Classifying patients and controls using multi-dimensional scaling and exploring the metric of semantic space

Steven Andrew Chance (steven.chance@clneuro.ox.ac.uk)
Neuroanatomy& Cognition Group (University of Oxford),Neuropathology,
Level 1, West Wing, John Radcliffe Hospital, Oxford, OX3 9DU, UK

Anthony C.D. James (tony.james@obmh.nhs.uk)
Highfield family and adolescent unit, Warneford Hospital, Oxford, UK

Rebecca Peet (rebecca.peet@hertford.ox.ac.uk)
Department of Statistics, University of Oxford, Oxford, UK

Geoffrey Nicholls (nicholls@stats.ox.ac.uk)
Department of Statistics, University of Oxford, Oxford, UK

Abstract

Multi-dimensional scaling (MDS) has been used to visualize the organization of semantic memory in normal control subjects and in psychiatric conditions such as schizophrenia and Alzheimer's disease. However, the potential for such techniques to classify subjects into diagnostic groups has not been realized. This study attempted to tackle this by developing classification statistics and by exploring the dimensional organization of semantic space using models with different underlying metrics. The test data were from controls and patients with early onset schizophrenia. The results indicated subtly altered semantic organization in schizophrenia, sufficient for novel classification statistics to correctly classify subjects as either patient or control with >80% accuracy.

Keywords: language; semantics; schizophrenia; verbal fluency; multi-dimensional-scaling; classification.

Introduction

Certain conditions, such as schizophrenia and Alzheimer's disease, are associated with disordered semantic memory (Huff et al, 1986; Phillips et al, 2004). The underlying alterations in memory organisation can be investigated using word generation tasks. Deficits include slower, disorganized generation of fewer exemplars from a semantic category. This data can then be analyzed with multidimensional scaling (MDS) methods (Rips et al, 1973) to visualize semantic networks and assess differences between patients and healthy control subjects (Chan et al, 1993). However, the refinement of these methods and application of additional data classification tools have not been sufficiently explored.

The present study emulates previous studies (Chan et al, 1993; Prescott et al, 2006) by analysing animal category fluency lists from schizophrenia patients and controls. We used MDS to identify axes in the data which represent the characteristics, or dimensions, people use to categorise animals (Chan et al, 1993). Typically, the two dominant dimensions characterizing semantic space for animals have been labeled as "size" and "domesticity" (a third dimension labeled "danger" may also be useful). A goal of this study was to identify aspects of the data which could be used to classify the individual as a patient or control.

A further issue to explore with MDS analysis is a possible diagnostic difference in the underlying metric of semantic space. Different metrics can be employed in MDS, probing differences in conceptual organisation (Shepard, 1987, Gardenfors, 2000). For example, various psychological spaces are usually better represented by the 'city-block' metric (Shepard, 1987) rather than the familiar Euclidean metric that has been used typically in MDS analyses (eg. Paulsen et al 1996). This is illustrated by a normal developmental shift. Older children and adults perceive dimensions such as high and tall, or big and bright, to be separable, whereas young children tend to confuse these concepts (Carey, 1978). This is seen as a developmental shift from a more Euclidean cognitive metric to the rigidly grid-like, separated dimensions of the city-block metric (Gardenfors, 2000). Such a shift may be disrupted in developmental disorders, particularly in early onset schizophrenia. The evidence that category boundaries are less clear in schizophrenia (beginning with Cameron, 1938) raises the prospect that the city-block metric may not provide a better fit for patients.

Semantic memory abnormalities appear to be worse in patients who had an earlier onset of illness (Paulsen et al, 1996) but adolescent patients with the early form of illness have not been adequately studied. The present study considers a group of adolescent subjects who have reductions in semantic fluency (as reported in Phillips et al, 2004). It was hypothesized that schizophrenia patients would exhibit differences in network organization and word

list generation that would enable statistical classification of subjects as either patient or control, based on their word list alone. Furthermore, a comparison of the metrics would show that control subjects use a city-block metric in the organisation of semantic space, whereas patients do not use such separable dimensions and therefore a Euclidean metric would be sufficient.

