Data-Driven Generation of Synthetic Behavioral Feature Vectors Modeling Children with Autism Spectrum Disorders

Kim Baraka
Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
INESC-ID / Instituto Superior Técnico
2744-016 Porto Salvo, Portugal
Email: kbaraka@andrew.cmu.edu

Francisco S. Melo
INESC-ID / Instituto Superior Técnico
Universidade de Lisboa
2744-016 Porto Salvo, Portugal
Email: fmelo@inesc-id.pt

Manuela Veloso
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: mmv@cs.cmu.edu

***Abstract*—Behavioral data on children with Autism Spectrum Disorders (ASD) are available thanks to standardized diagnostic tools, such as the Autism Diagnostic Observation Schedule (ADOS). This data can be of great use to enhance the learning and reasoning of agents interacting with children with ASD. However, the amount of such available data is limited and may not prove useful by itself to inform the algorithms of complex agents. To address this data scarcity problem, we present a method for generating synthetic behavioral data in the form of feature vectors characterizing a wide range of children with ASD. Our method relies on a thorough analysis and partition of the feature space based on a real dataset containing the ADOS scores of 279 children. We first analyze the real dataset using dimensionality reduction techniques, then introduce data-driven descriptors that partition the feature space into regions naturally arising from the data. We end by presenting a descriptor-based sampling method to generate synthetic feature vectors that successfully preserves the correlation structure of the real dataset.**

I. Introduction

Autism Spectrum Disorders (ASD) are a set of widespread developmental disorders usually affecting children at an early age. Individuals with ASD present a number of difficulties including communication, social interaction, sensory, and emotional challenges. Causes and mechanisms of ASD have been mainly studied from a developmental, neuropsychological [1], and genetic [2] perspective, but most ASD diagnosis tools to date strongly rely on *behavioral* features [3] [4] [5].

On the other hand, several research efforts have looked at introducing agents in ASD therapy sessions in the hopes of increasing the effectiveness of the treatment. These agents include robots [6], some of which reason and act using data [7], as well as other technological interfaces including avatars, virtual and augmented reality interfaces, and wearables [8].

There is great potential in utilizing behavioral data from individuals with ASD to enhance and personalize social interactions of agents with such individuals, especially given that the range of exhibited behaviors varies greatly between individuals. Such data can be used to create better models of interactions with different types of ASD individuals, and inform the reasoning and learning of an agent expected to interact with such a population. Unfortunately, such data are limited in number partly because they are difficult to obtain or share (high cost of diagnostic tests and privacy issues), and partly because the data can be high-dimensional and sparse. As a result, the real data need to be complemented by additional *synthetically generated data* in order to be useful for the training of complex AI and Machine Learning algorithms.

In this work, we extract domain knowledge from a limited number of real behavioral data points from children with ASD, and generalize from them to generate an arbitrary number of synthetic data points. We use the features of the *Autism Diagnostic Observation Schedule (ADOS)* [3], a standardized test to diagnose ASD by coding a wide range of behaviors in response to a set of activities. Compared to other existing tools, the very systematic aspect of ADOS makes it useful for computing applications such as agent modeling, but also learning and reasoning. The feature space associated with ADOS features is very large, and it prompts us to investigate methods for efficiently sampling from it synthetic data consistent with real data. But in order to do so, it is crucial to first understand patterns of the real data in order to generalize from it.

Synthetic data generation often includes constraining the properties of the data to be generated. In other words, it is useful to define regions of the feature space where we want the generated data to fall. In this paper, we introduce descriptors as a way to partition the feature space into a finite set of classes. These descriptors can then be used by our synthetic data generation algorithm to generate feature vectors in the different classes specified by the given descriptor.

The contributions of this paper are: (1) an *analysis* of the distribution of real ADOS data in the feature space using dimensionality reduction methods (section IV), (2) two new data-driven *descriptors* that take into account information unused by the existing ADOS descriptor, and partition the feature space into 4 classes (section V), and (3) a *sampling method* operating in the full feature space that generates synthetic data consistent with the correlation structure of the real dataset and constrained by a given descriptor (section VI).

II. Related work and background

A. Synthetic data generation

Synthetic data can be very useful in many applications to train, test or inform a wide range of algorithms. It is widely used to train fraud detection systems [9], but also information discovery systems [10], spatial microsimulation models to inform policy intervention [11], sensor network deployment [12], and much more. Some general purpose synthetic data generation methods also utilize ways of describing the data to be generated (e.g., the Synthetic Data Description Language (SDDL) [13]), while others are designed to generate very specific types of data in a rather unstructured way. In this paper, we focus on synthetic data sampling methods that preserve the correlation between features while incorporating constraints on the region of the feature space where that data will be generated.

