Topic Models for Automated Motor Analysis in Schizophrenia Patients

Talia Tron*[1,4] Yehezkel S. Resheff*[1,4] Mikhail Bazhmin[2] Abraham Peled[2,3] Alexander Grinsphoon[2,3] Daphna Weinshall[4]

***Abstract*— Wearable devices fitted with various sensors are increasingly being used for the automatic and continuous tracking and monitoring of patients. Only first steps have been taken in the field of psychiatric care, where long term tracking of patient behavior holds the promise to help practitioners to better understand both individual patients, and the disorders in general. In this paper we use topic models for unsupervised analysis of movement activity of schizophrenia patients in a closed ward setting. Results demonstrate that features computed on the basis of this analysis differentially characterize interesting sub-populations of schizophrenia patients. Positive-signs schizophrenia sub-population was found to have high *motor richness* and low *typicallity*, while negative-signs patients had low *motor richness* and lower *typicality*. In addition we design a classifier which correctly classified up to 80% of the clinical sub-population (f-score=0.774) based on motor features.**

I. INTRODUCTION

Motor peculiarities are an integral part of the schizophrenia disorder, both as aspects of the more general symptom repertoire, and in response to medications. To date, these symptoms are typically evaluated in a descriptive manner based on psychiatric rating scales such as the Positive and Negative Syndrome Scale (PANSS) [1], or targeted specifically using subjective clinical scales such as the Unified Dyskinesia Rating Scale (UDysRS) [2] and the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [3]. The lack of objective, quantitative methods for measuring these symptoms, and the insufficient conceptual clarity around them, may cause multiple interpretations of phenomenology, leading to low reliability and validity of diagnosis. In addition, the symptom evaluation process requires expert staff and availability of resources, and is therefore not done frequently enough to capture more subtle changes in spontaneous and drug-induced conditions. Clearly there is an urgent need for automatic monitoring and assessment tools.

The last decade has seen a steep rise in the use of wearable devices for medical applications in a range of fields, from human physiology [4] to movement disorders [5], [6] and mental health [7]. Accelerometers and gyroscopes, which are commonly embedded in smart-watches and other wearable devices, are now used to assess mobility and recognize activity. In a clinical setting, these sensors may be used in order to detect changes in high-level movement parameters, track their dynamics and correlate them with mental state.

Measures of activity such as step counts and overall activity, as well as changes thereof, have already been shown to effectively provide insights into the state of patients in a closed ward mental hospital setting [8]. Unsupervised behavioral mode analysis of sensor data, such as topic models, have previously been used in other domains to provide a high level description of behavior [9]. Here we combine these ideas and use topic models for unsupervised analysis of patient activity. These models allow a richer, qualitative description of behavior than the aforementioned measures. We demonstrate that features computed on the basis of topic model analysis differentiate sub-populations of patients.

II. MATERIALS AND METHODS

A. Study Design

27 inpatients from the closed wards at Shaar-Meashe MHC participated in the study. Most participants (21/27) were diagnosed with schizophrenia according to the DSM-5, 3 with paranoid schizophrenia, 2 with schizoaffective disorder, and one with psychotic state cannabinoids. Participants' age varied from 21 to 58 (mean 37.5), with course of illness varying from 0 (first hospitalization) up to 37 years (mean 16.9 years). Two of the patients dropped out of the study after less than a day due to lack of cooperation. The remaining 25 patients were followed for a period of three weeks on average (6-52 days).

The study was conducted in natural settings, where patients were *not* required to change any personal or medical procedure. On top of the normal care, every patient underwent an additional evaluation by a trained psychiatrist twice a week. The procedure included clinical evaluation of symptom severity using PANSS; Neurological Evaluation Scale (NES [10]) assessment was conducted as a control. In addition, continuous medication monitoring (type, dosage and frequency) by the clinical staff was observed.

