UNSUPERVISED LEARNING OF DISEASE SUBTYPES FROM CONTINUOUS TIME HIDDEN MARKOV MODELS OF DISEASE PROGRESSION

A Thesis
Presented to
The Academic Faculty

by

Amrita Gupta

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Computational Science and Engineering

Georgia Institute of Technology
December 2015

UNSUPERVISED LEARNING OF DISEASE SUBTYPES FROM CONTINUOUS TIME HIDDEN MARKOV MODELS OF DISEASE PROGRESSION

Approved by:

Professor James Matthew Rehg, Advisor
School of Interactive Computing
Georgia Institute of Technology

Professor Jimeng Sun
School of Computational Science and Engineering
Georgia Institute of Technology

Professor Theodore Hutman
Department of Psychiatry & Biobehavioral Science
University of California Los Angeles

Date Approved: 20 August 2015

To my parents,

Jayanta and Ratna Gupta.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor James Rehg, for his mentorship over the past two years. I also wish to thank my thesis committee members, Professor Ted Hutman and Professor Jimeng Sun, for their insights and suggestions during this study. Sincere thanks are also due to Dr. Agata Rozga for her input and interest in this project. I am also thankful to my collaborators Ted and Agata for sharing the UCLA-UC Davis autism dataset with me, and to Jimeng for granting me access to the ExactData dataset. Thank you to Yu-Ying Liu, for her guidance as I began working on this project and, crucially, for her development of the CT-HMM disease progression model that laid the foundation for my work. Thank you to Stephanie Tofighi for all that she does to ensure Jim's students have access to the resources they need. I am grateful to Nirav Shelat, for his support as I navigated my first few years in graduate school right up to the completion of this document. Lastly, I thank my mum and dad, who have given me the opportunity to pursue my dreams and the encouragement to drive me forward. I stand on the shoulders of giants.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The detection of subtypes of complex diseases has important implications for diagnosis and treatment. Numerous prior studies have used data-driven approaches to identify clusters of similar patients, but it is not yet clear how to best specify what constitutes a clinically meaningful phenotype. This study explored disease subtyping on the basis of temporal development patterns. In particular, we attempted to differentiate infants with autism spectrum disorder into more fine-grained classes with distinctive patterns of early skill development. We modeled the progression of autism explicitly using a continuous-time hidden Markov model. Subsequently, we compared subjects on the basis of their trajectories through the model state space. Two approaches to subtyping were utilized, one based on time-series clustering with a custom distance function and one based on tensor factorization. A web application was also developed to facilitate the visual exploration of our results. Results suggested the presence of 3 developmental subgroups in the ASD outcome group. The two subtyping approaches are contrasted and possible future directions for research are discussed.

CHAPTER I

INTRODUCTION

1.1 Data Analytics for Healthcare

Healthcare providers worldwide are faced with the formidable task of improving efficiency and quality of care while also lowering costs. This is especially needed in the US healthcare system, which suffers from exceptionally high per capita costs and underperforms in comparison to other developed countries on accessibility, equity, efficiency, care quality and health outcomes [10] (Figure 1). To begin addressing this problem, the US Congress passed the HITECH (Health Information Technology for Economic and Clinical Health) Act in 2009 to promote the adoption and "meaningful use" of electronic health records, paving the way for improved health information exchange and data management. Since then, the volume of healthcare data has risen dramatically, already reaching 150 exabytes (150×10^{18} bytes) in 2011 [40]. In order to leverage this vast resource, big data analytics tools are being developed to address several clinical tasks: screening, diagnosis, treatment, prognosis, monitoring and management [14].

Data analytics in healthcare can be implemented using either a batch processing approach or an online approach. The batch processing approach can be applied to study population health management by deriving actionable information from large-scale health data. For instance, it could allow researchers to assess the effect of different genetic or environmental risk factors on disease prevalence, or to monitor the effectiveness of drugs or interventions. At the same time, such insights can be applied at the individual level for precision medicine. By profiling patients to identify individual genetic, environmental and lifestyle characteristics, it may be possible to

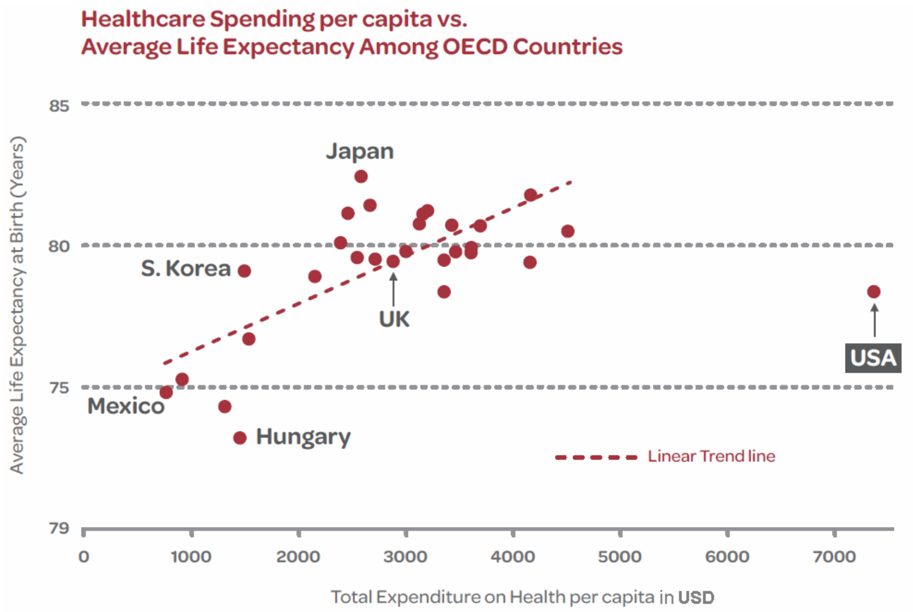

Figure 1: Discrepancy between US healthcare cost and quality of care. Source: [1]

perform more accurate disease risk assessment or to design more targeted treatment protocols. In this way, online data analytics could lead to the development of an evidence-based clinical decision support system.

1.2 Subtyping Complex Diseases

One task that lies at the junction of population- and individual-level health is the detailed study of complex diseases. Complex multi-system diseases are often highly heterogeneous in terms of clinical presentation, time course and outcome. Disease subtyping or phenotyping aims to identify homogeneous subgroups of affected individuals with a common set of clinical features or developmental course. It may be that these disease phenotypes arise from different underlying disease mechanisms or genetic variations, in which case the subgroups of affected individuals could be studied to identify which genes or physiological characteristics influence the disease process. This can give healthcare practioners a more precise way to characterize a specific disease and to identify treatment plans that target an appropriate pathway or component of the disease. At the population health level, subtyping can help identify risk factors for each type of the disease and develop diagnostic tools for early differentiation of the precise disease subtype. A meaningful categorization of complex diseases into subtypes could also enhance fundamental medical science by motivating genome-wide association and metabolomics studies.

1.3 Autism Spectrum Disorders

1.3.1 Overview

One disease that is well known for its heterogeneity is autism. Autism is a childhood neurodevelopmental disorder characterized by early impairment in social and communication skills and restricted, repetitive behaviors that can result in lifelong impairments. A recent report by the Centers for Disease Control and Prevention puts the prevalence of autism spectrum disorders (ASD) in the United States at 1 in every 68 children [23]. There is currently no known cure for ASD, and the long-term prognosis for children with autism is highly variable. Only about 12% of adults with autism achieve a high level of independence. It appears that the prognosis is largely influenced by the severity of the disorder, the extent of early intervention and the effects of co-morbid conditions. There is no known single cause of autism, and given the variation in symptoms and severity in affected individuals, it is very likely that there are multiple complex genetic and environmental factors involved in its etiology. This heterogeneity is also the reason for the classification of autism as a spectrum disorder, encompassing a range of linked conditions with similar symptoms, including Asperger syndrome and pervasive developmental disorder not otherwise specified (PDD-NOS). Currently, the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) published by the American Psychiatric Association treats ASD as one broad disorder with varying indices of severity. These considerations make ASD a prime candidate disease for subtyping efforts.

1.3.2 Subtyping Autism Spectrum Disorders

The clinical heterogeneity of autism complicates etiological investigation, diagnosis and detection, and the development of effective interventions. There is a large body of work dedicated to characterizing and explaining this heterogeneity, which we review briefly here.

Many early studies of autism aimed to discover behaviorally defined subgroups, with a particular focus on patterns of language skill development [2, 15, 36, 43]. Much of this work relied on prospective or retrospective cohort studies of children involving a combination of diagnostic and developmental assessments and the subsequent modeling of the resulting longitudinal data using latent class growth analysis and other related models. These empirical studies have identified anywhere between one and six subgroups on the basis of symptom severity and skill level trajectories.

With the increasing availability of genetic sequencing technology, a number of studies have shifted focus to the genetic underpinnings of the condition. For example, a study by Bruining et al. attempted to connect the Autism Diagnostic Interview-Revised behavioral characterization to genetic disorders known to be linked to an increased risk for autism [3]. Another study attempted to reveal genetically-meaningful phenotypes by clustering vectors constructed from a combination of behavioral, biomarker and genetic data [48].

Finally, a few studies explored the heterogeneity of autism using alternative descriptors of disease progression. Autism symptom severity levels have been linked to differences in diffusion weighted brain imaging data [47], while another study found that autism severity itself evolved over time [17]. Lastly, a very small number of studies have considered the variation in non-neurobehavioral comorbidities of autism, finding distinct subgroups characterized by either seizures, psychiatric disorders or multisystem disorders.

1.4 Goals of this Study

The aim of this study is to perform exploratory analysis on longitudinal healthcare data for the purpose of detecting disease subtypes with distinct temporal progression signatures. Specifically, we aim to investigate whether there are distinguishable patterns of early skill development in children with autism spectrum disorder compared to typical children.

The exploratory analysis is performed using a combination of data mining algorithms and visual analytics. Using a previously developed continuous-time hidden Markov model of disease progression as a basis for characterizing individual trajectories [34], we then compare the use of time-series clustering with tensor decomposition for subgroup identification. To the best of our knowledge, this study is the first instance of applying tensor factorization in autism research.

1.5 Concepts

1.5.1 Phenotype

Classically, a phenotype is a collection of *observable* traits in an individual (e.g. physiological properties, behavior) that result from the expression of the individual's genetic makeup. However, in the context of this thesis, we are concerned with the concept of a *disease phenotype*. Complex diseases typically show a great deal of variation in clinical presentation across affected individuals. Although the exact reasons for this heterogeneity are yet to be fully understood, some contributing factors include the extensive effect of multiple modifier genes, environmental factors, and complex interactions between them [35]. Even monogenic diseases, such as cystic fibrosis, which arise from the influence of a single gene, often display varying levels of penetrance and expressivity in carriers of the gene variant. The ability to differentiate complex diseases into clinically meaningful subgroups remains a much-sought-after goal of the medical community. It should be noted that the notion of what constitutes a clinically meaningful phenotype is flexible, and can encompass within-group similarities in symptoms, severity, temporal progression and co-morbidity occurrence, to name a few. In this study, we focus on similarities in the temporal progression of early skill development. It remains to be seen whether clustering individuals by their developmental trajectories will correspond to other grouping strategies, such as severity of impairment in one or more domains affected by autism spectrum disorder.

1.5.2 Panel Data

Multidimensional time-series data, referred to as panel data or longitudinal data, consist of multiple measurements in several dimensions for a set of individuals or entities over some period of time. This type of data is information-rich, as it captures both individual-specific time-series (obtained by restricting the data to those of one individual) as well as population-wide cross-sectional data (obtained by restricting the data to those taken at one time-point). Clinical data, which may be sourced from electronic health records, longitudinal cohort studies or ICU monitoring, often takes this form, since data are collected from multiple individuals at arbitrary times over the course of their treatment.

The development of disease progression models or subtyping methods using panel data is complicated by certain inherently challenging properties of these data, which include:

- **Discrete Observations:** The development of a disease or the evolution of a patient's state of health is a continuous process, but observations are often made at only discrete, sparse, irregularly-sampled time-points.

- **Irregular Sampling:** The visits or measurements are arbitrarily timed, and individual records may span time periods of different lengths that do not align in any obvious way.

- **Multiple Covariates:** The progression of a disease may be reflected in a multitude of variables, which may be difficult to identify without prior domain knowledge.

- **Missing Data:** Some relevant measurements may be missing at some timepoints or for certain individuals.

- **Population Heterogeneity:** Even within a carefully selected sample population, there is often a great deal of variation in disease trajectory and time-course, which may suggest the presence of disease subtypes, or may be due to other confounding effects.

Existing disease progression modeling approaches often handle these factors through aggregation in the data pre-processing step, but this can result in the masking of temporal effects.

1.5.3 Distance Measures

A distance function describes the distance between two points in a set. A distance measure should be symmetric, i.e. $d(x_1, x_2) = d(x_2, x_1)$, and should have a value of zero between identical points, i.e. $d(x_1, x_2) = 0$ if $x_1 = x_2$. If the distance measure also satisfies the triangle inequality, $d(x_1, x_2) + d(x_2, x_3) \geq d(x_1, x_3)$, it is also a metric. Since clustering aims to group similar elements together, the choice of distance measure between those elements has a profound effect on the clustering. Some commonly used distance measures include:

1. **Minkowski:** $d(x_a, x_b) = \left(\sum_{i=1}^{n} |x_{ai} - x_{bi}|^r\right)^{1/r}$ For $r = 2$, the above yields the familiar Euclidean distance. For $r = 1$, we have the Manhattan distance, and for $r = \infty$ we have the Chebyshev distance.

2. **Edit:** Minimum number of insert and delete operations needed to change one string into another.

3. **Jaccard:** The distance between two sets $d(A, B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|}$

1.5.4 Cluster Validation

The purpose of clustering is to compare items on the basis of some distance measure so that distinct groupings of the items become apparent. Two components of this "distinctness" are *compactness*, whereby items in each cluster are close to one another under the distance measure; and *separation*, by which we mean that the different clusters are far away from one another. There are different ways to define the distance between clusters of elements, e.g. distance between cluster centers, or between the closest elements of the clusters.

There are three families of methods for validating the output from clustering algorithms:

1.5.4.1 External Validation Measures

External criteria are applied when the user has access to some information that was not used to derive the cluster assignments, such as class labels. These are useful in cases where cluster analysis is used to assess whether a particular distance measure is effective for stratifying different categories of elements.

1.5.4.2 Internal Validation Measures

Internal validation criteria rely on inherent features of the data to quantify compactness and separation. These criteria are more applicable when the user is conducting cluster analysis for exploratory purposes.

