A Tutorial on Multiblock Discriminant Correspondence Analysis (MUDICA): A New Method for Analyzing Discourse Data From Clinical Populations

Lynne J. Williams
University of Western Ontario,
London, Ontario, Canada

Hervé Abdi
The University of Texas at Dallas,
Richardson, TX

Rebecca French
Southlake Regional Health Centre,
Newmarket, Ontario, Canada

Joseph B. Orange
University of Western Ontario,
London, Ontario, Canada

Purpose: In communication disorders research, clinical groups are frequently described based on patterns of performance, but researchers often study only a few participants described by many quantitative and qualitative variables. These data are difficult to handle with standard inferential tools (e.g., analysis of variance or factor analysis) whose assumptions are unfit for these data. This article presents *multiblock discriminant correspondence analysis* (MUDICA), which is a recent method that can handle datasets not suited for standard inferential techniques.

Method: MUDICA is illustrated with clinical data examining conversational trouble-source repair and topic maintenance in dementia of the Alzheimer's type (DAT). Seventeen DAT participant/spouse dyads (6 controls, 5 participants with early DAT, 6 participants with moderate DAT) produced spontaneous conversations analyzed for co-occurrence of trouble-source repair and topic maintenance variables.

Results: MUDICA found that trouble-source repair sequences and topic transitions are associated and that patterns of performance in the DAT groups differed significantly from those in the control group.

Conclusion: MUDICA is ideally suited to analyze language and discourse data in communication disorders because it (a) can identify and predict clinical group membership based on patterns of performance, (b) can accommodate few participants and many variables, (c) can be used with categorical data, and (d) adds the rigor of inferential statistics.

KEY WORDS: multiblock discriminant correspondence analysis, discourse, qualitative data analysis, discriminant analysis, inferential tool

In clinical research, it is necessary to identify the patterns of performance that discriminate clinical groups. Yet, speech and language researchers are often faced with the problem of finding a suitable method to address this question. This problem occurs because of two primary characteristics of speech and language data: First, the studies typically have few participants but many variables. Second, the studies frequently use categorical, or qualitative, variables. Because these types of data are not suitable for methods such as analysis of variance (ANOVA) or factor analysis, researchers often feel compelled to report only the frequency of occurrence of each variable rather than using inferential procedures. However, some recently developed data analysis methods can apply the rigor of statistical analysis to the questions and data specific to clinical research in our discipline.

In this article, we present *discriminant correspondence analysis* (DICA) and its extension, *multiblock discriminant correspondence analysis* (MUDICA), which are methods that can show relationships between clinical diagnostic groups described by categorical or qualitative variables. DICA and MUDICA also can identify diagnostic group membership based on patterns of performance. We illustrate DICA and MUDICA by analyzing a dataset collected to reveal the relationship between conversational trouble-source repair and topic maintenance in dementia of the Alzheimer's type (DAT).

DICA

DICA is used when variables are collected describing observations (e.g., participants, conversational dyads, etc.) obtained from a priori defined groups (e.g., control vs. clinical groups) and when it is necessary to (a) assess whether group membership explains some variance of the observations, (b) find out what variables are important to discriminate between the groups, and (c) predict group membership of new observations. The main idea behind DICA is to represent each a priori group, observation, and variable as a point on a map such that the positions of these points reflect the important features of the data.

In clinical research, groups are, in general, diagnostic groups. DICA is appealing for analyzing clinical data because the proximity between group points in the DICA maps represents their similarity, and the proximity between variable points represents their association (Abdi, 2007a).

In DICA, the "style" of a priori groups is examined because DICA is sensitive to the groups' relative use of all the variables rather than to the absolute number of occurrences of each variable. That is, DICA analyzes the variable frequencies for each group. (These frequencies are obtained by dividing each row entry by the sum of the variables for that row.)

Formally, DICA combines the features of discriminant analysis (see Klecka, 1980) and correspondence analysis (see Abdi & Williams, 2010c; Greenacre, 1984, 2007; see also Abdi & Valentin, 2007; Le Roux & Rouanet, 2010)[1] to perform a type of discriminant analysis appropriate for qualitative data. Because it is derived from correspondence analysis, which is a model-free technique, DICA is also model free and therefore does not impose parametric distributional assumptions such as normality or homogeneity of variance. In addition, DICA can handle data sets with few observations described by many qualitative

or quantitative variables. MUDICA, an extension of DICA, can be used to examine group performance on a subset of the variables included in the DICA analysis. All of these features make DICA—and, by extension, MUDICA—ideal tools for clinical research.

A Real-World Example: Trouble-Source Repair and Topic Maintenance in Dementia of the Alzheimer's Type (DAT)

Background

Individuals with DAT have problems participating in meaningful conversations, and these problems worsen as the disease progresses (Guendouzi & Müller, 2002; Orange & Colton-Hudson, 1998; Orange, Lubinski, & Higginbotham, 1996; Orange, Van Gennep, Miller, & Johnson, 1998; Watson, Chenery, & Carter, 1999). This deterioration in conversational ability creates more conversational breakdowns, which, in turn, require caregivers to spend more time and effort repairing these breakdowns (Orange et al., 1996, 1998).

As DAT progresses, conversational partners take on more responsibility for initiating and maintaining conversation as well as for negotiating trouble-source repairs. This results in differences in how individuals with DAT and their conversational partners signal breakdowns. For example, partners use repair initiators that require a specific response (e.g., "You played what?"), whereas individuals with DAT are less interactive and less specific in their requests for repair (e.g., "err"; Watson et al., 1999).

In addition to having more conversational breakdowns, DAT participants also experience topic management problems. Typically, they have difficulty introducing, changing, elaborating on, and maintaining the topic of conversation as compared with their conversational partners (Mentis, Briggs-Whittaker, & Gramingna, 1995). Consequently, DAT participants introduce and unexpectedly shift topics more frequently (Garcia & Joanette, 1994, 1997).

Because individuals with DAT have trouble with both conversation breakdown and topic management, it is expected that the amount of conversational breakdown is associated with the amount of topic management skills. But, curiously, there have been no systematic studies examining co-occurrence of conversational breakdown and topic maintenance. This led us to ask the following research questions: (a) Do conversational trouble-sources occur during topic transitions in DAT participant/spouse conversational dyads; and (b) Does dementia severity affect the pattern of conversational trouble-source repair at topic transitions?

---

[1]Correspondence analysis was developed by the French school of data analysis expressly for analyzing relationships in linguistic data.

Dataset

Diagnostic groups and participants. We examined trouble-source repair and topic transitions in three clinical groups: control (CTRL), early stage DAT (EDAT), and middle stage DAT (MDAT). DAT participants were diagnosed with probable DAT according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association criteria (McKahnn et al., 1984). We established the clinical stage based on scores from the Standardized Mini-Mental State Examination (Molloy, Alemayheu, & Roberts, 1991) and the Global Deterioration Scale (Reisberg, Ferris, De Leon, & Crook, 1982). Twelve participants made up 6 DAT participant/spouse conversational dyads in both the CTRL and MDAT groups, and 10 participants made up 5 DAT participant/spouse dyads in the EDAT group. (The sixth EDAT dyad was excluded from the analysis because the spouse produced no trouble-source repair sequences.) Spouses were selected as conversational partners because partner familiarity promotes natural conversation; ensures that individuals are familiar with one another's verbal, nonverbal, and idiosyncratic cues; and minimizes adjustments in communication style (Santo Pietro, 1994). Participant demographic characteristics are given in Table 1.

Procedure. Midday and evening mealtime conversations were videorecorded on a single day. The midday session was used to habituate the dyads to the recording equipment. All recording sessions occurred in the participants' homes in the location where the participants usually ate their meals (e.g., kitchen, dining room, family room). Participants were given no specific instructions on topics to discuss or on how to interact with one another. The examiner was not present during either recording session. Conversations ranged from 111 to 561 utterances ($M = 322.71$). All spouses reported that the recorded conversations were typical of their daily interactions.

Because we were interested in the pattern of communication between DAT participants and their spouses, we considered that dyads constituted the observations for our analysis. We used 36 variables to describe the trouble-source repair sequences produced by each participant. We used the same variables to code trouble-source repair sequences initiated by each member of the dyad, making a total of 72 variables per dyad. Note that the trouble-source repair sequences are not equally represented by all of the variables, which fall into three broad categories: (a) trouble-source repair, (b) topic, and (c) trouble-source repair sequencing. These variables and their definitions are shown in Appendix A.

Interrater reliability. Four raters recoded a random 16% of the conversational samples (as per Mentis et al., 1995). For trouble-source repair sequences at topic boundaries, percentage agreement with the original rater was 75%.

DICA

To perform DICA, we start with a data table in which each row represents an observation (i.e., a dyad) and each column represents a variable. In our example, the

Table 1. Demographics of participants, spouses, and conversational dyads.

Participants	Age (years) M	Age (years) SD	Education (years) M	Education (years) SD	SMMSE M	SMMSE SD	GDS M	GDS SD	DAS M	DAS SD
					CTRL dyads (*n* = 6)					
CTRL	69.54	5.72	13.17	2.32					114.83	9.81
Spouse	66.93	5.92	13.50	4.37					112.00	6.07
					EDAT dyads (*n* = 5)					
DAT	68.42	6.39	14.8	3.03	25.00	1.55	3.40	0.49		
Spouse	65.80	4.35	11.6	3.21					117.20	8.01
					MDAT dyads (*n* = 6)					
DAT	74.47	4.35	13.42	2.62	15.67	2.07	5.67	0.52		
Spouse	75.00	6.19	12.25	1.72					116.00	8.69