All procedures performed in the study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

B. Data Acquisition

Each participant was fitted with a smart-watch (GeneActiv[1]) with tri-axial accelerometer embedded sensors, the high

* T. Tron and Y.S. Resheff contributed equally to this work.
[1]The Edmond and Lily Safra center (ELSC) for Brain Science, Hebrew University, Jerusalem 91904, Israel talia.tron@mail.huji.ac.il
[2]Rappaport Faculty of Medicine, Technion Institute of Technology, Haifa 3200003, Israel
[3]Sha'ar Menashe Mental Health Center, Sha'ar Menashe 38706, Israel
[4]The Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel

[1]https://www.geneactiv.org/

frequency output ($50Hz$) of which was stored on memory cards. Data was collected continuously throughout the experiment for a total of 489 days, from 25 patients. In order to reduce noise introduced by the variability in patient activity, the analysis focused on fixed time windows corresponding to regular departmental daily activity: Occupational therapy time slots (10am-11am), lunch (12pm-1pm), and indoor free time (4pm-5pm). In addition, we calculated full day features (6am-10pm) and used night time features (10pm-6am) to evaluate sleep quality. Weekends were excluded from the analysis.

III. ANALYSIS

A. Revising clinical assessment

The 30-item Positive and Negative Syndrome Scale (PANSS) was reduced to a five-factor description (Positive, Negative, Disorganized/Concrete, Excited and Depressed), according to the consensus model suggested by Wallwork et al. [11], based on 25 previously published models and refined with confirmatory factor analysis (CFA). Only the positive and negative factors were used for further analysis.

Clinical observations show that changes in a patient's symptoms occur continuously on a daily basis [12]. We therefore interpolated the bi-weekly PANSS factor scores, to achieve smooth daily scoring of symptoms. This was done using the PCHIP 1-d monotonic cubic interpolation, resulting in 494 data points (originally 118).

Interpolated data points were used to classify clinical sub-populations of patients on a daily basis. Sub-populations included patients with "High positive" symptoms, "High negative", "High negative and positive", and "Low" level symptoms. The remaining intermediate data points were discarded from the classification. This sub-typing allowed us to explore how different motor features are expressed in different clinical manifestations. Clustering was done based on the percentile of the positive and negative factors, each axis separately (Fig. 1).

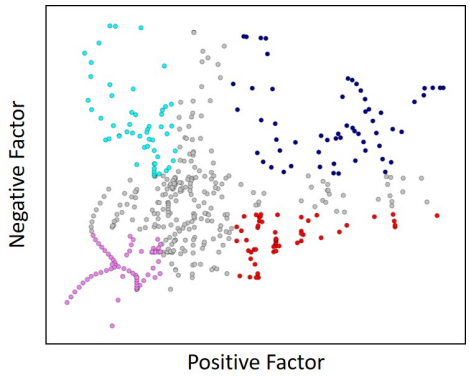

Fig. 1. Clinical sub-populations. Each data point represents the severity of the positive and negative factors for a specific patient in a specific day (N=494) based on the interpolated PANSS factors data. In the "Low" sub-population (magenta, N=65) both negative and positive symptoms lie in the bottom quartile, while in the "High negative and positive" (blue, N=59) both lie in the top quartile. The "High negative" sub-population (cyan, N=53) lies in the top vertical quartile with positive symptom values lower than median, while the "High positive" (red, N=57) lies in the top horizontal quartile with lower than median negative symptom values. The remaining data points (N=260) were classified as "Intermediate" (green).

B. Online computation of "patch features"

Topic model analysis requires the discretization of the continuous accelerometer signal both in time and in intensity, to produce word analogues – motor words. This mapping involves the creation of a code-book. The patch feature topic model procedure described in [9] contains a codebook generation stage where clustering (k-means) is applied to segments (a.k.a. patches) from the entire dataset. Given the larger dataset at hand, we designed an online greedy approximation to this procedure.

Specifically, the idea behind an online generation of code-book is to follow the way a dictionary would be created for a natural language corpus. The process proceeds with a single pass over the data. Each word is considered sequentially, and added to the dictionary on first encounter.

Since the words we are using describe the continuous accelerometer signal, we must also define what we mean by a word. Ideally, the dictionary should not be affected by small random changes in the signal. Additionally, since many behavioral modes are periodic to some extent, we would like the representation to allow wrap-around of patches. This would imply that the sequences of patches ABC and BCA, for example, have similar representation in the dictionary.