1.5.4.3 Relative Validation Measures

These measures compare different clustering measures with respect to one another, and is especially useful for comparing the effect of different input parameters to a clustering algorithm.

1.6 Importance and Scope of this Study

Computational models for disease subtyping have the potential to improve healthcare and clinical research. At the same time, the demand for data-driven descriptive and predictive tools in healthcare can drive the development of novel computational methodologies. Visual analytics can foster a collaboration between healthcare practice and computational research, whereby domain experts can better interpret and interact with computational models and provide feedback for further development. The visualization scheme presented in this work was developed for precisely that purpose, and it is hoped that it enables users, readers and analysts with another means to study multidimensional disease progression models.

Additionally, this work was conducted to gain a deeper appreciation for what longitudinal disease subtypes might look like. It is vital to carefully assess how different approaches to clustering either rely on being given an appropriate metric or implicitly learn their own metrics on the input space. Especially in unsupervised settings, where there is no "correct" answer, it is worthwhile to look beyond results and compare different methods based on their assumptions about the structure of the underlying process.

CHAPTER II

LITERATURE REVIEW

In view of the increasing volume and availability of digitized healthcare data, there have been numerous efforts to leverage this resource to produce actionable insights for the improvement of care. The knowledge discovery process consists of several steps, including data storage, cleaning and access (collectively referred to as *data warehousing*), feature selection, data mining, evaluation and visualization. Data mining has received the most attention from researchers due to its potential to automatically discover patterns of interest that can provide descriptive and predictive support to data analysis. More recently, visualization has also come to light as a valuable tool for finding, exploring, validating and communicating patterns and information. In this chapter, we survey some recent work in data mining and visual analytics with a specific focus on the medical task of disease subtyping.

2.1 Data Mining for Disease Subtyping

Numerous recent studies on the clinical heterogeneity of well-known diseases–such as asthma [51], heart failure [11] and autism–have promoted the hypothesis that many such diagnoses are not distinct diseases, but rather they are umbrella terms for multiple phenotypes with a common symptomatology. The presence of consistent groupings of clinical characteristics could point to different etiologies, and accordingly the resolution of complex diseases into subtypes has been an active area of research. Traditionally, disease subtyping efforts have been led by medical experts seeking to ask specific questions driven by their observations during clinical practice. While this direction has obvious benefits, such as expert-driven feature selection and outcome validation, it is poorly suited to handle the vast amounts of medical data at our disposal and can preclude the discovery of previously unknown medical concepts. The expert-driven approach may also be more vulnerable to bias towards conventional conceptions and methods. As such, many modern approaches to disease subtyping use *unsupervised learning* techniques from machine learning and computational statistics, which try to infer representations of the input without supervised target outputs or rewards. Most of the studies discussed in this section belong to this category.

2.1.1 Clustering

2.1.1.1 Partitional Clustering

In the search for subgroups of similar patients within a larger population, clustering is a natural choice of methodology, where even relatively simplistic clustering techniques have proved useful. A 2005 study by Lewis et al. explored heterogeneity in early-stage Parkinson's Disease (PD) by applying k-means clustering to raw data [33]. The data were collected at a single time point from a relatively small patient cohort (by current "big-data" standards), and comprised of carefully selected features to encompass a wide range of features associated with PD heterogeneity. Despite the simplicity of the method and the data, the clustering revealed four clinical phenotypes that were in good agreement with clinical knowledge. The authors attempted to validate their approach by tracking the assignment of patients to phenotypes while varying the number of clusters sought. They also found by inspection that the clusters varied in the rate of PD progression based on a computed "disease progression score". This work was one of the earliest efforts at data-driven disease subtyping and highlights the potential of clustering for this purpose, provided the input features are well chosen.

2.1.1.2 Hierarchical Clustering

In contrast to partitional clustering methods, one advantage of hierarchical clustering is that one does not need to provide the algorithm with the optimal number of clusters. A recent study by Yang et al. used unsupervised hierarchical clustering to search for clusters of non-motor symptoms in patients with early Parkinson's Disease [52]. The resulting clusters were validated by performing multiscale bootstrap resampling. They obtained three overall groups of patients with distinct patterns of symptoms that may be useful for diagnosing pre-motor Parkinson's Disease.

2.1.1.3 Longitudinal Data

Longitudinal data is information rich, but can be challenging to work with in its unprocessed form due to unequal observation periods, arbitrarily spaced observations, different variables observed at each time point, etc. Therefore, clustering using longitudinal clinical data calls for a careful choice about how to compare non-uniform time-series data. One common strategy for handling these issues is to aggregate the longitudinal data for each patient into a vector. This was the strategy adopted by Chen et al. who tried to subtype Down's syndrome, Crohn's disease and cystic fibrosis by severity [7]. They averaged the laboratory test values across all time points for each patient in their study to construct uniform disease-specific matrices of "clinarrays", and subsequently applied Pearson's correlation coefficient as a distance metric for hierarchical clustering. A study by Doshi-Velez et al. converted ICD-9 codes from the electronic health records of subjects with ASD into aggregate count vectors over 6-month blocks [12]. They subsequently applied hierarchical clustering and were able to distinguish 3 main patterns of ASD comorbidity trajectories.

2.1.1.4 Probabilistic Clustering

More recently, the problem of patient subgroup identification has been approached with more sophisticated clustering techniques based on probabilistic models rather than raw data or feature vectors constructed from raw data. Here, we assume a model for each cluster and find the models that best fit the data; subsequently, clustering is performed using model parameters or coefficients. Marlin et al. constructed a piecewise aggregated data matrix from the first 24 hours of physiological variable measurements from a pediatric intensive care unit, and subsequently developed a probabilistic clustering scheme based on a Gaussian mixture model [37].

One benefit of employing model-based clustering methods for time series is that issues like missing data, unequal observation intervals and uneven observation timing can be handled by choosing an appropriate model; this removes the need for aggregating or standardizing the data in a way that may alter the temporal information. Another probabilistic clustering model developed by Schulam et al. also adopts the mixture model approach, but with each subtype's temporal trajectory fit to B-splines [42].

2.1.2 Matrix and Tensor Factorization

In the past few years, there have been a few studies using matrix or tensor decomposition to reveal substructure in patient population data that could correspond to disease phenotypes. Such dimensionality reduction techniques have been of particular interest to the bioinformatics community, where principal component analysis, singular value decomposition, non-negative matrix factorization and independent component analysis have been used to cluster gene expression data to find endophenotypes [49]. Sun et al. developed a multi-view singular value decomposition to incorporate both genetic and clinical data [44]; by treating them as different views, the algorithm seeks clusters that have both marked clinical features and genetic markers.

Dimensionality reduction methods have also been applied to search for purely clinical phenotypes. Ho et al. developed a non-negative tensor factorization method to obtain candidate phenotypes from electronic medical record data [22]. The longitudinal data was pre-processed to construct a count tensor with three modes: patient mode, diagnosis mode and medication mode. For each patient, the counts are the number of co-occurrences of medications and diagnoses in visits over a two year observation window prior to some anchor event.

2.1.3 Growth Curve Analysis

There is a large body of work in trajectory modeling using growth curve analysis in the developmental sciences, which encompasses a variety of methods for studying inter-individual differences in intra-individual temporal patterns. These models are generalizations of linear regression designed to describe different modeling situations. Multilevel models (also referred to as hierarchical linear models) organize the independent variables on different levels to distinguish between individual-level and shared effects [46, 16, 53]. These techniques model between-cluster variation explicitly, while another set of methods, generalized estimating equations, estimates within-cluster similarity instead [20, 31]. Other approaches, like growth mixture models and latent class growth analysis, capture population heterogeneity by explicitly allowing for different unobserved subpopulations with different growth parameters [26, 30].

2.1.4 Other Miscellaneous Methods

While clustering, matrix factorization and growth curve analysis are some of the most common strategies for pattern mining with longitudinal clinical data, there have been numerous studies that utilize other machine learning methods for this goal. These include support vector machines, deep learning, temporal abstraction and factor analysis. A few of these are discussed here briefly, but a more complete overview of the applications of these methods can be found elsewhere [24].

2.1.4.1 SVM

In a recent study, Bruining [3] et al. used autism symptom profiles from subjects with genetic conditions linked to autism to investigate whether there are distinct patterns of behavioral symptoms associated with each condition. They trained a multi-class support vector machine to classify the genotype based on behavioral symptom scores, which performed with 67% classification accuracy using the leave-one-out cross validation method. Their analysis also showed that quality of social interactions contributed more to the SVM classifier than repetitive behaviors or communication deficiencies for differentiating each genotype's autism behavioral signature.

2.1.4.2 Deep Learning

Deep learning, which has had great success in unsupervised feature learning for object recognition, was recently applied to the problem of phenotype discovery from electronic medical records [32]. The authors converted raw uric acid measurements from patients with either gout or acute leukemia into continuous longitudinal probability densities which were then used to infer meaningful features. The resulting first-layer and second-layer feature sets were embedded in 2-d and the data for each subject were plotted using the diagnoses as labels. Both feature sets showed significant separation between the two phenotypes and also showed additional cluster structure, suggesting the presence of subtypes for each disease.

2.2 Visual Data Mining for Disease Subtyping

Another answer to the need for tools to interpret and analyze the overwhelming amount of clinical data is visual analytics. Visualization has the capacity to provide cognitive support in a way that can relieve information overload [4] and assist users in the performance of several tasks relevant to healthcare practice: combining information from multiple sources, analyzing several variables or individuals at once, and ultimately making decisions supported by data and clinical evidence.

The visualization community has recognized this opportunity, and there is a growing body of research dedicated to developing novel strategies and interfaces for effectively navigating complex clinical data. Moreover, combining automatic mining-based analysis with visual analysis can be argued to be more powerful than either approach used independently [27]. For example, data mining methods can discover patterns, but often fail to provide any context within which to interpret them, a limitation that can be addressed through visualization. Conversely, visual analysis without any confirmatory data analysis can also be misleading, and can quickly become difficult to interpret as more and more information is added to the display [19].

2.2.1 Visual Cluster Analysis

Another set of visualization tools relevant to disease subtyping are those developed to visualize the clusters themselves for interpretation and evaluation. One such tool, DICON [5], uses the well-known treemap visualization to produce icons for representing multidimensional clusters with the different color-coded features of each entity within the cluster grouped together. As a result, the icon gives a compact overview of the relative values and importance of different features for entities belonging to the cluster. The authors proposed the use of this visualization for clinical decision support by searching for similar patients to a target patient and then evaluating the quality of the resulting cohort [18]. Another recent work used a force-based network visualization to identify subgroups within a primary sample phenotype [50], demonstrating its utility for revealing links between similar patients or co-occurring symptoms that traditional pairwise correlation analyses have missed.

2.2.2 Temporal Progression Visualization

2.2.2.1 Population-level Trends

The visualization of disease progression is of particular value for managing chronic diseases, summarizing patient histories in a clinical setting, and understanding longitudinal cohort study data. MatrixFlow, developed by Perer and Sun, was one such disease progression visualization tool, in which closely-timed clinical events were represented in a pictorial adjacency matrix, highlighting co-occurrences between symptoms, diagnoses and medications [39]. Temporal trends were illustrated through the approach of small multiples, with multiple matrices for different time segments displayed side by side. MatrixFlow proved extremely successful for illustrating broad differences in the evolution of clinical events and symptoms between different cohort populations. However, the identification of intra-cohort subgroups with varying disease progression dynamics would prove challenging with this visualization scheme.

2.2.2.2 Population-derived Temporal Associations

Besides looking for overall trends, users performing visual analytics may be interested in correlations between early clinical observations and later events. This functionality is made available in the Visual Temporal Analysis Laboratory (ViTA-Lab) visual analytics environment, which supports an iterative workflow consisting of data mining and query-driven visual analytics [28]. In particular, the tool features a temporal association chart for visually exploring raw longitudinal data and probabilistic temporal associations between events. The authors showcased their tool by exploring data from a cohort of diabetic patients, starting with an overview of the distribution of the subjects across different albuminuria-level groups from year to year. Next, applying the data mining engine produced patterns linking HbA1C states to albuminuria levels indicative of renal damage; these patterns were further explored via temporal association charts to find the frequency of transitions in the cohort population. Although ViTA-Lab lacks the functionality to track individual trajectories, it is an appropriate tool for discovering temporal trends and links between clinical concepts at the population level.

2.2.2.3 Specific Clinical Event Sequences

Another approach to clinical event sequence pattern mining is to specify the event sequence of interest, use it to query patient data and subsequently search for noteworthy trends. This was the pipeline adopted by Gotz et al [19]. By applying temporal pattern mining algorithms to these event sequences and visualizing the resulting frequent patterns, the authors explored which intermediate event sequences were strongly linked to later outcomes of interest. This approach focuses heavily on finding variations in the effect of preconditions and intermediate event sequences on outcomes for patients with a common sequence of milestone events.

CHAPTER III

VISUAL ANALYTICS WITH FLUXMAP

One basic challenge in using data-driven methods to improve health care is the development of appropriate visualization techniques. Healthcare data is frequently large, unstructured and multidimensional, and therefore difficult to make sense of. Furthermore, with the increasing usage of off-the-shelf machine learning and statistics toolboxes, it is vital that information visualization methods are concurrently developed to ensure that the results of these analyses are easily interpretable by end-users. Otherwise, there is a danger that models that are fit to complex high dimensional datasets become “black box” analysis tools whose outputs cannot be clearly explained by health researchers. Finally, as healthcare analytics research produces increasingly advanced predictive models, it is necessary to help medical domain experts interact with these models to assess their validity and utility. Effective visualization can facilitate these tasks by providing an interface between users and data, making exploratory data analysis, hypothesis generation and model validation much faster. This is likely to lead to valuable insights that can provide clinical decision support to physicians and accelerate the discovery of new medical knowledge by researchers.

This chapter introduces FluxMap, an interactive visualization system that I developed for displaying and analyzing state-based disease progression models based on the continuous-time hidden Markov model (CT-HMM) framework. FluxMap is designed to visualize state and transition properties of the CT-HMM and supports the identification and tracking of sub-groups of patients over time. This chapter begins with a brief introduction to visual analytics, followed by a survey of past efforts in visualizing longitudinal healthcare data. Then, Section 3.2.1 briefly describes the the CT-HMM disease progression model, followed by an outline of FluxMap's design and features. The chapter concludes with some comments on visualizing longitudinal data from the UCLA-UC Davis study of early skill development in autism spectrum disorder [38, 41]. I would like to thank Agata Rozga and Ted Hutman for making this dataset available to me for this work.

