Gaussian Latent Dirichlet Allocation for Discrete Human State Discovery

Congyu Wu\textsuperscript{1}, Aaron Fisher\textsuperscript{2}, and David Schnyer\textsuperscript{1}

\textsuperscript{1}Department of Psychology, University of Texas, Austin
\textsuperscript{2}Department of Psychology, University of California, Berkeley

June 2022

Abstract

In this article we propose and validate an unsupervised probabilistic model, Gaussian Latent Dirichlet Allocation (GLDA), for the problem of discrete state discovery from repeated, multivariate psychophysiological samples collected from multiple, inherently distinct, individuals. Psychology and medical research heavily involves measuring potentially related but individually inconclusive variables from a cohort of participants to derive diagnosis, necessitating clustering analysis. Traditional probabilistic clustering models such as Gaussian Mixture Model (GMM) assume a global mixture of component distributions, which may not be realistic for observations from different patients. The GLDA model borrows the individual-specific mixture structure from a popular topic model Latent Dirichlet Allocation (LDA) in Natural Language Processing and merges it with the Gaussian component distributions of GMM to suit continuous type data. We implemented GLDA using STAN (a probabilistic modeling language) and applied it on two datasets, one containing Ecological Momentary Assessments (EMA) and the other heart measures from electrocardiogram and impedance cardiograph. We found that in both datasets the GLDA-learned class weights achieved significantly higher correlations with clinically assessed depression, anxiety, and stress scores than those produced by the baseline GMM. Our findings demonstrate the advantage of GLDA over conventional finite mixture models for human state discovery from repeated multivariate data, likely due to better characterization of potential underlying between-participant differences. Future work is required to validate the utility of this model on a broader range of applications.

\textsuperscript{*}Correspondence: congyu.wu@austin.utexas.edu
1 Introduction

In social science and medical research, a prevalent mode of data collection is measuring a set of potentially related variables repeatedly at different points in time from a cohort of distinct subjects. For example, when assessing mental wellbeing, psychologists often use a suite of survey questions to solicit the respondent’s in-the-moment experience of multiple mood aspects such as anxiety, sadness, and anger. In the medical context, cardiac, urinary, and serum measures are also extracted simultaneously to form a comprehensive evaluation of a patient’s symptoms and potential causes. These repeated multivariate measures are the backbone of both one-time diagnoses and continued monitoring of illnesses, therefore modeling methods for this type of data are important and applicable to a wide variety of domains. Clinical diagnosis entails categorizing a multivariate observation into a class, or discrete state, that represents a particular kind of illness or the absence of it. However, at any given time, each variable individually likely does not offer sufficient information to reach such categorization. It is often a combination of values across multiple symptom variables that allows doctors or scientists to confidently determine the type of health or behavioral issue. Due to the limited indicativeness of individual variables and the prevalent codependency between variables to form patterns, clustering multivariate observations into compound states is a crucial operation in clinical data science. Knowing the underlying state of a patient at the moment when a multivariate observation is taken may help healthcare providers correctly diagnose patients without being distracted by possibly spurious changes in individual variables in the sample.

Finite mixture models (FMMs) are a popular class of models used to tackle the problem of clustering momentary multivariate observations into states. Compared to partition-based clustering methods such as k-means, FMMs theorize a soft, probabilistic boundary between classes or components, which is especially fitting for psychophysiological and behavioral data. Because, as opposed to applications such as clustering plants or animals into species, boundaries between human behavioral and health states can be fuzzy. FMMs assume any data point is sampled from a mixture of underlying component distributions, which serve as clusters. The probability that a component distribution is to be sampled from is governed by its corresponding mixture weight. The parameters of such a model, including the mixture weights and the means and covariances of the component distributions, are then inferred iteratively based on data. Due to the shared mixture weights across all data used to fit an FMM, the model is best suited for clustering observations from a purportedly homogeneous source, such as from the same subject. When we collect repeated multivariate samples from a cohort of subjects, using FMM to infer underlying compound states becomes potentially problematic. On the one hand, if we fit an FMM to all data belonging to multiple distinct subjects, we are forcing different subjects’ observations to be sampled from one shared mixture, which is a potentially unrealistic assumption and will potentially result in unreliable clusters and weights. On the other hand, if we fit one FMM for each subject, the clusters and weights inferred are for each subject themselves and therefore lacks a mechanism to generalize across and
compare between subjects. Current work is lacking in assessing the potential commonality of the states obtained by individually-fitted FMMs.