Note. There were no significant differences on age, $F(2, 33) = 2.94$, $p > .05$, MSE = 41.08; education, $F(2, 14) = 0.07$, $p > .05$, MSE = 8.52; or the Dyadic Adjustment Scale, $F(2, 20) = 0.05$, $p > .05$, MSE = 69.39. All spouses and the control (CTRL) participants completed the Dyadic Adjustment Scale (DAS; cutoff score = 97; Spanier, 1976). SMMSE = Standardized Mini-Mental State Examination (Maximum score = 30; Molloy et al., 1991); GDS = Global Deterioration Scale (Scores range from 1 to 7; Reisberg et al., 1982); EDAT = participants in early stage dementia of the Alzheimer's type; DAT = participants with dementia of the Alzheimer's type; MDAT = participants in middle stage dementia of the Alzheimer's type.
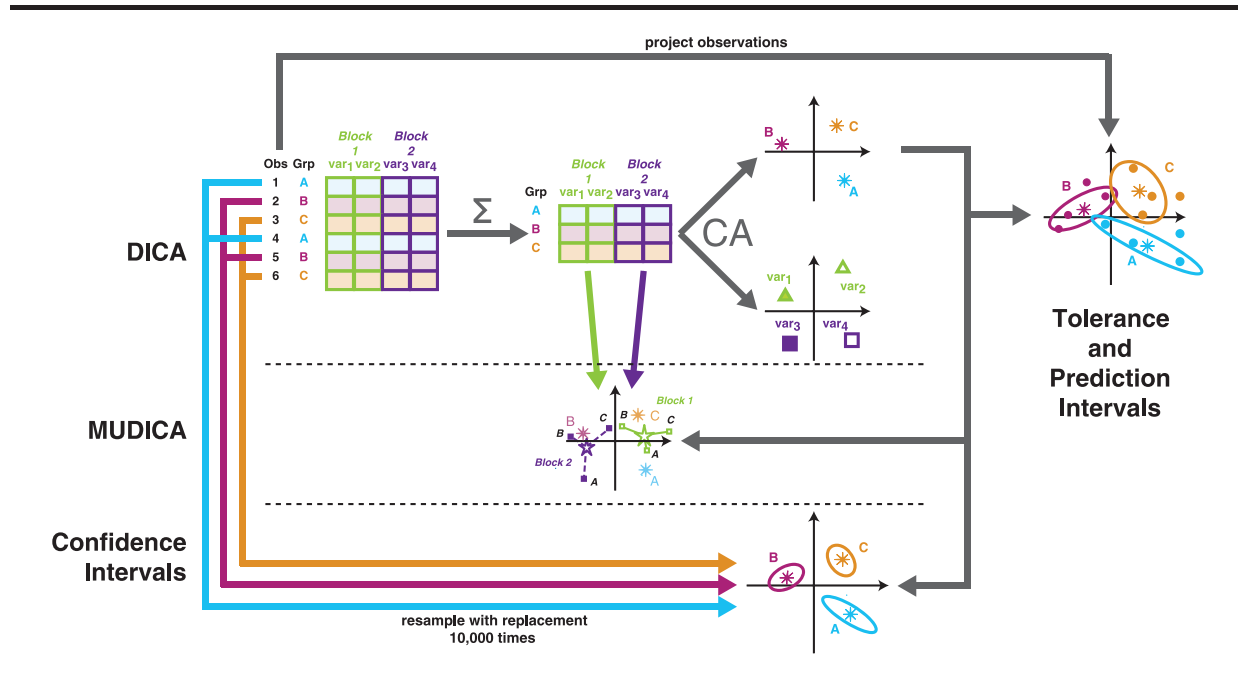
original data table is a contingency table in which each row is a dyad and each variable counts the number of occurrences of a given behavior. DICA per se is performed on a Group × Variable contingency table. This table gives the number of occurrences of each variable for all the observations of a given diagnostic group. (The Group × Variable contingency table for the trouble-source repair and topic dataset is shown in Appendix B; also see Figure 1.) Correspondence analysis (see Figure 1) is then applied to this Group × Variable contingency table. (*Correspondence analysis* is the equivalent of principal component analysis for qualitative data; see, e.g., Abdi & Williams, 2010c, for an introduction.) From the contingency table, correspondence analysis computes new variables called *factors,* which are combinations of the original variables. The values of these new variables are called *factor scores.* The scores of the first factor have the largest possible variance; therefore, the first factor "explains" the largest possible part of the variance of the data. The second factor is statistically independent of the first factor (i.e., these two factors are uncorrelated) and accounts for the largest possible amount of the remaining variance. The other factors are computed likewise. Correspondence analysis produces two sets of

factors scores—one for the rows and one for the columns of the tables—and, importantly, these two sets of factor scores have the same variance. To create a map of the groups and the variables, their factor scores are used as coordinates. This process is illustrated in Figure 1. These factor scores can also be interpreted geometrically as the *projections* of the groups and of the variables onto the factors. After the analysis of the Group × Variable contingency table has been performed, the original observations are then projected as points on the factor map. For each factor, the mean of the factor scores of the observations of a group is equal to the factor score of this group.

Interpreting DICA

Once the maps are generated, how are they interpreted? The points in the map represent the observations/groups and the variables. The factor scores of the groups and the factor scores of the variables have the same variance; therefore, observations/groups and variables can be plotted on the same map. However, the proximity between two points can be directly interpreted only when these two points belong to the same set (e.g., the proximity

Figure 1. Steps to run discriminant correspondence analysis (DICA): Step 1. From the original Observation (Obs) × Variable (Var) contingency table, take the sum of each variable within each a priori determined category (usually diagnostic groups [Grp] in clinical research). Step 2. Run correspondence analysis (CA) on the Category (diagnostic group) × Variable contingency table to produce the row and variable DICA maps. Steps to run a multiblock discriminant correspondence analysis (MUDICA): Step 3. Divide the Group × Variable contingency table into two or more blocks, each representing only a subset of the variables for all groups. Step 4. Project the blocks into the original DICA space as supplementary elements. Steps to add an inferential step to DICA and MUDICA: Step 5. Project the observations and the jackknifed observations into the DICA space as tolerance or prediction intervals. Step 6. Generate 10,000 samples with replacement from the original data using bootstrap resampling. Step 7. Project the 10,000 samples into the original DICA space as supplementary elements and as confidence intervals. Step 8. Trim tolerance, prediction, and confidence points to 95% intervals and replace them with ellipses.

of two variables can be interpreted directly, but not so for the proximity of a variable and a group). Accordingly, we show here the observations/groups and the variables in separate maps. The proximity of the points representing the observations/groups expresses their similarity, and the proximity of the variables represents their association (i.e., their correlation; see Abdi & Williams, 2010b). As previously mentioned, the interpretation of the proximity between observations/groups and variables is delicate, but a useful rule of thumb is that when a variable and an observation/group fall within the same quadrant of the map, this variable is more associated to this group than to the averages of the other groups.

What Are the Important Factors?

The importance of a factor is given by the amount of variance that it explains. This amount, called an *eigenvalue,* is represented by the Greek letter λ (lambda). In DICA (in contrast to principal component analysis), the eigenvalues are always smaller than 1. The importance of a factor is also expressed as the proportion or percentage of the total variance explained by a factor. This proportion is represented by the Greek symbol τ (tau). The maximum number of possible factors in DICA is 1 fewer than the number of groups or variables (whichever is smaller).

The procedure for interpreting the factors in DICA is similar to that used for principal component analysis (see Abdi & Williams, 2010e, for an example with principal component analysis). To find the important groups or variables, the first step is to compute for each group or variable an index called its *contribution* to the factor—or, simply, its *contribution*. This contribution gives the proportion of an eigenvalue accounted for by a given group or variable. Considering that the sum of the squared factor scores for a given factor equals the eigenvalue of that factor, a *contribution* is defined as

$$\text{contribution} = \frac{(\text{factor score of group or variable})^2}{\text{eigenvalue}}.$$

Groups or variables with greater than the average contribution are considered to be *important*. The average contribution is computed as

$$\text{average group contribution} = \frac{1}{\text{number of groups}}.$$

Also, because factor scores can be either positive or negative, contributions can be interpreted as positive or negative. This means that the groups or variables that contribute the most to each "end" (i.e., pole) of the factor in the map help determine what the factor represents.

As an example, DICA produced the maps displayed in Figures 2 and 3. Note that in Figure 3, the variables

are shown in separate maps because the number of variables makes interpretation difficult when they are all presented in one map. Keep in mind, however, that the variables are actually all in the same DICA space (i.e., they all come from the same map).

The DICA found two factors: Factor 1 (λ = .08), which accounts for 66% of the total variance, and Factor 2 (λ = .04), which accounts for the remaining 34% of the total variance. Because we have only two factors, the percentage of the variance accounted for by these two factors sums to 100. If we had more factors, the sum of the first two factors would be smaller than 100.
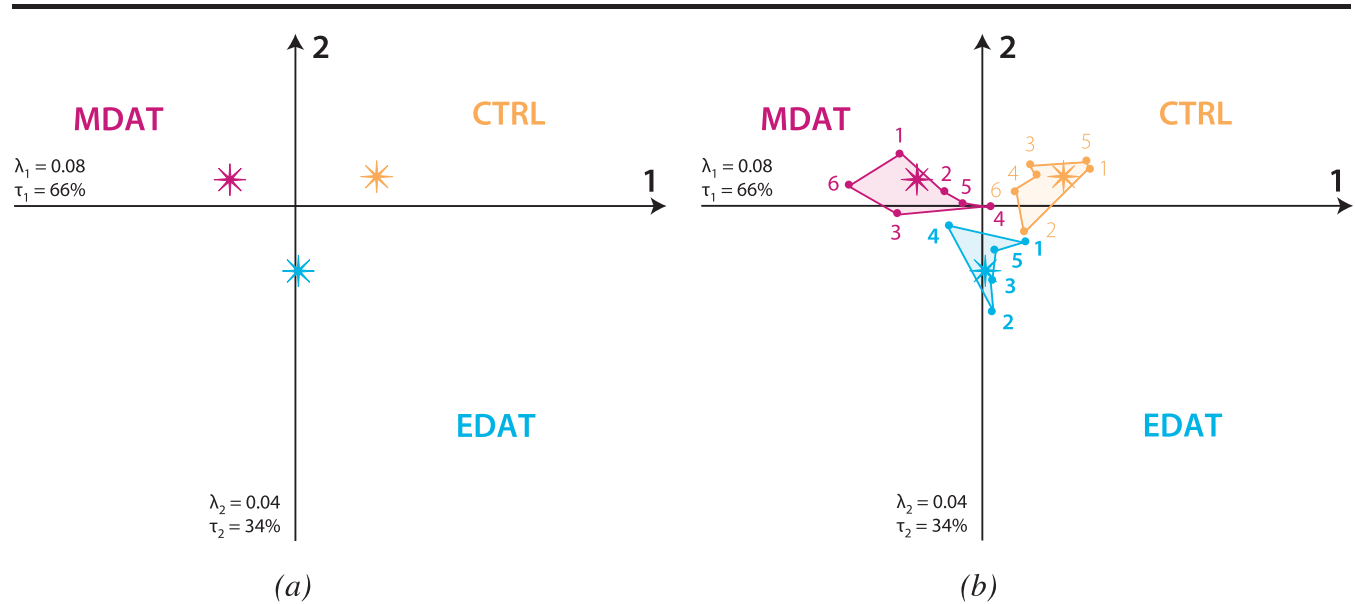
Now that the importance of the factors has been determined, a decision needs to be made regarding how many factors to keep. Deciding which factors to keep depends on their eigenvalues, the research hypothesis, and what makes sense given the dataset and the research questions. In our example, we have three groups, and because the maximum number of factors is the number of rows or columns minus 1 (whichever is smaller), we have only two factors in our example. Yet, in datasets with more than three groups, DICA finds, in general, more than two factors. When this occurs, a choice needs to be made regarding the number of factors to keep.

A *scree plot* is a useful tool to help determine which factors to keep. A scree plot is a line-segment plot showing the proportion of the total variance in the data accounted for by each factor. The factors are ordered from those accounting for the most to the least variance. When read from left to right across the *x*-axis, scree plots often show the point of separation between the "most important" and "least important" factors. This point of separation is often called the *elbow* (Abdi & Williams, 2010e). However, scree plots act only as guidelines. It is necessary to maintain a critical eye and decide on the number of factors based on what makes the most sense given the dataset.

Interpreting the Observations and Groups (Rows)

Interpreting the meaning of the factors is done in two stages: first by looking at the observations/groups and then by looking at the variables. The relative position of the observations/groups is shown in Figure 2. Recall that, for our example, the groups represent the responses of the CTRL, EDAT, and MDAT dyads. Looking at Factor 1 (plotted on the horizontal axis) in Figure 2a, we see that the CTRL and MDAT dyads are contrasted. For Factor 2 (plotted on the vertical axis), the EDAT dyad is contrasted with the CTRL and MDAT dyads. To confirm this interpretation, we look at the contributions of the three groups to each factor. From Table 2 we can see that, indeed, this interpretation is correct. The CTRL and MDAT groups contribute the most to Factor 1, whereas

Figure 2. DICA map of rows (observations) representing a priori diagnostic groups. (a) DICA map of the placement of the control (CTRL), early stage dementia of the Alzheimer's type (EDAT), and middle stage dementia of the Alzheimer's type (MDAT) groups along Factors 1 and 2. Factor 1 separates the CTRL and MDAT groups. Factor 2 separates the EDAT group from the CTRL and MDAT groups. (b) DICA map of the placement of the CTRL, EDAT, and MDAT groups with the DAT participant/spouse dyads projected into the DICA space as supplementary elements. The projections of the dyads confirm that Factor 1 separates the CTRL and MDAT groups. All subfigures are plotted at the same scale. High contributors to Factor 1 include CTRL and MDAT groups. High contributors to Factor 2 include the EDAT group. Factor 1: $\lambda_1 = 0.08$, $\tau_1 = 66\%$. Factor 2: $\lambda_2 = 0.04$, $\tau_2 = 34\%$. Note that in DICA λ is always smaller than 1.

the EDAT group contributes the most to Factor 2. Because the contributions for each factor always sum to 1, the average contribution of the three groups to a given factor is

$$\text{average group contribution} = \frac{1}{\text{number of groups}}$$
$$= \frac{1}{3} \approx 0.33.$$

Supplementary elements. Supplementary elements are a useful way to enrich the interpretation of the DICA factors. Supplementary elements are additional observations/groups (or variables) that are not included as part of the DICA computations. Rather, they are observations, groups, or variables that are projected into the DICA space after it has been computed. This shows where the observations, groups, or variables would have fallen had they been included in the analysis. For the sake of illustration, we projected the original dyads into the DICA space as supplementary elements (see Figure 2b). This shows the actual dispersion of the original dyads around the groups. The dispersion of the dyads confirms that Factor 1 differentiates the CTRL and MDAT groups.

