A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples

Grace O. Lawley[1], **Peter A. Heeman**[1], **Jill K. Dolata**[2], **Eric Fombonne**[3], **Steven Bedrick**[4]
[1]*Computer Science and Engineering*
[2]*Department of Pediatrics*, [3]*Department of Psychiatry*
[4]*Department of Medical Informatics and Clinical Epidemiology*
Oregon Health & Science University, Portland, OR, USA

Abstract

Topic distribution matrices created by topic models are typically used for document classification or as features in a separate machine learning algorithm. Existing methods for evaluating these topic distributions include metrics such as coherence and perplexity; however, there is a lack of statistically grounded evaluation tools. We present a statistical method for investigating group difference in the document-topic distribution vectors created by latent Dirichlet allocation (LDA). After transforming the vectors using Aitchison geometry, we use multivariate analysis of variance (MANOVA) to compare sample means and calculate effect size using partial eta-squared. We report the results of validating this method on a subset of the *20Newsgroup* corpus. We also apply this method to a corpus of dialogues between Autistic and Typically Developing (TD) children and trained examiners. We found that the topic distributions of Autistic children differed from those of TD children when responding to questions about social difficulties. Furthermore, the examiners' topic distributions differed between the Autistic and TD groups when discussing emotions and social difficulties. These results support the use of topic modeling in studying clinically relevant features of social communication such as topic maintenance.

1 Introduction

Throughout the course of a dialogue many different topics are traversed with varying frequencies, and many analytical tasks depend on the ability to meaningfully quantify or otherwise characterize these patterns. For example, a system designed to automatically summarize meetings might need to detect when a new topic has been introduced; in a clinical context, we might wish to characterize the topics discussed during a patient visit to facilitate some sort of downstream analysis involving clustering or classification.

Topic modeling techniques such as latent Dirichlet allocation (LDA; Blei et al., 2003) allow us to capture and quantify the topic distributions across a collection of language samples. Typical methods for evaluating the resulting topic distributions use intrinsic metrics such as within-topic coherence; however, to our knowledge there remains a shortage of methods for statistically comparing the topic distributions produced by a model.

The application of topic modeling methods in clinical research has become more common in recent years (Hagg et al., 2022; Boyd-Graber et al., 2017; Jelodar et al., 2019). While topic modeling approaches have advanced significantly over the last twenty years, evaluation methods have lagged behind (see Hoyle et al., 2021 for a recent survey of methods). Current metrics tend to focus on intrinsically assessing model performance (via perplexity on held-out data) or on attempting to measure the quality of the topics that a model produces using metrics based on constructs such as human interpretability of the topics themselves (sometimes referred to as "coherence"). In a clinical research setting, however, the topic distributions produced by a model are themselves often meant for use in meaningfully quantifying differences between clinical populations. In such a scenario, usefully evaluating the quality of a topic model's "fit", or comparing that "fit" to that of another model (perhaps trained via a different algorithm, or with a different choice of hyperparameters) becomes a question of *extrinsic* evaluation, as intrinsic metrics such as perplexity or coherence are unlikely to be sufficient. Additionally, in clinical research, topic models are typically one piece of a larger analytical puzzle, one which often depends on traditional hypothesis-driven inferential statistical approaches (rather than stand-alone evaluation or use, as is more typical with topic models in machine learning scenarios).

In this paper, we outline a statistical approach to explore and quantify group differences in topic

September 11–15, 2023. ©2023 Association for Computational Linguistics

distributions captured by topic models and demonstrate its application using LDA and two different corpora. First, we validate our method on the *20Newsgroup* corpus, a widely-used reference corpus for developing and evaluating topic modeling algorithms (Mitchell, 1997), by comparing topic distributions between groups of documents that we expect to be similar and groups that we expect to be different. Second, we use our method on a corpus of language samples of Autistic[1] and Typically Developing (TD) children. Based on previous clinical evidence, we expect the topic distribution vectors of Autistic children to differ from those of the TD children. Our proposed method allows for a robust and statistically meaningful evaluation of the output of a topic model in both clinical and non-clinical contexts.