To address this limitation, we propose and evaluate a probabilistic model called Gaussian Latent Dirichlet Allocation (GLDA) to combine the advantages of individually-fitted and globally-fitted FMMs and circumvent their respective shortcomings. GLDA was previously and separately introduced in the field of Natural Language Processing as a topic model, serving as a continuous extension of the discrete-input Latent Dirichlet Allocation. To the best of our knowledge, no existing work has investigated the applicability and benefit of the GLDA model for clustering multivariate psychophysiological data. As opposed to a probabilistic structure where data points are directly sampled from a global mixture of components, GLDA utilizes a Dirichlet prior to allow different subjects' observations to be sampled from each individual's own weighted mixture while at the same time keeping a global set of components. The ability for each individual to have their own mixture preserves idiographic differences, which is theoretically superior to fitting one FMM to data from inherently distinct subjects. A global, rather than local, set of components is desirable because clinically meaningful symptoms should be generalizable to different individuals. Practically, the global set of components allows observations from different subjects to be compared, which is not straightforwardly achievable by individually-fitted FMMs. Because of the more intricate dependency structure of GLDA, we hypothesize that the class components and the weights inferred by GLDA will be more meaningful, i.e., better predictive of clinically validated individual outcomes, than those obtained by traditional FMMs.

To test this hypothesis, we built our own implementation of GLDA using STAN, a probabilistic modeling and inference language, and applied it on two datasets that contain measures of participants’ psychology (from ecological momentary assessment or EMA) and physiology (from heart monitors) respectively. Both datasets contain repeated multivariate momentary samples from a cohort of participants and health outcomes clinically diagnosed for each participant. Detailed information about the two datasets is listed in Table 2. For each dataset, we fit a GLDA model to learn the individualized mixture weights and then used these weights to predict individual health outcomes using linear regression. As a baseline, we built a global Gaussian Mixture Model (GMM, a popular member of FMMs) for each dataset and use the realized class frequencies as class weights of each individual to predict the same health outcomes also using linear regression. We compared the significance of the linear regression models built with class weights learned by GLDA and GMM respectively.

Our results show that GLDA achieved predominantly greater significance and variance explained for both datasets. This finding indicates that for at least comparable types of cohorts and measures (i.e., psychophysiological), GLDA is indeed a superior clustering method to GMM for discovering states from repeated multivariate samples collected from different subjects. In the remainder of this article, we detail our experiments, present our results, and reflect on the model’s implications.
| Model       | Plate notation | Sampling process                                                                                                                                 |
|-------------|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| **GMM**     | ![Diagram](GMM_diagram.png) | Sample mixture weights \( \theta \sim \text{Dirichlet}(\alpha) \); For each \( k = \{1, \ldots, K\} \): Sample component mean \( \mu_k \sim \mathcal{N}(\mu_0, \lambda) \), Sample component covariance \( \sigma_k^2 \sim W^{-1}(\nu, \sigma_0^2) \); For each \( i = \{1, \ldots, N\} \): Sample class \( z_i \sim \text{Categorical}(\theta) \), Sample observation value \( x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \). |
| **LDA**     | ![Diagram](LDA_diagram.png) | For each \( m = \{1, \ldots, M\} \): Sample topic distribution \( \theta_m \sim \text{Dirichlet}(\alpha) \); For each \( k = \{1, \ldots, K\} \): Sample word distribution \( \phi_k \sim \text{Dirichlet}(\beta) \); For each \( i = \{1, \ldots, N\} \): Sample topic \( z_i \sim \text{Categorical}(\theta_m) \), Sample word \( w_i \sim \text{Categorical}(\phi_{z_i}) \). |
| **GLDA**    | ![Diagram](GLDA_diagram.png) | For each \( m = \{1, \ldots, M\} \): Sample mixture weights \( \theta_m \sim \text{Dirichlet}(\alpha) \); For each \( k = \{1, \ldots, K\} \): Sample component mean \( \mu_k \sim \mathcal{N}(\mu_0, \lambda) \), Sample component covariance \( \sigma_k^2 \sim W^{-1}(\nu, \sigma_0^2) \); For each \( i = \{1, \ldots, N\} \): Sample class \( z_i \sim \text{Categorical}(\theta_m) \), Sample observation value \( x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i}^2) \). |