B. Autism Diagnostic Observation Schedule (ADOS) structure

The ADOS is a state-of-the-art, standardized, semi-structured diagnostic tool for ASD used by therapists world-wide. It comprises 5 modules suitable for different language abilities and/or ages. Module 1 (Pre-verbal/Single Words) remains the main module used by therapists as an initial assessment of young children that aren't toddlers, which is why we focus on this particular module in this work. However, our methods can be applied to any of the ADOS modules as they possess a very similar structure.

Module 1 of the ADOS is composed of 10 standardized activities, ranging from unstructured activities such as 'Free Play' (where the child is left to freely play in the room) to highly structured activities such as a 'Response to name' activity (where the therapists calls the child's name at different degrees of intensity and observes the child's response). In a session where the ADOS is administered, the therapist performs the activities and records behaviors of interest throughout the session, in a time span of 40-60 min.

At the end of the session, the therapist codes the behaviors exhibited by the child throughout the whole session. There are a total of 29 ADOS features for different, usually exclusive, behavior types. Of these 29 features only 14 are used in the algorithm that returns the *overall total* score used for diagnosis. The overall total can be broken down into three subtotals: Communication (Comm.), Reciprocal Social Interaction (Soc.), and Restricted and Repetitive Behavior (RRB). Feature values are all remapped to a 0-2 integer scale before they are summed. From the overall total, one can compute a *comparison score* (integer-valued between 1 and 10) which serves as our measure for autism severity. In this paper, we make use of all 29 features with their full range, as summarized in Table I.

TABLE I
Summary of the ADOS module 1 features used

Feature name	Code	Range
Overall Level of Non-echoed Language	A1	1-4
Frequency of Vocalization Directed to Others	A2	1-3
Intonation of Vocalizations or Verbalizations	A3	1-2
Immediate Echolalia	A4	1-3
Stereotyped/Idiosyncratic Use of Words or Phrases	A5	1-3
Use of Other's Body to Communicate	A6	1-2
Pointing	A7	1-3
Gestures	A8	1-2
Unusual Eye Contact	B1	1-2
Responsive Social Smile	B2	1-3
Facial Expressions Directed to Others	B3	1-2
Integration of Gaze [etc.] During Social Overtures	B4	1-3
Shared Enjoyment in Interaction	B5	1-2
Response to Name	B6	1-3
Requesting	B7	1-3
Giving	B8	1-2
Showing	B9	1-2
Spontaneous Initiation of Joint Attention	B10	1-2
Response to Joint Attention	B11	1-3
Quality of Social Overtures	B12	1-3
Functional Play With Objects	C1	1-3
Imagination/Creativity	C2	1-3
Unusual Sensory Interest in Play Material/Person	D1	1-2
Hand and Finger and Other Complex Mannerisms	D2	1-2
Self-Injurious Behavior	D3	1-2
Unusually Repetitive Interests or Stereotyped Beh.	D4	1-3
Overactivity	E1	1-2
Tantrums, Aggression, Negative or Disruptive Beh.	E2	1-2
Anxiety	E3	1-2

III. Domain and dataset description

A. Domain definitions and notation

The **features** we use in this work are the 29 ADOS features, summarized in Table I. These include items related to A: Language and Communication, B: Reciprocal Social Interaction, C: Play, D: Stereotyped Behaviors and Restricted Interests, and E: Other Abnormal Behaviors. The values for these features, denoted f_i, $i = 1, \ldots, 29$ in the order listed in the table, are integers in the range shown in the third column of the table.

Since the aim of this paper is to generalize from real data to generate synthetic data, it is important to distinguish between:

- **Arbitrary feature vectors**, i.e., points in the feature space, denoted by $\mathbf{f} = (f_1, \ldots, f_{29})$,
- **Real data points**: $\mathbf{x} = \{x_i\}_{i=1}^{N}$ representing real subjects in our dataset, where $x_i = \mathbf{f^{(x_i)}} = (f_1^{(x_i)}, \ldots, f_{29}^{(x_i)})$ is a feature vector and N is the dataset size, and
- **Synthetic data points**: $\hat{\mathbf{x}} = \{\hat{x}_i\}_{i=1}^{M}$, representing simulated subjects and generated by our algorithm informed by real data, where $\hat{x_i} = \mathbf{f^{(\hat{x}_i)}} = (f_1^{(\hat{x}_i)}, \ldots, f_{29}^{(\hat{x}_i)})$ is a feature vector and M is the number of generated feature vectors.

B. Dataset description and preprocessing

Our dataset consists of the full ADOS Module 1 score set (f_1 through f_{29}) of children suspected of having an ASD (N=279). The data came from two sources: out of the 279 score sets, 212 were obtained from the National Database for Autism Research (NDAR)[1] and 67 were obtained from the Child Development Center (CDC) at the Garcia de Orta

[1]NDAR is a collaborative informatics system created by the National Institutes of Health (NIH) to provide a national resource to support and accelerate research in autism.