1.1 Topic Maintenance in ASD

Autism Spectrum Disorder (ASD) is a developmental disorder that is characterized by difficulties with social communication and restricted repetitive behavior (RRB) (American Psychiatric Association, 2013). These social communication difficulties sometimes include problems with topic maintenance (Baltaxe and D'Angiola, 1992; Paul et al., 2009), with Autistic children having more difficulty staying on topic than TD children. This difference may result in a signal that could be captured by a topic model as TD and ASD children would have different proportions of their speech assigned to different topics. In an effort to investigate this difference, we applied our statistical approach using LDA and a corpus of transcribed conversations between Autistic and TD children and trained examiners that were recorded during administration of a standard clinical assessment tool, the Autism Diagnosis Observation Schedule (ADOS, described further in section 3.2.1). Previous work with ADOS language samples (Salem et al., 2021; Lawley et al., 2023; MacFarlane et al., 2023) has shown that computational methods are able to capture a variety of differences in the language used by Autistic children from such dialogue samples, but to date have not focused on topic-level features. Our hypotheses for this experiment are two fold: (1) Autistic children will have different topic distributions than the TD children (i.e., talk about different topics than the TD children); (2) examiners will have similar topic distributions regardless of whether they are talking with Autistic children or TD children, as the ADOS task is designed (and examiners are trained) so as to ensure uniformity of delivery on the part of the examiner irrespective of the child's diagnostic status.

2 Statistical Motivation

LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics (Blei et al., 2003). Each document is treated as a bag-of-words (i.e., order does not matter) and is represented as a set of words and their associated frequencies. Given M documents and an integer k, LDA produces a $M \times k$ document-topic matrix (θ). LDA also produces a $k \times V$ topic-word matrix (β), where V is the total number of unique words across the entire corpus of documents. Since we will not be using the topic-word matrix in this analysis, from this point forward, we will use the phrases "LDA model" and "document-topic matrix" interchangeably.

In the document-topic matrix, each row represents a single document and each column represents one topic. The elements $(\theta_{1,1}, \dots, \theta_{i,j}, \dots, \theta_{M,k})$ are the estimated proportion of words in a document that were generated by a topic. From this matrix, each document can now be represented as a k-dimensional topic distribution vector.

These LDA-derived topic distribution vectors often serve as useful document representations for downstream analyses, such as a feature vectors for documentation classification or clustering. They are also commonly used as proxies for document content in more qualitative analyses of the composition of text corpora. To our knowledge, a statistical method for comparing topic distribution vectors between groups of documents has not yet been proposed.

One reason for this is due to the numerical properties of the resulting topic distribution vectors (each component θ_i is bounded between $\{0, 1\}$ with the further constraint of $\sum_{i=1}^{k} \theta_i = 1$), which render them unsuitable for use with many parametric statistical methods. This is an important limitation, because as previously mentioned, as the applications of topic modeling methods expand in clinical and behavioral research, the need for statis-

[1] We are using identity-first language (i.e., Autistic children) here instead of person-first language (i.e., children with Autism) as the former is the current preference among many Autistic individuals (Brown, n.d.).

tically based evaluation tools grows.

We realized that since the components in a topic distribution vector are proportions and all sum to one, they meet the definition of "compositional" data as formalized by Aitchison (1982), who also proposed a family of statistical approaches for such data. Compositional data are vectors of positive numbers that together represent parts of some whole: e.g., the demographic profile of a city or the mineral compositions of rocks.

There are three linear transformations that can be performed on compositional data: additive logratio (ALR), center logratio (CLR), and isometric logratio (ILR) transformation. The ILR transformation was introduced by Egozcue et al. (2003) in an effort to broaden the range of statistical methods that can be applied to compositional data by mapping compositonal data into real space. This transformation maps a composition from its original sample space (the D-part simplex) to the $D-1$ Euclidean space (ILR: $S^D \to \mathbb{R}^{D-1}$) with all metric properties preserved. Once the compositions are in $\mathbb{R}^{D-1}$, we are able to use classical multivariate analysis tools such as multivariate analysis of variance (MANOVA) to explore group differences (Egozcue et al., 2003; van den Boogaart et al., 2023).[2]

MANOVA is used to compare multivariate sample means and examines the effect of one discrete, independent variable on multiple continuous, dependent variables. For the analyses described in this paper, the independent variable is topic label when using the *20Newsgroup* corpus and diagnosis (ASD, TD) when using the clinical corpus. The dependent variables in both analyses are the various topic distribution probabilities in the document-topic matrix created by LDA: $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1, 2 \ldots, M$. It is important to note that a different discrete variable can be used as the independent variable, as long as it separates the documents into groups (e.g., author if modeling a corpus of newspaper articles); if one wished to incorporate multiple independent variables, one could could instead use MANCOVA. Since we used a k of 20 in both of our analyses and one dimension is removed during the ILR transformation, there are a total of 19 dependent variables.