Table 1: Comparison of the statistical dependencies and sampling processes of Gaussian Mixture Model (GMM), Latent Dirichlet Allocation (LDA), and Gaussian Latent Dirichlet Allocation (GLDA). Variables in smaller squares are pre-set hyperparameters while the ones in circles are sampled random variables. Note that GLDA’s structure has the top half of GMM and the bottom half of LDA.
2 Gaussian Latent Dirichlet Allocation

The statistical dependency (illustrated with plate notation) and sampling process of GLDA is provided in Table 1, together with those of GMM and LDA for a straightforward comparison. The three models are all unsupervised, generative, probabilistic clustering models. All three models require the researcher to supply the number of clusters $K$ to start running (as opposed to Dirichlet Process GMM [7], a variant of GMM that learns the number of clusters automatically). The GMM shown in Table 1 is a generalized, Bayesian version of the model; a non-Bayesian GMM would simply be a Bayesian GMM with uninformative priors (hyperparameters). Comparing the plate notations of GMM and GLDA, one should notice that in GMM there is one mixture weight vector $\theta$ for the sampling of all $z_i, i \in \{1, ..., N\}$, where $N$ is the total number of observations; whereas in GLDA, a separate mixture weight vector $\theta_m$ is sampled for each participant $m \in \{1, ..., M\}$ where $M$ is the total number of participants. Accordingly, the class and value of each observation belonging to participant $m$ would be sampled from the participant-specific mixture weights. Both GMM and GLDA define the component distributions as Gaussian distributions parameterized by mean $\mu_k$ and covariance $\sigma^2_k$, indicated by the identical top half of their plate notations. Because of their Gaussian components, both GMM and GLDA are better suited for continuous type samples.

