

Modeling Cognitive Trends in Preclinical Alzheimer’s Disease (AD) via Distributions over Permutations

Gregory Plumb^{1(✉)}, Lindsay Clark^{1,2}, Sterling C. Johnson^{1,2}, and Vikas Singh¹

¹ University of Wisconsin–Madison, Madison, USA
gplumb@wisc.edu

² William S. Middleton Memorial Veterans Hospital, Madison, USA
<https://github.com/GDPlumb/PermutationZScoreClustering>

Abstract. This paper presents an algorithm to identify subsets of subjects who share similarities in the context of imaging and clinical measurements within a cohort of *cognitively healthy* individuals at risk for Alzheimer’s disease (AD). In particular, we wish to evaluate how patterns in the subjects’ cognitive scores or PIB-PET image measurements are associated with a clinical assessment of risk of developing AD, image based measures, and future cognitive decline. The challenge here is that all the participants are *asymptomatic*, our predictors are noisy and heterogeneous, and the disease specific signal, when present, is weak. As a result, off-the-shelf methods do not work well. We develop a model that uses a probability distribution over the set of permutations to represent the data; this yields a distance measure robust to these issues. We then show that our algorithm produces consistent and meaningful groupings of subjects based on their cognitive scores and that it provides a novel and interesting representation of measurements from PIB-PET images.

1 Introduction

It is widely accepted that Alzheimer’s disease (AD) pathology, including amyloid and neurofibrillary tangles, begins to develop decades before cognitive decline reaches the stage of a clinical dementia diagnosis. Mild cognitive decline occurs several years preceding a clinical diagnosis of mild cognitive impairment or dementia [1, 3, 6]. Developing methods to reliably characterize biomarkers for AD within asymptomatic individuals during this preclinical stage of AD is essential for intervention trials. In this paper, we aim to characterize patterns in subjects’ PIB image measurements from eight regions of the brain and psychometric scores in order to identify a subgroup that is at highest risk for dementia [2, 4]. We analyze an asymptomatic, late middle-aged cohort that is at risk for Alzheimer’s disease due to parental history. We evaluate (a) whether such subgroups can

This research was supported by R01 grants EB022883, AG021155, AG027161, AG040396, UW CPCP and NSF CAREER award 1252725. Plumb was supported by NSF REU funding.

be reliably identified and (b) if the corresponding patterns are associated with risk of dementia and cognitive decline over time. We utilize the psychometric data (instead of imaging data alone) because it is less expensive and easier to collect, which makes it better suited for general screening. Later, we also provide evidence comparing our representation of the data (which is broadly applicable) to clustering schemes run on the native representations (such as z -scores).

Motivation: The goal of this paper is find a representation of the imaging/cognitive data (given as z -scores) that makes clustering much easier – we do not propose a “new” clustering algorithm. Recall that most representations of data produce clusters based on spatial proximity and a cluster is represented by an average value for each feature. One alternative, called sub-typing, uses patterns such as “has higher values for feature X than feature Y”. A natural way to model these patterns is to represent each example by a feature vector of *rankings* that sorts each feature, e.g., from best to worst. This process partitions the entire input space into regions corresponding to permutations. The key advantage of this perspective is that it abstracts various distributional issues away. However, it can be lossy because points near the separator are classified in a either-or manner. For example, consider the point $(0, 1, 0.9)$: this point is represented by the permutation $[1\ 3\ 2]$ but that representation loses the information that $[1\ 2\ 3]$ is probably almost as good a representation because the second and third features values are similar. We address this weakness by using *probability distributions over permutations* to model the data in order to get the sub-typing characterizations while reducing that information loss. Additionally, such a formulation is robust to missing/noisy data: missing data can be assumed to be equally likely to appear anywhere in the permutation and small perturbations probably do not change the underlying distribution much [8] (Fig. 1).