In the case that we do find a significant group difference, the next step is to find out the magnitude of the effect. After MANOVA, we can use partial eta-squared (η^2) to calculate effect size. Partial η^2 tell us what proportion of variance of the linear combination of the topics can be explained by the independent variable (Tabachnick and Fidell, 2013).

MANOVA is a compelling choice for this analysis for several reasons. As detailed above, it enables us to statistically determine whether the topic distributions learned by our topic model are significantly associated with our other variables of interest (group membership, etc.) under a conventional hypothesis-testing framework. Second, MANOVA allows us to calculate interpretable measurements of effect size, which in turn facilitate comparison between different models (even if they are trained using different modeling algorithms). Third, this framework enables us to incorporate additional covariates as independent variables (via upgrading to MANCOVA), in a way that a more traditional classification-centric downstream task would not. Lastly, MANOVA is a well-characterized and well-established statistical method and as such has numerous useful extensions; for example, it can be combined with post-hoc Roy–Bargmann stepdown procedure (Tabachnick and Fidell, 2013) which enables detailed statistical analysis of the relationship between individual topics (or combinations of topics) and our independent variable, thereby facilitating a far richer quantitative interpretation of our topic model's output than other methods. Note, however, that this would be slightly complicated under our protocol due to our use of ILR, which results in the loss of a dimension into a new feature space that is decoupled from the original topics learned by the model (but which preserves important semantic properties of the original feature space). In this work, we explore only the first two points mentioned, leaving the rest for future work.

3 Corpora

We demonstrate our approach on two separate corpora: a subset of the *20Newsgroup* corpus and a corpus of transcribed natural language samples of ASD and TD children.

3.1 *20Newsgroup* corpus

The *20Newsgroups* corpus is a collection of approximately 18,000 posts from twenty different Usenet

[2] Our ability to use MANOVA here is contingent on statistical assumptions that must be met before proceeding. These assumptions are discussed in more in detail in section 4.3.

newsgroups,[3] and is a classic and widely-used dataset for text classification and analysis (Mitchell, 1997). We used the version of the *20Newsgroups* corpus that is available through the Python library `scikit-learn` (Pedregosa et al., 2011). For this analysis, we used documents from the following topic labels: *comp.sys.ibm.pc.hardware*, *comp.sys.mac.hardware*, *rec.sport.baseball*, and *rec.sport.hockey*. Documents that contained less than 500 characters were omitted. All utterances were tokenized, converted to lowercase, and lemmatized (e.g., "troubling" and "troubles" both become "trouble"). Stop words and fillers (e.g., "uh-huh", "mmhmm", "hmm", etc.) were dropped.[4]

3.2 Clinical corpus

The data used to in our second analysis consists of transcribed natural language samples of 117 ASD children and 65 TD children between the ages of 4 and 15 years old. All participants were native English speakers and had an IQ of ≥ 70. Sample characteristics for all 182 participants are summarized in Table 1. Intellectual level was estimated using the Wechsler Preschool and Primary Scale of Intelligence, third edition (WPPSI-III; Wechsler, 2002), for children younger than 7 years old. For children 7 years and older, the Wechsler Intelligence Scale for Children, fourth edition (WISC-IV; Wechsler, 2003), was used. Language ability and pragmatic and structural language skills were estimated using the Children's Communication Checklist, version 2 (CCC-2; Bishop, 2003).

3.2.1 Language samples

The language samples are transcribed dialogues between the child and an examiner during the conversation activities in the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000). The ADOS is a semi-structured interview that is designed to provide opportunities to observe speech and behavior that are characteristic of ASD as defined by the DSM-IV-TR (American Psychiatric Association, 2000). All participants were administered the ADOS-2, Module 3, which is designed for children and adolescents with fluent speech. Sessions were scored using the revised algorithms (Gotham et al., 2009).