3.1 Visual Analytics

Visualization refers to the use of diagrams to represent information, as an aid for either communication or analysis. Visual representations can "amplify cognition" [6] by reducing working memory load and harnessing visual perception to quickly perform tasks like finding emergent patterns and outliers or comparing values. Visual analytics is an outgrowth of visualization that lies at the intersection of data mining, human-computer interaction and information visualization. It is typically associated with the coupling of *interactive* visual interfaces with computational analysis methods for *abstract* data, thereby facilitating an iterative process of data-driven pattern mining and user-driven querying and visual data mining. Visual analytics thus enables users to gain insight into complex data while also exploring and interpreting the results of data mining algorithms.

3.1.1 Representing Data

To generate a visualization, features from the data are mapped to certain graphical attributes such as the shape, size or color of a marker. Effective information visualization tools capitalize on the extremely fast first stage of human perception, known as pre-attentive processing. Certain features are processed unconsciously and in parallel in this stage, including:

- color

- shape
- size
- curvature
- orientation
- position
- grouping

These features can be used to encode different dimensions or properties of the data in a visualization scheme. While it may appear that there are nearly infinite possible combinations of these variables, allowing users to inspect multiple features simultaneously, there are several points one must be careful of. Although the aforementioned features are individually easily processed pre-attentively, this is not the case for certain combinations of those features, e.g. color and shape together. Furthermore, some encoding variables are more suitable for categorical features (e.g. shape) while others are appropriate for numerical data (e.g. size). And while these visual encoding variables can all be processed quickly, they vary greatly in terms of our ability to perceive them accurately; small variations in position or length are easy to distinguish, whereas small gradations in area or color are not.

3.1.2 Representing Models

Visualizing analytical models has the dual purpose of helping users gain an understanding of the model's rationale and assessing how much the model's output can be trusted. Though there has been relatively little study into how models ought to be visualized as opposed to data, the visualization design should give careful consideration to model representation, interaction and integration [45]. Selecting a graphical representation that conveys the structure of the model and the information it contains as completely as possible can elucidate the logic behind the model.

3.1.3 Interaction

While standalone visualizations go a long way towards helping users make sense of data, the ability to manipulate these visualizations makes exploratory data analysis much more flexible. Interaction allows the data mining process to be shaped along a user-driven line of inquiry, effectively combining expert knowledge with automated data analytics. In order to be effective, the interactive elements of a visual analytics system must provide functions that contribute to this adaptive "analytic dialogue".

Schneiderman's Visual Information Seeking Mantra of "overview first, zoom and filter, then details on demand" provides a widely acknowledged framework for designing interactions. The initial visualization provides a high-level view of the data, expediting the identification of clusters, patterns, outliers and other features of interest. Then, rather than viewing all of this information at once, the user may wish to modify the visualization to explore a particular question, by specifying the data being shown or the visualization scheme being applied. This can be achieved through features such as *filtering*, in which the displayed data subset satisfies some conditions that may isolate the effect of certain parameters on the observed phenomena; and *sorting*, where data points may be reordered along the display according to their values for one or more variables.

Once a visualization has been rendered, users may wish to manipulate the view in search of more details. This can include *selecting* a subset of the visible data by clicking it or hovering over it; *zooming* into one part to magnify it and display finer details more clearly; *panning* and *scrolling* to quickly scan for interesting patterns. With manipulations that essentially reduce the amount of visible data, a *focus plus context* approach is often adopted to help the user stay oriented with respect to the complete view. Multidimensional data is often visualized using multiple views, in which case selecting data from one view should highlight the corresponding data in the other views to facilitate comparison (*brushing*).

Advanced visual analytics systems may include more advanced interactive dynamics for capturing the analyst's inputs to the interface for future reference, through features such as annotation and interaction history. These higher-level design considerations are beyond the scope of this work, although interested readers can find more information in [21].

3.2 FluxMap

We will now describe FluxMap, a visual analytics tool developed to allow users to interactively explore disease progression as a co-evolution of multiple interacting factors. First, we review the continuous-time hidden Markov model used to model disease progression [34].

3.2.1 Continuous-Time Hidden Markov Model (CT-HMM)

We now provide a brief description of the CT-HMM [25, 34] and its application to autism progression modeling, although a more detailed discussion of the CT-HMM is provided in Section 4.1.2. We assume the progression of autism can be modeled as an unobserved Markov process between discrete hidden states. The state space is 4-dimensional, with each state characterized by range bands in chronological age, Mullen receptive language age-equivalent score, Mullen expressive language age-equivalent score and Vineland socialization age-equivalent score. Only transitions between adjacent states are permitted (e.g. we assume subjects cannot jump from a 6-month age-equivalent score to an 18-month age-equivalent score without first passing through the 12-month age-equivalent score), and bidirectional transitions are allowed in all dimensions except chronological age. The instantaneous transition intensities q_{rs}, representing the instantaneous probability of moving from state r to state s, are parameters learned by the model. Together, these transition intensities make up an instantaneous transition matrix Q.

Each individual i in the subject population has n_i irregularly timed visits with

Table 1: Attributes related to subject transitions through hidden states.

Attribute	**Values**	**Visual Encoding**
Outcome Group	ASD — TD	hue
Risk Group	High — Low	saturation
Chronological Age	6:6:36	position (horizontal)
Measurement Score	6:6:36	position (vertical)
State Visit Count	0:1:nSubjects	radius
Transition Count	0:1:nSubjects	thickness

observations $(o_{i,1}, ..., o_{i,n_i})$. These observations are generated conditionally on the hidden states according to emission probabilities $p(o|s)$, which are assumed to be distributed according to a multivariate Gaussian. The q_{rs} parameters are estimated using the Expectation-Maximization (EM) algorithm to maximize the likelihood of the observation sequences.

3.2.2 Visualization Design Considerations

3.2.2.1 Visual Encoding of Data

There is a vast amount of information about the CT-HMM that users may wish to explore. For example, analysts may be interested in the instantaneous transition matrix Q, or in how the transition probabilities evolve with time as $P(t) = e^{Qt}$, or the average dwell times associated with each state. It could be highly informative to compare these parameters between models trained separately on individuals with ASD or typical development. However, in the current iteration of FluxMap, we have focused on the exploration of *subject trajectories* through the state space. Table 1 shows how attributes are encoded.

3.2.2.2 Display

The FluxMap visualization scheme is comprised of two side-by-side panels inspired by the small multiples visualization concept introduced by Edward Tufte. In each panel, states are arranged on a 2D grid layout according to the states values in the chronological age dimension (horizontal axis) and one skill dimension (vertical axis). It should be noted that since the state space is 4D, each panel actually depicts the state space collapsed onto a plane. As a result, the state corresponding to a chronological age 6 months and receptive language age-equivalent score of 6 months contains all 4D states with those two values.

Each state is represented as a circular node whose radius is proportional to the number of visits to that state. If multiple patient subgroups are selected, the states are depicted as pie graphs to depict the proportion of measurements from each subgroup. Although the use of pie graphs is discouraged by the visualization community, this choice is still appropriate for use with a small number of patient subgroups as a means of identifying any states visited predominantly by one subgroup.

3.2.2.3 Interactions

A key component of the FluxMap visual scheme is its interactivity. The different ways of interacting with the visualization are detailed below:

- **Filtering**

  The user is able to change the set of subject trajectories being shown based on certain criteria. The outcome group can be set to display only ASD outcome subject, only typical subjects, or both. Similarly, the risk group can be set to display only high-risk individuals, only low-risk individuals, or both high- and low-risk subjects. Besides these standard filters, the user can also filter the visualized subject population based on some complementary analysis,such as clustering of subjects based on similar trajectories; such a use case will be demonstrated in Chapter 5.

- **Reconfiguring**

  The user can also specify which measurement is presented along the vertical axis as either receptive language, expressive language or socialization. Each of these options provides a different perspective to the exploratory analysis.

- **Brushing**

  By hovering over a state, the user can highlight the transitions into and out of that state. In-transitions are colored lighter than out-transitions to help differentiate them.

- **Linking**

  Hovering over a state in one panel shows the corresponding transitions in the second panel. For example, if we inspect receptive language with ASD subjects in the left panel and typical subjects in the right panel, hovering over the $(12, 12)$ state in either panel highlights the inward and outward transitions for each subject group.

- **Details-on-Demand**

  While hovering over a particular state, a tooltip appears, listing the number of subjects who visited that state, and what percentage of that number belong to each group.

- **Dynamic Querying**

  Clicking on a state causes only those trajectories which pass through that state to be displayed.

In order to explore and convey the results of the CT-HMM model in an easily interpretable way, we opted for an interactive web-based visualization scheme. The states are arranged on a 2D grid layout according to the states systolic and diastolic blood pressure ranges. Each state is represented as a circular node whose radius is proportional to the number of visits to that state.

This tool thus enables users to examine the disease progression model in the context of a particular set of features of interest. For instance, a user may wish to assess the effect of early plateauing in receptive language on subsequent expressive language development in subjects with ASD. To do this, the user selects the ASD outcome group in each panel, specifies receptive language in one and expressive language in the second panel, and selects an early-delay state in receptive language. This filters the visualization to show receptive language and expressive language development for the selected subjects. Hence this visualization scheme combines an intuitive presentation of the CT-HMM model predictive results with the capacity for some visual exploratory analysis.

3.2.3 Implementation and Software Design

FluxMap was implemented as a web application written in HTML/CSS. All views and interactions were programmed using JavaScript with the support of the jQuery, Ajax and d3 libraries. Data from the CT-HMM model was stored in JSON format. This architecture has the advantage that all data can be stored server-side, and therefore in compliance with most IRB's. Meanwhile, the derived measures used to create the visualizations, which are usually permissible for sharing, are transmitted to the client side for rendering in a web browser.

3.3 Exploratory Visual Analysis of UCLA-UC Davis Dataset

We conclude our discussion of the FluxMap visualization tool with a preliminary exploration of the UCLA-UC Davis dataset. In Figure 2, we use FluxMap to compare the hidden state trajectories of subjects with ASD and those with typical development. First, the trajectories of the ASD subgroup are much more varied than those of the typically developing subgroup. There are much stronger horizontal transitions for the ASD group (indicating plateauing of a skill) than for the typical group. Furthermore, there appears to be a particular tendency for plateauing in receptive language and socialization in the ASD subset.

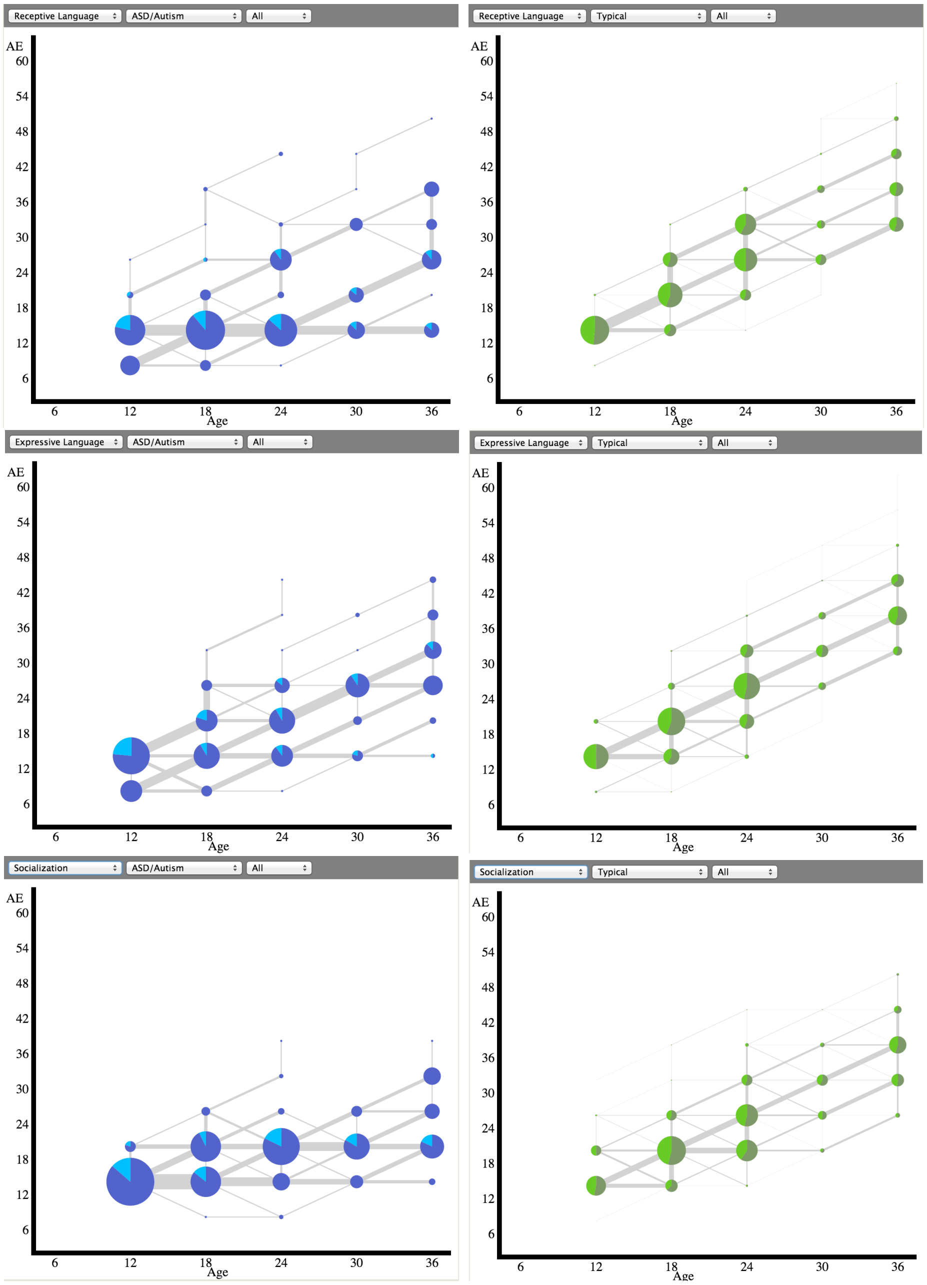

Figure 2: FluxMap: comparison of subjects with ASD (left column) or typical development (right column) in receptive (top) and expressive (middle) language and socialization (bottom).

CHAPTER IV

UNSUPERVISED PATTERN DISCOVERY

4.1 Preliminaries

4.1.1 Data

The data for this work were obtained with permission from a longitudinal cohort study of children who have siblings with autism spectrum disorder (high-risk group) or typical development (low-risk group). The study was led by Dr. Sally Ozonoff at the UCLA Center for Autism Research and Treatment in collaboration with the MIND Institute at UC Davis. Inclusion criteria for the high-risk group included having at least one older sibling with an autism spectrum disorder and without any genetic conditions linked to autism. Inclusion criteria for the low-risk group included the absence of developmental or learning disabilities in older siblings and no family history of autism spectrum disorders.