LDA, on the other hand, is only applicable to discrete data. It is primarily used for the NLP task of topic modeling for a collection of potentially thematically different documents filled with words, since words in human languages are typically considered disjoint units. A topic is simply a weight vector $\phi$ over a comprehensive collection of words, or the vocabulary, that sum up to 1. Intuitively, the semantic make-up of a topic is determined by the relative weights over different words. LDA assumes that for each document $m \in \{1, ..., M\}$ there is a specific mixture of $K$ topics (topic distribution), governed by a mixture weight vector $\theta_m$. To sample a word $w_i$ in document $m$, LDA first samples a topic $z_i \in \{1, ..., K\}$ according to the document-specific topic distribution $\theta_m$, then samples a word according to the topic-specific word distribution $\phi_{z_i}$. Such a document-topic-word hierarchy corresponds straightforwardly to the participant-class-observation hierarchy in GLDA. Note the identical structure of the bottom half of LDA’s and GLDA’s plate notation: just as each document has a separate topic distribution in LDA, each participant has a separate set of mixture weights in GLDA. The difference is in how the words/observations are sampled after a topic/class $z_i$ is sampled: in LDA a word is sampled from a categorical distribution parameterized by $\phi_{z_i}$ whereas in GLDA, each observation is sampled from a Gaussian distribution parameterized by $\mu_{z_i}$ and $\sigma^2_{z_i}$ (as the initial “G” suggests). Marrying the characteristics of GMM and LDA, GLDA allows the sampling of continuous values from participant-specific mixture weights over discrete states. We suspect this is an attractive structure for human state discovery from repeated multivariate psychophysiological samples.
| Dataset 1 (Psychology) | Dataset 2 (Physiology) |
|-----------------------|------------------------|
| **Subjects**          |                        |
| Source                | Clinical patients      | Clinical patients |
| Sex                   | 69% female             | 61% female        |
| Age                   | Mean age 38.5 years    | Mean age 39 years |
| Number $M$            | 45                     | 166                |
| **Samples**           |                        |
| Instrument            | EMA                    | MindWare Mobile   |
| Variables             | Self-report on 0-100 (4 | Hardware          |
|                       | times/day) of:         |                   |
|                       | rumination, worry,     |                   |
|                       | anger, irritability,   |                   |
|                       | anhedonia, hopelessness|                   |
|                       | depressed mood,        |                   |
|                       | avoidance of action,   |                   |
|                       | avoidance of people    |                   |
| Dimension $V$         | 10                     | 5                  |
| Total $N$             | 5,076                  | 25,394             |
| **Ground truth**      |                        |
| Depression            | HRSD score (Hamilton   |                   |
|                       | Rating Scale for      |                   |
|                       | Depression, 52 pts     |                   |
| Anxiety               | HAMA score (Hamilton   |                   |
|                       | Anxiety Rating Scale,  |                   |
|                       | 56 pts                 |                   |
| Stress                | DASS-S score (Depression Anxiety and Stress Scale–Stress, 42 pts) |

Table 2: Summary of the two datasets
3 Data

As mentioned in Section 1, we use two datasets to validate GLDA. Dataset 1 contains a total of 5,076 time-stamped Ecological Momentary Assessment (EMA) samples of ten mental health items from a group of 45 participants (69% female; mean age 38.5 years, s.d. 13.8 years). The participants were asked to self-report four times a day for at least 30 days consecutively their momentary experience of (1) rumination, (2) worry, (3) fear, (4) anger, (5) irritability, (6) anhedonia, (7) hopelessness, (8) depressed mood, (9) avoidance of action, and (10) avoidance of people on a scale of 0-100. In addition, participants' trait depression, anxiety, and stress were diagnosed pre-study using the following clinical inventories: for depression, the Hamilton Rating Scale for Depression (HRSD, 17 questions, 0-52 points) \( \text{[9]} \); for anxiety, the Hamilton Anxiety Rating Scale (HAMA, 14 questions, 0-56 points) \( \text{[8]} \), and; for stress, the stress subscale from the Depression Anxiety and Stress Scale (DASS-S, 14 questions each rated on 0, 1, 2, and 3 with a total score of 42 for each outcome) \( \text{[10]} \). The DASS inventory also contains a depression (DASS-D) and an anxiety (DASS-A) subscale; however, these were eschewed in favor of the HRSD and HAMA scores as the official ground truth for depression and anxiety.

Dataset 2 contains 25,394 physiological samples of 5 physiological measures from 166 participants (61% female; mean age 39 years, s.d. 13.8 years), a superset of the group of 45 in Dataset 1. The same depression, anxiety, and stress diagnosis scores from HRDS, HAMA, and DASS-S are available for each of the 166 participants as well. Continuous electrocardiogram (ECG) and impedance cardiography data were collected at 500 Hz using the MindWare Mobile Hardware during a clinical interview of each participant. From these data we extracted five physiological variables for use in this study: (1) heart rate (hr), (2) respiratory sinus arrhythmia (rsa), (3) pre-ejection period (pep), (4) respiration rate (resp_rate), and (5) respiration amplitude (resp_amp). Each of these five variables forms a time series.