There is one remaining challenge that we have neglected so far: most subjects in our dataset are middle aged and do not show significant AD-like brain atrophy patterns yet. We therefore want to augment our representation with a distance measure that, within clustering, encourages subjects/examples to “follow the leader”. That is, we will assign a subject with a weakly defined pattern to the same cluster as a subject with a clearly defined pattern as long as the two patterns are consistent with one another. To do so, we introduce the notion of a *concentration distance metric*. If U is a discrete probability space and p_1, p_2 are two probability distributions over U such that p_1 is non-zero on X_1 and p_2 is non-zero on X_2 , then a distance is a concentration distance metric if it

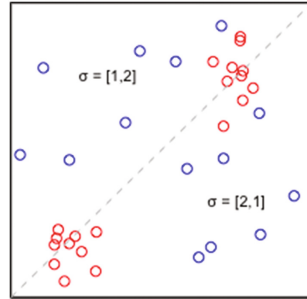


Fig. 1. For a ranking data model, the dashed line ($y = x$) partitions the plane into two sections corresponding to permutations. The red points clearly display spatial clustering, but are poorly separated by ranking geometry. Conversely, the blue points display little spatial clustering but are grouped well by ranking.

penalizes differences between $p_1(x)$ and $p_2(x)$ more heavily for $x \in (X_1 \cap X_2)^c$ than for $x \in X_1 \cap X_2$. Intuitively, this means that the distance metric encourages similarity between distributions where one is more concentrated than the other.

Contributions: We propose a model that is well-suited to handle many of the challenges described above and in other datasets with a weak signal. By modeling the data as permutations, our model achieves (1) a sub-typing characterization and (2) abstracts many distributional issues away. By using probability distributions over the permutations, it becomes (3) highly tolerant to noisy and missing data and mitigates the information loss inherent in representing data with a single permutation. Finally, by exploiting the group structure of permutations, \mathbb{S}_n , of these probability distributions, we give a distance metric that makes the clustering procedure (4) well adapted to data where the signal is weak.

2 Identifying Patterns in Sets of Distributions over \mathbb{S}_n

At a high level, we have to compare different features which may be on very different scales (e.g. cognitive scores, image measurements) to one another to determine which is “better”. This is accomplished by converting each feature into a z -score (zero mean, unit variance) which places an individual’s measurement relative to the population. For interpretability, the z -score for some features may be multiplied by -1 so that a positive score is a “good” result. Then each subjects’ representative permutation can be found by sorting their z -scores. For example, the z -scores $(1.2, -0.5, 0)$ are represented by the permutation $[2\ 3\ 1]$ because the first feature has the “best” value and the second has the “worst”.

Constructing a Distribution over \mathbb{S}_n : If $X = (x_1, x_2, \dots, x_p)$ is our feature vector (z -scores), then representing X as a permutation, σ , may not be robust because the underlying features may be noisy. Additionally, x_i may be missing for some subjects. To address these issues, we instead express X as distribution in the space of permutations, centered on σ . To obtain the distribution, we perturb each feature and consider a set of normal random variables: $N_i \sim \mathcal{N}(x_i, \gamma)$. We then sample $d_i \sim N_i$ (which can be thought of as a sample drawn from z -score space) and then increment the number of times that that the permutation representing those perturbed z -scores has been seen. We repeat this sampling procedure many times and then normalize the counts to produce a probability distribution over permutations. This sampling procedure approximates the probability distribution of the ordering of the random variables N_1, N_2, \dots, N_p . An example of this conversion can be seen in Fig. 2. Finally, we must choose γ used to construct the normal variables. If it is too small, the model tends towards representing each subject as a single permutation. If it is too large, the probability distributions becomes flat and we lose all discriminative power. We choose this parameter using a 5-fold cross validation, clustering on the training set, making assignments to the test clusters using a 1-NN classifier, and then measuring the consistency of the (full) set of assignments between folds. *Remark:* Most applications will use a distribution over the top- k features rather than the

entire distribution because it is rare that all features are relevant. This representation is consistent with our distance metric and reduces the runtime to be polynomial, as long as k is constant.