Hospital in Alameda, Portugal[2]. The average age is 67.7 months ($SD = 39.6$ months), with a minimum age of 20 months and a maximum age of 236 months. Part of the dataset doesn't have gender information, but for the 147 data points that have it, the Male-to-Female ratio is 38:49.

This type of data presents some challenges, outlined below:
- Data are *discrete*, which makes it harder to generate synthetic data that are consistent with real data (for example, sampling synthetic data points according to a correlation model is straightforward with the Gaussian assumption, but without it, it is not).
- Data are *ordinal*, which means that traditional parametric methods might not be suitable for this type of data.
- Data are *noisy*. Although the ADOS is a standardized test, there is some level of subjectivity in the coding as different therapists may assign different values for a same set of observations during an administration session.
- Data are sometimes *incomplete*. The dataset has missing entries (`NaN`'s) in some features for some of the subjects (28 of 8091 entries were missing, mostly for feature A3).
- The dataset includes data from both the first version of the test, ADOS-G [14] and the second version, ADOS-2 [3], contributing to additional noise. The main differences between the two versions are: (1) coding (feature A6 coding rules were slightly revised and 4 out of 29 have their range changed from a 3-pt to 4-pt scale), (2) some extra features added in ADOS-2 (those were neglected), and (3) a different algorithm for computing total scores (we used the most recent).

Values of 7 and 8 in the database, corresponding to special circumstances or insufficient observations, were directly treated as `NaN`'s, which we replaced with randomly sampled values in the allowed range for the missing feature value. Even though more advanced methods of dealing with missing data such as matrix completion [15] could be used, this simple method ensures that the correlation structure of the data, the basis for our synthetic data generation algorithm, is maintained after preprocessing, at the cost of adding some noise.

IV. Visualizing and analyzing real ADOS data

In this section, we present two methods for visualizing the feature vectors of our dataset using a lower-dimensional, human-readable representation. The first one uses a Self-Organizing Map (SOM) that maps the data points to a 2D space and the second one exploits groupings of features arising in the ADOS total computation algorithm and maps data points to a 3D space. These visualizations are useful to assess whether or not clusters or low-density regions are present in the data, and to inform our method, presented in section VI, for generating new descriptors based on the real data.

A. Dimensionality reduction using a 2D Self-Organizing Map

A self-organizing map (SOM) is a neural network that learns, in an unsupervised way, an alternative, low-dimensional, representation of high-dimensional data in the form of a 2D map consisting of interconnected neurons preserving the topology of the original data [16]. We trained on our dataset an SOM consisting of 25 neurons connected in a 5-by-5 hexagonal map. With each neuron, there is an associated position in the map space.

A common visualization method of an SOM is through the unified distance matrix (U-matrix) [17], which computes the distance between a neuron and its neighbors in the map space. The left part of Fig. 1 shows the U-matrix for our trained SOM, where brighter regions correspond to more clustered regions and darker regions correspond to low-density regions. Our U-matrix suggests that there aren't clearly separated clusters in the data, but rather some low-density regions.

The trained network maps input feature vectors to the closest neuron in the map space. The right part of Fig. 1 shows a histogram of the number of data points being mapped to each neuron for our dataset. This plot confirms the intuition we got from the U-matrix that there are low-density regions in the dataset rather than clearly separated clusters.

As a final note, we justify our choice of the main SOM hyperparameter, namely its size. Even though no systematic validation of our size choice was performed, it was chosen as a result of experimenting with different sizes, as a tradeoff between overfitting and generalization power, especially in relation to the resulting sample hit histogram, which gives us an idea of the probability distribution across the map. It is worth mentioning that no clear cluster separation was found even with larger SOM sizes.

B. Dimensionality reduction using ADOS subtotals (3D)

Even though all 29 features are coded by the therapist during an ADOS session, only 14 are used for the computation of a total score on which a diagnosis is made to assess the subject's autism severity. The *ADOS algorithm* generates an overall total from a feature vector, which can then be transformed into a severity score or a severity class. The algorithm consists in first categorizing the subject into either the 'no words' or the 'some words' category according to the value of feature A1. Then, feature values are remapped to a range of 0-2, and a subset of the features are summed to form three subtotals, namely the *Communication (Comm.) subtotal* (range: 0-6), the *Reciprocal Social Interaction (Soc.) subtotal* (range: 0-16), and the *Restricted Repetitive Behaviors (RRB) subtotal* (range: 0-8), with slight differences for the 'no words' and 'some words' categories. The overall total is the sum of these three subtotals. The algorithm was revised in ADOS-2 to increase the robustness of the diagnosis [3]. Since both the original and revised algorithm operate on the same features, we can safely use the more robust algorithm [3] to compute those subtotals.