Interpreting the Variables (Columns)

The relative position of the variables is shown in Figures 3a through 3f. Recall that, for convenience, the

variables are shown in different maps but they are all part of the same DICA space (and come from the same map). When there are so many variables, determining which ones are the most important to a given factor can be difficult when only looking at the map. Note that in DICA, variables that rarely occur contribute more to the factors because in correspondence analysis, rare variables have much importance and, therefore, could uniquely define a factor. When there are a lot of variables (or groups), it is helpful to start with the variables having large contributions and then return to the map(s). Determining the importance of a variable to a factor is done in the same way as is done for the groups. That is, the important variables have contributions larger than the average contribution[2] (i.e., 1 ÷ number of variables).

When contributions are used to select the important variables, it can be seen that Factor 1 reveals the following contrasts: (a) DAT-initiated trouble-source repair sequences at topic introductions versus DAT-initiated trouble-source repair at two utterances following topic introductions and spouse-initiated trouble-source repair sequences at topic introductions (see Figure 3a); (b) DAT- and spouse-initiated reintroduced topics and subtopics versus spouse-initiated new topics and subtopics (see Figure 3d); (c) DAT discourse and other trouble sources

---

[2]The contributions of the variables are available for download at http://www.utdallas.edu/~herve.

Figure 3. DICA map of variables. Note that all variables are represented in the same DICA space and have been separated for easier viewing. (a) Trouble sources at topic, subtopic, and social unit boundaries. (b) Number of trouble sources and trouble-source repair sequence complexity. (c) Trouble sources at topic shifts and changes. (d) Trouble sources at new and reintroduced topics. (e) Trouble-source types. The sub-subscript represents the dyad member who initiated the trouble source. (f) Trouble-source resolution success. DAT_X = trouble-source repair sequence initiated by DAT participant. SP_X = trouble-source repair sequence initiated by spouse. T = topic boundary. ST = subtopic boundary. SU = social unit boundary. X_{+1} = 1 utterance after boundary. X_{+2} = 2 utterances after boundary. $X_{>2}$ = more than 2 utterances after boundary. $X_{X_{DAT}}$ = trouble source by DAT participant. $X_{X_{SP}}$ = trouble source by spouse. Disc = discourse trouble source. Phon = phonological trouble source. Sem = semantic trouble source. Cog = cognitive trouble source. Other = other trouble source. High contributors to Factor 1 include cognitive ($DAT_{Cog_{DAT}}$), other ($DAT_{Other_{DAT}}$), and discourse ($DAT_{Disc_{DAT}}$) trouble source produced by the individual with DAT in DAT participant-initiated trouble-source repair sequences as well as trouble source at topic shifts (SP_{Shift}) and trouble source at new topic introductions ($SP_{T(New)}$) in spouse-initiated trouble-source repair sequences. High contributors to Factor 2 include semantic trouble source produced by the individual with DAT ($DAT_{Sem_{DAT}}$) in DAT participant-initiated trouble-source repair sequences as well as trouble source at topic introduction (SP_T), trouble source at subtopic introduction plus 1 utterance (SP_{ST+1}), trouble source at social unit introduction plus 1 utterance (SP_{SU+1}), trouble source at greater than 2 utterances after topic introduction ($SP_{T>2}$), and trouble source at topic shift (SP_{Shift}) in spouse-initiated trouble-source repair sequences.
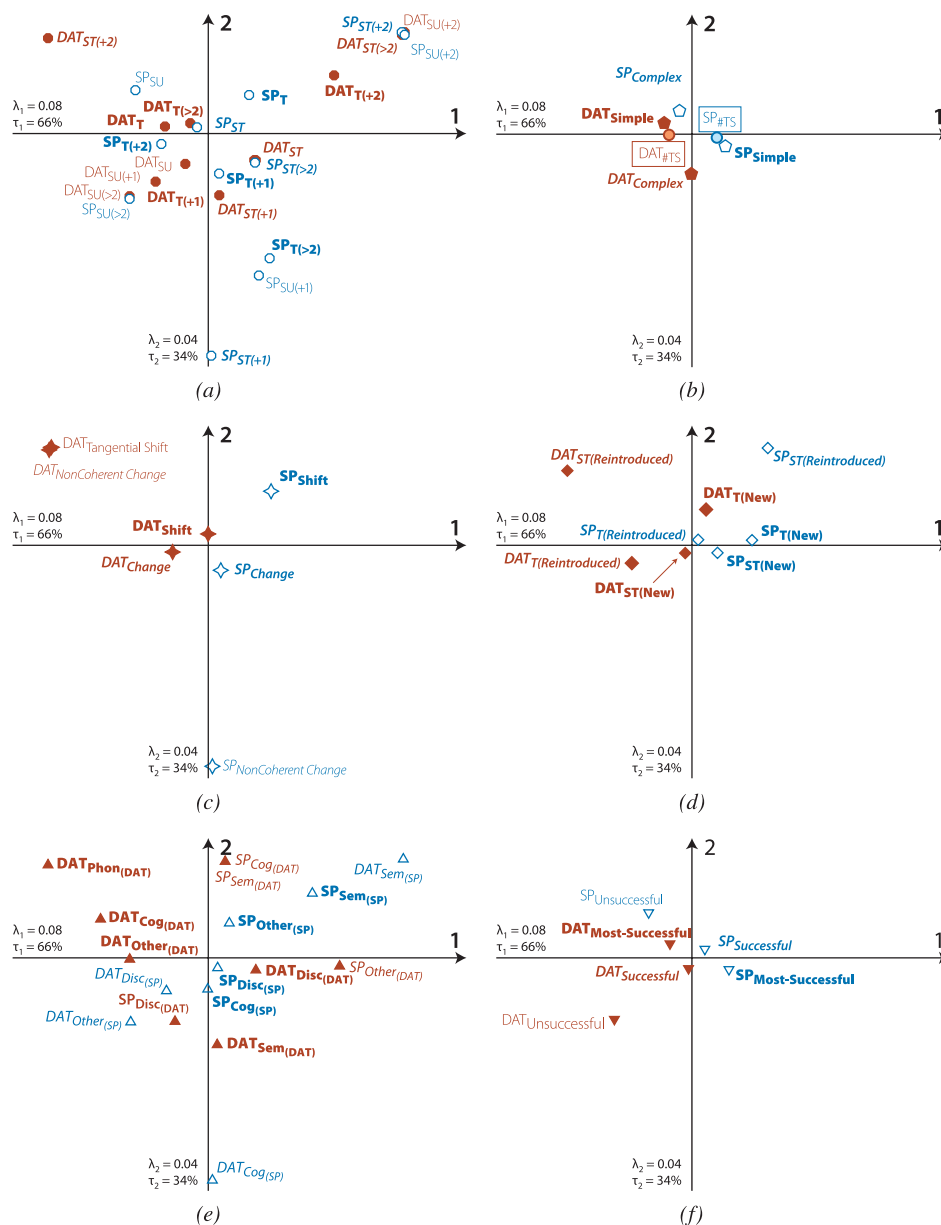
Table 2. Contributions of the groups to Factors 1 and 2.

Group	Factor 1	Factor 2
CTRL	0.544	0.145
EDAT	0.000	0.694
MDAT	0.456	0.161

versus DAT cognitive trouble source in DAT-initiated trouble-source repair sequences (see Figure 3e); and (d) new topics versus topic shifts in spouse-initiated trouble-source repair sequences (see Figures 3d and 3c, respectively). All together, this suggests that Factor 1 represents the difference in performance between the individuals with DAT and their spouses, with the individuals with DAT producing more discourse and other types of trouble-source repair sequences at both new and reintroduced topics.

From the map in Figure 3, we also see that Factor 2 contrasts the following variables: (a) DAT-initiated trouble-source repair sequences at one utterance following topic introductions and spouse-initiated trouble-source repair sequences at two utterances following topic introduction versus DAT-initiated trouble-source repair sequences at two utterances following topic introductions and spouse-initiated trouble-source repair sequences at topic introductions, at two utterances following topic introductions, and at one utterance following social unit introductions (see Figure 3a); (b) DAT-initiated new and reintroduced topics, tangential topic shifts, and spouse-initiated reintroduced topics and topic shifts versus spouse-initiated topic changes and noncoherent topic changes (see Figures 3c and 3d); (c) DAT semantic trouble source in DAT-initiated trouble-source repair sequences, DAT cognitive trouble source in spouse-initiated trouble-source repair sequences, and spouse discourse trouble source in DAT-initiated trouble-source repair sequences versus DAT cognitive trouble source in DAT-initiated trouble-source repair sequences, spouse semantic, and other trouble sources in spouse-initiated trouble-source repair sequences (see Figure 3e). All together, this suggests that Factor 2 represents instances when both the individuals with DAT and their spouses encounter trouble sources. The individual with DAT had more difficulty at topic transitions and following their spouse's reintroduction of a previous topic. In contrast, the spouses had more difficulty one utterance past the topic transition, suggesting that they are not understanding the DAT individual's utterance.