Collecting Structural Information of these Distributions:

Each subject is now represented as a probability distribution over permutations; next, we will construct a concentration distance metric. To do so, we need information about the structure of the set of permutations. Collecting this information from the distributions directly is inefficient, so we use the Fourier transform for groups. This is accomplished using Clausen’s fast Fourier transform algorithm for the \mathbb{S}_n [5] which works by breaking \mathbb{S}_n into cosets corresponding to \mathbb{S}_{n-1} recursively until it gets to a single permutation (a coset for \mathbb{S}_1). Figure 3 shows an example of this structure for \mathbb{S}_3 . Further, this structure is nicely encoded into the coefficients of the Fourier transform.

Harmonic analysis on \mathbb{S}_n is defined via the notion of *representations*. A matrix valued function $\rho : \mathbb{S}_n \rightarrow \mathbb{C}^{d_\rho \times d_\rho}$ is said to be a d_ρ dimensional representation of the symmetric group if $\rho(\sigma_2)\rho(\sigma_1) = \rho(\sigma_2\sigma_1)$ for any pair of permutations $\sigma_1, \sigma_2 \in \mathbb{S}_n$. A representation ρ is said to be *reducible* if there exists a unitary basis transformation which simultaneously block diagonalizes each $\rho(\sigma)$ matrix into a direct sum of lower dimensional representations. If ρ is not reducible, then it is said to be *irreducible*. Irreducible representations are the elementary building blocks of all of \mathbb{S}_n ’s representations. A complete set of inequivalent irreducible representations are denoted by \mathcal{R} . The Fourier transform of a function $f : \mathbb{S}_n \rightarrow \mathbb{C}$ is then defined as the sequence of matrices $\hat{f}(\rho) = \sum_{\sigma \in \mathbb{S}_n} f(\sigma)\rho(\sigma)$ $\rho \in \mathcal{R}$. The inverse transform is $f(\sigma) = \frac{1}{n!} \sum_{\rho \in \mathcal{R}} d_\rho \text{tr}[\hat{f}(\rho)\rho(\sigma)^{-1}]$ $\sigma \in \mathbb{S}_n$. Much practical interest in Fourier transform can be attributed to various properties of the irreducible representations, such as conjugacy and unitarity. Additional details of the fast Fourier transform for \mathbb{S}_n and its irreducible representations are available in [7].

Distance matrix: We now have the structural information of our probability distributions encoded nicely by the Fourier transforms. Let $q = (q_1, q_2, \dots, q_d)$ be a partition of n ; the weight that we use for the component of the Fourier transform corresponding to q is $q_1!$ which matches the

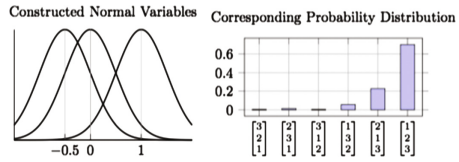


Fig. 2. An example of conversion from z -scores (left) to a probability distribution based over permutations (right) for the z -scores $(-0.5, 0.0, 1.0)$ with $\gamma = 0.5$. Note that most of the probability still places the third feature as the “best”

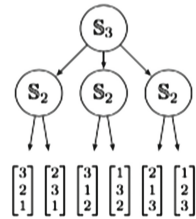


Fig. 3. Left coset tree for \mathbb{S}_3 showing its members as leaves. Note that the last element is fixed by each \mathbb{S}_2 coset.

size of the corresponding coset. These weights essentially make matching the probability in the \mathbb{S}_k cosets much more important than in the \mathbb{S}_{k-1} cosets. Importantly, a normal distance metric, such as the Hilbert-Schmidt norm, fails to satisfy properties of a concentration metric. Consider an example with five subjects whose movie preferences were $s_1 = (1, 0, 0, 0)$, $s_2 = (0, 1, 0, 0)$, $s_3 = (0, 0, 1, 0)$, $s_4 = (0, 0, 0, 1)$, $s_5 = (1, 0, 1, 0)$. If we consider a value of 1 to be an indicator that the person watched the movie and a 0 to mean that they have not, it is clear that s_1 and s_3 likely have similar preferences to s_5 while s_2 and s_4 do not. We want our distance metric to accurately capture this information. A comparison of our distance metric to a normal one can be seen in Fig. 4.