Audio files were transcribed by a team of trained transcribers who were blind to participants' diagnostic status and intellectual abilities. Transcription was completed following modified Systematic Analysis of Language Transcripts (SALT) guidelines (Miller and Iglesias, 2012). Both the child and examiner speech were transcribed.

For this analysis, we used the transcribed dialogues from the four ADOS conversation activities: *Emotions*; *Social Difficulties and Annoyance*; *Friends, Relationships, and Marriage*; *Loneliness*. These activities were chosen for this analysis because of their conversational structure and naturalistic dialogue. Other ADOS activities, such as *Description of a Picture* and *Telling a Story From a Book*, were omitted. For each conversation activity, examiners are instructed to ask the child a series of questions, such as "What do you like doing that makes you feel happy and cheerful?" and "Do you have some friends? Can you tell me about them?". We followed same text preprocessing steps as described in section 3.1.

4 Methods

Figure 1 shows an example workflow for our method using LDA and a k of 5. All analyses were completed using the statistical programming language R (R Core Team, 2020). LDA models were estimated using the the `topicmodels` package (Grün and Hornik, 2011). The ILR transformation was performed using the `compositions` package (van den Boogaart et al., 2023). Box's M Test was performed using the `heplots` package (Friendly et al., 2022) and partial eta-squared was calculated using the `effectsize` package (Ben-Shachar et al., 2020). Our code for the *20Newsgroup* analysis is available online.[5]

4.1 *20Newsgroup*

Using the documents from four different topics, we fit a single LDA model with a k value of 20. After transforming the topic distribution vectors using the ILR transformation, we performed seven MANOVA tests. First, we compared the topic distributions between the broader *comp.sys.** and *rec.sport.** categories, where the former is composed of the documents from *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* and the latter of those from *rec.sport.baseball* and *rec.sport.hockey*.

[3] Usenet was an early internet-based network of hierarchically-organized discussion groups where users could post messages about a given topic.

[4] We used the lexicon of stop words provided in the tidytext package (Silge and Robinson, 2016).

[5] https://github.com/gracelawley/lawley-sigdial-2023

	ASD *(n = 117, 98 males)*				TD *(n = 65, 37 males)*				
	min	*max*	*mean*	*s.d.*	*min*	*max*	*mean*	*s.d.*	*p*
Age in years	4.54	15.6	10.03	2.82	4.21	14.5	8.22	2.83	<.001
IQ	72	138	102.19	15.77	90	147	116.94	12.37	<.001
ADOS SA	3	19	9.18	3.48	0	8	0.95	1.47	<.001
ADOS RRB	0	8	3.59	1.53	0	2	0.45	0.64	<.001
ADOS Total	7	24	12.77	3.73	0	10	1.40	1.79	<.001
CCC-2 Pragmatic	1.5	10.8	4.96	1.69	7.5	15.8	12.05	1.73	<.001
CCC-2 Structural	1	12	7.01	2.29	8.5	15	11.73	1.57	<.001
CCC-2 GCC	45	103	75.13	11.0	87	143	115.18	12.09	<.001

Table 1: Demographic and clinical sample characteristics. Abbreviations: ADOS = Autism Diagnostic Observation Schedule; SA = Social Affect; RRB = Restricted and Repetitive Behavior; CCC-2 = Children's Communication Checklist, version 2; GCC = Global Communication Composite.

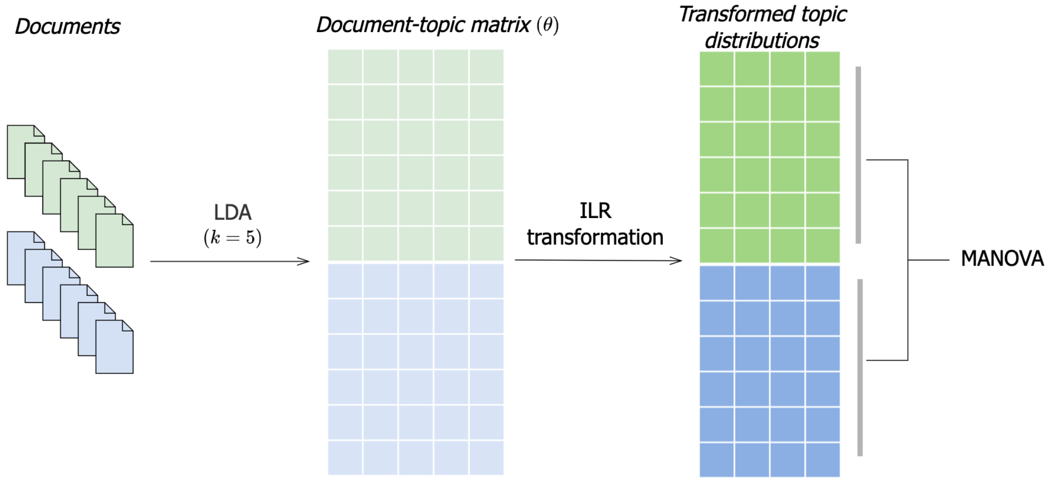
Documents
Document-topic matrix (θ)
Transformed topic distributions
LDA
($k = 5$)
ILR transformation
MANOVA