Putting It All Together

Now that the factors have been defined in terms of both the observations/groups and the variables of the original data set, it is necessary to integrate the information from both the rows and the columns to interpret the factors. For Factor 1, individuals with DAT produce

discourse and other trouble sources when a topic or subtopic is reintroduced. In contrast, spouse-initiated trouble-source repair sequences occur when there is a shift to a new topic, whereas DAT-initiated trouble-source repair sequences occur when there is a cognitive trouble source. Because the first set of variables falls to the left of the origin of the axes, one can say that cognitive and other trouble sources at reintroduced topics and subtopics characterize the performance of the MDAT group. The second set of variables, falling to the right of the origin, characterizes the performance of the CTRL group. For Factor 2, the occurrence of trouble sources at topic boundaries other than topic introductions (i.e., greater than two utterances after a topic introduction and one utterance past a subtopic or social unit introduction) characterizes the performance of the EDAT group, whereas the second set of variables is more characteristic of the CTRL and MDAT groups' performances.
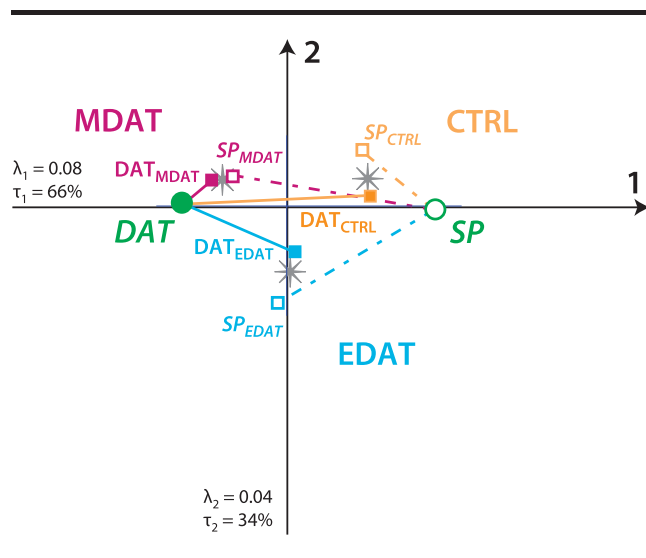
Analyzing "Blocks" of Variables: Multiblock Discriminant Correspondence Analysis (MUDICA)

Thus far, DICA characterizes the patterns of performance of the three diagnostic groups. However, researchers often want to know more about some subsets of the variables. A common example would be to examine differences in discourse performance on language- versus memory-related variables. In our example, we wanted to analyze the differences in patterns of performance between the trouble-source repair sequences initiated by the DAT participants and by their spouses because we expected that they would play different roles in their conversations. To analyze these differences, we used a MUDICA. To perform MUDICA, we divided the Group × Variable contingency table into two blocks or subtables (see Figure 1): One block represented the trouble-source repair sequences initiated by the DAT participants, and the other block represented the trouble-source repair sequences initiated by the spouses. We computed the average performance of the DAT participants and of their spouses. Then, we projected the averages of the two blocks into the original factor space as supplementary elements. This procedure is outlined in Figure 1.

In our example, MUDICA produced the map shown in Figure 4. The map(s) produced by MUDICA are interpreted as the observations/groups and variable maps previously described. Factor 1 separates the two blocks. Note that trouble-source repair sequences initiated by the spouses closely resemble the performance of the CTRL group in the original DICA.

In our example, we also could have projected the three types of variables as blocks (i.e., trouble-source repair, topic, and sequencing variables). The decision on

Figure 4. MUDICA. The blocks representing trouble-source repair sequences initiated by the DAT participant (DAT) and the spouses (SP) were projected into the DICA space as supplementary elements.

which blocks to project depends on the research question of interest.

From Descriptive to Inferential

So far, DICA—and, by extension, MUDICA—provides a description of the patterns of variables that differentiate the diagnostic groups. However, DICA does not assess the quality or the robustness of this description. This assessment requires an inferential step, which determines the quality of the discrimination for the observations used to create the discriminative model, the quality of the discrimination for new observations, and the significance of the DICA results. To determine the quality of group assignment, an R^2 is computed, and an examination of how the DICA model assigns observations to groups is conducted. For purposes of determining the significance of the DICA model, confidence intervals (CIs) are computed around the groups.

The quality of DICA group assignment is determined in three ways. First, an examination is conducted regarding how much variance of the observations is explained by their groups. Second, the accuracy of the assignment both within the sample and for new observations is examined. Finally, the separability of the groups is examined.

Quality of Group Assignment: R^2 and the Permutation Test

As a way of evaluating the quality of group assignment, an R^2 is computed, which expresses the proportion of variance explained by group membership. In MUDICA, the *total variance* is the sum of the distances between all

points and the grand *barycenter* (i.e., the origin in the DICA/MUDICA maps). In a way analogous to ANOVA— for which the sum of squares total is equal to the sum of the between-groups sum of squares and the within-groups sum of squares—MUDICA decomposes the total variance (i.e., total inertia) into between-groups and within-groups variance. Specifically, R^2 is computed as

$$R^2 = \frac{\text{between-groups variance}}{\text{total variance}}.$$

R^2 takes values between zero and 1 and is a squared correlation coefficient. If group membership is random, then R^2 is close to zero. If group membership is systematic, R^2 is close to 1. Therefore, a large R^2 value indicates that observations can be reliably assigned to the groups (Abdi & Williams, 2010a).

To assess significance of R^2, a *permutation test* is used; this test randomly assigns observations repeatedly to the groups and computes the R^2 values associated with each random group assignment. This gives a probability distribution for R^2 under the null hypothesis (i.e., when the assignment of the observations to the groups is random), and this distribution can then be used to derive p values. When the null hypothesis is rejected, the assignment of the observations to their groups is not due to chance. In our example, R^2 is .75 ($p < .001$). This confirms that group assignment is reliable.

Fixed Effect Model: Accuracy and Separability of Group Assignment Within the Sample

How well DICA classifies observations within the sample is called a *fixed effect model*.

Accuracy: Fixed Effect Confusion Matrix

To assign an observation to a group, DICA computes the distance between this observation and all groups and then assigns this observation to the closest group. Table 3 shows the DICA's assignment of the dyads within our sample to the CTRL, EDAT, and MDAT groups.

Within our sample, the observations are well classified by the DICA. It correctly assigned 5 of 6 CTRL

Table 3. Fixed effect model: Discriminant correspondence analysis (DICA) assignment of dyads within the sample to the CTRL, EDAT, and MDAT groups.

Assigned group	Actual group		
	CTRL	**EDAT**	**MDAT**
CTRL	5	0	0
EDAT	1	5	1
MDAT	0	0	5

dyads, all 5 EDAT dyads, and 5 of 6 MDAT dyads. The misclassified dyads in the CTRL and MDAT groups were both assigned to the EDAT group. This misclassification arises because of the variability of the EDAT and MDAT groups.

Separability: Tolerance Intervals

The accuracy of the assignment of the observations to their groups can be expressed graphically by *tolerance intervals* (see Figure 1). Tolerance intervals are computed so that they encompass a given proportion of the observations. In two dimensions, these intervals have the shape of ellipses and are often called *tolerance ellipsoids*. For example, a 95% tolerance interval indicates the range in which 95% of the observations from a given group fall, and, similar to the confusion matrix, the tolerance intervals represent how well the DICA assigns observations to the groups. In the tolerance interval map, when two groups do not overlap, they are separable (Abdi, Dunlop, & Williams, 2009).

For our example, we computed the 95% tolerance intervals (see Figure 5a). From the display, it is evident that the tolerance intervals for all three groups overlap. Therefore, the EDAT, MDAT, and CTRL groups are not separable. However, the small amount of overlap indicates that the DICA model shows good performance.

Random Effect Model: Accuracy of Group Assignment of New Observations

Knowing how well DICA assigns observations within our sample is important. However, from the standpoint of clinical researchers, the real interest lies in classifying new clients or participants whose diagnosis is unknown.

Although the fixed effect model gives a good representation of the variability within the sample, the fixed effect tolerance intervals overestimate the separability of groups in the population. This is because the same observations are used both to develop and to test the model, and this can result in gross overestimation of the predictive performance of the model (Cureton, 1950; Kriegeskorte, Simmons, Bellogowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009). Correct estimation of the model's performance requires the use of different sets of observations to build and test the DICA model. This amounts to assigning *new* observations to groups and corresponds to a *random effect model.*

Accuracy: Random Effect Confusion Matrix

To evaluate the performance of DICA for a random effect model, we used a *jackknife* (i.e., a "leave one out") procedure (see Appendix C and Abdi & Williams, 2010d). The jackknife removes each observation in turn and recalculates the DICA without this observation.

The removed observation is then projected into the DICA space as a supplementary element, and the distance between that observation and each group is computed. The removed observation is then assigned to its closest group. Table 4 shows the random effect model assignment of new dyads to the CTRL, EDAT, and MDAT groups.

For the random effect model, observations from the CTRL group are well classified by the DICA—which,

Figure 5. (a) 95% tolerance intervals: Tolerance intervals represent the range in which 95% of each of the CTRL, EDAT, and MDAT groups' scores should fall in the population. Because the tolerance intervals for all three groups overlap, the CTRL, EDAT, and MDAT groups cannot be reliably separated. (b) 95% prediction intervals: Prediction intervals represent the range in which 95% of new CTRL, EDAT, and MDAT scores should fall in the population. Because the prediction intervals for all three groups overlap, the CTRL, EDAT, and MDAT groups cannot be reliably separated. (c) 95% confidence intervals (CIs): CIs represent the range in which the population parameter will fall 95% of the time. Because the CTRL group's confidence ellipse does not overlap with the EDAT and MDAT groups' ellipses, the CTRL group is significantly different from the EDAT and MDAT groups at the *p* < .05 level.
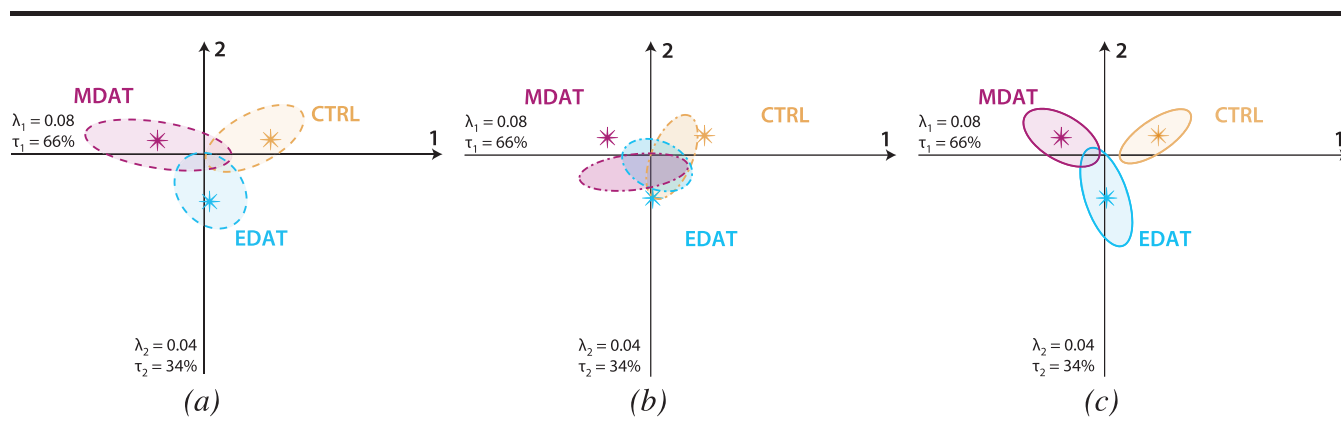
Table 4. Random effect model: DICA assignment of new dyads to the CTRL, EDAT, and MDAT groups.

Assigned group	Actual group		
	CTRL	**EDAT**	**MDAT**
CTRL	5	2	0
EDAT	1	2	4
MDAT	0	1	2

similar to that for the fixed effect model, correctly assigned 5 of 6 dyads (the misclassified dyad was assigned to the EDAT group). DICA, however, performed less well in assigning new observations to the EDAT and MDAT groups. DICA correctly classified 2 of the dyads to the EDAT and MDAT groups, respectively. From the 3 misclassified dyads in the EDAT group, 2 were assigned to the CTRL group and 1 was assigned to the MDAT group. The 4 misclassified dyads from the MDAT group were all assigned to the EDAT group. The group assignment using the random effect model suggests that it is unlikely that a new CTRL dyad will be misclassified as an MDAT dyad (and vice versa). However, both the CTRL and the MDAT dyads could potentially be misclassified as an EDAT dyad. In addition, a new EDAT dyad could be assigned to any of the diagnostic groups. The random effect model supports the interpretation of increased variability in trouble-source repair at topic boundaries in the EDAT and MDAT groups.

Separability: Prediction Intervals

As in the fixed effect model, the separability of the groups in a random effect model can be examined through computation of a *prediction interval* for each group (see Figure 1). A 95% prediction interval indicates the range in which 95% of the population should fall and, therefore, represents how well the DICA assigns *new* observations to the groups. In the prediction interval map, when two groups do not overlap, their corresponding populations are separable (Abdi et al., 2009).