The details described above are summarized in Table 2. The commonality between, as well as the reasons we choose these two datasets, are three fold. First, they fit the type of data in question in this article: repeated, multivariate samples from a cohort of distinct subjects. Second, inferring discrete states from these “atomic” variables is a meaningful practice for automated diagnostic purposes \( \text{[4]} \). Last but not least, ground truth about each participant is available, in both cases trait depression, anxiety, and stress, allowing the validation of the model-learned mixtures of discrete states against the participants’ actual clinical outcomes.

4 Experiments

We used STAN, a programming language for probabilistic modeling and inference to specify our GLDA models. We found one existing implementation of GLDA (in Python) for topic modeling \( \text{https://pypi.org/project/gaussianlda/} \), taking
in words as input and involving the conversion of words into real numbers using a word-embedding application. We created our own implementation to directly process continuous data. The STAN code is provided in the Supplementary Material. The number of $K$, the number of discrete states, or clusters, is the only parameter that needs to be set by the researcher. Hyperparameters were initiated by the STAN program automatically to have neutral or unit values. The rest of the input parameters are directly read from the data (see Table 2), which include: (1) $V$, the number of variables, or dimension, in the multivariate samples; (2) $M$, the number of distinct participants in the cohort, and; (3) $N$, the total number of observations.

The next step is fitting the model to data. We standardized each variable within each participant so that they follow a standard normal distribution (Z-score; 0 mean and 1 standard deviation). This operation preserves the interrelations between variables but ensures that the discrete states discovered by our model reflect momentary deviations from intra-individual mean values.

For a simple, demonstrative clustering model, we set $K = 3$, and ran the STAN program to find the participant-specific mixture weights $\theta_m$ (each $\theta_m$ is a $K$-length weight vector that sums up to 1), as well as the mean $\mu_k$ ($V$-length vector) and covariance matrix $\sigma_k$ ($V \times V$ matrix) of each cluster $k$. The mean and covariance effectively defined each cluster as a distribution over the multiple variables. We configured the STAN program to run a 1000-iteration Markov Chain Monte Carlo chain including an initial 200 warm-up iterations and calculated the mean values of the output parameters from all the iterations following the warm-up. Based on our observation, it safely takes <100 iterations to warm up so convergence has long been reached when the sampling algorithm stops running and results are returned.

We repeated the process described above for both datasets. We then explored the correlation between the participants’ clinically diagnosed depression, anxiety, and stress levels with the mixture weights inferred by GLDA on the EMA data (Dataset 1) and physiological data (Dataset 2). We built linear regression models of HRSD depression score, HAMA anxiety score, DASS stress score, using the GLDA-learned participant-specific mixture weights. In doing so we aimed to validate our GLDA-discovered discrete states because intuitively they should be significantly correlated with the results from the established inventories. Returning significant coefficients would support the validity of our inferred clusters and their weights. A lack of correlation between machine learned discrete states from EMAs and the one-time assessment from inventories would indicate that the learned states are not capturing significant variance in the participants’ mental health status. A strong correlation, on the other hand, would validate the discovered classes and their individual-specific weights.

Finally, to determine whether and to what extent the GLDA-discovered classes may be more meaningful than regular GMM, we built a 3-class GMM on all observations in each dataset, producing one set of mixture weights over component distributions. We computed the class membership of each observation and then the class proportions for each participant as the frequency of each class in each participant divided by the number of observations from the participant.
Linear regression models of HRSD, HAMA, and DASS-S scores were then built using the individual-specific class proportions as predictors. We compared the results of these GMM-based linear models with those of the GLDA-based ones to identify which method may produce more meaningful or predictive classes and weights.