Fig. 2(a) shows the data points in the 3D *ADOS subtotal space*, where each axis corresponds to one subtotal. Consistent with our SOM analysis, we observe that the data points do not form clearly separated clusters, but rather present some low density regions.

[2]These ADOS scores are part of a database kept for statistical purposes. All data is anonymous; only age and gender were collected from the sample for biographical characterization.

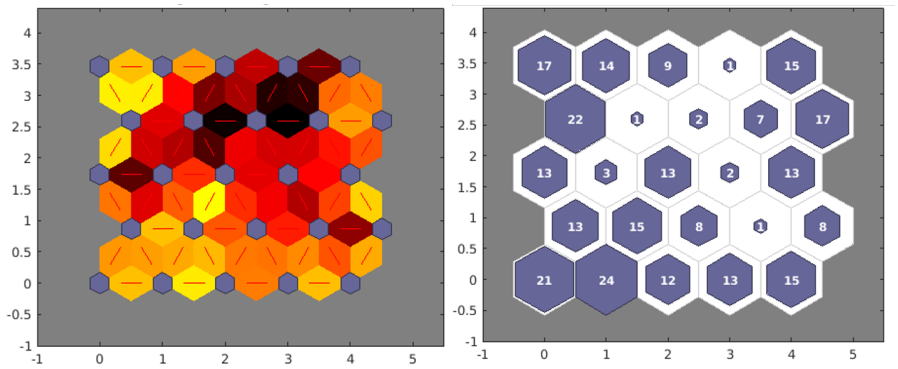

Fig. 1. 2D visualization of the real dataset using a Self-Organizing Map (SOM). **Left**: U-Matrix showing distance between neighboring neurons. The darker the color, the more separated the connected neurons. **Right**: Sample hit histogram showing the number of data points mapping to each neuron. The U-matrix and Sample Hit plots suggest that there are no clearly defined clusters, but rather low-density regions in the input space. SOM parameters: # epochs for training (1,000); distance type (link distance); initial input space covering (100); initial neighborhood size (3).

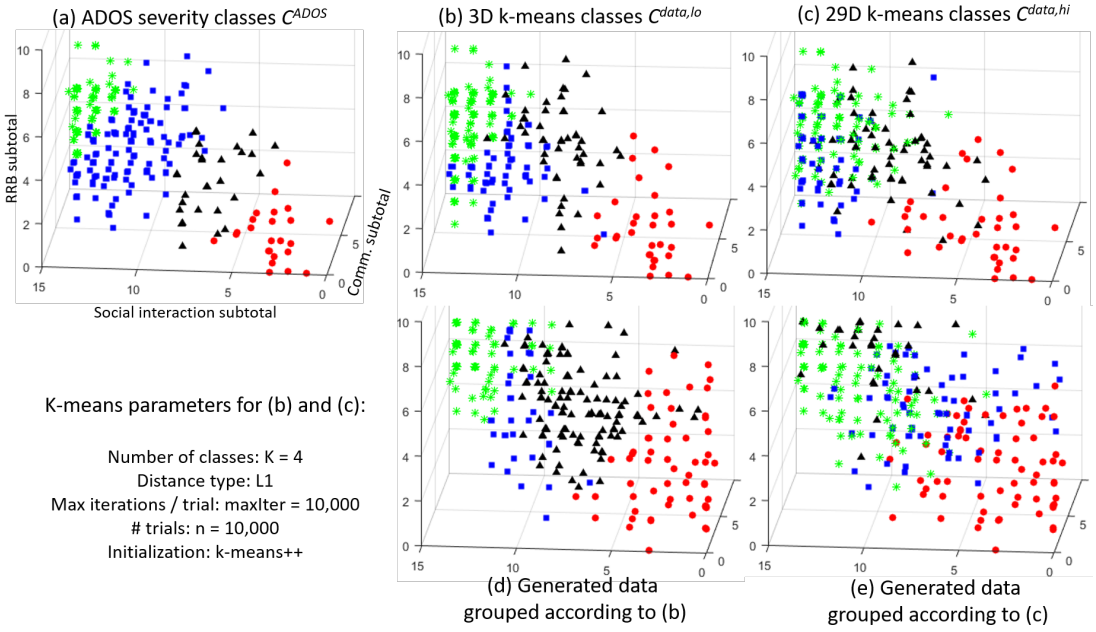

Fig. 2. **(a)** Real data grouped according to descriptor D^{ADOS} (red circles: Minimal to No Evidence, black triangles: Low, blue squares: Moderate, green asterisks: High). **(b)** Real data grouped according to K-means run in the ADOS subtotal space (descriptor $D^{\mathrm{data,lo}}$). **(c)** Real data grouped according to K-means run in the full feature space (descriptor $D^{\mathrm{data,hi}}$) and visualized in the ADOS subtotal space. **(d)** ADOS subtotal space visualization of synthetic feature vectors generated according to the sampling method described in section VI, and grouped according to $D^{\mathrm{data,lo}}$. **(e)** Same as (d) but grouped according to $D^{\mathrm{data,hi}}$. **Notes**: Axes labels shown in (a) and consistent in all other plots. Very small perturbations were applied to the visualized data to be able to distinguish points that are exactly overlapping.