Spectral Clustering: Given our distance measure, we perform a simple clustering using the algorithm described in [9]. There are two main benefits to using spectral analysis for obtaining the clusters. The first is that the space in which our distance metric lies is not well studied at all and, as a result, the assumptions, such as convexity, that normal clustering methods often make may not be reasonable. Also, we do not expect many subjects in our dataset to show a disease specific signal, so not ‘normalizing’ the cluster size is helpful. Importantly, this is a standard method and most of the sensitivity in our results (to be discussed shortly) is due to our representation and distance metric.

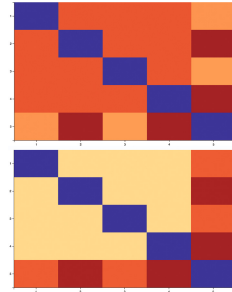


Fig. 4. The computed distance matrices for this example using our concentration distance metric (above) and a Hilbert-Schmidt norm (below). As we can see, the normal distance metric actually has the opposite effect of what is desirable; it makes s_5 more dissimilar from s_1 and s_3 .

3 Experimental Evaluations

Datasets: The larger of our two primary datasets ($n = 1211$) is comprised of test scores on eight psychometric exams: Rey Auditory Verbal Learning Test Long-Delay Free Recall (RAVLT), Letter Fluency (CFL), Stroop Color-Word Interference condition (Stroop), Boston Naming Test (BNT), Trailmaking Test Part B (TMT B), Brief Visuospatial Memory Test-Revised Delayed recall (BVMT), WMS-R Logical Memory Delayed recall (LM), and WAIS-R Digit Sybmol (DS). These scores were adjusted for demographic information (age, gender, and literacy score) using linear regression so that the data modeled cognitive phenotypes independent of demographic information. This data was collected every two years and the cognitive slopes for these variables

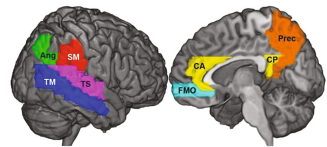


Fig. 5. The eight regions with PIB data

are computed as the average rate of change across the study (6+ years). The smaller dataset ($n = 183$) consisted of imaging measures of regional beta-amyloid plaque burden from [C-11] PiB-PET scans (PIB) for eight brain regions (Fig. 5), global atrophy, and total white matter lesion volume. We also have cerebrospinal fluid (CSF) biomarkers of amyloid (amyloid-beta 42) and neural injury (total tau and phosphorylated tau) for this group. Each subject also has demographic data (age, gender, total years of education) and a clinical consensus diagnosis of either cognitively normal (CN) or early mild cognitive impairment (eMCI).

Overall design and evaluation criteria: Our goal was to evaluate whether a clustering procedure, using either of our datasets, could group the subjects into clusters which associate with longer term cognitive decline trends. *Distance metric:* For each dataset, we will use two different distance metrics. The first will be a normal ℓ_2 norm and the second will be the sub-typing based concentration measure that we defined. *Stability:* In order to choose γ for the concentration metric, we defined a procedure to measure the stability of the clusters to perturbations in the initial dataset. If a clustering algorithm is finding well-defined clusters, it will be relatively stable. Conversely, if the clusters are poorly separated, the clusters will change dramatically when the input data is changed. Clearly, we should expect that stability is a prerequisite to finding meaningful clusters. *Evaluation:* The clusters will be evaluated in terms of the percentage of subjects who are eMCI in each cluster (a clinical classification of risk for AD which is used because most of the sample are middle-aged and asymptomatic) and future cognitive decline. Additionally, the clusters defined using the psychometric data will also be compared on the imaging metrics; this serves as an additional validation because the relationship between biomarkers of AD and dementia are more direct. For the psychometric data, our method found results associated with both the risk assessment and cognitive decline; normal methods do not find consistent clusters. For the imaging data, our method found clusters associated with the clinical assessment of risk for AD; normal methods' results are only associated with future cognitive decline. These comparisons suggest that our representation has advantages over simpler methods.