Methods

Subjects

The schizophrenia subjects were 36 patients with adolescent onset of illness between 12 and 18 diagnosed with DSM-IV criteria. Most patients had experienced only one psychotic episode. A further 31 non-psychiatric comparison subjects matched for age and education were recruited, with suitable exclusion criteria for both groups. Written informed consent was obtained from the subjects and their parents. Of the patients, 21 exhibited some negative symptoms and 6 were thought disordered.

Procedure

As reported previously for these patients (see Phillips et al, 2004 for more details on procedure) all subjects were examined with either the full version of the Wechsler Intelligence Test for children – III-R or, if older than 16 years, the Wechsler Adult Intelligence Test – Revised. The animal fluency test analysed here was administered by a neuropsychologist. Subjects were asked to name as many animals, excluding fish and birds, as possible in a period of 60 seconds. Any repetitions and non-animals were removed from the lists. The fifteen most common words across both groups were used to generate semantic network representations using multidimensional scaling (MDS). Timing data was not available.

Data Analysis

MultiDimensional Scaling: MDS embeds data on the differences between objects in a mathematical space with a number of dimensions determined by the user. The result is a map of the objects whose positions reflect their statistical difference (similar objects are close together). The objects in this case are words, whose differences are calculated from their distance apart in the lists of animals generated by each subject. For example, in the list: "dog, cat, elephant, tiger..", cat is 1 place from dog and 2 places from tiger.

The "Fluent" software was used to create a dissimilarity matrix for each diagnostic group based on the word list distances. The software controls for differences in list length by the mean cumulative frequency (MCF) method described in Prescott et al (2006). The resulting dissimilarity matrix for controls and the matrix for patients were used for MDS analysis.

Metric scaling was applied using automated routines in the "R" statistics software package. The selection of the number of dimensions in which to represent semantic space is governed by a measure of the stress value, which reflects the degree to which the resulting word map is a good summary of the actual distance data. 5% stress = a good fit, 10% stress = fair, 20% = poor. Low dimensional (2D) fits were used as is standard practice for visualisation, although a 4D fit was the best and this was used for the more subtle comparison of different metrics.

Roughness Statistic

The dissimilarities between animals were used to condense the information contained in word lists down to a single value representing the 'roughness' of an individual list. In order for the statistic to be comparable between patients and controls, the dissimilarites used were those calculated for controls. The 'roughness' statistic, r, for a list is defined to be the average dissimilarity between adjacent items in the list. For a list of length n it is calculated as follows:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n-1} D_{a_i, a_{(i+1)}}$$

Where ai is the animal in position i of the list, and D is the MCF dissimilarity measure.

Variability statistic

Another statistic formulated using the dissimilarities calculated for controls is the 'variability' in the 'roughness' of each list. The 'variability' statistic, v, for a list is defined to be the sample variance of the dissimilarity between adjacent items in the list. For a list of length n it is calculated as follows:

$$v = \frac{1}{(n-2)} \sum_{i=1}^{n-1} (D_{a_i, a_{(i+1)}} - r)^2$$

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a statistical classification tool. LDA was applied to the data in order to predict the diagnosis, patient or control, of an individual based upon their animal word list. LDA was applied to the data with roughness and variability as the observable variables.

Alternative metrics

An alternative metric scaling solution was calculated using non-Euclidean, City-Block MDS. Most published analyses of clinical samples assume that the semantic space underlying the solution is Euclidean. However, as described above, psychological space is often better represented by a non-Euclidean metric such as the 'city-block' metric

(Arabie, 1991). The metric is so-called because the distance between concepts is measured as if restricted to a grid-like system of roads (hence 'city-block' or 'Manhattan' metric) rather than "as the crow flies" in Euclidean space. The apparent suitability of this metric is due to the psychological tendency to make categorical decisions based on orthogonal conceptual dimensions (such as height and width) which are not arbitrary and interchangeable.