To compute the prediction intervals, we again used the jackknife procedure on the original observations and projected the jackknifed observations in the DICA space. The 95% prediction intervals are shown in Figure 5b. Similar to the groups in the fixed effect model, the EDAT, MDAT, and CTRL groups overlap and cannot be reliably separated. This confirms the fixed effect findings that the groups cannot be reliably separated. However, note that in a random effect model, prediction intervals—unlike tolerance intervals—are not centered around the sample means. The difference between the group mean and the mean of the corresponding prediction interval reflects the "bias of the estimation" because a group mean is only an estimator of the population mean.

Determining Significance of Group Differences: Bootstrap CIs

Finally, the determination needs to be made as to whether the group differences in the DICA model are statistically significant. To do this, a 95% CI is computed for each group. The American Psychological Association recommends using CIs over standard null hypothesis testing because CIs specify the range of values that likely includes the population parameter of interest (American Psychological Association, 2001; Wilkinson & the Task Force on Statistical Inference, 1999). If CIs are computed from samples taken repeatedly from the population, a population sampling distribution can be created. A certain percentage of the samples from the population (usually set a priori to 95%) will contain the parameter (Easton & McColl, n.d.). The population sampling distribution can be estimated by sampling within our sample using the *bootstrap,* a nonparametric resampling technique used to estimate sampling distributions (Efron & Tibshirani, 1993; Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005).

For the bootstrap, a large number of new samples of the diagnostic groups (usually 1,000 or 10,000) are created. The new samples for the CTRL, EDAT, and MDAT groups have the same number of observations as the original groups and are obtained by sampling the dyads with replacement within their respective diagnostic groups. When resampling with replacement is conducted, each observation is put back into the sample after it has been drawn; therefore a given observation can be drawn several times, once, or not at all (see Figure 1).

We resampled within the CTRL, EDAT, and MDAT groups 10,000 times. We then projected the observations from the bootstrapped samples into the original DICA space as supplementary elements (see Figure 1). The resulting CIs are shown in Figure 5c. To make the graph easier to read, the dispersion of the dyads around each diagnostic group is represented by a 95% confidence ellipse centered on the group. The confidence ellipses are read the same way as CIs. That is, when two confidence ellipses do not overlap, they represent different populations, and the corresponding groups can be declared significantly different at the $\alpha = .05$ level (Abdi et al., 2009).

As shown in Figure 5c, the 95% confidence ellipse for the CTRL group did not overlap with the ellipses from the EDAT and MDAT groups. Therefore, the CTRL group can be considered significantly different from the EDAT and MDAT groups ($p < .05$). However, there was no significant difference between the EDAT and MDAT groups because their CIs overlap.

Discussion

Overall, the DICA found that topic boundaries did affect trouble-source repair sequences in spousal

conversational dyads in which one member has probable DAT. In the early stage, DAT dyads had more difficulty with subtopic and social unit boundaries after a topic transition. This suggests a reduction in the cognitive flexibility required to shift and/or change topics and confirms the findings of Mentis et al. (1995) and Garcia and Joanette (1994, 1997). In middle-stage DAT, dyads began to show cognitive trouble sources when a topic or subtopic was reintroduced. This suggests that topic and subtopic reintroduction may function as a type of event boundary and also suggests that individuals with DAT may experience more difficulty due to the increased memory load inherent in boundaries/transitions (Speer & Zacks, 2005). As such, individuals with middle-stage DAT may have trouble recalling previously introduced information and have trouble using shared knowledge with their spouse. However, these differences between the EDAT and MDAT groups may be more a matter of degree than type because the groups could not be reliably separated.

These findings have important implications for understanding the nature of conversation breakdowns in DAT and their association with topic management. Understanding these associations is crucial for developing successful communication enhancement education and training programs and testing empirically based enhancement strategies for spousal caregivers of individuals with DAT (Savundranayagam, Hummert, & Montgomery, 2005).

In this article, we discussed DICA, which is a method that describes patterns of performance of a priori determined diagnostic groups. Using DICA, the relationships between groups and between variables can be displayed in two maps. With the addition of a hierarchical or multi-block component (MUDICA), it is also possible to visualize the groups' performance(s) on blocks representing subsets of variables. Furthermore, inferential steps can be added that show the reliability of the analysis through confidence and tolerance intervals. These inferential steps help determine whether the a priori designated groups significantly differ and how well the DICA model categorizes old and new observations. As such, DICA provides an ideal method to analyze language and discourse data in communication disorders research, especially datasets with few observations described by a large number of qualitative and quantitative variables.

Software

DICA and MUDICA are based on correspondence analysis, which is implemented by most statistical packages such as SAS (PROC CORRESP) and SPSS/PASW (the CATEGORIES model). Therefore, in principle, any standard package can be used as long as the data are correctly preprocessed (see Appendix C for technical details).

The freely available package (R) incorporates several libraries dedicated to correspondence analysis (e.g., ca, FactoMineR, and ade4, which also incorporates DICA). A library (written by Derek Beaton) specially written to implement the MUDICA analyses described here is available from the home page of the second author (www.utd.edu/~herve). Finally, the set of MATLAB programs and the dataset used for this study are also available from the home page of the second author.

Acknowledgments

References

Abdi, H. (2007a). Discriminant correspondence analysis. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 270–275). Thousand Oaks, CA: Sage.

Abdi, H. (2007b). Distance. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 280–284). Thousand Oaks, CA: Sage.

Abdi, H. (2007c). Eigen-decomposition: Eigenvalues and eigenvectors. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 304–308). Thousand Oaks, CA: Sage.

Abdi, H. (2007d). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 907–912). Thousand Oaks, CA: Sage.

Abdi, H., Dunlop, J. P., & Williams, L. J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage, 45,* 89–95.

Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 651–657). Thousand Oaks, CA: Sage.

Abdi, H., & Williams, L. J. (2010a). Barycentric discriminant analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 64–75). Thousand Oaks, CA: Sage.

Abdi, H., & Williams, L. J. (2010b). Coefficients of correlation, alienation and determination. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 171–180). Thousand Oaks, CA: Sage.

Abdi, H., & Williams, L. J. (2010c). Correspondence analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 267–278). Thousand Oaks, CA: Sage.

Abdi, H., & Williams, L. J. (2010d). Jackknife. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 655–660). Thousand Oaks, CA: Sage.

Abdi, H., & Williams, L. J. (2010e). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2,* 433–459.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10,* 94–96.

Easton, V. J., & McColl, J. H. (n.d.). *Statistics glossary*. Retrieved from http://www.stats.gla.ac.uk/steps/glossary/index.html.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman.

Garcia, L. J., & Joanette, Y. (1994). Conversational topic-shifting analysis in dementia. In R. L. Bloom, L. K. Obler, S. De Santi, & J. S. Ehrlich (Eds.), *Discourse analysis and applications: Studies in adult clinical populations* (pp. 161–183). Hillsdale, NJ: Erlbaum.

Garcia, L. J., & Joanette, Y. (1997). Analysis of conversational topic shifts: A multiple case study. *Brain and Language, 58,* 92–114.

Gottman, J. M. (1979). Time-series analysis of continuous data in dyads. In M. E. Lamb, S. J. Suomi, & G. R. Stephenson (Eds.), *Social interaction analysis: Methodological issues* (pp. 207–229). Madison, WI: University of Wisconsin Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London, England: Academic Press.

Greenacre, M. J. (2007). *Correspondence analysis in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Guendouzi, J., & Müller, N. (2002). Defining trouble-sources in dementia: Repair strategies and conversational satisfaction in interactions with an Alzheimer's patient. In F. Windson, M. L. Kelly, & N. Hewlett (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 15–30). Mahwah, NJ: Erlbaum.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. In G. P. McCabe & D. S. Moore (Eds.), *Introduction to the practice of statistics* (pp. 14.1–14.70). New York, NY: W. H. Freeman Company.

Klecka, W. R. (1980). *Discriminant analysis*. Thousand Oaks, CA: Sage.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12,* 535–540.

Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. Thousand Oaks, CA: Sage.

McKahnn, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34,* 939–944.

Mentis, M., Briggs-Whittaker, J., & Gramingna, G. D. (1995). Discourse topic management in senile dementia of the Alzheimer type. *Journal of Speech and Hearing Research, 38,* 1054–1066.

Molloy, D., Alemayheu, E., & Roberts, R. (1991). Reliability of a Standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination. *American Journal of Psychiatry, 148,* 102–105.

Orange, J. B., & Colton-Hudson, A. (1998). Enhancing communication in dementia of the Alzheimer's type. *Topics in Geriatric Rehabilitation, 14,* 56–75.

Orange, J. B., Lubinski, R., & Higginbotham, D. (1996). Conversational repair by individuals with dementia of the Alzhiemer's type. *Journal of Speech and Hearing Research, 39,* 881–895.

Orange, J. B., Van Gennep, K. M., Miller, L., & Johnson, A. M. (1998). Resolution of communication breakdown in dementia of the Alzheimer's type: A longitudinal study. *Journal of Applied Communication Research, 26,* 120–138.

Reisberg, B., Ferris, S., De Leon, M., & Crook, T. (1982). The Global Deterioration Scale for assessment of primary degenerative dementia. *American Journal of Psychiatry, 139,* 1136–1139.

Santo Pietro, M. (1994). Assessing the communication styles of caregivers of patients with Alzheimer's disease. *Seminars in Speech and Language, 15,* 236–254.

Savundranayagam, M. Y., Hummert, M. L., & Montgomery, R. J. V. (2005). Investigating the effects of communication problems on caregiver burden. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 60,* S48–S55.

Shewan, C. (1988). The Shewan Spontaneous Language Analysis (SSLA) system for aphasic adults: Description, reliability, and validity. *Journal of Communication Disorders, 21,* 103–138.

Spanier, G. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and Family, 38,* 15–28.

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language, 53,* 125–140.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274–290.

Watson, C. M., Chenery, H. J., & Carter, M. S. (1999). An analysis of trouble and repair in the natural conversations of people with dementia of the Alzheimer's type. *Aphasiology, 13,* 195–218.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Contact authors:

Lynne J. Williams, who is now with the Kunin-Lunenfeld Applied Research Unit, Rotman Research Institute, Baycrest, 3560 Bathurst Street, Toronto, Ontario M6A 2E1, Canada. E-mail: lwilliams@klaru-baycrest.on.ca.

Hervé Abdi, The University of Texas at Dallas, GR4.1, 800 West Campbell Road, Richardson, TX 75080-3021. E-mail: herve@utdallas.edu.

Appendix A (p. 1 of 2). Definitions of variables.

Trouble-Source Repair Variables

We coded trouble-source repair sequences using a modified version of the trouble-source repair variables described by Orange, Lubinski, and Higginbotham (1996); Orange and Colton-Hudson (1998); and Orange, Van Gennep et al. (1998).

The term *trouble-source repair variables* refers to problems related to speaking, hearing, attending, or understanding. They represent and identify problems in the interaction between conversation partners. The term *trouble sources* relates to an incongruence of the intent and the understanding of a speaker and a listener and may result from difficulties in the output of the speaker or mishearings by the listener (Orange & Colton-Hudson, 1998, p. 62).

Trouble-Source Type

<u>Semantic (Sem)</u>: Disturbances related to lexical access, word recall, word retrieval, and accurate or known word use.

<u>Discourse (Disc)</u>: Difficulties relating to the listener's apparent comprehension of topic content (e.g., maintenance, change, accuracy, initiation), shared knowledge (e.g., clarity and relevance), and cohesion (e.g., referencing problems).