We achieve both these goals by using a discretized version of the signal and wrap-around equivalence classes. We use a SAX-like method [13] to encode each patch into a string. The process is as follows: Each interval on the time-axis is replaced by the mean value in the interval. Next, these point-values are replaced by a letter (discretized) according to their value. The output of this process is a string of length $\frac{patch-size}{interval-size}$ over a pre-determined alphabet.

Algorithm 1 Online codebook creation

```
codebook ← empty list
for each patch in the dataset do
    patch_word ← SAX_representation(patch)
    if patch_word (or equivalent) not in codebook then
        append patch_word to the codebook
    end if
end for
return codebook
```

The procedure resulted in 150K distinct words which described the entire dataset, distributed much as would be expected from a text corpus (see top panel in Fig. 2).

C. Topic Modeling over Motor Words

Latent Dirichlet Allocation (LDA) is a widely used topic model, with origins in natural language processing, and applications in many domains ranging from music modeling to motion of cars. On top of their traditional purpose of finding hidden semantic structures in data, these models have been shown to be useful for detecting surprising (or novel) events [14], [15].

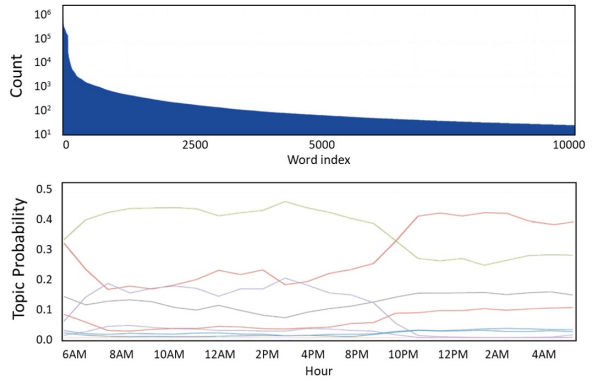

Fig. 2. Top - the motor-word frequency histogram for the entire dataset, truncated after the most common 10K words. These pseudo-words demonstrate the long-tail scale-free property characteristic of a natural language. Bottom- the daily topic distribution vector averaged over all patients.

Data was divided into blocks of 15 minutes of continuous signal; these serve as documents for the topic model, each represented as a histogram over the motor words as described above. The LDA process provides as output both a distribution over topics for each of the documents, and a distribution over words for each of the topics. Subsequently, a specific time window of a specific patient is characterized by a probability vector over the topics. The bottom panel in Fig. 2 demonstrates the distribution of the 10 topics used here over all patients and days. We can see that topic 6 (green) and topic 10 (purple) are typically prominent during the day, while other topics are more likely to occur during the night or throughout.

D. Topic Features

The advantage of using a data-driven unsupervised representation is that its features, unlike the supervised energy and step-count measures [8], are not directly connected with the intensity of the motor signal. Instead, this representation captures the *quality* of motor behavior in a given period of time. For example, a very repetitive behavior can be expressed by a low number of unique 'motor-words' in a specific time window. This allows us to compare patients behavior to themselves and to others in different activity windows, and thus be able to measure 'typicality' of the behavior for instance. Three Features were calculated based on topic models, separately for each data point (namely for each patient on each day, and each of the predefined activity windows described in section II-B):

1) Motor Richness: The normalized distinct word count per activity window.

$$\text{Motor Richness} = \frac{\text{distinct word count in window}}{\text{window length}}$$

This measure represents the range of motor activity repertoire. A low score implies that the patient repeatedly performed similar movements, while a high score corresponds to the use of many different movement patterns.

2) Consistency:

$$\text{Consistency} = 1 - D_{KL}(v \parallel \bar{v})$$

where v denotes the topic distribution over the time window, and $\bar{v}$ the mean topic distribution for the patient in the same window over all measured days. D_{KL} denotes the Kullback-Leibler divergence. This score measures how regular the patient's motor behavior is in the given time window.