Consequently, following an established approach (Gardenfors, 2000;Johannesson, 1996), the stress values of the solutions were compared to determine the better fit using two different metrics with increasing numbers of dimensions up to the most desirable 4D fit. Optimization was performed by starting the procedure from multiple (hundreds of thousands) of initial configurations.

Results

The usual conceptual dimensions of animal "size" and "domesticity" did not form clear orthogonal axes for either word map (see Figures 1 and 2). However, the different dimensions, size, domesticity, and "danger" appeared much more correlated for patients than for controls (see metric comparisons and separability of dimensions, below).The word maps illustrated show the estimated binary transition line between one side of a conceptual dimension, and the other, for each of three dimensions (size, domesticity, and danger).

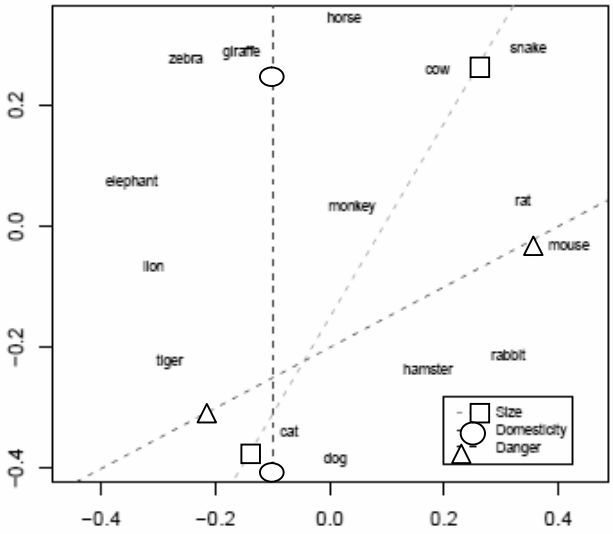

Figure 1: MDS for controls (squares, circles, and triangles simply differentiate boundary lines in gray scale image)

Controls represented size on a diagonal from bottom right to upper left. No consistent representation of the size dimension could be discerned in the patient map. However, there were interpretable groupings in both maps. For controls – household pets were on the lower right side of the map, large farm animals (horse, cow) to the upper center, and African/zoo animals on the left. For patients, the groupings were also logical although zebra overlapped with the farm animals and monkey was in a relatively extreme position. Overall, the patient semantic map did not differ markedly from that of controls, although it did have a slightly more idiosyncratic organization.

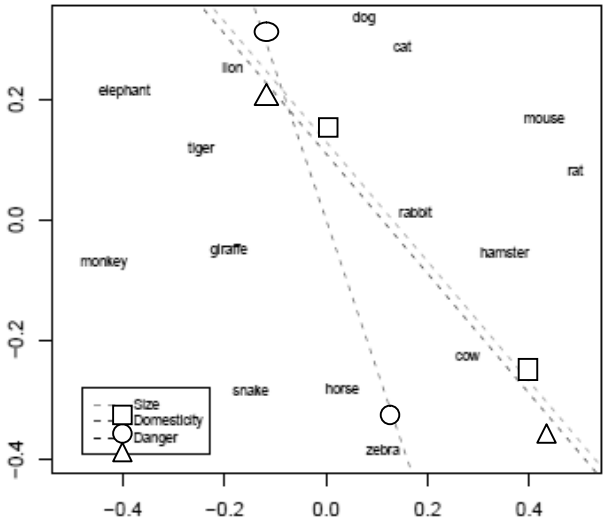

Figure 2: MDS for schizophrenia patients (squares, circles and triangles differentiate boundary lines in gray scale image and do not represent data points)

MDS enables wordlists to be shown visually. Wordlists for patients and controls are shown in Figure 3 using the points given by the MDS mapping for controls, to allow direct comparison.