Figure 1: Example workflow for the described statistical approach described to explore and quantify group differences in topic distributions captured by topic models.

We hypothesize that the topic distributions between these groups will be very different. Second, we compared topic distributions between subcategories: *comp.sys.ibm.pc.hardware* vs. *comp.sys.mac.hardware*; *rec.sport.baseball* vs. *rec.sport.hockey*. We hypothesize that these groups will also be different, but not as different as the previous comparison. Third, we compared the topic distributions within each of the four topics by randomly splitting each topic into two groups (e.g., *rec.sport.baseball.1* vs. *rec.sport.baseball.2*). Since the documents are from the same topic, we hypothesize that there will be no difference between the topic distributions. For all of the above MANOVA tests, the independent variable is the topic label and the dependent variables are the topic probability values from the document-topic vectors.

4.2 Clinical corpus

Since our plan involves analyzing the child and examiner speech separately, we created two separate LDA models: one containing only the child speech and one containing only the examiner speech. In both models, we define a document as all words said by a speaker during a single ADOS conversation activity. Since there are four activity types, within each model each child-examiner conversation is associated with four, distinct documents.

We used a k value of 20 for both models. This decision was informed by prior knowledge of the type and quantity of questions the examiners are instructed to ask during the ADOS conversation activities. Hyperparameter estimation was done using the variational expectation-maximization (VEM) algorithm with a starting α value of $50/k$ (Grün and Hornik, 2011; Griffiths and Steyvers, 2004).

For each of our MANOVA tests, the independent variable is diagnosis (either ASD or TD) and the dependent variables are the topic probability values from the document-topic vectors. Since we used a k of 20 in our analysis and one dimension was lost during the ILR transformation there are 19 dependent variables. The null hypothesis is that the multivariate means of the ASD and TD groups are equal.

4.3 MANOVA assumptions

Before proceeding further with MANOVA, there are multiple assumptions that must be met (Tabachnick and Fidell, 2013). First, each combination of independent and dependent variables should be multivariate normally distributed. Since there are more than 20 observations for each dependent × independent variable combination the Multivariate Central Limit Theorem holds so we can assume the multivariate normality assumption holds.

Second, dependent variables should have a linear relationship with each group of the independent variable. This assumption was initially not met since each topic distribution vector summed to 1. However after performing the ILR transformation described in section 2, this is no longer the case.

Third, variance-covariance matrices for dependent variables should be equal across groups. This can be tested using Box's M test (Box, 1949), which tests the null hypothesis that the matrices are equal. For our data, Box's M test yielded p-values of $p < 0.001$ for each topic for the *20Newsgroups* documents and also for each conversation activity for both child and examiner speech, and thus this assumption (of equal covariance matrices) was not met. However, MANOVA is robust to unequal covariance matrices when Pillai's criterion is used (Tabachnick and Fidell, 2013; Pillai, 1955), and as such we are able to proceed .

Lastly, there should be no extreme outliers in the dependent variables. Extreme outliers can be identified by calculating the Mahalanobis distance for each observation and then performing a chi-squared test (using $df = k - 1$) to calculate the corresponding p-values. The null hypothesis is that the observation is not an outlier. We repeated analyses with identified outliers excluded and saw no difference in results. The results presented here are with these outliers included.