Subjects entered the study at either 4, 6, 12 or 18 months of age, and were reassessed at 6, 12, 18, 24 (30 for subset), and 36 months of age, or within two weeks of the target age. At each visit, the Mullen Scales of Early Learning (MSEL) and the Vineland Adaptive Behavior Scales (VABS) assessments were administered.

Table 2: Distribution of subjects between risk-groups and outcome-groups.

	High-risk	Low-risk	Other	Total
ASD	27	3	2	32
Non-ASD	107	89	0	196
Total	134	92	2	228

4.1.1.1 Mullen Scales of Early Learning

The Mullen Scales of Early Learning are a tool for assessing learning abilities in children between 1 and 69 months of age. The MSEL provides age-equivalent scores in four areas: visual reception, fine motor, expressive language and receptive language. These four scores can be combined into an overall Early Learning Composite. Only the expressive language and receptive language scores were available for the work in this thesis.

4.1.1.2 Vineland Adaptive Behavior Scales

The Vineland Adaptive Behavior Scales measure adaptive functioning in individuals from birth to adulthood and produce age-equivalent scores. The assessment covers five domains: communication, socialization, daily living skills, motor skills and maladaptive behaviors. Only socialization, communication, daily living and motor scores were available for this work.

4.1.2 CT-HMM Disease Progression Model

A continuous-time hidden Markov model of disease progression serves as the basis for the disease trajectory clustering in subsequent sections of this work. The model is formulated as follows. For an individual i with n_i irregularly-timed visits at times $(t_{i,1}, \dots, t_{i,n_i})$ we have observations $(\vec{o}_{i,1}, \dots, \vec{o}_{i,n_i})$. These observations are assumed to be generated from a set of hidden states S with conditional emission probability $p(o|s)$.

The state space is formulated in order to let us model the progression of a disease as a co-evolution of multiple factors. In the autism disease progression model, the subject age, Mullen receptive language age-equivalent score, Mullen expressive language age-equivalent score and Vineland socialization age-equivalent score are the state space dimensions, giving rise to a 4-dimensional state space. Each state is then defined as having a specified range of these three variables.

The individual moves through different hidden states following a Markov process, with transitions between states and the transition times governed by transition intensities q_rs between all pairs of states r and s. The transition intensities together form a transition intensity matrix Q, whose the diagonal entries are set to $q_{rr} = -\sum_{r \neq s} q_{rs}$ so that the rows sum to 0. Then, the transition probability matrix $P(t) = e^{Qt}$ gives the probabilities of the individual's state membership t units of time in the future.

The model parameters, that is the transition intensities and the emission probabilities, can be found by maximizing the likelihood of the data with the *EM algorithm*. In the *E* or *expectation* step, the Viterbi algorithm is used to find the best state sequence for each individual's observation sequence, given the current model parameters. Then, in the *M* or *maximization* step, the model parameters are updated to maximize the likelihood for the previous *E* step. Once the model parameters stabilize after alternating between the *E* and *M* steps, the most likely state sequence for each subject can be obtained via the Viterbi algorithm once more.

4.2 Methods

A number of variables are available for each subject at the conclusion of the Viterbi decoding algorithm, summarized in Table 3. We consider the most probable hidden state sequence, `instant_state_seq`, to be an encoding of the subject's co-evolution in different skills over time, and use this variable to compare the development trajectories of different subjects. Given two sequences of hidden states, $(s_{i,1}, \ldots, s_{i,m_i})$ and $(s_{j,1}, \ldots, s_{j,m_j})$, where m_i and m_j denote the number of hidden states in the sequences belonging to i and j respectively, we need to quantify the distance between these trajectories. It should be noted that m_i and m_j need not be equal to the number of visits n_i and n_j recorded for each individual, since an individual may pass through multiple unobserved states between visits.

Table 3: Subject Variables after Viterbi Decoding.

Variable	**Description**
ID	subject ID number
visit_list	cell of visit data for each visit
num_visit	number of visits
ori_state_seq	visit data encoded in states
ori_dur_seq	dwell time at each observed state
viterbi_prob	probability associated with Viterbi decoded path
instant_state_seq	Viterbi decoded path through hidden states
instant_dur_seq	dwell time at each hidden state
instant_trvisit_seq	indicator variable encoding true visits

4.2.1 Time-Series Clustering

4.2.1.1 Unidimensional Sequence Alignment

The most straightforward strategy is to treat the problem as an alignment of unidimensional sequences (states) of multidimensional information. The distance between two K-dimensional states $s_1 = (v_{1,1}, v_{1,2}, \ldots v_{1,k})$ and $s_2 = (v_{2,1}, v_{2,2}, \ldots, v_{2,k})$ is taken to be the L_1 norm between the states' centered parameter values:

$$d(s_1, s_2) = \sum_{k=1}^{K} |v_{1,k} - v_{2,k}|$$

The distance between two sequences of states, then, can be measured as the average distance between overlapping states. The number of overlapping states is dependent on the lengths of the state sequences and the alignment method selected. Two possible alignment strategies were explored:

1. anchoring the sequence alignment by the age dimension of the states, and

2. sliding one sequences along the other to search for the minimum-distance alignment with at least 3 overlapping states

The reasoning behind the first choice was to use a distance function that would yield early skill development patterns with chronological age as a reference; refer Figure 3

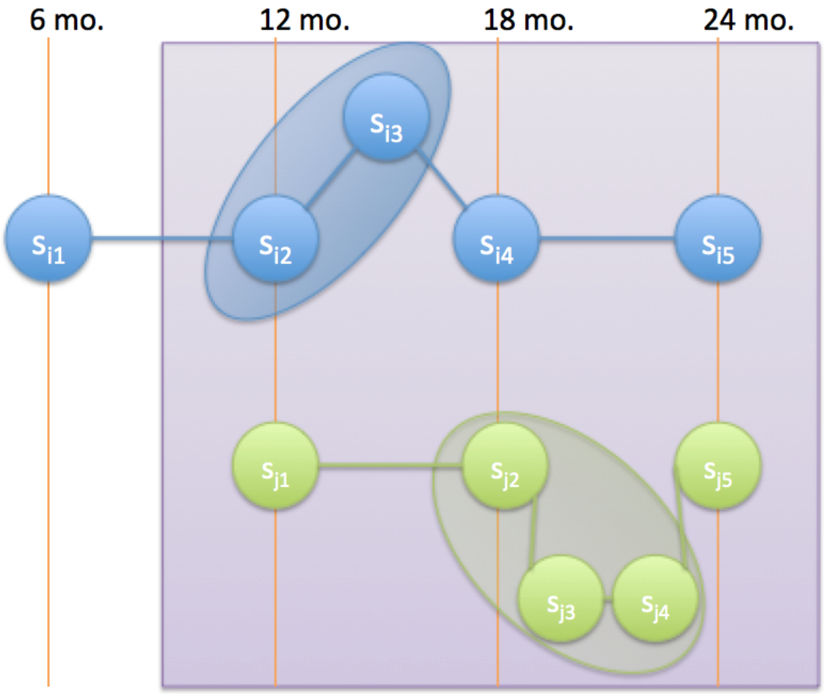

Figure 3: Two hidden state sequences $(s_{i1}, \ldots, s_{i5})$ and $(s_{j1}, \ldots, s_{j5})$ anchored by chronological age. When multiple hidden states are traversed between visits, their values are averaged (represented by ovals above), weighted by the dwell time in each hidden state. States are compared only within the window of overlap (purple square above) to compute the average difference between coinciding states.

for details. The second alignment strategy was devised in an effort to highlight more general dynamics in skill acquisition.

4.2.1.2 Clustering

A symmetric pairwise distance matrix was constructed using the distance measure described above. This distance matrix D was used to perform agglomerative hierarchical clustering analysis, in which each sequence initially belongs to its own cluster and subsequently, clusters that are the closest together are combined at each step. There are multiple methods for computing the distance between clusters of elements, including:

- **single linkage:** shortest distance between any pair of elements in the clusters

- **complete linkage:** furthest distance between any pair of elements in clusters

- **average linkage:** average distance between elements in the two clusters

The relative merit of these three linkage criteria for representing the distances computed in D was evaluated using the cophenetic correlation coefficient, (CPCC). The *cophenetic distance* between two elements is the distance between the subclusters joined at the linkage step where the two elements are first connected. The pairwise cophenetic distances between elements can be used to construct a cophenetic matrix C. These distances are represented in dendrograms as the height at which the link is drawn. The cophenetic correlation coefficient is then the correlation between the cophenetic distance matrix C and the original distance matrix D.

The dendrogram derived from the best linkage criterion as reflected by the CPCC was pruned to obtain clusters for further exploration and evaluation.

4.2.1.3 Cluster Validation

Cluster validation was performed through a combination of visual inspection and internal validation measures.

- **Distance Matrix Visualization** The pairwise distance matrix was reordered so that elements assigned to the same cluster were grouped together, and the reshuffled distance matrix was visualized.

- **Dunn Index Computation** The Dunn Index is a well-known internal validation measure that aims to identify compact, well-separated clusters. It is the ratio between the smallest inter-cluster distance and the largest within-cluster distance, defined as follows:

$$DU_k = \min_{i=1,\dots,k} \left\{ \min_{j=1,\dots,k} \left(\frac{diss(c_i, c_j)}{\max_{m=1:k} diam(c_m)} \right) \right\}$$

- **FluxMap Visualization** Lastly, the clusters are visualized using FluxMap.

4.2.2 Tensor Methods for Data Mining

4.2.2.1 Overview

Dimensionality reduction via matrix or tensor decomposition is an alternative to distance function-dependent methods for unsupervised clustering. A tensor is a multidimensional array. The *order* of a tensor is the number of dimensions, also called modes or ways. An Nth-order tensor is called *rank-one* if it can be written as the outer product of N vectors:

$$\mathcal{X} = \vec{a}^{(1)} \circ \vec{a}^{(2)} \circ \ldots \circ \vec{a}^{(N)}$$

where each element of the tensor is given by:

$$x_{i_1,i_2,\ldots,i_N} = a^{(1)}_{i_1} a^{(2)}_{i_2} \ldots a^{(N)}_{i_N}$$

4.2.2.2 CP Decomposition

The CANDECOMP/PARAFAC (CP) tensor decomposition can be thought of as a higher-order extension of singular value decomposition (SVD). SVD can be formulated as the decomposition of a matrix into a weighted sum of rank-one matrices, that is:

$$X = U\Sigma V^T = \sum_i \sigma_i \vec{u}_i \circ \vec{v}_i$$

Analogously, the CP decomposition factorizes an Nth-order tensor into a sum of rank-one Nth-order component tensors. For a third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, the CP decomposition is of the form:

$$\mathcal{X} \approx \sum_{r=1}^{R} \vec{a}_r \circ \vec{b}_r \circ \vec{c}_r$$

where $\vec{a}_r \in \mathbb{R}^I$, $\vec{b}_r \in \mathbb{R}^J$ and $\vec{c}_r \in \mathbb{R}^K$. It should be noted that there is no known polynomial time algorithm for determining the rank of a tensor (the problem is NP-hard), so it is non-trivial to select the number of components R into which the tensor should be factorized.

One widely-used method for computing the CP tensor decomposition given a set number of components R is through a technique known as alternating least-squares (ALS) [29]. This approach assumes that random variation in the data follows a Gaussian distribution and accordingly optimizes the decomposition using a Euclidean distance-based objective function [8].

However, for tensors consisting of count data, assuming a Gaussian likelihood model for randomness does not provide a good description of the data. Instead, we wish to perform a decomposition that assumes the tensor elements are Poisson-distributed. It can be shown that performing maximum likelihood estimation under the assumption of i.i.d. Poisson variables is equivalent to minimizing the generalized Kullback-Leibler divergence [9]

$$\begin{aligned} x_{\mathbf{i}} &\sim Poisson(m_{\mathbf{i}}) \\ \Rightarrow P(x_{\mathbf{i}} = k) &= \frac{e^{-m_{\mathbf{i}}} m_{\mathbf{i}}^k}{k!} \end{aligned}$$

Thus the likelihood of the tensor $\mathcal{X}$ is:

$$\begin{aligned} \mathcal{L}(\mathcal{X}|\hat{\mathcal{X}}) &= \prod_{\mathbf{i}} \frac{e^{-m_{\mathbf{i}}} m_{\mathbf{i}}^{x_{\mathbf{i}}}}{x_{\mathbf{i}}!} \\ \Rightarrow log\mathcal{L}(\mathcal{X}|\hat{\mathcal{X}}) &= \sum_{\mathbf{i}} \left(-m_{\mathbf{i}} + x_{\mathbf{i}} log(m_{\mathbf{i}}) - log(x_{\mathbf{i}}!)\right) \\ &= \sum_{\mathbf{i}} \left(-m_{\mathbf{i}} + x_{\mathbf{i}} log(m_{\mathbf{i}})\right) + \sum_{\mathbf{i}} \left(-log(x_{\mathbf{i}}!)\right) \\ &= -\sum_{\mathbf{i}} \left(m_{\mathbf{i}} - x_{\mathbf{i}} log(m_{\mathbf{i}})\right) + c \end{aligned}$$

Maximizing the log-likelihood is equivalent to minimizing $\sum_{\mathbf{i}} \left(m_{\mathbf{i}} - x_{\mathbf{i}} log(m_{\mathbf{i}})\right)$, the KL divergence. This can be achieved using a technique known as alternating Poisson regression (APR) [8].

4.2.2.3 Data Tensor Construction

We constructed a multidimensional array with five modes: subject, age at visit, receptive language level, expressive language level, and socialization. For each subject, every state in the Viterbi path inferred based on the CT-HMM model discussed in 4.1.2 was encoded into the data tensor. Two different encoding schemes were utilized.

- **Scheme 1**

  Each state encountered in the Viterbi path was counted in the data tensor based on its central values. For example, if subject number 5 visited a state with the values $(12, 6, 6, 6)$–that is, 12 months chronological age and an age-equivalent score of 6 months in all 3 measurements–this state is encoded as $\mathcal{X}_{5,2,1,1,1} = 1$.

