Detection of Aytipical Neurodevelopment in Infants Using PCA and K-Means Clustering

Mansi Sakarvadia
mansisak@live.unc.edu

Nathaniel Fulmer
nfulmer@live.unc.edu

Advised by: Dr. Martin Styner

COMP562, UNC Fall 2020

1 Introduction

Our study aims to understand the disparities in neurological development between different infant populations based on volumetric measurements of the sub-cortical regions of the brain and various demographic data. For our study, mothers were recruited from UNC Hospitals and Duke University Medical Center and enrolled in the Early Brain Development Study at UNC (PI Gilmore). Our analysis begins with data collected from the children of these mothers, starting at infancy and continuing at age one and two.

The primary sources of information are brain scans, demographic information collected from the parents, and developmental milestone assessment scores (i.e. Mullen Scales of Early Learning Assessment). As with most studies involving human subjects, many cases are missing information from one or more of these sources. Information about each patient's brain structure comes from analysis of T1 and T2 MRI scans performed on the patient at different ages. All included subjects have a scan from infancy; however, only some have brain scans at one and two years.

Our long term investigation focuses on identifying associations between the structures of the brain and potential atypical neurological development. This paper attempts to divide the population into subgroups with unique neurodevelopmental trajectories. Complicating attempts to find associations among the infant population are the lack of diagnosis data and missing assessment information. This obfuscates our ability to separate patients into "typical" and "atypical" groups and was a major motivating factor in choosing to apply unsupervised machine learning methods.

2 Methods

To reduce dimensionality and identify the areas with the greatest amount of variation between subjects, we performed principal component analysis. We then applied k-means clustering to characterize cohorts within the population, with particular attention paid to potentially atypically developing groups.

2.1 Pre-Processing

The data set we used was originally composed of 651 subjects with 114 features. Initially, we dropped any features that repeated data (ex: weeks of gestation captures a similar value as days of gestation) and converted categorical data to a numerical representation. Eleven categorical features were converted to numerical representation and the meaningfulness of this conversion is subject to further discussion and evaluation. Next, we excluded any features that had over 10% of subjects missing that particular value. Then, we dropped any patients that were missing values for the remaining features. This left us with a data set composed of 646 subjects with 45 numerically descriptive features.

2.2 Principal Component Analysis

In order to reduce the dimensionality of the data and improve our likelihood of getting meaningful results when trying to identify subgroups within the population, we decided to use principal component analysis (PCA) as an initial pre-process. Since our data has dramatically different scale values (i.e. volumes have magnitudes ranging from the hundreds to thousands while age has a magnitude in the tens), we applied Sci-kit Learn's built-in scaling function to the data, while Sci-kit Learn's PCA function[1] centers the data automatically.

For our PCA, we tried a several different numbers of components. The two main decisions were

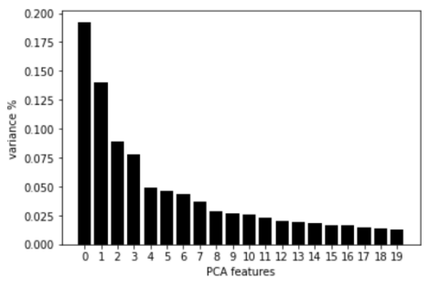
- Deciding the optimal trade-off between reconstruction accuracy and dimensionality reduction
- Deciding which PCA model to use for reducing our data before k-means clustering

The assessment of the validity of any given PCA was determined from our testing data set, a subset of 100 randomly chosen cases that were removed from the training data. Once the PCA was generated from the training data, we applied the PCA model to the testing data and then inverted the fit. Our "loss" in recreation accuracy was calculated as the mean of the squared difference between the original test data and the inverted fit of the PCA on the test data.

We generated PCA models for our data with 5, 10, 15, 20, 25, 30, and 35 principle components. For each potential number of components, we conducted three trials where the training set was randomly split into three groups of 149 cases. We created a PCA model for each group of cases with a total of nine models generated for every previously listed number of principal components.

While fitting these models, we kept track of the most heavily weighted feature for those components that had an explained variance ratio of more than 10%. From the 63 models we generated in total, the Mother's language was the most heavily weighted feature in 9 components, gestational age at birth in 10, gestational birth order in 10, thalamus volume in 15, grey matter volume in 17, and age when MRI was taken in 34. The relevance and potential meaningfulness of these features is examined in the results section.