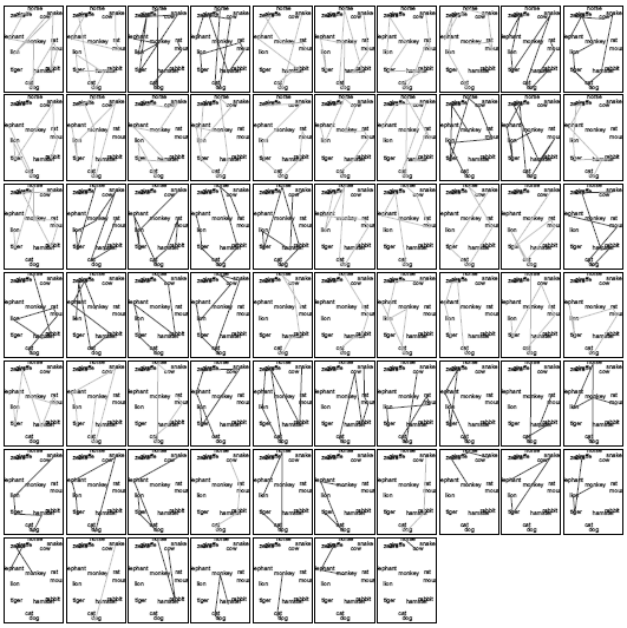

Figure 3: Animal wordlists plotted as networks on the MDS mapping for controls. Details are not visible, but the matrix is shown to enable the reader to visualise the method. Wordlists are displayed in order of decreasing list length. Light gray indicates that the list belongs to the controls. Dark gray indicates that the list belongs to the patients.

An important observation is that, although wordlist routes through semantic space for controls tended to appear more disorganised than for patients, this was not the case. The 'messy' appearance of control lists was generally caused by the fact that controls are more inclined than patients to exhaust a particular category of animals, such as small, domestic, safe animals, before branching out into another category.

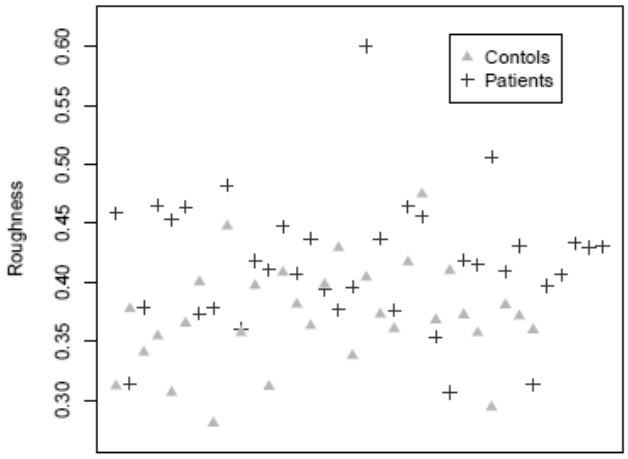

Figure 4: Roughness scores

Figure 4 shows the roughness scores for schizophrenia patients and controls. The roughness scores for patients appear to have a greater spread than those for controls. This suggests that the two sets of roughness scores follow different distributions for the different groups. A two sample Kolmogorov-Smirnov test for this gives a p-value of 0.0017, confirming the observation. This makes roughness an appropriate variable for use with classification tools.

Classification

Leave-One-Out Cross-Validation was used to assess the quality of the classification procedure, the result being that this procedure classified 81% of individuals correctly (see Figure 5).

As observed already, the typical patient list length is shorter than controls and, as might be expected, the incorporation of list length assists classification. The emphasis of the present study was to attempt to identify components of semantic organization itself that could be used to classify subjects, regardless of list length. However, it may be noted that further classification testing indicated that the inclusion of list length largely superceded the variability statistic. Meanwhile the roughness statistic continued to provide useful additional classification information.

This LOO-CV could be improved for our particular application if, when removing control points from the LDA, we also removed the control's word list, recomputed the dissimilarities for controls and hence adapted the roughness and variability scores. However, this would be computationally expensive, and the difference in the estimate of the number of correctly classified individuals would be small. Therefore, for our purpose, the simpler version of LOO-CV was applied.