<u>Cognitive (Cog)</u>: Related to memory impairment; previously discussed information is reintroduced as if not previously mentioned or discussed.

<u>Phonological (Phon)</u>: Mispronunciations, slips of the tongue, and poor knowledge of rules for sound combinations.

<u>Morphological/Syntactical (Morph)</u>: Disturbances in grammatical and syntactical rule systems (e.g., time and possession markers, agreement, and word order problems).

<u>Other</u>: Problems that cannot be unambiguously classified (e.g., abandoned/incomplete utterances, unrepaired utterances, no indication in repair initiation to nature of trouble source, direct repetitions without paralinguistic/nonverbal adjustments).

Trouble-Source Repair Sequence Complexity

The term *trouble-source repair sequence complexity* refers to whether or not there are embedded or secondary trouble sources in the trouble-source repair sequence.

<u>Simple</u>: Trouble-source repair sequence consists of a single trouble source, repair initiator, and repair or a trouble source and repair without a repair initiator.

<u>Complex</u>: Trouble-source repair sequence contains a primary trouble source and one or more embedded or secondary trouble sources.

Degree of Resolution Success

The term *resolution variables* refers to the outcome of the repairing process. Resolutions describe how trouble sources are overcome by participants.

<u>Most successful</u>: A single trouble source, repair initiator, and repair or a single trouble source and repair with no repair initiator; partners continue with the conversation on topic or move appropriately to a new topic.

<u>Successful</u>: More than one repair initiator and repair are used to successfully repair trouble source(s); partners continue the conversation on topic or appropriately move to a new topic.

<u>Unsuccessful</u>: More than one repair initiator and repair are used to attempt to repair trouble source(s); trouble sources are not repaired; may result in continuation of the conversation on topic or may result in abrupt and inappropriate turn-taking, topic shift/change, or termination of the conversation.

Topic Variables

We coded topic information using a modified version of the taxonomies described by Garcia and Joanette (1994, 1997) and Mentis et al. (1995).

Topic Types

<u>Global topic</u>: Each utterance within a topic sequence expresses the central concept of theme being addressed based on shared background knowledge of the interlocutors; may be stated explicitly or implicitly.

<u>Subtopic</u>: An associated but distinct concept or theme that is related back to the global topic.

<u>Topic unit (T)</u>: A set of continuous utterances relating to the same global topic without being separated by introduction or renewal (reintroduction) of another global topic (see "global topic").

<u>Subtopic unit (ST)</u>: A set of continuous utterances appearing to relate to the same subtopic without being separated by introduction or renewal (reintroduction) of another subtopic or global topic (see "subtopic").

<u>Social unit (SU)</u>: An utterance or sequence of utterances that addresses an element within the immediate social context and fulfills a social convention (e.g., politeness: "Would you like some butter?").

Topic Introductions

<u>New topic (T{New})</u>: A topic that has not occurred previously in the conversation.

<u>Reintroduced topic (T{Reintroduced})</u>: A topic that has occurred previously in the conversation unrelated to the prior topic.

<u>New subtopic (ST{New})</u>: A subtopic that has not previously occurred in the conversation but remains connected to the global topic.

<u>Reintroduced subtopic (ST{Reintroduced})</u>: A subtopic that has previously occurred in the conversation but is unrelated to the prior global topic or subtopic.

Manner of Topic Introduction

<u>Change</u>: The content of the new or reintroduced topic is not derived from the prior topic sequence.

<u>Shift</u>: The topic sequence under discussion is the source for the introduction of a new topic.

<u>Noncoherent change</u>: The (re)introduction of a topic in the absence of an established topic boundary or utterance signaling a topic transition.

<u>Tangential shift</u>: The topic sequence under discussion is used to lead the conversation in an irrelevant or confusing direction.

Trouble-Source Sequencing Variables

We examined co-occurrence of trouble-source repair variables and topic variables using methods modified from Gottman (1979). For each dinnertime conversation, we orthographically transcribed all utterances and placed them in sequential order. We then mapped the trouble-source repair variables and the topic variables onto the sequenced utterances and examined where they occurred in relation to each other. We coded utterances using Shewan's (1988) definition, which states that an *utterance* is a complete idea or thought expressed in connected words and is differentiated from other utterances on the basis of content, intonational contour, and/or pausing.

Trouble Source at Topic Boundary

At topic boundary (subscript T): Trouble source occurs at topic transition.

At topic boundary +1 utterance (T{+1}): Trouble source occurs one utterance after topic transition.

At topic boundary +2 utterances (T{+2}): Trouble source occurs two utterances after topic transition.

At topic boundary + more than 2 utterances (T{>2}): Trouble source occurs more than two utterances after topic transition.

Trouble Source at Subtopic Boundary

At subtopic boundary (subscript ST): Trouble source occurs at subtopic transition.

At subtopic boundary +1 utterance (ST{+1}): Trouble source occurs one utterance after subtopic transition.

At subtopic boundary +2 utterances (ST{+2}): Trouble source occurs two utterances after subtopic transition.

At subtopic boundary + more than 2 utterances (ST{>2}): Trouble source occurs more than two utterances after subtopic transition.

Trouble Source at Social Unit Boundary

At social unit boundary (subscript SU): Trouble source occurs at social unit transition.

At social unit boundary +1 utterance (SU{+1}): Trouble source occurs one utterance after social unit transition.

At social unit boundary +2 utterances (SU{+2}): Trouble source occurs two utterances after social unit transition.

At social unit boundary + more than 2 utterances (SU{>2}): Trouble source occurs more than two utterances after social unit transition.

Table B1. Group × Variable contingency.

	Individual with DAT						
Group	#TS	T_{intro}	$T_{intro+1}$	$T_{intro+2}$	$T_{intro>2}$	ST_{intro}	$ST_{intro+1}$	$ST_{intro+2}$
CTRL	31	14	1	6	3	2	1	0
EDAT	38	20	5	1	3	2	2	0
MDAT	56	37	5	1	5	1	1	1

	Individual with DAT						
Group	$ST_{intro>2}$	SU_{intro}	$SU_{intro+1}$	$SU_{intro+2}$	$SU_{intro>2}$	T_{new}	$T_{reintroduced}$
CTRL	3	0	3	3	0	1	0
EDAT	0	2	2	0	1	8	21
MDAT	0	2	2	0	1	17	31

	Individual with DAT					
Group	Change	Shift	Noncoherent change	Tangential shift	St_{new}	$ST_{reintroduced}$
CTRL	12	12	0	0	27	1
EDAT	19	10	0	0	32	1
MDAT	28	15	1	4	38	13

	Individual with DAT					
Group	$TS_{discourse_{DAT}}$	$Ts_{cognitive_{DAT}}$	$TS_{semantic_{DAT}}$	$TS_{phonological_{DAT}}$	$TS_{other_{DAT}}$	$TS_{discourse_{spouse}}$
CTRL	28	2	3	0	3	3
EDAT	22	5	8	0	10	8
MDAT	15	22	3	1	20	9

Group	Individual with DAT					
	$TS_{cognitive_{spouse}}$	$TS_{semantic_{spouse}}$	$TS_{other_{spouse}}$	TS_{simple}	$TS_{complex}$	$Res_{most\text{-}successful}$
CTRL	0	1	0	24	7	22
EDAT	1	0	1	26	12	22
MDAT	0	0	1	47	9	39

Group	Individual with DAT		Spouse				
	$Res_{successful}$	$Res_{unsuccessful}$	TS	T_{intro}	$T_{intro+1}$	$T_{intro+2}$	$T_{intro>2}$
CTRL	9	0	47	29	4	1	3
EDAT	11	5	40	11	6	2	7
MDAT	12	5	39	20	4	3	0

Group	Spouse							
	ST_{intro}	$ST_{intro+1}$	$ST_{intro+2}$	$ST_{intro>2}$	SU_{intro}	$SU_{intro+1}$	$SU_{intro+2}$	$SU_{intro>2}$
CTRL	4	0	1	2	1	1	1	0
EDAT	4	3	0	2	1	3	0	1
MDAT	6	0	0	1	4	0	0	1

Group	Spouse						
	T_{new}	$T_{reintroduced}$	Change	Shift	Noncoherent change	ST_{new}	$ST_{reintroduced}$
CTRL	22	15	16	21	0	40	4
EDAT	13	13	20	5	1	35	0
MDAT	10	17	16	11	0	32	2

Group	Spouse					
	$TS_{discourse_{DAT}}$	$TS_{cognitive_{DAT}}$	$TS_{semantic_{DAT}}$	$TS_{other_{DAT}}$	$TS_{discourse_{spouse}}$	$TS_{cognitive_{spouse}}$
CTRL	1	1	1	2	29	4
EDAT	4	0	0	1	30	6
MDAT	3	1	1	0	31	5

Group	Spouse					
	$TS_{semantic_{spouse}}$	$TS_{other_{spouse}}$	TS_{simple}	$TS_{complex}$	$RES_{most\text{-}successful}$	$RES_{successful}$	$RES_{unsuccessful}$
CTRL	7	13	39	8	35	9	3
EDAT	1	6	34	6	31	7	2
MDAT	2	12	27	12	23	9	7

Note. In complex trouble-source repair sequences, individual trouble sources may come from the individual with DAT or the spouse, regardless of who initiated the trouble-source repair sequence. TS = trouble source; T = superordinate topic; intro = topic introduction; +1 = plus one utterance; +2 = plus two utterances; >2 = plus greater than 2 utterances; ST = subordinate topic; SU = social unit; Res = trouble-source resolution success.

Multiblock discriminant correspondence analysis (MUDICA) extends discriminant correspondence analysis (DICA) to take into account the fact that the data matrix is structured in multiblocks. The first step of MUDICA is a DICA, which is followed by a specific analysis that incorporates multiblocks. The goal of DICA is to predict group membership of observations that are described by nominal variables (or by variables which represent the amount of some quantity). For MUDICA, the variables are also partitioned into blocks of variables, and the interest is to analyze the effect of these blocks on group membership.

Notations

We have I observations each described by J variables. The values of the variables for the observations are stored in an I by J data matrix denoted **X**. The observations of **X** are partitioned into N *a priori* groups of interest, with I_n being the number of observations of the nth group (and, so, $\sum_n^N I_n = I$). The columns of matrix **X** can be arranged in K a priori blocks (or subtables). The number of columns of the kth block are denoted J_k (and, so, $\sum_k^K J_k = J$). So, the I by J matrix **X** can decomposed into N by K blocks as

$$
\mathbf{X} = n \begin{array}{c} \\ 1 \\ \vdots \\ n \\ \vdots \\ N \end{array}
\begin{array}{ccccc} 1 & \cdots & k & \cdots & K \\ \left[\begin{array}{ccccc}
\mathbf{X}_{1,1} & \cdots & \mathbf{X}_{1,k} & \cdots & \mathbf{X}_{1,K} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{X}_{n,1} & \cdots & \mathbf{X}_{n,k} & \cdots & \mathbf{X}_{n,K} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\mathbf{X}_{N,1} & \cdots & \mathbf{X}_{N,k} & \cdots & \mathbf{X}_{N,K}
\end{array}\right] \end{array} \tag{1}
$$

where $\mathbf{X}_{n,k}$ is an I_n by J_k matrix, which corresponds to the nth group and the kth block. The elements of **X** are assumed to be positive or zeros, and it is assumed that there are no empty rows or columns (i.e., rows or columns with only zero values).

Notations for Groups (Rows)

Y is used to denote the I by N design matrix for the groups describing the rows of **X**: $y_{i,n} = 1$ if row i belongs to group; n, $y_{i,n} = 0$ otherwise.

Notations for Blocks (Columns)

Z is used to denote the J by K design matrix for the blocks from the columns of **X**: $z_{j,k} = 1$ if column j belongs to block k, $z_{j,k} = 0$ otherwise.