- **Scheme 2**

  The *change* in measurements between adjacent states in the Viterbi path is encoded, with the change described as either improvement, stagnation or decline. For example, if subject 5 visits state $(12, 6, 6, 6)$, followed by $(12, 6, 12, 12)$ and then $(18, 6, 12, 6)$, the first transition is encoded as an increment to $\mathcal{X}_{5,2,2,1,1}$ (`rlang` stagnation→ 2, `elang` improvement → 1, `soc` improvement → 1). The second transition is encoded as an increment to $\mathcal{X}_{5,2,2,2,3}$ (`rlang` stagnation→ 2, `elang` stagnation → 2, `soc` decline → 3).

This resulted in a count tensor representing the age-score combinations or transitions observed over the cohort population, with all other elements set to zero.

4.2.2.4 Factorization

The tensor factorization was done with the help of the `cp_apr` function available through Sandia Corporation's MATLAB Tensor Toolbox (2015). As mentioned previously, it is difficult to know beforehand the number of component rank-one tensors to provide the CP-APR algorithm. First, we explored giving the CP-APR algorithm between 2 and 6 components and inspecting the resulting phenotypes.

We also attempted to empirically find an acceptable number of components R by comparing the fit of several decompositions as R was varied from 40 to 160. The fit of each model to the input tensor was compared based on a combination of the final log-likelihood and the least-squares fit, which roughly quantifies the proportion of the data described by the CP decomposition. The value of R that produced the highest log-likelihood and fit was chosen to fit a CP model.

4.2.2.5 Scoring and Clustering

With the final CP model determined, we needed to derive some way to assign a distance score to each pair of subjects based on the extracted components. Again, two methods were explored:

- **Projection Method:**

  In the first method, inspired by [22], each subject's individual data tensor was projected onto each of the R components to produce a normalized membership vector. These membership vectors were then used to cluster the subjects using k-means clustering.

- **Subject Factor Matrix Method:**

  The second method, inspired by [13], was based on the interpretation of each component r in the decomposition. For component r, the rth column of the subject factor matrix $\mathbf{a}_r$ captures a group of subjects; the rth column of the receptive language factor matrix captures receptive language activity that the subjects in $\mathbf{a}_r$ show; the same is true for the expressive language and socialization factor matrices. The time factor matrix captures the times at which these associations occur. The scores assigned to each subject $\mathbf{a}_r$ essentially denote the extent to which they contribute to the rth component. The subject factor matrix can thus be used to cluster the subjects using k-means clustering.

CHAPTER V

RESULTS

5.1 Data Preprocessing

Although data was available for a total of 228 infants from the UCLA-UC Davis dataset, a number of the subjects had only one or two visits in their record. Since these subjects had little to no temporal data, they were excluded from subsequent analysis, and only subjects with at least 3 visits were retained. The remaining 188 subjects were distributed across risk and outcome groups as shown in Table 4.

5.2 Hierarchical Clustering of State Sequences

5.2.1 Linkage Criteria

After the pairwise distance matrix between subjects was constructed from their trajectory distances, agglomerative hierarchical clustering was performed using three linkage criteria: single-linkage, average-linkage and complete-linkage. The the cophenetic correlation coefficient was used to identify the linkage strategy which produced the dendrogram that best reflected the relative distances in the distance matrix. This was done separately for the ASD outcome group, the non-ASD outcome group and the entire cohort all together. The average-linkage criterion performed consistently well for all three partitions of the data, while single-linkage performed the worst.

Table 4: Distribution of remaining subjects between risk-groups and outcome-groups after preprocessing.

	High-risk	Low-risk	Other	Total
ASD	23 (27)	3 (3)	2 (2)	28 (32)
Non-ASD	87 (107)	73 (89)	0 (0)	160 (196)
Total	110 (134)	76 (92)	2 (2)	188 (228)

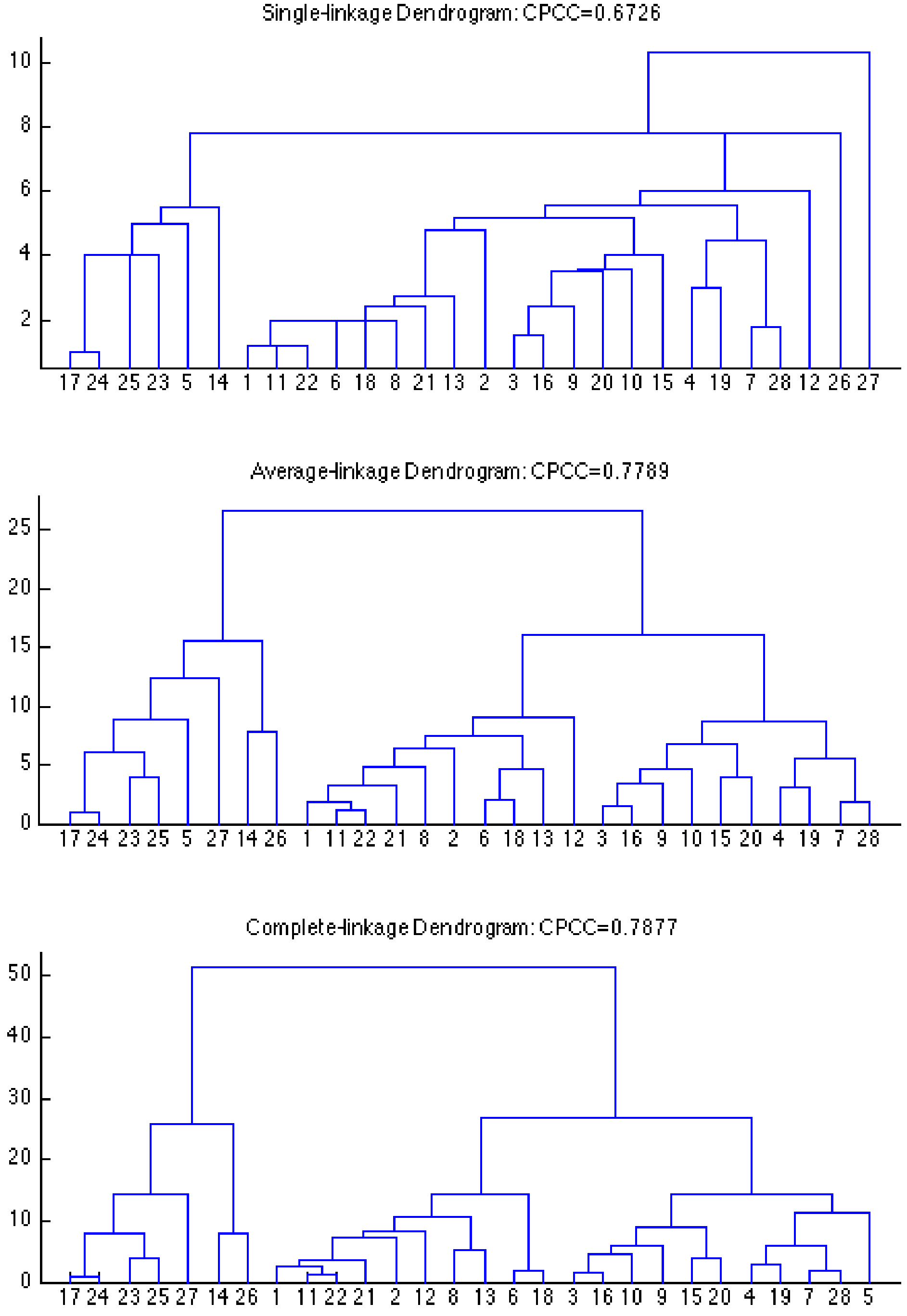

Figure 4: The effects of linkage criterion on dendrogram representation for ASD outcome class. Inter-sequence distance computed using the chronological age-anchored alignment. The closer the cophenetic correlation coefficient is to 1, the better the dendrogram represents the distances between data elements.

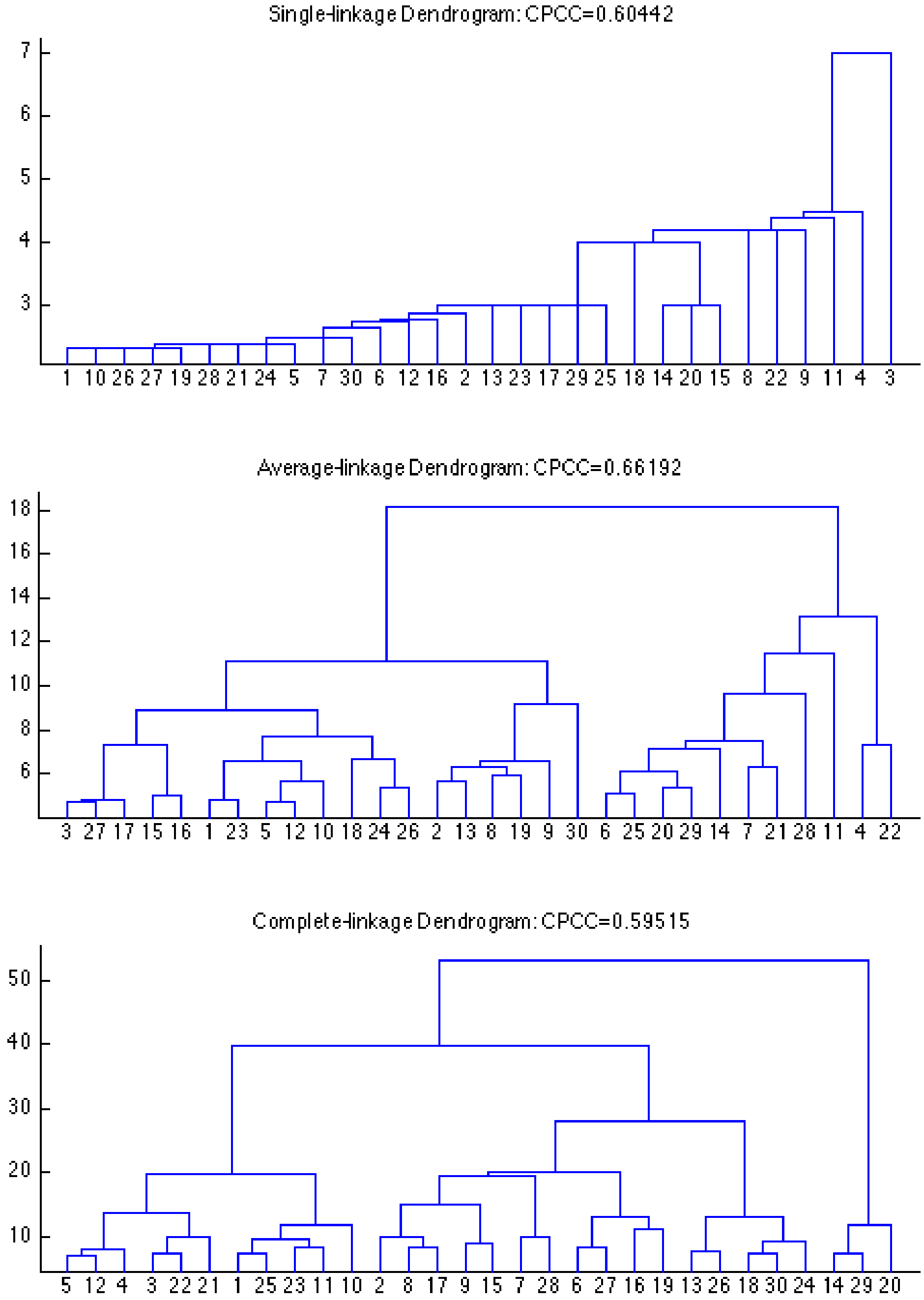

Figure 5: Different linkage criteria for dendrograms representing non-ASD outcome class. Inter-sequence distance computed using the chronological age-anchored alignment.

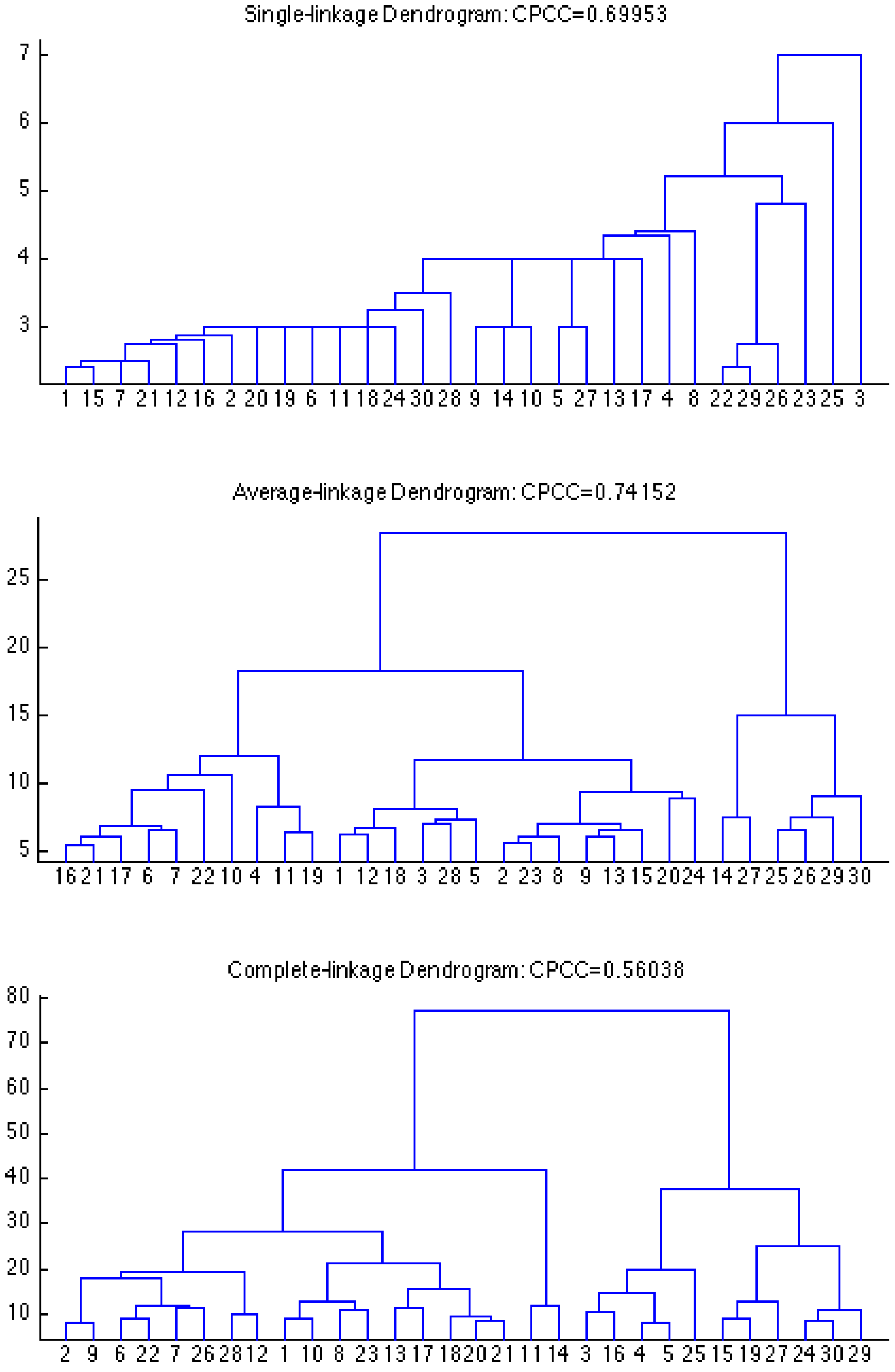

Figure 6: Different linkage criteria for dendrograms representing all outcomes.