V. Descriptors for Partitioning the High-Dimensional Feature Space

In this section, we introduce descriptors as a tool to partition the large feature space into classes. As a result, we are able to characterize large sets of 'similar' feature vectors by grouping them under the same class.

A. Descriptor formalism and existing descriptor

We define a descriptor D as a function mapping any feature vector $\mathbf{f}$ to a class index k in $\{1,\ldots,K\}$. A descriptor partitions the feature space into K classes C_k, such that

$$C_k = \{\mathbf{f} | D(\mathbf{f}) = k\},\ k = 1,\ldots,K.$$

In this paper, we consider three types of descriptors: ADOS-based D^{ADOS}, low-dimensional data-driven $D^{\mathrm{data,lo}}$, and high-dimensional data-driven $D^{\mathrm{data,hi}}$.

ADOS-based descriptor (D^{ADOS}): The ADOS algorithm produces an overall total, as explained in section IV-B, which is then converted to a comparison score (1-10), thresholded to form 4 severity classes. This function from feature space to class index can be thought of as a descriptor with 4 corresponding classes, namely 'Minimal to No Evidence' (C_1^{ADOS}), 'Low' (C_2^{ADOS}), 'Moderate' (C_3^{ADOS}), and 'High' (C_4^{ADOS}). We overlaid the class information according to this descriptor on the scatter plot of Fig. 2(a).

B. Generating new descriptors from the real data

The existing descriptor D^{ADOS} classifies subjects into 4 classes, designed to capture a scale of different autism severities. These classes are useful for diagnosing and informing decisions such as whether or not the subject needs therapy. However, from a behavioral modeling point of view, this descriptor may be neglecting important behavioral aspects that are not directly related to a one-dimensional scale of autism severity. More specifically, there are two limitations to the existing descriptor, as we explain next.

First, two subjects can have the same overall totals but very different subtotals (for instance, one subject might have a very high RRB subtotal and a very low Comm. and Soc. subtotal and another might have medium values on all subtotals). In this case, it is not clear whether or not it is natural to group them under the same class. Second, although only 14 out of the 29 features have been identified as having enough predictive power when it comes to the autism severity, the remaining 15 features carry behavioral information that might be useful in the behavioral model. Also, the calculation of totals involves remapping which reduces the resolution of some features by lumping values of 2 and 3 in one category.

In order to address these limitations, we use a data-driven approach to generate new descriptors obtained through *clustering* of the real data points. Even though we established in section IV the absence of clearly separated clusters in the data, clustering algorithms effectively define regions of the feature space using the distribution of the data across that space. To address the first limitation, we perform clustering in the 3D ADOS subtotal space to generate descriptor $D^{\mathrm{data,lo}}$. To address the second, we perform clustering in the full feature space to generate descriptor $D^{\mathrm{data,hi}}$.

There exist many types of clustering algorithms, broadly categorized as density-based, distribution-based, connectivity-based, and centroid-based methods. Density-based clustering [18] assumes large density differences within and between clusters, which from our SOM analysis is not a reasonable assumption. Distribution-based clustering (e.g. using Expectation-Maximization over a Gaussian Mixture Model) [19] assumes we know the distribution of the data, which is not a practical assumption since such domain knowledge is hard to approximate. Connectivity-based clustering [20] is not robust to noise and outliers, which makes it not suited for our noisy dataset. Therefore, we perform clustering on our data using a simple K-means [21] (centroid-based approach) with parameters summarized in Fig. 2. The tendency of the algorithm to partition the data into equally-sized regions makes it desirable for our purposes. We select as our number of classes $K = 4$ (similar to D^{ADOS}). We use L1 distance as our distance function since we are dealing with discrete features. An analysis of the resulting class centroids is presented below.

Low dimensional data-driven descriptor ($D^{\mathrm{data,lo}}$): This descriptor maps feature vectors to the index of a class obtained through running K-means on the real data points in the 3D ADOS subtotal space. The resulting centroids for classes $C_1^{\mathrm{data,lo}}$ through $C_4^{\mathrm{data,lo}}$ are $(1, 3, 1.5)$, $(3, 8, 5)$, $(5, 11, 3)$, and $(6, 13, 5)$ (vector order is Comm., Soc., RRB). Fig.2(b) shows the real data points grouped according to this descriptor. Comparing this partition to that of D^{ADOS} shows that $D^{\mathrm{data,lo}}$ captures differences in the RRB subtotal not reflected in the 'Moderate' class region of D^{ADOS}.