3) Typicality: the entropy of the topic distribution vector.

$$\text{Typicality} = H[v] = -\sum_{i=1}^{10} v_i \log(v_i)$$

Low entropy implies that a small number of topics can capture the activity. High entropy implies that the observed activity is a mixture of many topics. We name this measure *typicality* since typical activity should be captured by one (or a few) topics, thus producing low-entropy topic distributions [15].

The manifestation of each feature in clinical sub-populations was tested using one-way ANOVA separately for each of the activity windows. In addition, a learning algorithm for automated sub-population classification was designed and evaluated.

E. Classification Algorithm

Classification was carried out using a two-step algorithm based on linear support vector machines (SVM) and decision trees classifiers, in order to distinguish between different sub-populations (described in III-A) based on motor features (Algorithm 2). The algorithm was trained to discriminate sub-populations, and specifically classify "High positive" vs. "High negative" and "Low" vs. "High positive and negative".

In the first step, individual classifiers were trained for each activity window separately (lunch, occupational therapy, free-time, day, night, and all). In the second step, the probabilistic output of the 6 time-specific first-stage classifiers was used to train a second, daily-model, which determined the clinical category.

Feature selection was done based on the ANOVA f-values of each individual feature on train data. These were calculated separately for each activity window, and the same features were used also for testing.

Algorithm 2 Two stage algorithm for patient sub-type classification based on activity in time-windows.

1: **for all** time-windows w_i **do**
2: train base classifier c_i on w_i and target y
3: **end for**
4: **for all** time-windows w_i **do**
5: $\hat{y}_i \leftarrow$ prediction of c_i on w_i
6: **end for**
7: train final classifier c on the set of first-stage predictions $\hat{y}_i$ and target y

The algorithm was evaluated in a leave-one-out framework, where in each iteration a different observation (specific

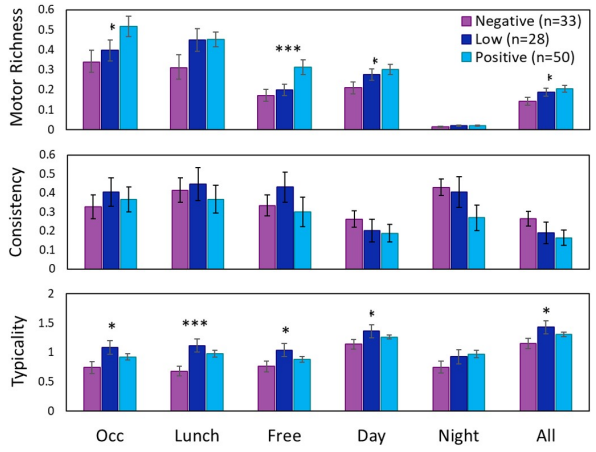

Fig. 3. ANOVA results for topic features in 3 clinical sub-populations (see Fig. 1): "High positive" (denoted *Positive*), "High negative" (denoted *Negative*), and "Low". The analysis was repeated separately for each time window (X-axis). *Motor richness* was highest in the *Positive* sub-population (cyan) and lowest in the *Negative* sub-population (magenta). This was most significant during free time, but was also true for all other activity windows (p-values between 0.05-0.07 are marked by half an asterisk). *Typicality* was generally highest in the *Low* sub-population, and lowest in the *Negative* sub-population, with the most significant difference during lunch time. No significant group different was found for *Consistency* although it was lowest in the *Positive* sub-population in all activity windows.

patient in a specific day) was left out and the model was trained on the remaining data and tested on the left out sample. To avoid possible contamination of test data (leakage) due to observation interpolation, when using an interpolated point as the test, all actual observations it was based upon were excluded from the train data.

IV. RESULTS

A. Motor Activity in different Clinical Sub-populations

Fig. 3 summarizes the results of subjecting all features to ANOVA analysis. *Motor richness* is consistently highest for the "High positive" sub-population, and lowest for the "High negative" sub-population, with the "Low" sub-population somewhere in the middle. This indicates that patients with active positive symptoms tend to have a higher variety of motor activities, while negative symptoms are expressed in poorer movement repertoire. The trend was evident in all activity windows but was only found significant during free time ($F = 5.09$, $p = 0.0077$).