Table 5: Quality of clusters derived from different cuts of the ASD outcome dendrogram.

‘maxclust’	Dunn Index
2	0.3039
3	0.2026
4	0.3291
5	**0.3611**
6	0.3611
7	0.3472
8	0.3472

5.2.2 Cluster Evaluation: ASD Subset

In the interest of searching for longitudinal subtypes of early skill progression in autism spectrum disorders, we focus on the ASD outcome population subset for the analysis in this section.

5.2.2.1 Number of Clusters

The average-linkage dendrogram was pruned using the MATLAB `cluster` function with parameter `‘maxclust’` varied between 2 and 8. The `‘maxclust’` argument specifies the maximum number of clusters to be returned by a horizontal cut in the tree. For each subsequent clustering, the Dunn Index was computed, shown in Table 5.

The 5-cluster solution appeared to be the best out of these choices. However, the smallest three clusters (with 1, 2 and 5 elements respectively) were merged together in a single cluster in the 3-cluster solution. Since the 1- and 2-member clusters are of little interest on their own, we looked at the 3-cluster solution in spite of its lower Dunn Index.

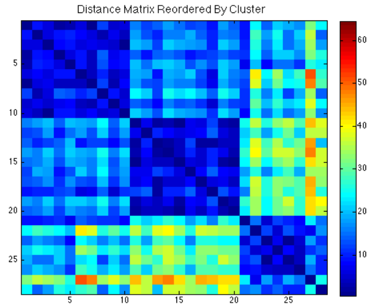

Figure 7: ASD outcome group distance matrix reshuffled to reflect three clusters.

5.2.2.2 Three-Cluster Solution

The average-linkage dendrogram for the ASD outcome group was pruned to create three clusters of subjects. The pairwise distance matrix was reordered by cluster membership and visualized. This clustering result had a very modest Dunn index of 0.2026, indicating relatively low compactness relative to the cluster separation. Nevertheless, the distance matrix visualization immediately highlighted one very distinct cluster, as well as the other two somewhat similar clusters.

The hidden state trajectories for subjects belonging to each cluster were visually explored using FluxMap, revealing distinct trends in receptive language and expressive language development for each group. One cluster (the top cluster in Figure 8) is characterized by early plateaus in receptive language followed by later recovery, coupled with delayed but steady expressive language growth. A second cluster appears to have a much stronger tendency for stagnation in all three measured skills. The final group, which is actually a combination of the three smallest clusters from the 5-cluster solution, shows much more varied development patterns, but overall appears to show steady receptive and expressive language development with slower gains in socialization.

The time-series clustering approach to autism subtyping established a shape-based distance measure in order to find trajectories with similar shapes. As such, it is

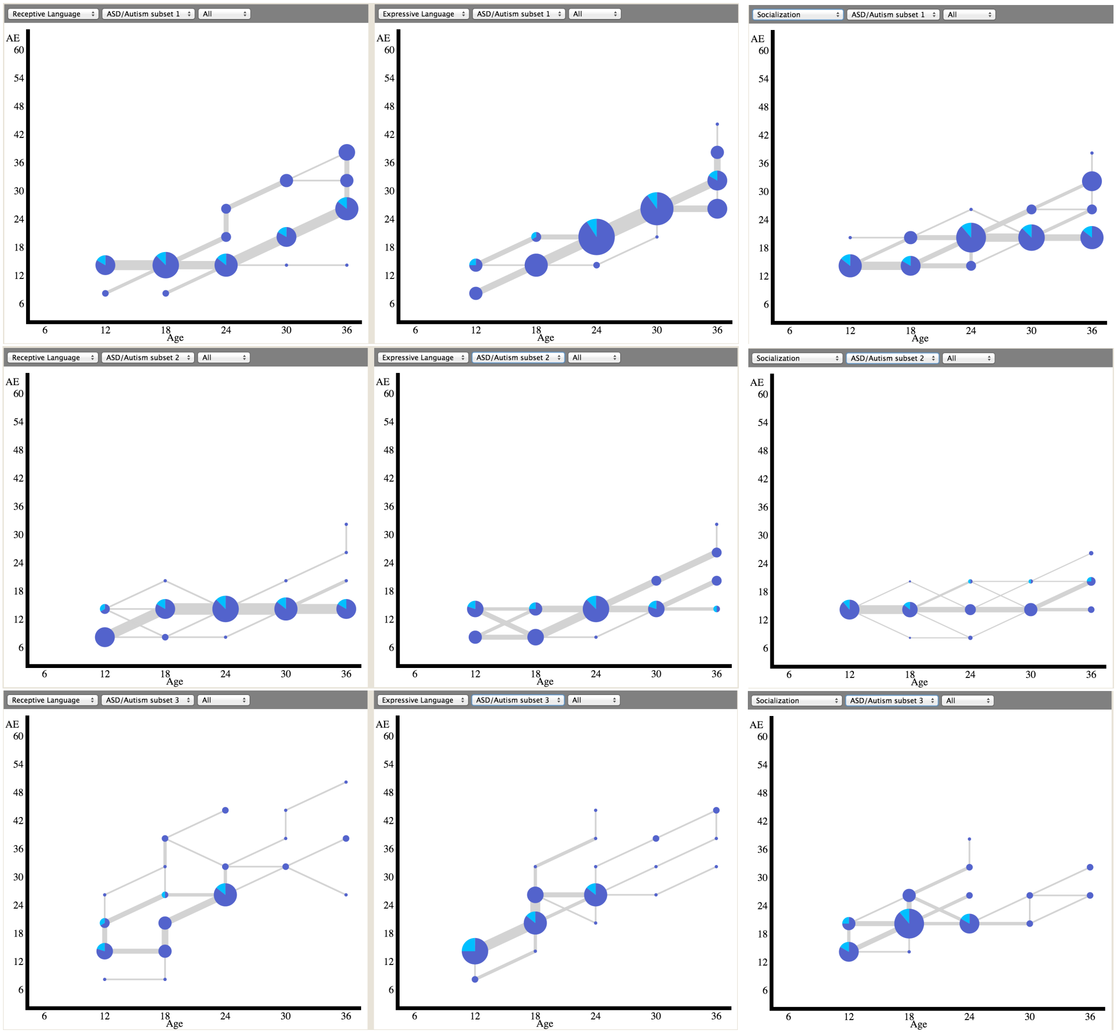

Figure 8: ASD outcome group clusters 1 (top), 2 (middle) and 3 (bottom), trends in receptive language (left), expressive language (center) and socialization (right) development, obtained from hierarchical clustering.

unsurprising that the clusters obtained with this method had fairly clear trends in skill development. The need for a custom distance measure appropriate to the domain can be seen as advantageous in that it allows domain expertise to be incorporated into the notion of meaningful clusters. However, it can also be challenging to define an appropriate distance function, especially when the data is multivariate.

5.3 CP Decomposition

As we saw previously, shape-based time-series clustering relies on the user to define a distance measure that captures shapes in the time series. Subsequently, it is assumed that under this distance measure, clusters will have small within-cluster distances and large between-cluster differences in shape. However, it is not guaranteed that an arbitrarily-imposed distance measure will achieve this. It may be desirable to convert the feature space into one that emphasizes the *variance* in the dataset. This was the reason for exploring tensor factorization applications, which we discuss next.

5.3.1 Initial Guesses for R

The number of R components was varied between 2 and 8 and inspected. However, these phenotypes were difficult to interpret, had a large degree of overlap, and varied greatly from run to run even with the same value for R. For these reasons, we decided to try higher values of R.

5.3.2 Model Selection

The number of components R given to the CP-APR algorithm to fit a decomposition to was varied between 40 and 160 to assess the effect of R on the resulting log likelihood and least-squares fit of the CP model to the data tensor. This was done for data tensors constructed using the ASD subset data according to both schemes 1 and 2 described in Section 4.2.2.3. The best fit CP model for the scheme 1 data tensor had a log-likelihood of -159.9 and a least-squares fit of 0.908 for $R = 160$, which was a

Table 6: Least-squares fit and log-likelihood for CP models for the scheme 2 data tensor.

R	40	80	120	160
μ_{FIT}	0.6090	0.8162	0.9824	0.9830
σ_{FIT}	0.0521	0.0603	0.0394	0.0380
μ_{LL}	-158.4	-132.6	-125.4	-125.4
σ_{LL}	8.5029	3.6469	0.8944	0.8944

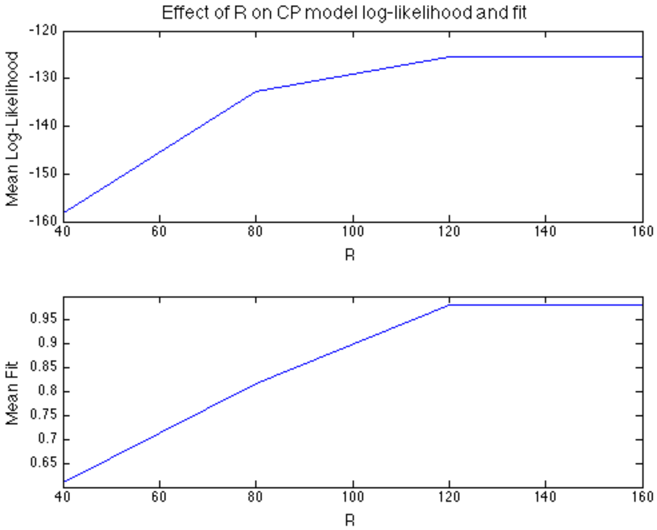

Figure 9: The number of rank-one components given to the CP-APR algorithm was varied to compare the fit of each resulting model to the input tensor.

poorer fit than was achieved for the scheme 2 transition data tensor, shown in Table 6. It should be noted that the final least-squares fit is considered only as an estimate of the model quality, since CP-APR optimizes the Kullback-Leibler divergence rather than least-squares cost.

These experiments favored larger values of R, with the log-likelihood and least-squares fit plots leveling off at $R = 120$. Subsequently, a CP model with $R = 120$ was fit to a tensor constructed from the ASD-outcome class data using transition data (scheme 2), resulting in a model with a final least-squares fit of `0.999987` and final log-likelihood of `-125.455`.

5.3.3 Cluster Evaluation: ASD Subset

5.3.3.1 Projection Method

The phenotype membership vectors of the ASD outcome group were clustered by k-means with varying values of k between 2 and 10. The quality of each of these clusterings was evaluated using the average silhouette coefficient; these results suggested that the optimal number of clusters was either 2 or 8. However, the 2-cluster solution had one "cluster" with a single subject. The 8-cluster solution also had two such clusters, and two more with just two subjects each. The dominant 4 clusters are viewed through FluxMap as shown in Figure 11.

The phenotype membership vectors were also used to create a pairwise distance matrix based on the L_1. The dendrogram obtained from hierarchical clustering with complete linkage is shown in Figure 12. Note the negative cophenetic correlation coefficient, indicating that the dendrogram is actually a poor representation of the pairwise distance matrix.

5.3.3.2 Subject Factor Matrix Method

We applied k-means clustering to the subject factor matrix, again varying the values of k between 2 and 10. The quality of each of these clustering solutions was evaluated using the average silhouette coefficient, plotted in Figure 13. The optimal number of clusters indicated by the graph was $k = 3$, but this led to a relatively uninteresting solution in which two clusters had one subject each.

The subject factor matrix was used to construct a pairwise distance matrix using the L_1 distance metric, which was then used to perform a hierarchical clustering with complete linkage. The resulting dendrogram is shown in Figure 14, and the two main clusters obtained from this approach are viewed through FluxMap in Figure 15.

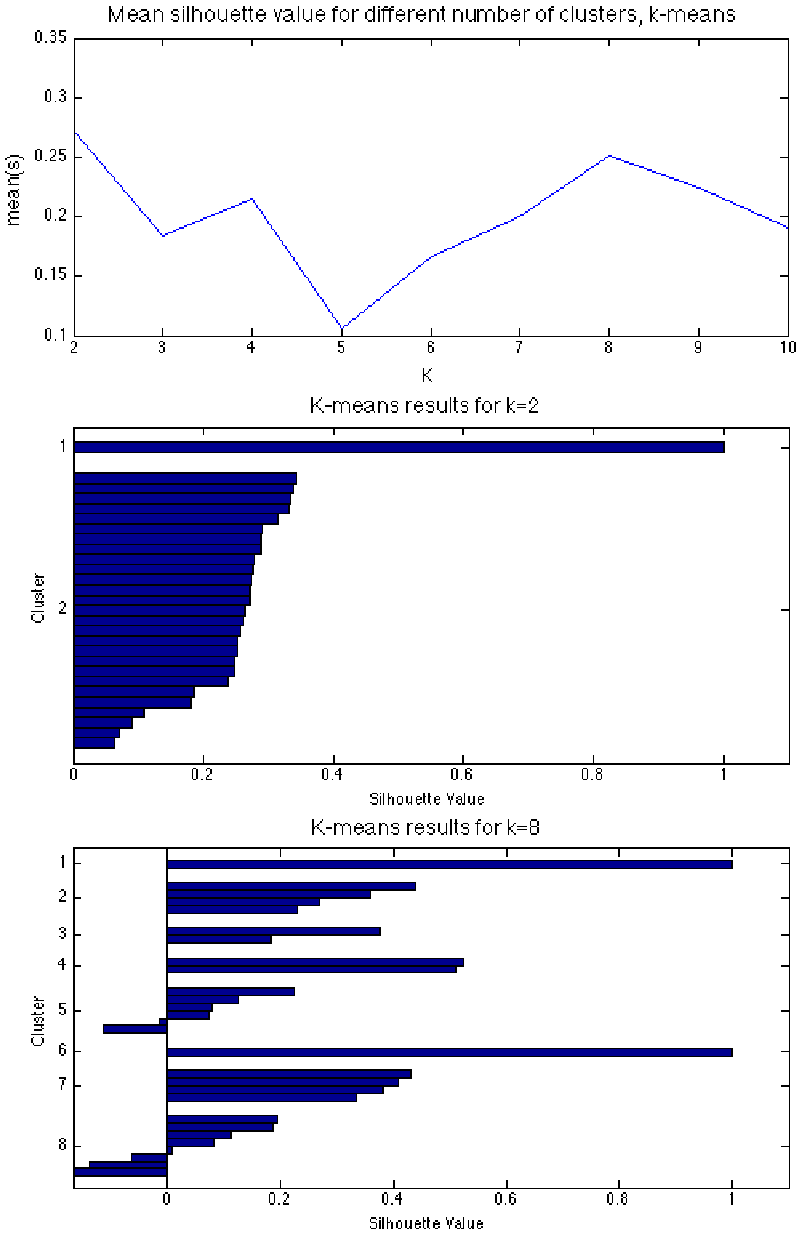

Figure 10: (Top) Average K-means silhouette coefficient vs. K, projection method. (Middle) K-means silhouette for K=2. (Bottom) K-means silhouette for K=8.