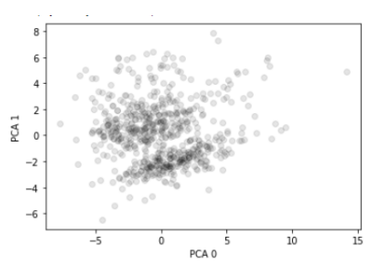
Evaluating the loss of the nine models for each number of components, we recognized a general trend of: 5 components had 50% loss, 10 components had 35% loss, 15 components had 24% loss, 20 components had 15% loss, 25 had 8% loss, 30 had 4% loss, and 35 had 1% loss, with variations of +/- 2% for any specific model. We decided that using 20 components with around 15% loss allowed us to maintain acceptable levels of accuracy in recreation while ensuring that each of our components was able to explain more than 1% of the variation in the whole data set (as shown below). Then, we chose the PCA model that had the minimal amount of loss from the nine generated models.



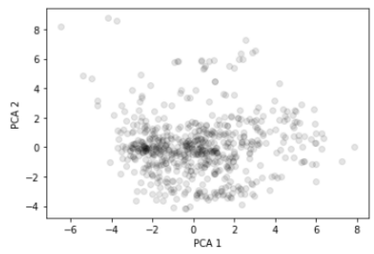
2.3 K-means Clustering

After reducing the dimensionality of our data, we aimed to determine if cohorts exist within our infant population with potentially different neurodevelopment trajectories based on varying neurological markers.

When graphing the 0th and 1st principle component or the 1st and 2nd principle component from our chosen PCA model against one another, it becomes apparent that subgroups do exist within our data.

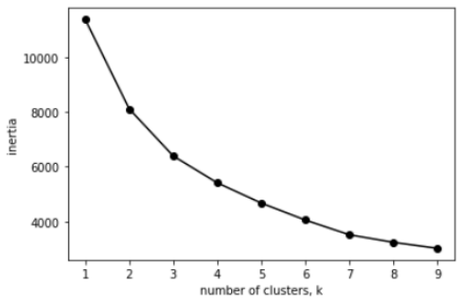


(a) 0th vs 1st component



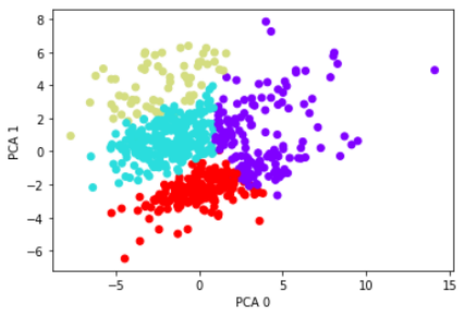
(b) 1st vs 2nd component

For this reason, we applied Sci-kit learn's K-means clustering algorithm[1] to our PCA transformed data. We determined the ideal number of clusters by analyzing the inertia (within-cluster sum-of-squares) of our K-means algorithm when fitted to our data with 1 - 9 clusters.



When choosing our number of clusters we had to make sure we picked a good balance between over-fitting the data and minimizing the inertia. For this reason, we choose to proceed with four clusters because that is when the inertia begins to level off (as seen in the figure above).

After clustering, we were able to conduct sub-group analysis and analyze the differences in each of the infant cohorts, discussed in the results section. Below is a visualization of the four clusters among the 0th and 1st principle components.



3 Results and Conclusions

3.1 PCA

Six features were repeatedly heavily-weighted in our principal components: age at the date of the MRI scan, grey matter volume, gestational age at birth, birth order, and thalamus volume. A reasonable indicator that our dimensionality reduction appropriately captures the major sources of variation in our data is the high frequency of age at MRI being a heavily weighted feature. This matches with the relevant neurological literature, as it is widely recognized that the brain is in rapid levels of development throughout the first months of infancy[2], and therefore, one would expect large amounts of variation to be caused by this age difference. The importance of grey matter is likely associated with age at MRI, since grey and white matter undergo large amounts of change during infancy[3]. Higher gestational age at birth is associated with increased wellness factors and better developmental progress[4], so early gestational birth age may be relevant in recognizing atypical neurodevelopment. Furthermore, gestational birth order affects brain maturation since twins have very different early-life development than singletons given the limitations of space and nutrients for multiple fetuses in the uterus [5].

The more interesting finding, and a potential source of future investigation, is the implication of thalamus volume as an important bio-marker. The thalamus is a structure in the center of the brain that acts as a relay center for sensory and motor information[6]. This part of the brain is relatively small compared to other sections, so its ability to capture large amounts of variation between subjects could be a very fruitful avenue of exploration. A particularly interesting route of study is the role the thalamus plays in sensory and motor control[7].