As expected, *typicality* is highest for the "Low" sub-population, consistently over all activity windows. The lowest *typicality* is observed in the "High negative" sub-population, indicating that the motor activity of these patients is less similar to the common motor behavior. The biggest group difference was found over lunch time ($F = 7.48$, $p = 0.00090$) but it was also significant during occupational therapy ($F = 4.39$, $p = 0.015$), free time ($F = 3.38$, $p = 0.037$) and throughout the day ($F = 3.78$, $p = 0.026$). No group difference was found for *consistency*, although it was lower in the "High positive" sub-population in all activity windows.

B. Classification Results

For "High positive" vs. "High negative" classification, the best results were achieved using linear SVM for the first stage (window-based model, see Algorithm 2) with top-5 selected features, and decision tree for the second (daily) stage. The algorithm correctly classified 78% of the"High negative" observations, and 58% of the "High positive" observations. All together the mean precision was 0.651 and mean recall was 0.654 on test data (f-score=0.652).

Slightly better results were achieved for the "Low" vs. "High positive and negative" classification, using linear SVM for both stages and top-5 selected features. Here the algorithm correctly classified 81% of the "Low" observations and 70% of the "High positive" observations. All together the mean precision was 0.757 and mean recall was 0.748 on test data (f-score=0.774).

REFERENCES

[1] S. R. Kay, A. Flszbein, and L. A. Opfer, "The positive and negative syndrome scale (panss) for schizophrenia.," *Schizo. bulletin*, vol. 13, no. 2, p. 261, 1987.

[2] C. G. Goetz, J. G. Nutt, and G. T. Stebbins, "The unified dyskinesia rating scale: presentation and clinimetric profile," *Mov. Dis.*, vol. 23, no. 16, pp. 2398–2403, 2008.

[3] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. Stebbins, *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Mov. dis.*, vol. 23, no. 15, pp. 2129–2170, 2008.

[4] J. Staudenmayer, S. He, A. Hickey, J. Sasaki, and P. Freedson, "Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements," *Journal of Applied Physiology*, vol. 119, no. 4, pp. 396–403, 2015.

[5] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian, and W. Grundfest, "Implementation of an iphone for characterizing parkinson's disease tremor through a wireless accelerometer application," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 4954–4958, IEEE, 2010.

[6] A. Wagner, N. Fixler, and Y. S. Resheff, "A wavelet-based approach to moniotring parkinson's disease symptoms," *International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[7] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM Int Joint Conf on Pervasive and Ubiquitous Computing*, pp. 3–14, ACM, 2014.

[8] T. Tron, Y. S. Resheff, M. Bazhmin, A. Peled, and D. Weinshall, "Real-time schizophrenia monitoring using wearable motion sensitive devices," in *MobiHealth*, Springer, 2016.

[9] Y. S. Resheff, S. Rotics, R. Nathan, and D. Weinshall, "Topic modeling of behavioral modes using sensor data," *International Journal of Data Science and Analytics*, vol. 1, no. 1, pp. 51–60, 2016.

[10] R. W. Buchanan and D. W. Heinrichs, "The neurological evaluation scale (nes): a structured instrument for the assessment of neurological signs in schizophrenia," *Psychiatry research*, vol. 27, no. 3, pp. 335–350, 1989.

[11] R. Wallwork, R. Fortgang, R. Hashimoto, D. Weinberger, and D. Dickinson, "Searching for a consensus five-factor model of the positive and negative syndrome scale for schizophrenia," *Schizophrenia research*, vol. 137, no. 1, pp. 246–250, 2012.

[12] S. Arieti, *Interpretation of schizophrenia*. Basic Books (AZ), 1974.

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, ACM, 2003.

[14] U. Shalit, D. Weinshall, and G. Chechik, "Modeling musical influence with topic models," in *International Conference on Machine Learning*, pp. 244–252, 2013.

[15] A. Hendel, D. Weinshall, and S. Peleg, "Identifying surprising events in videos using bayesian topic models," in *Asian Conference on Computer Vision*, pp. 448–459, Springer, 2010.