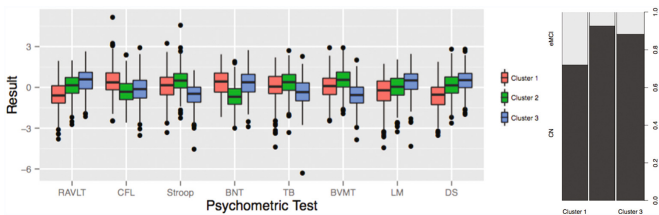


Fig. 6. (Left) Results of covariates used in clustering, grouped by test. This immediately allows us to see the patterns that define the clusters; for example, cluster 1 had lower scores on RAVLT and cluster 3 had higher scores. (Right) We see that Cluster 1 has a significantly higher percentage of eMCI subjects.

Psychometric Scores and ℓ_2 norm: This configuration yielded clusters whose assignments differed by 35–45% between folds in the 4-fold stability procedure. In fact, the signal is so weak that a full suite of baselines (hierarchical, k -means with numerous distance measures, and spectral clustering) produce results that are not meaningful. Given that the results are either highly unstable or produce small (<5 subject) clusters for a variety of methods, it is reasonable to conclude that these distance measures are not producing well separated clusters.

Psychometric Scores and Concentration metric: This metric identified three clusters and the assignments differed by only 10–15% between folds when three clusters were desired. This immediately suggests that this metric is better suited for the dataset. *Cluster Characteristics:* The cluster characteristics in terms of the psychometric tests and risk for developing AD are summarized in Fig. 6. It is important to observe that, by design, the three clusters did not differ on demographic information including age, sex, or literacy estimates. But the clusters differed significantly in the composition of subjects with a eMCI diagnosis ($p < 0.001$). Generally, we found that Cluster 1 exhibited significantly worse performance on measures of verbal episodic memory (RAVLT) and attention/processing speed than Clusters 2 and 3, Cluster 3 had worse performance than Clusters 1 and 2 on measures of executive functioning and visual memory, and subjects in cluster 2 had worse performance on one measure of language. *Cognitive Slopes:* Clusters 1 and 3 both exhibited more negative slopes on the measures that they performed worse on. This is not surprising given that lower cognitive performance may be an indicator of an early start of cognitive decline. However, this result does support the longitudinal stability of these clusters despite the progressive nature of the disease; starting with lower values and having a more negative slope means that the signal within the cluster is getting stronger. Interestingly, Cluster 2's performance on BNT increased more quickly than for the other clusters. This group may be a normally aging group showing normal practice effects at repeated visits over time. *Image Features:* Due to fewer data points available for imaging variables, we grouped clusters 1 and 3 together and compared the pooled data to cluster 2. The analysis was done using ANCOVA with age and gender as covariates. Clusters 1 and 3 had greater amyloid burden (PIB) in the superior and middle regions of the temporal lobe bilaterally (p values between 0.03 and 0.04). These clusters also had greater global atrophy and higher levels of both ttau/ab42 and ptau/ab42 in the CSF, profiles typical of AD. There were no significant differences for the other PIB measurements or total lesion white matter volume.

Summary: These findings suggest that the *order of severity of measures* can reveal meaningful groupings in such data, where alternatives yield unsatisfactory results [12]. Cluster 1 demonstrated the profile consistent with those found in preclinical AD, including verbal memory and attention deficits [11]. Additionally, this cluster also shows a greater rate of decline on verbal memory measures, and had a higher proportion of eMCI subjects. This suggests that this cluster is likely at highest risk for AD. Cluster 3 showed lower scores on measures of executive functioning, as well as greater decline in speed and executive functioning

measures over time. It is possible these individuals are at higher risk for a non-AD related cognitive disorder, such as vascular cognitive impairment. Finally, cluster 2 showed worse performance on language measures, but improved over subsequent visits. This may correspond to a normal aging group.