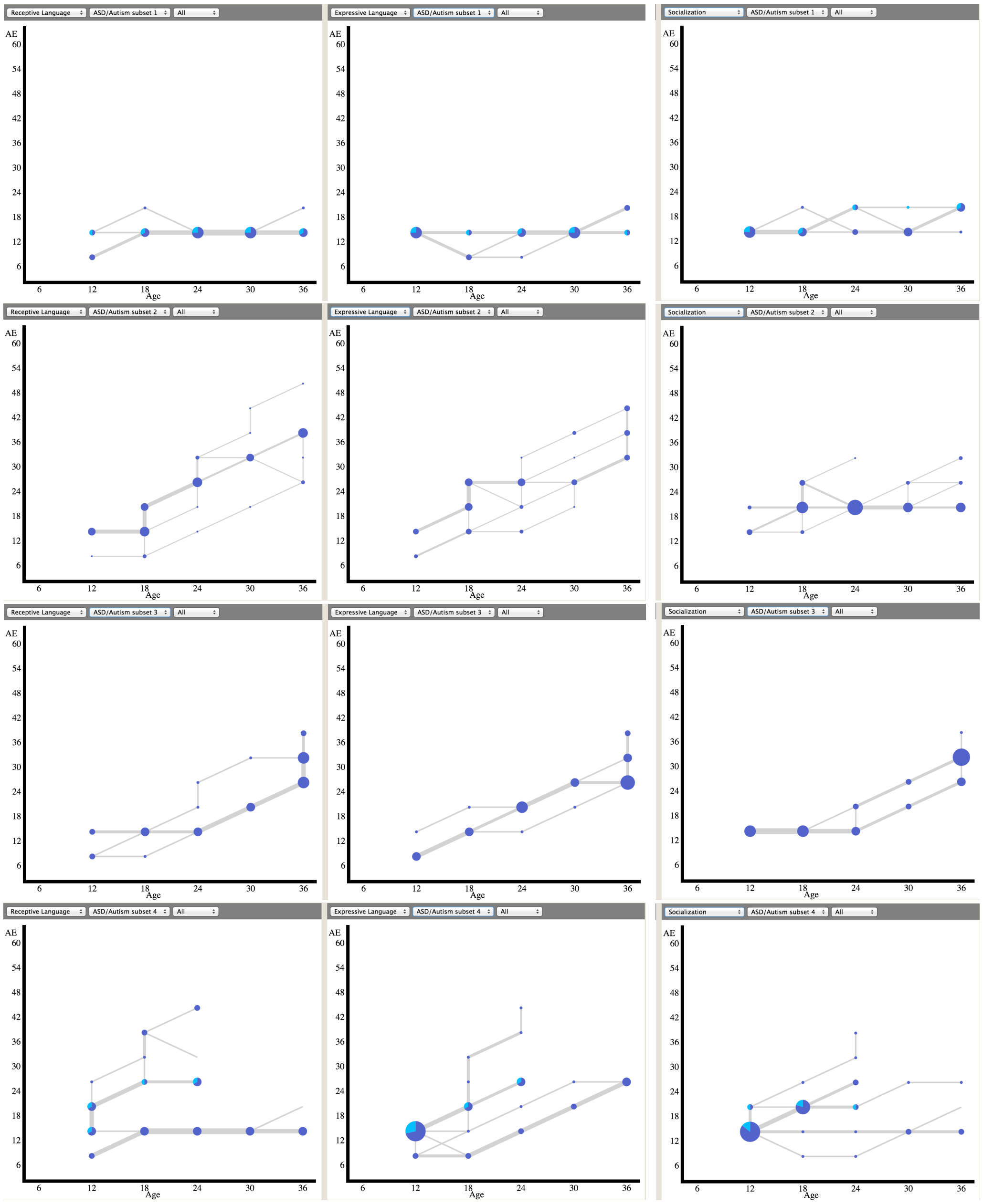

Figure 11: Dominant 4 clusters from the 8-cluster solution using phenotype membership vectors.

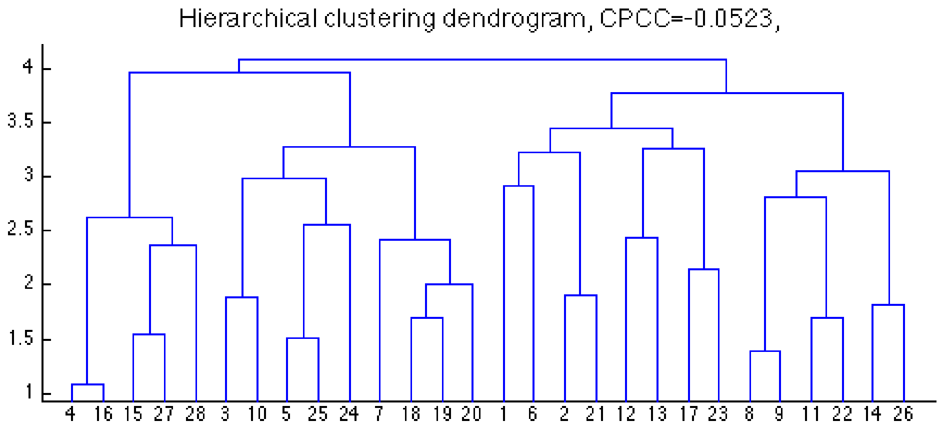

Figure 12: Hierarchical clustering of ASD subjects based on their projections onto the CP components.

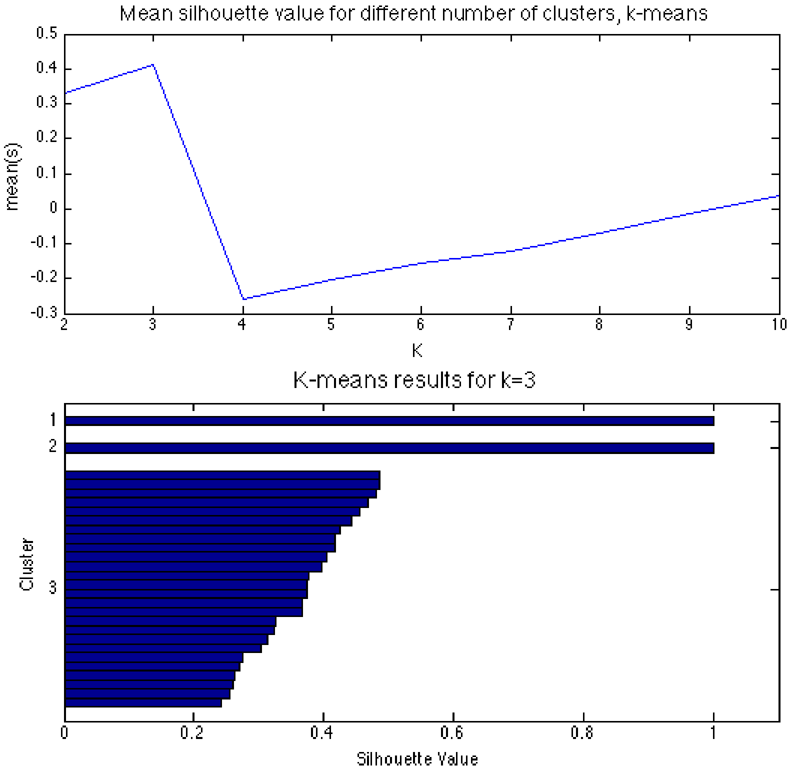

Figure 13: (Top) Average K-means silhouette coefficient vs. K, factor matrix method. (Bottom) K-means silhouette for K=3.

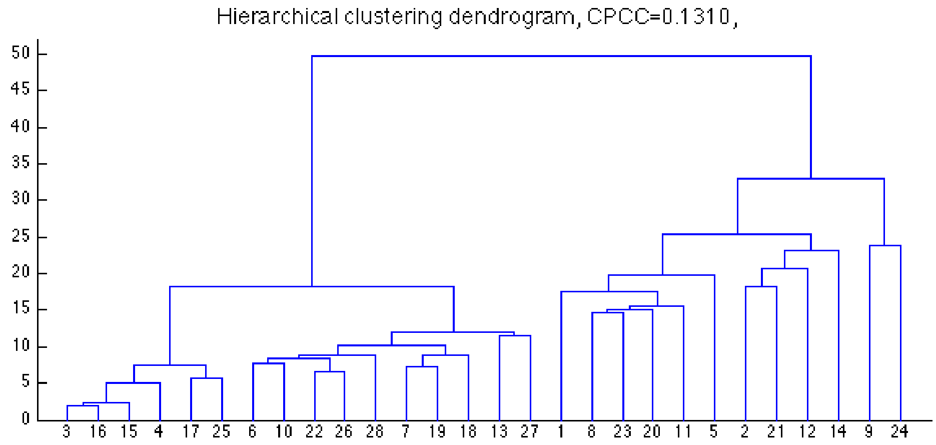

Figure 14: Hierarchical clustering of ASD subjects based on their subject factor matrix weights.

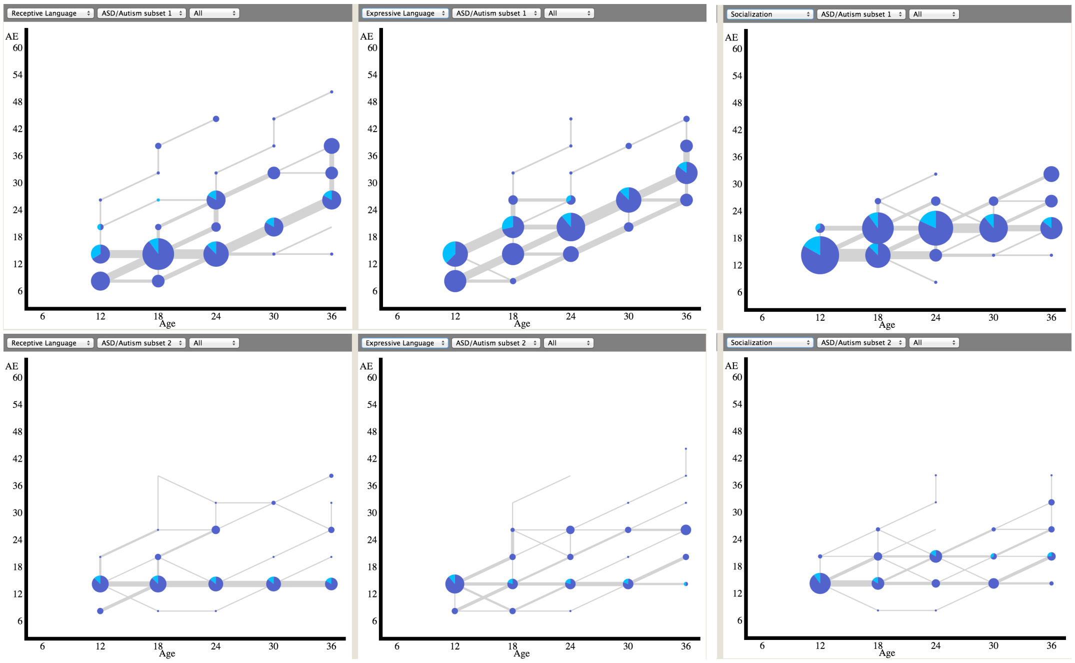

Figure 15: Two dominant clusters from factor matrix hierarchical clustering.

CHAPTER VI

DISCUSSION

This chapter reviews the results from my experiments in the context of previous work, highlights the contributions of this work, discusses its limitations and outlines possible areas for future work.

6.1 Summary of Results

This research addressed a common problem in clinical research: the identification of subgroup populations with similar characteristics to help define meaningful phenotypes. This translates directly to an unsupervised learning or clustering task in machine learning. We are specifically interested in finding subgroups with distinct temporal patterns in longitudinal data. Traditionally, past efforts to find longitudinal patient subtypes have relied on fitting the observation data with mixture models where each mixture is characterized by a learned cluster trajectory function. While these methods have proved valuable for extracting prototypical disease trajectories from population data, the descriptive power of the prototypes relies heavily on the choice of model or function for disease progression. Much of this prior work begins by assuming that clusters are indeed present and then fits the cluster parameters and assignments to maximize the likelihood of the training data.

In this study, we adopt the reverse approach of first training a disease progression model for the overall population and then subsequently searching for longitudinal clusters in individual trajectories as understood by the overall model. That is, the disease progression model does not assume the presence of any subgroups. Disease progression was modeled as the co-evolution of multiple variables using a multidimensional continuous-time hidden Markov model, which provided a detailed representation of the dynamics of disease evolution across the study population. Next, two different clustering pipelines were explored to find subsets of similar subjects based on each subject's Viterbi-decoded path of most likely states. The results of each clustering approach were compared through a use-case of characterizing the heterogeneity in early skill development patterns for subjects with an autism spectrum disorder.

The clustering results from each pipeline varied considerably. This was not unexpected, since each approach utilized different ways of representing relationships between different variable dimensions. The differences are further examined in Section 6.2.3.

In the first approach, the comparison between two subjects was based on the mean cumulative difference in receptive language, expressive language and social scores at a common age. Hierarchical clustering yielded a five-cluster solution with two large, distinct clusters that together accounted for 20 of the 28 subjects. Through the FluxMap visualization tool, it was evident that each of the larger clusters was characterized by prominent temporal patterns in receptive and expressive language skill. These patterns agreed with the well-documented early delays in language associated with autism spectrum disorder. If these patterns are observed at a larger sample size, they could be valuable for predicting early on the likelihood of an ASD diagnosis at a later age.

