http://dx.doi.org/10.1093/biostatistics/kxp008`.

Svante Wold, Paul Geladi, Kim Esbensen, and Jerker Öhman. Multi-way principal components-and pls-analysis. *Journal of chemometrics*, 1(1):41–56, 1987.

Tao Wu and Mark Hallett. The cerebellum in parkinsons disease. *Brain*, 136(3): 696–709, 2013.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1): 187–193, 2012. URL `http://dblp.uni-trier.de/db/journals/pami/pami34.html#XuCM12`.

Allen Y Yang, Shankar S Sastry, Arvind Ganesh, and Yi Ma. Fast l1-minimization algorithms and an application in robust face recognition: A review. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1849–1852. IEEE, 2010.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3): 1125–1165, 2011.

M K Stephen Yeung, Jesper Tegnr, and James J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–6168, Apr 2002. doi: 10.1073/pnas.092576199. URL `http://dx.doi.org/10.1073/pnas.092576199`.

Ron Zass and Amnon Shashua. Nonnegative sparse pca. In *Advances in Neural Information Processing Systems*, pages 1561–1568, 2006.

Tong Zhang and David Johnson. A robust risk minimization based named entity recognition system. CONLL '03, pages 204–207, 2003.

Tuo Zhang, Lei Guo, Kaiming Li, Changfeng Jing, Yan Yin, Dajiang Zhu, Guangbin Cui, Lingjiang Li, and Tianming Liu. Predicting functional cortical rois via dti-derived fiber shape models. *Cerebral Cortex*, 22(4):854–864, 2012.

Luping Zhou, Yaping Wang, Yang Li, Pew-Thian Yap, Dinggang Shen, and ADNI. Hierarchical anatomical brain networks for mci prediction: Revisiting volumetric measures. *PLoS One*, 6(7):e21935, 2011.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Xi-Nian Zuo, Ting Xu, Lili Jiang, Zhi Yang, Xiao-Yan Cao, Yong He, Yu-Feng Zang, F Xavier Castellanos, and Michael P Milham. Toward reliable characterization of functional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space. *Neuroimage*, 65:374–386, 2013.