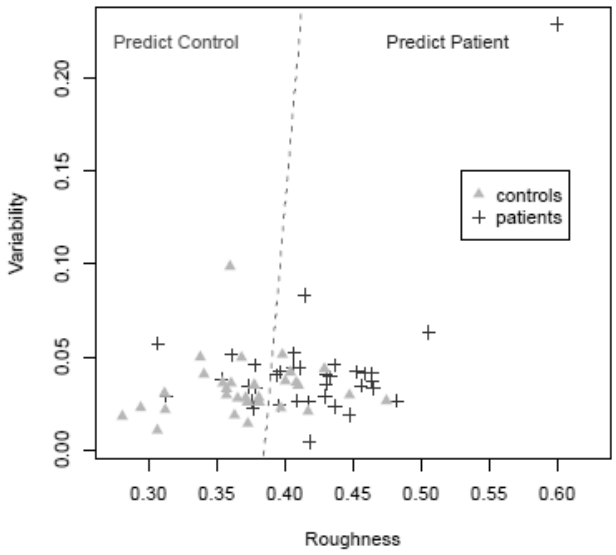

Figure 5: LDA using roughness and variability to predict patient/control status

Metric comparison

Generating distance measures with a city-block metric, using automated routines, resulted in an improved fit of the scaling solution for control subjects. However, in schizophrenia, the improvement was reduced when the number of dimensions increased to 4 dimensions.

The stress value of the Euclidean scaling solution for controls at 4-D was 10.3% (R2= 0.97) (borderline "poor"/"fair"), for patients it was 9.7% (R2= 0.98) ("fair"). The Euclidean solutions for patients and controls were not improved with further optimisation, but the fully optimized city-block stress was lower for both diagnoses (control stress = 7%, patient stress = 7.6%). Overall, the Euclidean solution was better for patients than for controls, and the patients benefited less from the city-block solution than controls. However, comparing optimization between metrics is challenging since the process of optimization is technically easier for city-block than Euclidean.

Discussion

Previous MDS studies have tended to assess older patients with chronic schizophrenia (eg. Chen et al, 2000) or dementia (Chan et al, 1993), concluding that the deficit is due to semantic store degradation. The present study found reduced fluency but fairly mild abnormality of semantic organisation early in the illness. It may be argued that as these patients represent an early onset group with recent diagnosis, anomalies of semantic organisation in these cases are unlikely to be due to the progressive degradation of memory storage caused by long illness duration and medication. However, adolescent patients generally represent a more severe form of illness with worse outcome and less exposure to medication. A solution to the apparent contradiction is that reduced fluency causes the appearance of a bigger difference in semantic organisation due to shorter word list length. Unlike several other studies (seeStorms et al, 2003) our stress values for patients indicated a better Euclidean fit for the MDS solutions than controls. We suggest this is, at least partly, due to the Prescott et al (2006) list length correction.

Further analysis using linear discriminant analysis illustrated that two novel statistical derivations from the word lists: "roughness" and "variability", are useful for classification of subjects into controls or patients with >80% accuracy. The addition of list length to classification algorithms renders the inclusion of the variability statistic largely redundant. However, the roughness statistic appear to provide significant additional information enabling correct classification. For further contrasts between diagnoses a qualitative semantic network assessment revealed that although wordlists for controls tend to appear more disorganised than those for patients, this is not the case. The 'messy' appearance is because controls are more inclined than patients to exhaust a particular category of animals, such as small, domestic, safe animals, before branching out into another category.

The nature of further abnormal organization in schizophrenia may be interpreted by the difference in the metric of semantic space. The present results demonstrate that semantic space is similar to various psychological spaces that are normally better represented by the 'city-block' metric than the commonly used Euclidean metric (Arabie, 1991). The sharp-cornered form of the city-block metric better models the natural tendency to discriminatory learning in which the orthogonal axes reflect the way conceptual dimensions (such as width and height, or domesticity and size) are not arbitrary and interchangeable as they are in the Euclidean metric. While we found that the city-block metric was always best for control subjects, the results were more ambiguous for patients suggesting less separated conceptual dimensions in schizophrenia.