DICA

The first step of DICA is to compute the N by J matrix of the total of each group. This matrix is called **S**, and it is computed as

$$
\mathbf{S} = \mathbf{Y}^\top \mathbf{X}. \tag{2}
$$

The grand total of **S** is denoted s_{++} (i.e., $s_{++} = \mathbf{1}^\top \mathbf{S1}$). From matrix **S**, a matrix of barycentric row *profiles* denoted **R*** is computed, and it is computed as

$$
\mathbf{R}^* = \text{diag}\{\mathbf{S1}\}^{-1}\mathbf{S} \tag{3}
$$

where the diag operator transforms a vector into a diagonal matrix when applied to a vector and extracts the vector of the diagonal elements when applied to a matrix. A row of **R*** is a profile because it is made of non-negative numbers whose sum is equal to one. When transformed into profiles, two rows can be compared independently of their overall level. The *masses* of the

barycenters are proportional to the sum of the corresponding groups. Specifically, the N by 1 group mass vector, denoted **b**, is computed as

$$
\mathbf{b} = \mathbf{S1} \times s_{++}^{-1}. \tag{4}
$$

The diagonal barycenter mass matrix is obtained from the barycenter mass vector as

$$
\mathbf{B} = \text{diag}\{\mathbf{b}\}. \tag{5}
$$

The "grand barycenter," denoted **c**, is the overall barycenter of matrix **R***; it is computed as

$$
\mathbf{c} = \mathbf{S}^\top \mathbf{1} \times s_{++}^{-1}. \tag{6}
$$

The weights of the columns are inversely proportional to their frequency. The weights are stored in a J by 1 vector denoted **w**, and the corresponding J by J diagonal matrix is denoted **W**. Specifically, **W** and **w** are computed as

$$
\mathbf{W} = \text{diag}\{\mathbf{c}\}^{-1} \quad \text{and} \quad \mathbf{w} = \text{diag}\{\mathbf{W}\}. \tag{7}
$$

Correspondence Analysis of the Barycenter Matrix

The **R*** matrix is then analyzed using correspondence analysis. Specifically, the first step of the analysis is to center the matrix **R*** in order to create a matrix of centered profiles. This matrix denoted **R** is computed as

$$
\mathbf{R} = \mathbf{R}^* - \mathbf{1c}, \tag{8}
$$

(with **1** being an N by 1 vector of 1s). Then, the matrix **R** is analyzed with a generalized singular value decomposition under the constraints provided by the matrices **B** (masses for the N groups) and **W** (weights for the columns; see Abdi, 2007d; Abdi & Williams, 2010e; Greenacre, 1984) as

$$
\mathbf{R} = \mathbf{P}\Delta\mathbf{Q}^\top \quad \text{with} \quad \mathbf{P}^\top\mathbf{BP} = \mathbf{Q}^\top\mathbf{WQ} = \mathbf{I}, \tag{9}
$$

where Δ is the L by L diagonal matrix of the singular values (with L being the number of nonzero singular values), and **P** (respectively, **Q**) being the N by L (respectively, J by L) matrix of the left (respectively, right) generalized singular vectors of **R** (the singular vectors are also called *eigenvectors,* and the the squared singular values are also called *eigenvalues*; see Abdi, 2007c, for details).

Row Factor Scores

The N by L matrix of factor scores for the groups is obtained as

$$
\mathbf{F} = \mathbf{P}\Delta = \mathbf{RWQ}. \tag{10}
$$

The variance of the columns of **F** is given by the square of the corresponding singular values (i.e., the eigenvalue denoted λ; these are stored in the diagonal matrix Λ). This can be shown by combining Equations 9 and 10 to give

$$
\mathbf{F}^\top\mathbf{BF} = \Delta\mathbf{P}^\top\mathbf{BP}\Delta = \Delta^2 = \Lambda. \tag{11}
$$

Column Factor Scores (Loadings)

In correspondence analysis, the roles of the row and the columns are symmetrical: They can be represented in the same map because they have the same variance. Therefore, the columns are described by *factor scores,* which

can also be interpreted as loadings. Column factor scores are used to identify the variables important for separation between groups. In DICA, the column factor scores (cf. Equations 9 and 11) are computed as

$$\mathbf{G} = \mathbf{WQ\Delta}. \tag{12}$$

Projection of the Observations in the Discriminant Space

The I rows of matrix $\mathbf{X}$ can be projected (as "supplementary" or "illustrative" elements) onto the space defined by the factor scores of the barycenters. The first step is to transform $\mathbf{X}$ into a matrix of centered row profiles that are called $\mathbf{L}$ and are computed as

$$\mathbf{L} = \left(\operatorname{diag}\{\mathbf{X1}\}^{-1}\mathbf{X}\right) - \mathbf{1c} \tag{13}$$

(with $\mathbf{1}$ being an I by 1 vector of 1s). Then, from Equations 9 and 10, we find that matrix $\mathbf{WQ}$ is a projection matrix. Therefore, the I by L matrix $\mathbf{H}$ of the factor scores for the rows of $\mathbf{X}$ can be computed as

$$\mathbf{H} = \mathbf{LWQ} = \mathbf{LG\Delta}^{-1}. \tag{14}$$

These projections are barycentric because the weighted average of the factor scores of the rows of a group gives the factor scores of the group. Specifically, if $\mathbf{M}$ is defined as the mass matrix for the observations as

$$\mathbf{M} = \operatorname{diag}\{\mathbf{m}\} = \operatorname{diag}\{\mathbf{X1} \mathbf{x} \, s_{++}^{-1}\}. \tag{15}$$

Then the factor scores of the barycenters are the barycenter of the factor scores of the projections of the observations. This is shown by first computing the barycenters of the row factor scores as (cf. Equation 3)

$$\overline{\mathbf{H}} = \operatorname{diag}\{\mathbf{YM1}\}^{-1}\mathbf{YMH}, \tag{16}$$

and then plugging in Equation 14 and developing. Taking this into account, Equation 10 gives

$$\overline{\mathbf{H}} = \operatorname{diag}\{\mathbf{YM1}\}^{-1}\mathbf{YMXWQ} = \mathbf{RWQ} = \mathbf{F}. \tag{17}$$

Quality of the Prediction

The *performance,* or *quality,* of the prediction of a discriminant analysis is assessed by predicting the group membership of the observations and by comparing predicted with actual group membership. The pattern of correct and incorrect classifications can be stored in a *confusion matrix,* in which the columns represent the actual groups and the rows represent the predicted groups. At the intersection of a row and a column is the number of observations from the column group assigned to the row group.

The performance of the model can be assessed for the observations used to compute the groups: this is known as the *fixed effect model.* In addition, the model's performance can be estimated for new observations (i.e., observations not used to compute the model): This is known as the *random effect model.*

Fixed Effect: Old Observations

The *fixed effect model* predicts the group assignment for the observations used to compute the barycenters of the groups. To assign an observation to a group, the first step is to compute the distance between this observation and all N groups. Then, the observation is assigned to the closest group.

Several possible distances can be chosen, but a natural choice is the Euclidean distance computed in the factor space (Abdi, 2007b). If $\mathbf{h}_i$ is used to denote the vector of factor scores for the ith observation, and if $\mathbf{f}_n$ is used to denote the vector of factor scores for the nth group, then the squared Euclidean distance (in the factor space) between the ith observation and the nth group is computed as

$$d^2(\mathbf{h}_i, \mathbf{f}_n) = (\mathbf{h}_i - \mathbf{f}_n)^\top (\mathbf{h}_i - \mathbf{f}_n). \tag{18}$$

(Note that the Euclidean distance in the factor space is equivalent to the so called "chi-squared" distance in the original space.) Obviously, other distances are possible (e.g., Mahalanobis distance; see Abdi, 2007b, for more details), but the Euclidean distance has the advantage of being "directly read" on the map.

Tolerance intervals. The quality of the group assignment of the actual observations can be displayed using tolerance intervals. A *tolerance interval* encompasses a given proportion of a sample or a population. When displayed in two dimensions, these intervals have the shape of an ellipse and are called *tolerance ellipsoids.* For DICA, a group tolerance ellipsoid is plotted on the group factor score map. This ellipsoid is obtained by fitting an ellipse that includes a given percentage (e.g., 95%) of the observations. Tolerance ellipsoids are centered on their groups, and the overlap of the tolerance ellipsoids of two groups reflects the proportion of misclassifications between these two groups.

Random Effect: New Observations

The *random effect model* evaluates the quality of the assignment of *new* observations to groups. This estimation is obtained, in general, by using cross-validation techniques that partition the data into a *learning set* (used to create the model) and a *testing set* (used to evaluate the model). For DICA, a variation of this approach is used: In the *jackknife* (a.k.a. "leave one out") approach, each observation is taken out from the dataset, in turn, and then is projected onto the barycenter factor space computed from the remaining observations. This projection is then used to predict its group membership from the distances between the projected observation and the barycenters. In DICA, the only pre-processing needed to project an observation consists of the transformation of this observation into a profile. This transformation does not require estimating parameters from the learning set, and this guarantees that the prediction of the left-out observation is random.

The assignment of an observation to a group follows the same procedure as for a fixed effect model: The observation is projected onto the group factor scores, and the observation is assigned to the closest group. Specifically, $\boldsymbol{\ell}_i$ denotes the profile vector for the ith observation, and the following matrices obtained without the ith observation are denoted as (1) $\mathbf{X}_{-i}$, (2) $\mathbf{R}_{-i}$, (3) $\mathbf{B}_{-i}$ and (4) $\mathbf{W}_{-i}$ and refer to (1) the $I-1$ by J data matrix, (2) the N by J barycenter matrix, (3) the N by N mass matrix, and (4) the J by J weight matrix. All of these matrices are obtained using $I-1$ instead of I observations. Then, the generalized eigendecomposition of $\mathbf{R}_{-i}$ is obtained (cf. Equation 9) as

$$\mathbf{R}_{-i} = \mathbf{P}_{-i}\mathbf{\Delta}_{-i}\mathbf{Q}^\top_{-i} \quad \text{with} \quad \mathbf{P}^\top_{-i}\mathbf{W}_{-i}\mathbf{P}_{-i} = \mathbf{Q}^\top_{-i}\mathbf{B}_{-i}\mathbf{Q}_{-i} = \mathbf{I}. \tag{19}$$

The matrices of row and column factor scores denoted as $\mathbf{F}_{-i}$ and $\mathbf{G}_{-i}$ are obtained (cf. Equations 10 and 19) as

$$\mathbf{F}_{-i} = \mathbf{P}_{-i}\mathbf{\Delta}_{-i} = \mathbf{R}_{-i}\mathbf{W}_{-i}\mathbf{Q}_{-i} \quad \text{and} \quad \mathbf{G}_{-i} = \mathbf{W}_{-i}\mathbf{Q}_{-i}\mathbf{\Delta}_{-i}. \tag{20}$$

The jackknifed projection of the ith observation, denoted $\tilde{\mathbf{h}}_i$, is obtained (cf. Equation 14) as

$$\tilde{\mathbf{h}}_i = \boldsymbol{\ell}_i \mathbf{W}_{-i} \mathbf{Q}_{-i} = \boldsymbol{\ell}_i \mathbf{G}_{-i} \boldsymbol{\Delta}_{-i}^{-1}. \tag{21}$$

Distances between the ith observation and the N groups can be computed (cf. Equation 18) with the factor scores. The observation is then assigned to the closest group. Note that the jackknife procedure assumes that there are no columns with only one nonzero entry. If there is such a column, a "division by zero error" would be created when the nonzero observation is jackknifed.