5 Results

5.1 Dataset 1

Figure 1 shows the GLDA- and GMM-inferred clusters’ means over the ten, momentarily evaluated mood items. The three classes learned by both methods correspond to an unwell state (solid line; high Z-scores across all variables), a well state (dotted line), and a middle state (dashed line). The middle state from both GLDA and GMM stays on the negative side over all ten variables, indicating a mildly positive experience. No clusters appear to contain low and high values in different variables simultaneously (i.e., lines not crossing), likely due to the negative valence of all the EMA items evaluated. Two differences
Table 3: Results of individual-level linear regression models of clinically evaluated depression, anxiety, and stress, using mixture weights of the three classes learned by GLDA and the realized class proportions learned by GMM. Emojis 😅, 😌, and 😊 represent the unwell, middle, and well states visualized in Figure 1 respectively. Significant pairs of class weight/proportion and mental illness score are shown in bold and marked according to its significance level (**: 0.001 ≤ p < 0.01; *: 0.01 ≤ p < 0.05; : 0.05 ≤ p < 0.1)

between the cluster means are visible. First, the unwell state by GMM is more severe than its GLDA counterpart by a significant margin. Second, in the GMM result, outside of the separation observed in avoidance of action and avoidance of people, the mean values of the middle state nearly overlap with those of the well state; whereas the middle state defined by GLDA exhibits better separation from both the well and the unwell states. Overall, the cluster means from the two methods are more similar than they are different.

In Table 3 we list the p-values and adjusted R squared values of the linear regression models predicting HRSD, HAMA, and DASS-S scores with individualized mixture weights by GLDA and realized class proportions by GMM from the three-class clustering analysis described in Section 4. Each set of p-values and adjusted R squared values come from a univariate linear regression between a pair of clustering weight/proportion and a ground truth variable. Evidently, the only significant pairs (at 0.05 level of confidence) are between the weight of the GLDA-learned unwell state ( zakładka 😅) and each of the clinical scores. The strongest correlation is between the GLDA-learned unwell state weight and DASS-S stress score: the former accounted for 15.1% of the variance in the latter. The GMM-learned unwell state proportion is not significantly correlated with clinical scores, neither are the weights/proportions of the middle and the well states ( zakładka 😌 and 😊) from either methods. This result supports the hypothesis that the class components and the weights inferred by GLDA are more meaningful, i.e., better predictive of clinically validated individual outcomes, than those obtained by GMM. This result also conforms to the intuition that the unwell state should be correlated with clinical ground truth of mental illness due to the
Figure 2: Mean values of the three clusters (solid, dashed, and dotted lines) over the five physiological sensor measures in Dataset 2, learned by GLDA and GMM. The five measures, from left to right on the X-axis, are heart rate (hr), respiratory sinus arrhythmia (rsa), pre-ejection period (pep), respiration rate (resp_rate), and respiration amplitude (resp_amp).

shared underpinning of negative affect.

5.2 Dataset 2

Figure 2 shows the GLDA- and GMM-inferred clusters’ means over the five physiological measures from MindWare Mobile Hardware. The three classes discovered by both methods again have similar shapes but also exhibit key differences. There is a conspicuous high heart rate (hr) and high respiratory amplitude (resp_amp) class from both methods (dotted line); but GLDA associates high respiratory sinus arrhythmia (rsa) with the state whereas GMM does not. The other two, relatively lower arousal states (solid and dashed lines) from GLDA and GMM differ in the combination of high and low values in all variables except pre-ejection period (pep), whose Z-score is right at zero for both methods.

In Table 4, we list the p-values and adjusted R squared values of linear regression models predicting HRSD, HAMA, and DASS-S scores with the GLDA-learned mixture weight and the GMM-learned realized proportion of the high
Table 4: Results of individual-level linear regression models of clinically evaluated depression, anxiety, and stress, using the GLDA-learned mixture weight and the GMM-learned realized proportion of the high heart rate (hr), high respiration amplitude (resp_amp) state, visualized by dotted line in Figure 2.