These results suggest that the developmental shift (see Gardenfors, 2000), from a Euclidean cognitive metric to the more separable dimensions of the city-block metric, has not completed successfully in early onset schizophrenia.Goldstone and Barsalou (1998) have described the development of reasoning about dimensions: "evidence suggests that dimensions that are easily separated by adults, such as the brightness and size of a square, are treated as fused together for children… [they] have difficulty identifying whether two objects differ on their brightness or size even though they can easily see that they differ in some way. Both differentiation and dimensionalization occur throughout one's lifetime."

In conclusion, the alteration of semantic networks in schizophrenia appears to be more subtle than that indicated by some previous studies (Paulsen et al, 1996). The results are consistent with the suggestion that larger differences identified in previous studies are partly due to statistical confounds such as differences in list length (Prescott et al, 2006) and sub-optimal solutions. However, using the additional innovative techniques for data exploration and classification shown here, fitting a low dimensional (2-D) Euclidean word map to the data reveals differences in semantic organisationbetween patients and controls. The passage through semantic space revealed by MDS visualization of control word lists appears "messy" because controls are more inclined than patients to exhaust a particular category grouping of animals (eg. small, domestic), before exploring another sub-cluster. Although list length differentiates controls from patients, two alternative statistical derivations from the word lists: "roughness" and "variability", can be used to classify subjects into controls or patients with >80% accuracy. A further novel finding is the indication of less separated conceptual dimensions in schizophrenia compared to controls.

References

Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika*, 56, 567-87.

Cameron, N. (1938). Reasoning, regression and communication in schizophrenics.*Psychological Monographs*, 50(1), 1–33.

Carey, S. (1978). The child as a word learner.In M. Halle, J. Bresnan& Miller G. (Eds.),*Linguistic theory and psychological reality*. MIT press, Cambridge, Massachusetts

Chan, A.S., Butter, N., Salmon, D. P. &McGuire, K.A. (1993).Dimensionality and clustering in the semantic network of patients with Alzheimer's disease.*Psychology and Aging*, 3, 411-419.

Chen, R. Y. L., Chen, E. Y. H., Chan, C. K. Y., Lam, L. C. W. &Lieh-Mak, F. (2000). Verbal fluency in schizophrenia: reduction in semantic store. *Australian and New Zealand Journal of Psychiatry*, 34, 43-48.

Gardenfors, P. (2000).*Conceptual spaces: the geometry of thought*. MIT press, Cambridge, Massachusetts.

Goldstone, R. L. &Barsalou, L.W. (1998).Reuniting perception and conception.*Cognition*, 65, 231-262.

Huff, F. J., Corkin, S. &Growdon, J. H. (1986). Semantic Impairment and Anomia in Alzheimer's Disease. *Brain and Language*, 28 (2), 235-249.

Johannesson, M. (1996).Obtaining psychologically motivated spaces with MDS. Lund: *Lund University Cognitive Studies*, 45.

Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K. &Jeste, D. V. (1996).Impairment in the semantic network in schizophrenia.*Psychiatry Research*, 63, 109-121.

Phillips, T.J., James, A.C., Crow, T. J. &Collinson, S.L. (2004). Semantic fluency is impaired but phonemic and design fluency are preserved in early-onset schizophrenia. *Schizophrenia Research*, 70, 215-22.

Prescott, T.J., Newton, L.D., Mir, N.U., Woodruff, P.W. R. & Parks, R.W. (2006). A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia.*Neuropsychology*, 20(6), 685-699.

Rips, L. J., Shoben, E. J. & Smith, E. E. (1973) Semantic Distance and the Verification of Semantic Relations.*Journal of Verbal Learning and Verbal Behaviour*, 12, 1-20.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science.*Science*, 237, 1317-23.

Storms, F., Dirikx, T., Saerens, J., Verstraeten, S.& De Deyn P.P. (2003).On the use of scaling and clustering in the study of semantic deficits.*Neuropsychology*, 17, 289-301.