5 Results

The first part of our analysis was to demonstrate the application of our approach on the *20Newsgroup* corpus, a popular corpus for topic modeling. The results for the MANOVA tests are reported in Table 2. There was a significant difference between the topic distributions from the *comp.sys.** and *rec.sport.** categories, $F(19, 1710) = 414.240$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.82$. Between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* subcategories, topic distributions were significantly different, $F(19, 795) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.26$. Topic distributions were also significantly different between the *rec.sport.baseball* and *rec.sport.hockey* subcategories, $F(19, 895) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.57$. When comparing topic distributions within each topic (by randomly splitting the documents into two groups), there were no significant differences found.

For the second part of our analysis, we compared the children's topic distribution vectors between diagnostic groups (ASD, TD). The results of the MANOVA tests for each ADOS conversation activity for child speech are reported in Table 3. The children's topic distributions were significantly different between the Autistic and TD children within the *Social Difficulties and Annoyance* activity, $F(19, 169) = 2.055$, $p = 0.0083$, with a large effect size, partial $\eta^2 = 0.19$. There was no significant group difference in topic distributions within the other three conversation activities (*Emotions*; *Friends, Relationships, and Marriage*; *Loneliness*). To address potential Type I error from multiple comparisons, p-values can be evaluated using a Bonferroni adjusted α of 0.0125. When evaluating the results using the adjusted α of 0.0125, the significant result within the *Social Difficulties and Annoyance* conversation activity remains.

Lastly, the results of the statistical analyses performed on the examiner speech are reported in Table 4. The examiners' topic distributions differed significantly between ASD and TD groups within three of the four conversation activities examined: *Emotions*, $F(19, 175) = 2.235$, $p = 0.0035$, with a large effect size, partial $\eta^2 = 0.20$; *So-*

topics	n	df	Pillai	approx. F	df$_1$	df$_2$	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							
comp.sys.ibm.pc.hardware	219	1	0.020	0.460	19	427	0.976	0.02
"	228							
comp.sys.mac.hardware	198	1	0.044	0.840	19	348	0.659	0.04
"	170							
rec.sport.baseball	206	1	0.041	0.903	19	403	0.579	0.04
"	217							
rec.sport.hockey	247	1	0.029	0.738	19	472	0.780	0.03
"	245							

Table 2: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

		df	Pillai	approx. F	df$_1$	df$_2$	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 3: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

cial Difficulties and Annoyance, $F(19, 174) = 3.858$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.30$; *Friends, Relationships, and Marriage*, $F(19, 176) = 1.833$, $p = 0.0224$, with a large effect size, partial $\eta^2 = 0.17$. There was no significant difference between groups for the *Loneliness* conversation activity. A Bonferroni adjusted α of 0.0125 can be used to address potential Type I error from multiple comparisons. With this adjusted α, a significant group difference within the *Emotions* and *Social Difficulties and Annoyance* activities remains; however, the previous group difference within *Friends, Relationships, and Marriage* is no longer significant.

6 Discussion

The Autistic children and TD children had significantly different topic distributions for one of the four conversation analyzed: *Social Difficulties and Annoyance*. We expected to observe a group difference in all four of the conversation activities instead of only one. Incorporating additional participant-level information such as IQ and age or examining other measures of conversational reciprocity such as the length and complexity of utterances may help shed some light as to why a group difference was only seen in one of the four activities analyzed. In addition, further investigation into sampling context differences between the conversation activities is needed before conclusions can be drawn. This finding illustrates the value of our proposed statistical approach, in that we have numerous ways we could incorporate these additional covariates into our analysis in quantitatively useful ways within the same statistical framework.

The examiners' topic distributions differed significantly between the ASD and TD groups for two of the four activities: *Emotions* and *Social Difficulties and Annoyance*. This is surprising as our initial hypothesis was there would not be any sig-

		df	Pillai	approx. F	df$_1$	df$_2$	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 4: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

nificant group differences for the examiners' topic distributions. ADOS examiners are instructed to cover the same questions for each child, regardless of diagnosis, and are trained to a high standard of consistency and repeatability, as the assessment is meant for clinical use. Since one goal of the conversation activities is to foster a dialogue, the examiner would likely avoid actions that could discourage the child from conversing and sharing their interests. It may be the case that the examiners are mirroring the topics introduced by the children during the activities and those topics are being picked up by the topic distributions created by LDA.[6] This could be explored in the future by investigating pairwise group differences.