|                | HRSD (Depression) | HAMA (Anxiety) | DASS-S (Stress) |
|----------------|-------------------|----------------|-----------------|
| GLDA high-hr   | 0.012* 3.4%       | 0.047* 1.9%    | **0.001** 5.8%  |
| GMM high-hr    | 0.107 1.0%        | 0.260 0.2%     | 0.015* 2.9%     |

heart rate (hr), high respiration amplitude (resp_amp) state (dotted line in Figure 2). The weights/proportions associated with the other two classes (solid and dashed lines in Figure 2) did not return any significant result with clinical mental health outcomes, so we do not include them in Table 4. Once again, the GLDA-learned mixture weight achieved significant correlation with all three clinical scores, with DASS-S stress being the strongest. Proportion of the GMM-inferred high-hr class was also significantly correlated with DASS-S, but not with depression and anxiety. Between the two correlations with DASS-S, GLDA provided better statistical prediction than GMM (GLDA $p = 0.001$ vs. GMM $p = 0.015$; the second to last column of Table 4). This result once again supports the hypothesis that the GLDA-inferred discrete states and their mixture weights are more meaningful compared to the realized class proportions learned by GMM. This result also suggests that high heart rate and high respiratory amplitude are two physiological measures that are robustly predictive of depression, anxiety, and especially stress.

6 Discussion

In this section we reflect on two implications of the GLDA model, each lending itself to follow-up research questions and future work. The first reflection is on the conditions under which GLDA may be a superior model to GMM (as demonstrated in this article) and the conditions under which it may not be the case. What allowed GLDA to discover discrete states from momentary psychophysiological measures whose mixture weights are more significantly correlated with clinical mental health outcomes in this paper? Our hypothesis is that it is because of the inherent psychophysiological differences between the participants in the two datasets we used. It is indeed likely that negative affect (Dataset 1) and cardiovascular health (Dataset 2) are experienced differently by different individuals, allowing the individually separate mixture weight sampling mechanism of GLDA (see Table 1) to better reflect this reality. Contrarily, if there are no purportedly intrinsic between-participant differences, then GLDA’s additional assumption would be unnecessary and it would ostensibly achieve results that are no better or worse than GMM. Given the higher model complexity
in GLDA, in that case using GMM may be the more desirable (economical) choice. Future work is needed to validate GLDA on repeated multivariate samples that measure largely homogeneous processes between participants in a cohort and compare the result with GMM. Such work will further pinpoint the conditions under which GLDA is a more advisable approach compared to regular finite mixture models.

The second reflection is about using the GLDA model to infer not only individual level characteristics (such as mixture weights focused on in this article) but also moment-to-moment transition between discrete states within each participant’s observations. In addition to calculating the individual-level mixture weights over clusters, we can also compute the probability that each observation from each participant belongs to each of the $K$ clusters using participant-specific mixture weights $\theta_m$, means $\mu_k$, and covariance matrices $\Sigma_k$.

The formula for this probability is:

$$P(z_{n,k} = 1| x_n) = \frac{\theta_{m,k} N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \theta_{m,j} N(x_n|\mu_j, \Sigma_j)},$$

where $z_{n,k}$ is a binary variable that takes value 1 if the $n$-th observation $x_n$ (which belongs to participant $m$) belongs to cluster $k$ and 0 otherwise. $N(x_n|\mu_k, \Sigma_k)$ represents the density value of the multivariate Gaussian distribution $N(\mu_k, \Sigma_k)$ at $x_n$. The output from this step $P(z_{n,k} = 1| x_n)$ where $n = 1, 2, ..., N$, provides observation-level, or momentary insight. Intuitively, cluster $\text{argmax}_k P(z_{n,k} = 1| x_n)$ is assigned the cluster to which observation $x_n$ belongs. The class membership of each observation, which is timestamped, reveals how each participant transitions between different discrete states as time progresses. Figure 3 shows such a time series for an example participant in Dataset 1. One can see that most of the moments that fall in the unwell state gather around an earlier part (2015-10-08 to 2015-10-16) of the study period. The further research question we may ask here is whether those momentary unwell states indicate any temporary behavioral change or health variation?
Ground truth for such momentary fluctuation is not available in the datasets used in this article but we defer to future work to validate the utility of these momentary states for predictive, real-time monitoring. Increasingly available, dense, and multi-modal passive sensing data from mobile, wearable, and environmental sensing devices \cite{11} should provide plenty of evidence for in-the-moment objective ground truth of an individual’s daily behavior and health.