In the second subtyping approach, subjects were first compared to a set of "phenotypes" defined by rank-1 tensors. Each component rank-1 tensor captured some feature of the full data tensor, as well as those subjects associated with that feature and the changing strength of the association over time. The projection or contribution of each subject's data across these components recasts the subject's state space trajectory as a time-varying mixture of components. However, the components we obtained captured relatively low-level features, such as a particular link direction at a particular time point. As a result, one particular component could feature multiple subjects whose trajectories looked overall very different. This led to clusters with some overlap in temporal patterns of skill development, suggesting that the tensor-based subtyping approach needs to be refined.

6.2 Contributions

6.2.1 Novel Methods in Autism Research

To the best of our knowledge, this work is the first use of a CT-HMM to describe and visualize the progression of early language and social skill development in autism. It is also the first application of tensor decomposition methods to autism research.

6.2.2 Visualization Support for Longitudinal Subtyping

The FluxMap visualization scheme provides users with an interface through which they can interact with the results of the CT-HMM model and the results of the clustering.

6.2.3 Comparison of Clustering Paradigms

In the time-series clustering approach, we devised a custom distance function to compare two hidden-state sequences. This distance function was domain-specific, subjective, and by no means perfect:

- The decision to anchor sequences by one dimension, chronological age, reflected the importance of the age at which certain key skills appear to the understanding of developmental patterns.

- Computing the cumulative difference in each dimension effectively weights each dimension as equally important. This may or may not be the case, but cannot be validated without additional external criteria.

Despite these considerations, the clustering results were clearly interpretable, indicating that using a well-chosen custom distance measure for a specific application can still yield useful outcomes.

In the tensor decomposition approach, the original data tensor is decomposed into a set of 120 components. The subject factor matrix of the decomposition gives the extent to which each component features different subjects. Clustering this matrix resulted in two broad clusters with overlapping temporal patterns in each skill measure. However, the first cluster could be approximately described as having a tendency for early delay in receptive language followed by later improvement, and steady expressive language development. In contrast, the second cluster appeared to feature more consistent plateauing in receptive and expressive language.

The subjects' distribution across the tensor components was also estimated by projecting each individual's data onto the 120 component tensors. This approach yielded an 8-cluster solution, from which the dominant 4 clusters are shown in Figure 11. The first cluster here also shows prominent plateauing in receptive language, expressive language and socializaton. The second cluster is characterized by steady improvements in language, but a lack of improvement in socialization. The third cluster again agrees with the first group found using the factor matrix, with early delays in `rlang` followed by improvement, and steady improvement in `elang`. Furthermore, this group seems to show a concurrent early delay in socialization followed by later improvements after 24 months of age. The last cluster shows highly varying temporal trends in all three dimensions.

6.3 Limitations

One significant limitation of this work is the small subject population of just 32 subjects, only 28 of whom had enough data to be included in these analyses. It is unlikely that such a small cohort adequately captures all of the variation in temporal development trajectories associated with autism spectrum disorders. In addition, the clustering methods explored here would benefit from having more subjects to “reinforce” common patterns, if present.

Another limitation of this work is the lack of external validation measures for a more thorough evaluation of clustering solutions. Especially in the case of longitudinal disease subtyping, it is important to assess the impact of disease trajectory on prognosis or final outcome.

6.4 Future Directions

The first steps for extending on this work should be to address the aforementioned limitations by analyzing a larger cohort population, preferably with additional features for external validation such as a severity score at diagnosis or some other outcome of interest. This would help confirm the existence of longitudinal autism progression patterns, and investigate possible links between these development patterns and outcomes.

It would also be valuable to explore other tensor construction strategies. One possibility could be to encode the ratio between age-equivalent scores and chronological age. Another possible extension would be to explore the use of tensor factorization methods for temporal link prediction as discussed in [13]. Each component r of the CP decomposition we obtained has a time mode encoding the strength of the association between subjects in $\mathbf{a}_r$ and the component r activity pattern in the skill modes. These component-wise temporal profiles may be useful for predicting changes in skill level into future time steps.

REFERENCES

[1] "2012 - the year in healthcare charts." `http://www.forbes.com/sites/danmunro/2012/12/30/2012-the-year-in-healthcare-charts/`. Accessed: 2015-07-13.

[2] Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., Welch, K., and Pickles, A., "Patterns of growth in verbal abilities among children with autism spectrum disorder.," *Journal of consulting and clinical psychology*, vol. 75, no. 4, p. 594, 2007.

[3] Bruining, H., Eijkemans, M., Kas, M., Curran, S. R., Vorstman, J., and Bolton, P. F., "Behavioral signatures related to genetic disorders in autism," *Mol Autism*, vol. 5, no. 11, 2014.

[4] Caban, J. J. and Gotz, D., "Visual analytics in healthcare–opportunities and research challenges," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 260–262, 2015.

[5] Cao, N., Gotz, D., Sun, J., and Qu, H., "Dicon: Interactive visual analysis of multidimensional clusters," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2581–2590, 2011.

[6] Card, S. K., Mackinlay, J. D., and Shneiderman, B., *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[7] Chen, D. P., Weber, S. C., Constantinou, P. S., Ferris, T. A., Lowe, H. J., and Butte, A. J., "Clinical arrays of laboratory measures, or clinarrays, built from an electronic health record enable disease subtyping by severity," in *AMIA Annual Symposium Proceedings*, vol. 2007, p. 115, American Medical Informatics Association, 2007.

[8] Chi, E. C. and Kolda, T. G., "Making tensor factorizations robust to non-gaussian noise," *arXiv preprint arXiv:1010.3043*, 2010.

[9] Chi, E. C. and Kolda, T. G., "On tensors, sparsity, and nonnegative factorizations," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1272–1299, 2012.

[10] Davis, K., Stremikis, K., Squires, D., and Schoen, C., "Mirror, mirror on the wall," *The Commonwealth Fund*, 2014.

[11] De Keulenaer, G. W., Brutsaert, D. L., Borlaug, B. A., Redfield, M. M., and others, "Systolic and diastolic heart failure are overlapping phenotypes within the heart failure spectrum," *Circulation*, vol. 123, no. 18, pp. 1996–2005, 2011.

[12] Doshi-Velez, F., Ge, Y., and Kohane, I., "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis," *Pediatrics*, vol. 133, no. 1, pp. e54–e63, 2014.

[13] Dunlavy, D. M., Kolda, T. G., and Acar, E., "Temporal link prediction using matrix and tensor factorizations," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 2, p. 10, 2011.

[14] Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., and Tabar, V. K., "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4434–4463, 2014.

[15] Fountain, C., Winter, A. S., and Bearman, P. S., "Six developmental trajectories characterize children with autism," *Pediatrics*, vol. 129, no. 5, pp. e1112–e1120, 2012.

[16] Gamliel, I., Yirmiya, N., Jaffe, D. H., Manor, O., and Sigman, M., "Developmental trajectories in siblings of children with autism: Cognition and language from 4 months to 7 years," *Journal of Autism and Developmental Disorders*, vol. 39, no. 8, pp. 1131–1144, 2009.

[17] Gotham, K., Pickles, A., and Lord, C., "Trajectories of autism severity in children using standardized ados scores," *Pediatrics*, vol. 130, no. 5, pp. e1278–e1284, 2012.

[18] Gotz, D., Sun, J., Cao, N., and Ebadollahi, S., "Visual cluster analysis in support of clinical decision intelligence," in *AMIA Annual Symposium Proceedings*, vol. 2011, p. 481, American Medical Informatics Association, 2011.

[19] Gotz, D., Wang, F., and Perer, A., "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data," *Journal of biomedical informatics*, vol. 48, pp. 148–159, 2014.

[20] Hanley, J. A., Negassa, A., Forrester, J. E., and others, "Statistical analysis of correlated data using generalized estimating equations: an orientation," *American journal of epidemiology*, vol. 157, no. 4, pp. 364–375, 2003.

[21] Heer, J. and Shneiderman, B., "Interactive dynamics for visual analysis," *Queue*, vol. 10, no. 2, p. 30, 2012.

[22] Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., and Sun, J., "Limestone: High-throughput candidate phenotype generation via tensor factorization," *Journal of biomedical informatics*, vol. 52, pp. 199–211, 2014.

[23] Home, C., "Prevalence of autism spectrum disorder among children aged 8 yearsautism and developmental disabilities monitoring network, 11 sites, united states, 2010," 2010.

[24] Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., Geissbuhler, A., and others, "Clinical data mining: a review," *Yearb Med Inform*, vol. 2009, pp. 121–133, 2009.

[25] Jackson, C. H. and others, "Multi-state models for panel data: the msm package for r," *Journal of Statistical Software*, vol. 38, no. 8, pp. 1–29, 2011.

[26] Jung, T. and Wickrama, K., "An introduction to latent class growth analysis and growth mixture modeling," *Social and Personality Psychology Compass*, vol. 2, no. 1, pp. 302–317, 2008.

[27] Keim, D. A., Mansmann, F., Oelke, D., and Ziegler, H., "Visual analytics: Combining automated discovery with interactive visualizations," in *Discovery Science*, pp. 2–14, Springer, 2008.

[28] Klimov, D., Shknevsky, A., and Shahar, Y., "Exploration of patterns predicting renal damage in patients with diabetes type ii using a visual temporal analysis laboratory," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 275–289, 2015.

[29] Kolda, T. G. and Bader, B. W., "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[30] Landa, R. J., Gross, A. L., Stuart, E. A., and Bauman, M., "Latent class analysis of early developmental trajectory in baby siblings of children with autism," *Journal of Child Psychology and Psychiatry*, vol. 53, no. 9, pp. 986–996, 2012.

[31] Landa, R. J., Gross, A. L., Stuart, E. A., and Faherty, A., "Developmental trajectories in children with and without autism spectrum disorders: the first 3 years," *Child development*, vol. 84, no. 2, pp. 429–442, 2013.

[32] Lasko, T. A., Denny, J. C., and Levy, M. A., "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PloS one*, vol. 8, no. 6, p. e66341, 2013.

[33] Lewis, S., Foltynie, T., Blackwell, A., Robbins, T., Owen, A., and Barker, R., "Heterogeneity of parkinsons disease in the early clinical stages using a data driven approach," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 3, pp. 343–348, 2005.

[34] Liu, Y.-Y., Ishikawa, H., Chen, M., Wollstein, G., Schuman, J. S., and Rehg, J. M., "Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden markov model," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp. 444–451, Springer, 2013.

[35] Lobo, I., "Same genetic mutation, different genetic disease phenotype," *Nature Education*, vol. 1, no. 1, p. 64, 2008.

[36] Lord, C., Luyster, R., Guthrie, W., and Pickles, A., "Patterns of developmental trajectories in toddlers with autism spectrum disorder.," *Journal of consulting and clinical psychology*, vol. 80, no. 3, p. 477, 2012.

[37] Marlin, B. M., Kale, D. C., Khemani, R. G., and Wetzel, R. C., "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 389–398, ACM, 2012.

[38] Ozonoff, S., Iosif, A.-M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., Rogers, S. J., Rozga, A., Sangha, S., Sigman, M., and others, "A prospective study of the emergence of early behavioral signs of autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 49, no. 3, pp. 256–266, 2010.

[39] Perer, A. and Sun, J., "Matrixflow: temporal network visual analytics to track symptom evolution during disease progression," in *AMIA annual symposium proceedings*, vol. 2012, p. 716, American Medical Informatics Association, 2012.

[40] Raghupathi, W. and Raghupathi, V., "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.

[41] Rozga, A., Hutman, T., Young, G. S., Rogers, S. J., Ozonoff, S., Dapretto, M., and Sigman, M., "Behavioral profiles of affected and unaffected siblings of children with autism: Contribution of measures of mother–infant interaction and nonverbal communication," *Journal of autism and developmental disorders*, vol. 41, no. 3, pp. 287–301, 2011.

[42] Schulam, P., Wigley, F., and Saria, S., "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[43] Siller, M. and Sigman, M., "Modeling longitudinal change in the language abilities of children with autism: parent behaviors and child characteristics as predictors of change.," *Developmental psychology*, vol. 44, no. 6, p. 1691, 2008.

[44] Sun, J., Bi, J., and Kranzler, H. R., "Multi-view singular value decomposition for disease subtyping and genetic associations," *BMC genetics*, vol. 15, no. 1, p. 73, 2014.

[45] Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., and Sommerfield, D., "Visualizing data mining models," *Information visualization in data mining and knowledge discovery*, vol. 24, 2001.

[46] Toth, K., Munson, J., Meltzoff, A. N., and Dawson, G., "Early predictors of communication development in young children with autism spectrum disorder: Joint attention, imitation, and toy play," *Journal of autism and developmental disorders*, vol. 36, no. 8, pp. 993–1005, 2006.

[47] Tunc, B., Ghanbari, Y., Smith, A. R., Pandey, J., Browne, A., Schultz, R. T., and Verma, R., "Punch: Population characterization of heterogeneity," *NeuroImage*, vol. 98, pp. 50–60, 2014.

[48] Veatch, O., Veenstra-VanderWeele, J., Potter, M., Pericak-Vance, M., and Haines, J., "Genetically meaningful phenotypic subgroups in autism spectrum disorders," *Genes, Brain and Behavior*, vol. 13, no. 3, pp. 276–285, 2014.

[49] Wang, H.-M., Hsiao, C.-L., Hsieh, A.-R., Lin, Y.-C., and Fann, C. S., "Constructing endophenotypes of complex diseases using non-negative matrix factorization and adjusted rand index," *PloS one*, vol. 7, no. 7, p. e40996, 2012.

[50] Warner, J. L., Denny, J. C., Kreda, D. A., and Alterovitz, G., "Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization," *Journal of the American Medical Informatics Association*, pp. amiajnl–2014, 2014.

[51] Wenzel, S. E., "Asthma phenotypes: the evolution from clinical to molecular approaches," *Nature medicine*, vol. 18, no. 5, pp. 716–725, 2012.

[52] Yang, H.-J., Kim, Y. E., Yun, J. Y., Kim, H.-J., and Jeon, B. S., "Identifying the clusters within nonmotor manifestations in early parkinson's disease by using unsupervised cluster analysis," *PloS one*, vol. 9, no. 3, p. e91906, 2014.

[53] Young, G. S., Rogers, S. J., Hutman, T., Rozga, A., Sigman, M., and Ozonoff, S., "Imitation from 12 to 24 months in autism and typical development: a longitudinal rasch analysis.," *Developmental psychology*, vol. 47, no. 6, p. 1565, 2011.