7 Conclusion

In this paper we presented a novel application of existing statistical methods to evaluate the document-topic distribution vectors created by topic models in order to investigate group differences. By treating the document-topic distribution vectors as compositional data (Aitchison, 1982), we are able to use the ILR transformation (Egozcue et al., 2003) to map the vectors from their original sample sample, the D-part simplex, into the $D-1$ Euclidean space (ILR: $S^D \rightarrow \mathbb{R}^{D-1}$). Once in $\mathbb{R}^{D-1}$, we are able to use classical multivariate analysis tools such as MANOVA (Egozcue et al., 2003).

When applied to an LDA model fitted to the *20Newsgroups* corpus, our method successfully identified that the topic distributions for documents from different categories (computer hardware vs. sports) and also documents from related subcategories (PC hardware vs. Macintosh hardware; baseball vs. hockey) were significantly different. The effect size, measured with partial η^2, also varied across these comparisons, with the effect size being the largest when comparing computer hardware vs. sports and smallest when comparing Macintosh vs. PC hardware. Furthermore, our method did not find that topic distributions are significantly different when comparing groups of documents from the same category.

We also demonstrate the application of this method using LDA and a corpus of child-examiner dialogues of Autistic and TD children, where prior clinical research gave us reason to expect to find group differences. We found that the topic distributions of Autistic and TD children were significantly different during one of the four ADOS conversation activities examined. This result aligns with prior clinical research that Autistic children often have difficulties with topic maintenance in a conversational context. Interestingly, we also found that examiners' topic distributions were significantly different whether they were conversing with an Autistic child or a TD child for two of the four ADOS conversation activities examined. This may indicate that although the examiners are trained to ask the same set of questions irrespective of diagnosis status, tangential topics introduced by the child during the conversation may be mirrored by the examiner and thus are reflected in the associated topic distributions.

There are a few points about the statistical approach outlined in this paper that should be highlighted. Although we demonstrate this method using the document-topic distribution matrix created by LDA, this method can be extended to any topic modeling algorithm that outputs a topic distribution that can be treated as a composition. We decided to use LDA here as it is a well-established technique that has been extended and built upon many times over since it was first introduced in 2003. Another important point to highlight is that, although not shown in here, this analysis has the potential to be extended further with a post-hoc Roy-

[6] An anonymous reviewer brought to our attention that interviewers have been found to adjust their conversational patterns when speaking to patients with other cognitive conditions, such as Alzheimer's disease (Nasreen et al., 2021).

Bargmann step down procedure to explore how much each topic (or combination of topics) contributes to the significant effect of the independent variable (Tabachnick and Fidell, 2013). However, as previously mentioned, the loss of a dimension during the ILR transformation would need to be addressed first. Overall, the statistical approach presented in this paper represents a very promising direction for methods of making topic models more interpretable in a quantitative way, beyond human inspection of topics. In the future we would like to extend this specific analysis to include additional participant-level, independent variables (e.g., age, sex, IQ) by using multivariate analysis of covariance (MANCOVA). Since social communication skill level can vary throughout the ASD spectrum (Tager-Flusberg and Kasari, 2013), we would also like to look at differences within the ASD group by exploring within group variance metrics. We would also like to explore the use of other methods of topic modeling, beyond LDA, for this application.

As the application of topic modeling methods continues to grow into areas such as clinical and behavioral research, so does the need for statistically based methods for evaluation and comparison. Our hope is that the statistical approach described in this paper contributes to bridging that gap by focusing on improving evaluation metrics for existing topic modeling methods.

Limitations

There are several limitations of this analysis that should be mentioned. First, the decision to set k to 20 was specific to the particular clinical discourse corpus used. Our decision was informed by of the type and quantity of questions the examiners are instructed to ask during the ADOS conversation activities; however, it may not always be possible to choose a value for k using existing knowledge of the corpus. Second, as mentioned in section 2, after performing the ILR transformation we lose one dimension from our original topic model's output and go from k to $k-1$ elements in each vector. A consequence of this is that there is no direct mapping between dimensions of the ILR-transformed $\mathbb{R}^{k-1}$ vector and the original k topics after the transformation, though the new dimensions retain the information contained in the original data (as shown by their ability to be used via MANOVA). Depending on the nature of the analysis that one is conducting, this may or may not be an issue; it was not during the present analysis, since we were interested in the overall topic distributions of each document (rather than in specific document-topic associations) but this may not always be the case. A possible direction for future work would be to draw further upon statistical methods from compositional spaces to assist with this issue.