3.2 K-means

We were able to identify four subgroups in our data, and analyze the means of variables of interest with regards to these subgroups. A few findings were of particular interest:

- The Mullen composite score at two years of age for two of the groups was lower by about half a standard deviation (std is about 15) than the two other groups (approximately 100 vs 109). The Mullen Scales of Early Learning are administered at age one and two and measure fine motor, verbal response, and expressive-receptive language abilities of children. Lower performance on this assessment can potentially indicate atypical cognitive development in the future.

- One of the groups is, on average, at a higher risk for atypical development because their mothers have a higher frequency of diagnosis of some form of Psychosis (mood disorder, schizophrenia, or bipolar disorder) and the research indicates an increased risk for these children of developing some form of psychosis themselves[8].

These groups are of particular interest for future longitudinal studies in regards to both cognitive and psychological development. If neurological or demographic features can be identified as markers of future atypical development, appropriate therapies could be introduced at a young age. Studies have shown that the earlier treatment, therapy, or intervention is conducted, the better the quality of life is for that individual [9,10].

4 Citations

[1]@articlescikit-learn, title=Scikit-learn: Machine Learning in Python, author=Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., journal=Journal of Machine Learning Research, volume=12, pages=2825–2830, year=2011

[2] Normal maturation of the neonatal and infant brain: MR imaging at 1.5 T. A.J. Barkovich, B.O. Kjos, D.E. Jackson, Jr, and D. Norman Radiology 1988 166:1, 173-180

[3] Hüppi, P.S., Warfield, S., Kikinis, R., Barnes, P.D., Zientara, G.P., Jolesz, F.A., Tsuji, M.K. and Volpe, J.J. (1998), Quantitative magnetic resonance imaging of brain development in premature and mature newborns. Ann Neurol., 43: 224-235. https://doi.org/10.1002/ana.410430213

[4]Knickmeyer, R. C., Kang, C., Woolson, S., Smith, J. K., Hamer, R. M., Lin, W., Gerig, G., Styner, M., Gilmore, J. H. (2011). Twin-singleton differences in neonatal brain structure. Twin research and human genetics : the official journal of the International Society for Twin Studies, 14(3), 268–276. https://doi.org/10.1375/twin.14.3.268

[5] Deanne K. Thompson, Claire E. Kelly, Jian Chen, Richard Beare, Bonnie Alexander, Marc L. Seal, Katherine Lee, Lillian G. Matthews, Peter J. Anderson, Lex W. Doyle, Alicia J. Spittle, Jeanie L.Y. Cheong, Early life predictors of brain development at term-equivalent age in infants born across the gestational age spectrum, NeuroImage, Volume 185, 2019, Pages 813-824, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2018.04.031.

[6] Sherman, S. M., Guillery, R. W. (2006). Exploring the thalamus and its role in cortical function (2nd ed.). MIT Press.

[7] Infant functional thalamocortical connectivity Hilary Toulmin, Christian F. Beckmann, Jonathan O'Muircheartaigh, Gareth Ball, Pumza Nongena, Antonios Makropoulos, Ashraf Ederies, Serena J. Counsell, Nigel Kennea, Tomoki Arichi, Nora Tusor, Mary A. Rutherford, Denis Azzopardi, Nuria Gonzalez-Cinca, Joseph V. Hajnal, A. David Edwards Proceedings of the National Academy of Sciences May 2015, 112 (20) 6485-6490; DOI: 10.1073/pnas.1422638112

[8] Fisher, H. L., McGuffin, P., Boydell, J., Fearon, P., Craig, T. K., Dazzan, P., ... Murray, R. M. (2014). Interplay between childhood physical abuse and familial risk in the onset of psychotic disorders. Schizophrenia bulletin, 40(6), 1443-1451./Hyperactive Difficulties," J. Abnorm. Child Psychol., p. 17, 2002.

[9] E. J. S. Sonuga-Barke et al., "Nonpharmacological Interventions for ADHD: Systematic Review and Meta-Analyses of Randomized Controlled Trials of Dietary and Psychological Treatments," Am. J. Psychiatry, vol. 170, no. 3, pp. 275–289, Mar. 2013, doi: 10.1176/appi.ajp.2012.12070991.

[10] W. Bor, M. R. Sanders, and C. Markie-Dadds, "The Effects of the Triple P-Positive Parenting Program on Preschool Children with Co-Occurring Disruptive Behavior and