High-dimensional data-driven descriptor ($D^{\mathrm{data,hi}}$): This descriptor maps feature vectors to the index of a class obtained through running K-means on the real data points in the full feature space. We analyze the resulting class centroids[3] by looking at the (sample) variance for the centroid features as well as the (sample) correlation between the centroid features. The highest variance, corresponding to features that vary most across the 4 class centroids, occurs for feature A7: 'Pointing' followed by feature A1: 'Overall Lev. of Non-echoed Lang.'. The lowest variance occurs for features E3: 'Anxiety' and D3: 'Self-Injurious Beh.' where all 4 values are zero for both E3 and D3. The most negative correlation (-0.4263) between pairs of features occurs for pairs (E1:'Overactivity', A3:'Intonation of Voc. or Verb.') and (E1:'Overactivity', A5:'Stereotyped/Idiosyncratic Use of Words or Phrases'). However, since both A3 and A5 had a particularly significant number of `NaN` values replaced by random values, this last result might not be very accurate. On the other hand, many features had a correlation of 1 across class centroids, indicating that it is more common to have similar trends in different feature values across classes as opposed to inversely related trends.

Fig.2(b) shows the real data points rendered in the ADOS subtotal space for easy visualization and grouped according to $D^{\mathrm{data,hi}}$. Even though this descriptor still somehow encodes severity, it also captures specific differences that seem to vary more intensely across subjects such as pointing behaviors and use of language. Some overlapping points in the 3D space are even mapped to different classes, validating the fact that D^{ADOS} neglects important features for behavioral modeling.

VI. Generalizing from real data: a synthetic data sampling method

In this section, we discuss a method for generating feature vectors consistent with the correlation structure of our real dataset, for a given class specified by a descriptor.

A. Correlation analysis

In our correlation analysis, we only consider pairwise correlations between features. Fig. 3(a) shows the Spearman correlation matrix for our features. The Spearman's rank correlation coefficient is a non-parametric measure (real number between -1 and 1) capturing how well a monotonic function can be used

[3]The full centroids for $C_1^{\mathrm{data,hi}}$ through $C_4^{\mathrm{data,hi}}$ are:
(1,1,0,0,0,0,0,1,0,0,1,1,0,0,1,1,0,1,0,1,1,2,0,0,0,0,1,0,0)
(1,1,2,1,2,0,1,1,2,1,1,1,1,0,1,1,2,2,0,1,1,1,1,1,0,1,0,0,0)
(3,2,2,1,2,1,3,2,2,2,2,2,1,1,1,1,2,2,0,2,2,3,2,2,0,2,1,0,0)
(3,2,1,2,1,2,3,2,2,2,2,2,2,2,2,2,2,2,2,2,3,1,1,0,1,2,1,0)

to describe the relation between two random variables (the sign indicates whether this function is increasing or decreasing) [22]. It is a well suited measure of correlation when dealing with ordinal variables like our features.

We observe that many features are positively correlated, which is not surprising given that neurological, developmental, and genetic [23] causal relationships have been found to explain a large set of behaviors in autistic subjects. However, some features seem to have a low or even slightly negative correlation with other features, especially in the D (Stereotyped Behaviors and Restricted Interests), and E (Other Abnormal Behaviors) categories.

B. Unconstrained generation of synthetic feature vectors

Our aim is to generate synthetic data such that correlations between the features are maintained. Since the features are discrete and ordinal, such a task is non-trivial. There exist methods to sample ordinal correlated data such as the Gaussian copula [24], the binary conversion, and the mean mapping [25] methods. In this work, we opt for the mean mapping method, which gave best results for our application. The method takes as an input the target correlation matrix and the marginals for each feature, and performs the following steps [25]:

1) For the given marginals, compute the quantiles assuming an underlying Gaussian model for each feature.
2) Estimate a corresponding correlation matrix in continuous space, where ordinal variables are replaced with the underlying Gaussian variables. This step involves interpolating a function over a regular grid of computed probabilities to estimate correlation coefficients.
3) Sample normal data according to the estimated corresponding correlation matrix and cut the samples according to the computed quantiles to get back ordinal data.

The resulting samples asymptotically achieve the target correlation structure and marginals.

Fig. 3(b-d) shows the sample correlation matrix for an increasing number of generated feature vectors, along with the associated RMS errors (rmsE) and maximum absolute errors (maxE) with respect to the target matrix (Fig. 3(a)). As a heuristic to enforce uniform generation across descriptor classes, we set our target marginals to uniform distributions over the feature ranges. The generated feature vectors for $M = N = 279$ are plotted in Fig. 2(d-e), and grouped according to $D^{\text{data,lo}}$ and $D^{\text{data,hi}}$ respectively for comparison.