Ethics Statement

This study was approved by the Oregon Health & Science University IRB (Protocol #531) and all research was performed in accordance with their relevant guidelines and regulations.

Acknowledgements

This work was supported in part by the National Institute on Deafness and Other Communication Disorders of the NIH under Awards R01DC012033 (PI: Dr. E. Fombonne) and R01DC015999 (PIs: Dr. S. Bedrick & G. Fergadiotis).

References

John Aitchison. 1982. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177.

American Psychiatric Association. 2000. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed., text rev.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed.

Christiane A. M. Baltaxe and Nora D'Angiola. 1992. Cohesion in the discourse interaction of autistic, specifically language-impaired, and normal children. *Journal of Autism and Developmental Disorders*, 22(1):1–21.

Mattan S. Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815.

Dorothy V. M. Bishop. 2003. *The Children's Communication Checklist, version 2 (CCC-2)*. Pearson, London.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

George E. P. Box. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Lydia Brown. n.d. Identity-First Language. Autistic Self Advocacy Network (ASAN). `https://autisticadvocacy.org/about-asan/identity-first-language/` Last accessed on 2023-05-10.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. 2003. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35:279–300.

Michael Friendly, John Fox, and Georges Monette. 2022. *heplots: Visualizing Tests in Multivariate Linear Models*. R package version 1.4-2.

Katherine Gotham, Andrew Pickles, and Catherine Lord. 2009. Standardizing ADOS Scores for a Measure of Severity in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 39(5):693–705.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Lauryn J. Hagg, Stephanie S. Merkouris, Gypsy A. O'Dea, Lauren M. Francis, Christopher J. Greenwood, Matthew Fuller-Tyszkiewicz, Elizabeth M. Westrupp, Jacqui A. Macdonald, and George J. Youssef. 2022. Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review. *Journal of Medical Internet Research*, 24(11):e33166.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Grace O. Lawley, Steven Bedrick, Heather MacFarlane, Jill K. Dolata, Alexandra C. Salem, and Eric Fombonne. 2023. "Um" and "Uh" usage patterns in children with autism: Associations with measures of structural and pragmatic language ability. *Journal of Autism and Developmental Disorders*, 53:2986–2997.

Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H. Cook, Bennett L. Leventhal, Pamela C. DiLavore, Andrew Pickles, and Michael Rutter. 2000. The Autism Diagnostic Observation Schedule, Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223.

Heather MacFarlane, Alexandra C. Salem, Steven Bedrick, Jill K. Dolata, Jack Wiedrick, Grace O. Lawley, Lizbeth H. Finestack, Sara T. Kover, Angela John Thurman, Leonard Abbeduto, and Eric Fombonne. 2023. Consistency and reliability of automated language measures across expressive language samples in autism. *Autism Research*, 16(4):802–816.

J. Miller and A. Iglesias. 2012. SALT: Systematic analysis of language transcripts [Research version]. *Middleton, WI: SALT Software*.

Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. *Frontiers in Computer Science*, 3.

Rhea Paul, Stephanie Miles Orlovski, Hillary Chuba Marcinko, and Fred Volkmar. 2009. Conversational Behaviors in Youth with High-functioning ASD and Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 39(1):115–125.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

K. C. S. Pillai. 1955. Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1):117–121.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alexandra C. Salem, Heather MacFarlane, Joel R. Adams, Grace O. Lawley, Jill K. Dolata, Steven Bedrick, and Eric Fombonne. 2021. Evaluating atypical language in Autism using automated language measures. *Scientific Reports*, 11(1):10968.

Julia Silge and David Robinson. 2016. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, 1(3).

Barbara G. Tabachnick and Linda S. Fidell. 2013. *Using Multivariate Statistics*, 6th edition. Pearson.

Helen Tager-Flusberg and Connie Kasari. 2013. Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research*, 6(6):468–478.

K. Gerald van den Boogaart, Raimon Tolosana-Delgado, and Matevz Bren. 2023. *compositions: Compositional Data Analysis*. R package version 2.0-6.

David Wechsler. 2002. WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence - 3rd ed. *San Antonio, TX: Psychological Corporation*.

David Wechsler. 2003. WISC-IV: Wechsler Intelligence Scale for Children. *San Antonio, TX: Psychological Corporation*.