7 Conclusion

In this article we proposed and validated an unsupervised probabilistic model, Gaussian Latent Dirichlet Allocation (GLDA), for the problem of discrete state discovery from repeated, multivariate psychophysiological samples collected from multiple, inherently distinct, individuals. Psychology and medical research heavily involves measuring potentially related but individually inconclusive variables from cohorts of participants to derive diagnosis, necessitating clustering analysis. Traditional probabilistic clustering models such as Gaussian Mixture Model (GMM) assume a global mixture of component distributions, which may not be realistic for observations from different patients. The GLDA model borrows the individual-specific mixture structure from a popular topic model Latent Dirichlet Allocation (LDA) in Natural Language Processing and merges it with the Gaussian component distributions of GMM to suit continuous type data.

We implemented GLDA using STAN (a probabilistic modeling language) and applied it on two datasets, one containing Ecological Momentary Assessments (EMA) and the other heart measures from electrocardiogram and impedance cardiograph. Both datasets contained clinically diagnosed depression (Hamilton Rating Scale for Depression), anxiety (Hamilton Anxiety Rating Scale), and stress scores (Depression Anxiety and Stress Scale–Stress) for each participant studied. We found that, in both datasets, the GLDA-learned class weights achieved significantly higher correlations with the clinical mental health outcomes than the class proportions discovered by GMM (Sections 5.1 and 5.2). Our findings point to an advantage of GLDA over conventional finite mixture models for human state discovery from repeated multivariate data in psychology and medical research, likely due to better characterization of potential underlying between-participant differences.

Future work is required to validate the utility of this model on a broader range of applications. The purpose of this planned work is twofold, as discussed in Section 6 to elucidate the conditions under which GLDA is and is not an ideal choice for human state discovery from repeated multivariate cohort samples, and to extend the analyses presented in this article from individual-level to momentary, observation-level for real-time health and behavior informatics. Moreover, since the main purpose of this study is to validate GLDA against regular GMM, we chose a singular number of clusters $K = 3$ to be simplistic yet realistic; future work can be dedicated to evaluating the sensitivity of model fit and predictive power to different choices of number of clusters.
Funding

This work was supported by Whole Communities—Whole Health, a research grand challenge at the University of Texas at Austin.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[3] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, 2015.

[4] Aaron J Fisher and Hannah G Bosley. Identifying the presence and timing of discrete mood states prior to therapy. *Behaviour research and therapy*, 128:103596, 2020.

[5] Aaron J Fisher, Esther Howe, and Zoe Y Zong. Unsupervised clustering of autonomic temporal networks in clinically distressed and psychologically healthy individuals. *Behaviour Research and Therapy*, 154:104105, 2022.

[6] Aaron J Fisher, John D Medaglia, and Bertus F Jeronimus. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27):E6106–E6115, 2018.

[7] Dilan Görür and Carl Edward Rasmussen. Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010.

[8] MAX Hamilton. The assessment of anxiety states by rating. *British journal of medical psychology*, 1959.

[9] Max Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960.

[10] Peter F Lovibond and Sydney H Lovibond. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour research and therapy*, 33(3):335–343, 1995.
[11] Congyu Wu, Hagen Fritz, Sepehr Bastami, Juan P Maestre, Edison Thomaz, Christine Julien, Darla M Castelli, Kaya de Barbaro, Sarah Kate Bearman, Gabriella M Harari, et al. Multi-modal data collection for measuring health, behavior, and living environment of large-scale participant cohorts. *GigaScience*, 10(6):giab044, 2021.