C. Generating synthetic feature vectors according to descriptors

As mentioned previously, descriptors are a convenient tool to specify the type of synthetic data an algorithm generates. They can be thought of as high-level controls a potential user can specify. To incorporate the constraint on the type of data to be generated, we present a simple alteration of the original mean mapping algorithm, Descriptor-Based Mean Mapping Sampling (DBMMS), to sample feature vectors belonging to a specific class C_i^l defined by descriptor D^l. The algorithm relies on the idea of rejection sampling, by which only the generated samples that fall under a given descriptor are accepted, otherwise they are rejected. This method does not depend on the choice of descriptor or feature choice.

Algorithm 1 shows a pseudocode of the algorithm for centroid-based descriptors such as $D^{\text{data,lo}}$ and $D^{\text{data,hi}}$. Parameters of the mean mapping method `params`, including the corresponding correlation matrix and quantiles, only need to be computed once before starting the sampling process. If the closest centroid index $\hat{k}$ to the generated feature vector is equal to the desired centroid index, the sample is accepted; otherwise it is rejected. The effect of the descriptor constraint on the generated data points in Fig. 2(d-e) would simply be to filter out any data point outside the specified class.

Algorithm 1 Descriptor-Based Mean Mapping Sampling (DBMMS) algorithm to generate M synthetic feature vectors in class C_k^l, according to descriptor D^l represented as a list of centroids $\mathbf{c^l} = (c_1^l, \ldots, c_K^l)$, targeting correlation matrix `corr` and feature marginals `marginals`

procedure DBMMS($M, \mathbf{c^l}, k$, corr, marginals)
  $\hat{\mathbf{x}} \leftarrow \emptyset$
  $m \leftarrow 0$
  params $\leftarrow$ MeanMappingParams(corr, marginals)
  while $m < M$ **do**
    $\hat{x} \leftarrow$ MeanMapping(params,1)
    $\hat{k} \leftarrow \text{argmin}_{k \in \{1,\ldots,K\}} |\hat{x} - c_k^l|$
    if $\hat{k} = k$ **then**
      $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} \cup \{\hat{x}\}$
      $m \leftarrow m + 1$
  return $\hat{\mathbf{x}}$

VII. Conclusion and Future Work

This work aimed at devising a method to generalize from real behavioral data on children with ASD to consistently generate synthetic data. Our dataset consisted of the ADOS Module 1 features for 279 children suspected of having an ASD. We began by analyzing the data to assess the distribution of the data points in the feature space using two dimensionality reduction methods. The first one used a Self-Organizing Map to visualize the data with a learned 2D representation, and the second one used the 3D space defined by the ADOS subtotals. We moved on to introduce two new data-driven descriptors of the feature space using a K-means clustering algorithm. The resulting classes capture subtle variability that the existing ADOS descriptor neglects. Finally, we present a descriptor-based sampling method which preserves the correlation structure of the data. The method is a modification of the mean mapping algorithm for generating correlated ordinal values.

In the future, we would like to use the generated synthetic data to train algorithms for a social agent interacting with children with ASD. The feature values can be transformed into a reward function which, in combination with a transition model, could potentially be used in the context of model-based reinforcement learning approaches to learn optimal policies,