Prediction intervals. To display the quality of the prediction for new observations, *prediction intervals* are used. To compute these intervals, the first step is to project the jackknifed observations onto the original complete factor space. There are several ways to project a jackknifed observation onto the factor score space. Here, we proposed a two-step procedure. First, the observation is projected onto the jackknifed space and is reconstructed from its projections. Then, the reconstituted observation is projected onto the full-factor score solution. Specifically, a jackknifed observation is reconstituted from its factor scores (cf. Equations 9 and 21) as

$$\tilde{\boldsymbol{\ell}}_i = \tilde{\mathbf{h}}_i \mathbf{Q}_{-i}^{\top}. \tag{22}$$

The projection of the jackknifed observation is denoted $\hat{\mathbf{h}}_i$ and is obtained by projecting $\tilde{\boldsymbol{\ell}}_i$ as a supplementary element onto the original solution. Specifically, $\hat{\mathbf{h}}_i$ is computed as

$$\begin{aligned} \hat{\mathbf{h}}_i &= \tilde{\boldsymbol{\ell}}_i \mathbf{W} \mathbf{Q} && \text{(cf. Equation 10)} \\ &= \tilde{\mathbf{h}}_i \mathbf{Q}_{-i}^{\top} \mathbf{W} \mathbf{Q} && \text{(cf. Equation 22)} \\ &= \boldsymbol{\ell}_i \mathbf{W}_{-i} \mathbf{Q}_{-i} \mathbf{Q}_{-i}^{\top} \mathbf{W} \mathbf{Q} && \text{(cf. Equation 21).} \end{aligned} \tag{23}$$

Note that $\hat{\mathbf{h}}_i$ can also be computed from the column factor scores as

$$\hat{\mathbf{h}}_i = \boldsymbol{\ell}_i \mathbf{G}_{-i} \boldsymbol{\Delta}_{-i}^{-2} \mathbf{G}_{-i}^{\top} \mathbf{W}_{-i} \mathbf{G} \boldsymbol{\Delta}^{-1}. \tag{24}$$

The quality of the predicted group assignment of the observations as a random model can be displayed using prediction intervals. A *prediction interval* encompasses a given proportion of the predicted elements of a sample or a population. When displayed in two dimensions, these intervals have the shape of an ellipse and are called *prediction ellipsoids*. For DICA, a group prediction ellipsoid is plotted on the group factor score map. This ellipsoid is obtained by fitting an ellipse, which includes a given percentage (e.g., 95%) of the predicted observations. Prediction ellipsoids are not necessarily centered on their groups—in fact, the distance between the center of the ellipse and the group represents the estimation bias. Overlap of two prediction intervals directly reflects the proportion of misclassifications for the "new" observations.

Quality of the Group Separation
R2 and Permutation Test

To evaluate the quality of the discriminant model, we use a coefficient inspired by the coefficient of correlation. Because DICA is a barycentric technique, the total variance (i.e., the *inertia)* of the observations to the grand barycenter (i.e., the barycenter of all groups) can be decomposed into two additive quantities: (1) the inertia of the observations relative to the

barycenter of their own category and (2) the inertia of the group barycenters to the grand barycenter.

Specifically, if $\bar{\mathbf{f}}$ denotes the vector of the coordinates of the grand barycenter (i.e., each component of this vector is the average of the corresponding components of the barycenters), the *total inertia*—denoted $\mathcal{I}_{\text{Total}}$—is computed as the sum of the squared distances of the observations to the grand barycenter (cf. Equation 18):

$$\mathcal{I}_{\text{Total}} = \sum_i^I m_i d^2(\mathbf{h}_i, \bar{\mathbf{f}}) = \sum_i^I m_i (\mathbf{h}_i, \bar{\mathbf{f}})^{\top}(\mathbf{h}_i - \bar{\mathbf{f}}). \tag{25}$$

In correspondence analysis, the grand barycenter is the center of the space. Therefore, $\bar{\mathbf{f}} = 0$, and so Equation 25 reduces to

$$\mathcal{I}_{\text{Total}} = \sum_i^I m_i \mathbf{h}_i^{\top} \mathbf{h}_i. \tag{26}$$

The inertia of the observations relative to the barycenter of their own category is abbreviated as the *inertia within*. It is denoted $\mathcal{I}_{\text{Within}}$ and is computed as

$$\mathcal{I}_{\text{Within}} = \sum_n^N \sum_{i \text{ in group } n} m_i d^2(\mathbf{h}_i, \mathbf{f}_n) = \sum_n^I \sum_{i \text{ in group } n} m_i (\mathbf{h}_i - \mathbf{f}_n)^{\top}(\mathbf{h}_i - \mathbf{f}_n). \tag{27}$$

The inertia of the barycenters to the grand barycenter is abbreviated as the *inertia between*. It is denoted $\mathcal{I}_{\text{Between}}$ and is computed as

$$\begin{aligned} \mathcal{I}_{\text{Between}} &= \sum_i^I b_n \times d^2(\mathbf{f}_n, \bar{\mathbf{f}}) = \sum_n^N b_n \times d^2(\mathbf{f}_n - \bar{\mathbf{f}}) \\ &= \sum_n^N b_n \times (\mathbf{f}_n - \bar{\mathbf{f}})^{\top}(\mathbf{f}_n - \bar{\mathbf{f}}) = \sum_n^N b_n \times \mathbf{f}_n^{\top} \mathbf{f}_n. \end{aligned} \tag{28}$$

So, the additive decomposition of the inertia can be expressed as

$$\mathcal{I}_{\text{Total}} = \mathcal{I}_{\text{Within}} + \mathcal{I}_{\text{Between}}. \tag{29}$$

This decomposition is similar to the familiar decomposition of the sum of squares in the analysis of variance. This suggests that the intensity of the discriminant model can be tested by the ratio of between inertia by total inertia, as is done in analysis of variance and regression. This ratio is denoted R^2 and is computed as

$$R^2 = \frac{\mathcal{I}_{\text{Between}}}{\mathcal{I}_{\text{Total}}} = \frac{\mathcal{I}_{\text{Between}}}{\mathcal{I}_{\text{Between}} + \mathcal{I}_{\text{Within}}}. \tag{30}$$

The R^2 ratio takes values between 0 and 1. The closer to 1, the better the model. The significance of R^2 can be assessed by permutation tests, and confidence intervals (CIs) can be computed using cross-validation techniques such as the jackknife (see Abdi & Williams, 2010d).

Confidence Intervals

The stability of the position of the groups can be displayed using CIs. A CI reflects the variability of a population parameter or its estimate. In two dimensions, this interval becomes a *confidence ellipsoid*. The problem of

estimating the variability of the position of the groups cannot, in general, be solved analytically, and cross-validation techniques need to be used. Specifically, the variability of the position of the groups is estimated by generating bootstrapped samples from the sample of observations. A *bootstrapped sample* is obtained by sampling with replacement from the observations (recall that when sampling with replacement, some observations may be absent, and some others maybe repeated). The "bootstraped barycenters" obtained from these samples are then projected onto the discriminant factor space and, finally, an ellipse is plotted such that it comprises a given percentage (e.g., 95%) of these bootstrapped barycenters. When the CIs of two groups do not overlap, these two groups are "significantly different" at the corresponding alpha level (e.g., $\alpha = .05$). In DICA, the bootstrap can be performed directly in the factor space by sampling the elements of matrix **H** and projecting their weighted means onto the factor space.

Partial Projection

Each of the K blocks can be projected in the common solution. The procedure starts by rewriting Equation 9 to show the blocks (recall that **R** is the matrix of deviations to the grand barycenter; see Equation 3):

$$\mathbf{R} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\top} = \mathbf{P}\boldsymbol{\Delta}[\mathbf{Q}_1, \ldots, \mathbf{Q}_k, \ldots, \mathbf{Q}_K]^{\top} = \sum_{k}^{K} \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}_k^{\top}, \qquad (31)$$

where $\mathbf{Q}_k$ is the kth block (comprising the J_k columns of **Q** corresponding to the J_k columns of the kth block). In addition, $\mathbf{R}_k$ is denoted as the kth block of the profile matrix, $\mathbf{G}_k$ as the kth block of the column factor scores, and $\mathbf{W}_k$ as the diagonal matrix corresponding to the weights of the kth block. In addition, the weight of the kth block is denoted as w_k and is defined as

$$w_k = \frac{\text{trace}\{\mathbf{W}\}}{\text{trace}\{\mathbf{W}_k\}} \qquad (32)$$

(where the "trace" operator gives the sum of the diagonal elements of a matrix). Then, Equation 10 is rewritten to get the projections for the kth block as

$$\mathbf{F}_k = w_k \mathbf{R}_k \mathbf{W}_k \mathbf{Q}_k = w_k \mathbf{R}_k \mathbf{G}_k \boldsymbol{\Delta}^{-1}. \qquad (33)$$

Note that it can be shown that **F** is the barycenter of the K blocks by rewriting Equations 31 and 33. Specifically, it is found that

$$\mathbf{F} = \sum_{k}^{K} \mathbf{R}_k \mathbf{W}_k \mathbf{Q}_k = \sum_{k}^{K} w_k^{-1} \mathbf{F}_k. \qquad (34)$$

The coordinates of a block for the observations are obtained by projecting the kth block of the observation profile matrix, denoted $\mathbf{L}_k$, as supplementary elements. This is obtained from Equation 33 as

$$\mathbf{H}_k = w_k \mathbf{L}_k \mathbf{G}_k \boldsymbol{\Delta}^{-1}. \qquad (35)$$

Contribution to the Inertia of a Dimension

Recall from Equation 11 that for a given dimension, the variance of the factor scores of all the N or J columns of matrix **R** is equal to the eigenvalue of this dimension. To identify the important groups, variables, or blocks of variables, the strategy is to compute the proportion of an element in the total (i.e., the eigenvalue).

Contribution of a group or variable to a dimension. The *contributions,* denoted ctr, of barycenter n to factor ℓ and of column j to factor ℓ are obtained, respectively, as

$$\text{ctr}_{n,\ell} = \frac{b_n f_{n,\ell}^2}{\sum_n b_n f_{n,\ell}^2} = \frac{b_n f_{n,\ell}^2}{\lambda_\ell} \quad \text{and} \quad \text{ctr}_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\sum_j c_j f_{j,\ell}^2} = \frac{c_i g_{j,\ell}^2}{\lambda_\ell}. \qquad (36)$$

Contributions help locate the observations important for a given factor. An often used rule of thumb is to consider that the important contributions are larger than the average contribution, which is equal to the number of elements (i.e., $\frac{1}{I}$ for the barycenters and $\frac{1}{J}$ for the columns). A dimension is then interpreted by opposing the positive elements with large contributions to the negative elements with large contributions.

Inertia and contribution to the inertia of a block. Because each block comprises a set of columns, the contribution of a block to a dimension can be expressed as the sum of this dimension squared factor scores of the columns of this block. Precisely, the inertia for the kth table and the inertia for the ℓth dimension are computed as

$$\mathcal{I}_{\ell,k} = \sum_{j \in J_k} c_j g_{\ell,j}^2. \qquad (37)$$

Note that the sum of the inertia of the blocks gives back the total inertia:

$$\lambda_\ell = \sum_k \mathcal{I}_{\ell,k}. \qquad (38)$$

The *contribution* of a block to a dimension is simply the sum of the contribution of its columns. Specifically, if $\text{ctr}_{\ell,k}$ denotes the contribution of the kth block to the ℓth dimension, it is computed as

$$\text{ctr}_{\ell,k} = \sum_{j \in J_k} \text{ctr}j\ell^2. \qquad (39)$$

Note that the sum of the inertia of the blocks is equal to 1.