PIB Imaging Measurements: We also compare the ℓ_2 norm to our concentration metric for clustering the subject based on their PIB imaging data. With the ℓ_2 norm, the assignments were very stable for producing two clusters, but highly unstable for three. The concentration metric produced consistent assignments for either two or three clusters. For the sake of comparison, we analyze the two cluster results for both metrics. *Cluster Characteristics:* For the ℓ_2 norm, the clusters differed significantly on all of PIB measurements and the average values in Cluster 1 were lower than those of Cluster 2. The clusters produced by the concentration metric actually did not differ on all of the measurements; Cluster 2 had higher values in the anterior cingulate region and lower values in supramarginal and superior temporal regions. This difference is expected due to a degree of information loss in our distance measure. *Risk of AD:* Our metric found that Cluster 2 was at higher risk than Cluster 1 ($p = 0.039$) while the ℓ_2 metric did not ($p = 0.87$). Interestingly, this suggests that simply having elevated measurements does not correspond an increased risk but having larger measurements in some regions than in others does. *Cognitive Slopes:* For the ℓ_2 norm, subjects from Cluster 1 experienced slower decline on the Stroop and DS tests and a slower increase on CFL than subjects from Cluster 2; further, their LM scores increased while Cluster 2's decreased. This suggests that broadly elevated PiB levels are associated with steeper cognitive decline, but regional PiB levels (e.g., elevated frontal, but lower temporal measurements) were not.

4 Conclusions

This paper demonstrates that using a subtype based distance measure can (a) find well separated clusters in data (psychometric tests) that normal distance metrics cannot and (b) that it finds different clusters than normal distance metrics (PIB data). The significance of the sub-typing distance metric is demonstrated by the fact that regional differences in PiB levels (which is exactly what the sub-typing metrics consider) are associated with risk of AD and not with future cognitive decline and that generally increased PIB levels exhibit the opposite pattern. We propose a harmonic analysis based algorithm that operates on the permutation-based representation of z -scores derived from the native feature-vector representation of the subjects [10]. We show how a concentration metric based distance derived via a Fourier transform of a probability distribution over the permutations reveals interesting structure in the data, and enables a follow-up clustering scheme to identify scientifically meaningful clusters, that are associated with a clinical assessment of developing AD, future cognitive decline, and differences on image based features.

References

1. Backman, L., Jones, S., Berger, A.K., et al.: Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. *Neuropsychology* **19**(4), 520–531 (2005)
2. Blacker, D., Lee, H., Muzikansky, A., et al.: Neuropsychological measures in normal individuals that predict subsequent cognitive decline. *Arch. Neurol.* **64**(6), 862–871 (2007)
3. Clark, L.R., Racine, A.M., Kosciak, R.L., et al.: Beta-amyloid and cognitive decline in late middle age: findings from the Wisconsin registry for Alzheimer's prevention study. *Alzheimers Dement.* **12**, 805–814 (2016)
4. Clark, L.R., Schiehser, D.M., Weissberger, G.H., et al.: Specific measures of executive function predict cognitive decline in older adults. *J. Int. Neuropsychol. Soc.* **18**(1), 118–127 (2012)
5. Clausen, M.: Fast generalized fourier transforms. *Theor. Comput. Sci.* **67**(1), 55–63 (1989)
6. Jack, C.R., Knopman, D.S., Jagust, W.J., et al.: Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12**(2), 207–216 (2013)
7. Kondor, R.: Group theoretical methods in machine learning. Ph.D. thesis, Columbia University (2008)
8. Kondor, R., Howard, A., Jebara, T.: Multi-object tracking with representations of the symmetric group. In: *AISTATS* (2007)
9. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: *NIPS*, pp. 849–856 (2002)
10. Plumb, G., Pachauri, D., Kondor, R., Singh, V.: S_n FFT: a Julia toolkit for fourier analysis of functions over permutations. *J. Mach. Learn. Res.* **16**, 3469–3473 (2015)
11. Sperling, R.A., Aisen, P.S., Beckett, L.A., et al.: Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**(3), 280–292 (2011)
12. Young, A.L., Oxtoby, N.P., Huang, J., et al.: Multiple orderings of events in disease progression. *Inf. Process. Med. Imaging* **24**, 711–722 (2015)