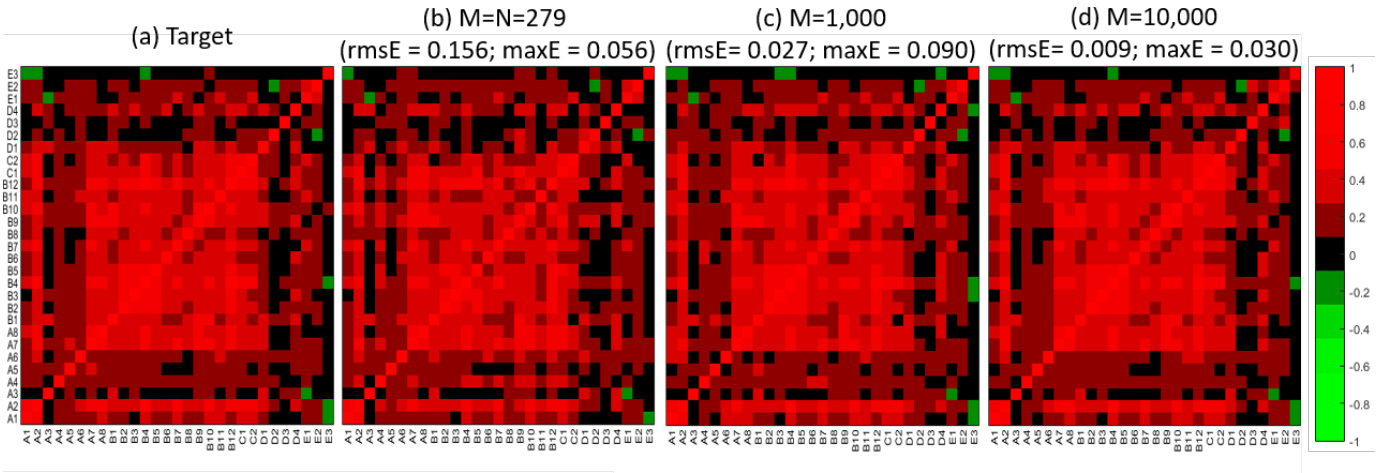

Fig. 3. Spearman correlation matrix for: (a) the feature vectors from the real dataset; (b-d) feature vectors from a generated dataset sampled according to the mean mapping method of (b) the same size as the real dataset, (c) of size 1,000, and (d) of size 10,000.

hence providing adaptability to each child's therapy needs. Alternatively, the generated data could be used as part of a behavioral simulator of children with ASD in order to train apprentice therapists to administer the ADOS in simulation.

ACKNOWLEDGMENTS

This research was partially supported by the CMUP-ERI/HCI/0051/2013 grant and national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013. We would like to thank the National Database for Autism Research (NDAR) and the CDC at Hospital Garcia de Orta for granting us access to the data. We would also like to thank Marta Couto, Anabela Farias, and the INSIDE project for assisting us with some of the data used in this research. The views and conclusions of this document are those of the authors only.

REFERENCES

[1] M. Sigman, S. J. Spence, and A. T. Wang, "Autism from developmental and neuropsychological perspectives," *Annu. Rev. Clin. Psychol.*, vol. 2, pp. 327–355, 2006.

[2] R. Muhle, S. V. Trentacoste, and I. Rapin, "The genetics of autism," *Pediatrics*, vol. 113, no. 5, pp. e472–e486, 2004.

[3] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of autism and developmental disorders*, vol. 37, no. 4, p. 613, 2007.

[4] A. Le Couteur, C. Lord, and M. Rutter, "The autism diagnostic interview-revised (adi-r)," *Los Angeles, CA: Western Psychological Services*, 2003.

[5] E. Schopler, R. J. Reichler, and B. R. Renner, *The childhood autism rating scale (CARS)*. Western Psychological Services Los Angeles, CA, 2002.

[6] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research," *Annual review of biomedical engineering*, vol. 14, pp. 275–294, 2012.

[7] D. J. Feil-Seifer, "Data-driven interaction methods for socially assistive robotics: validation with children with autism spectrum disorders," Ph.D. dissertation, University of Southern California, 2012.

[8] J. A. Kientz, M. S. Goodwin, G. R. Hayes, and G. D. Abowd, "Interactive technologies for autism," *Synthesis Lectures on Assistive, Rehabilitative, and Health-Preserving Technologies*, vol. 2, no. 2, pp. 1–177, 2013.

[9] E. Lundin, H. Kvarnström, and E. Jonsson, *A Synthetic Fraud Data Generation Methodology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 265–277.

[10] P. J. Lin, B. Samadi, A. Cipolone, D. R. Jeske, S. Cox, C. Rendon, D. Holt, and R. Xiao, "Development of a synthetic data set generator for building and testing information discovery systems," in *Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on*. IEEE, 2006, pp. 707–712.

[11] D. M. Smith, G. P. Clarke, and K. Harland, "Improving the synthetic data generation process in spatial microsimulation models," *Environment and Planning A*, vol. 41, no. 5, pp. 1251–1268, 2009.

[12] Y. Yu, D. Ganesan, L. Girod, D. Estrin, and R. Govindan, "Synthetic data generation to support irregular sampling in sensor networks," *GeoSensor Networks*, pp. 211–234.

[13] J. E. Hoag, *Synthetic Data Generation: Theory, Techniques and Applications*. ProQuest, 2008.

[14] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[15] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[16] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[17] A. Ultsch, *Self-Organizing Neural Networks for Visualisation and Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 307–313.

[18] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[20] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, p. 364, 1977.

[21] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[22] G. W. Corder and D. I. Foreman, *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014.

[23] P. Chaste, M. Leboyer *et al.*, "Autism risk factors: genes, environment, and gene-environment interactions," *Dialogues Clin Neurosci*, vol. 14, no. 3, pp. 281–92, 2012.

[24] L. Madsen and D. Birkes, "Simulating dependent discrete data," *Journal of Statistical Computation and Simulation*, vol. 83, no. 4, pp. 677–691, 2013.

[25] S. Kaiser, D. Träger, and F. Leisch, "Generating correlated ordinal random values," 2011.