1 Background

Mobile health, also known as mHealth, becomes more and more important in psychiatry. Smartphones now offer the technology to gather rich behavioral data on patients and deliver information about how symptoms vary in time, space, and across social dimensions (Torous et al. [2015]). Among mental health patients, there is a great interest in monitoring symptoms with mobile apps (Vangeepuram et al. [2018]). The digital techonologites are used to monitor patients behavior over time through collection of the long-term objective data continuosly and automaticly. These data vary from activity data such as GPS, distance travaeled to social communication data such as text messages (eg, the number of outgoing SMS) and phone calls (eg, the number of incoming and outgoing calls, duration outgoing calls). Indeed, using both self-reported information collected by the self-rating questionnaires and the smartphone sensor data offers a unique opportunity to describe the person in terms of his or her lifestyle and behavior at the physical, cognitive, and environmental level (Cornet and Holden [2018], Torous et al. [2018]).

Researchers hope to gain insight in the user's clinical outcome through analysis of the variety of personal data. More studies are needed to establish the relationship between mobile measurements and clinical symptoms. This mobile phone-based approach may provide an estimation of physiological and mental state. In addition, it may be valueable to predict the changes in clinical states and investigate causal inferences about state changes in patients (eg, those with affective disorders) Dogan et al. [2017]

Multiple statistical methods were developed for the analysis of time series data of mental health. Time series methods like autoregressive model(AR) are commonly used to analyze mood scores or depression rating scale (Holmes et al. [2016]). Moore et al. [2012]used exponential smoothing and Gaussian process regression to forecast depression in bipolar disorder. Nevertheless, when meeting with the unbounded count data and categorical outcomes, the regular time series model that are established on continuous outcomes with continuous distribution are not suitable (Zucchini et al. [2017]). Multiple machine learning and VARX algorithms are also used on psychological multivariate time series data (Tuarob et al. [2017], Reece et al. [2017]). Poisson regression model is also a useful approach. Kolliakou et al. [2020] use this approach to quantify the impact of daily levels of relevant Twitter posts to the mental health crisis episodes.

Also, for the mobile-communication count data, instead of being interested in the average count of mental patients' calls and texts for instance, we hope to find some meausure which allows us some access to the underlying variable we truly care about: the clinical state or the outcome of these patients. Accordingly, a state-space model like the Hidden Markov Model, which uses observable data to estimate the status of a latent, or hidden, variable over time, may provide useful insights (Reece et al. [2017]). A Hidden Markov Model can be used to detect changes between different clinical states of patients over time. Hidden Markov model can also accommodate both over dispersion and serial dependency by allowing the probability distribution of each observation to depend on the unobserved state of a Markov chain(Zucchini et al. [2017]).

In past few years, although HMM have been widely used in economics, bioinformatics, and engineering (e.g. MacDonald and Zucchini [1997], Durbin et al. [1998],Rabiner [1989]), it has not been well-integrated within psychological research, especially for mobile-communication

health data research. Alafchi et al. [2018] proposed an approach to forecast new cases of bipolar disorder sing poisson hidden markov model based on new BD cases.

Moreover, there are some challenges for the mobile-communication data.

1. Missing data are common in time series. It is expected that some missing values in self-reported data and the recording of clinical outcome might have impacts on the modeling. Imputation can be created for individuals having missing values on the observed variables and outcomes.

2. The time series data of mental patients might be non-stationary. NS-HMM(Non-stationary Hidden Markov Model) was introduced to capture state duration behaviour by defining a set of dynamic transition probability parameters(Mor et al. [2020]).

3. There are excess zeros from count data(phone calls or texts) which are hard to be exlained by simple models. The multivariate zero-inflated model may be useful to deal with this problem.

A Mixed Effects Hidden Markov Model (Altman [2007], Maruotti and Rocci [2012])can capture the information of both mobile-comminication count data and the clinical outcome of mental patients. In order to deal with the above issues, we need to add more structures of the models.

2 Objective

The primary goal of this project is to build a hidden markov models for zero-inflated count data and self-reported categorical data to predict the clinical state of next time point through estimation of the transition probability of clinical states.

Next, we need to use simulations to evaluate the prediction performance of different hidden markov models, such as basic HMM, HMM dealing with missing data, HMM for non-stationary time series and other time series model such as GLM, ARMA.

Furthermore, the final goal is to obtain the causal effect between mobile communication time series data and the clincial outcomes.

3 Method

3.1 The basic Multivariate Discrete Hidden Markov Model

This method is motivated by Zhang et al. [2010] and Ip et al. [2013]. Let y_{ijj} denote the discrete response of subject i at occasion t on outcome $j, i = 1, \ldots, N; j = 1 \ldots, J; t = 1, \ldots, T$. We assume that each outcome indicates the level of response $k, k = 1, \ldots K_j$, from an individual as measured by an "item," such as a survey question. Specifically, $\mathbf{y}_{it} = (y_{it1}, \ldots, y_{itJ})'$ denotes the response vector of subject i at occasion t, and $\mathbf{y}_i = (\mathbf{y}'_{i1}, \ldots, \mathbf{y}'_{iT})'$ represents the complete response pattern of subject i. Additionally, $z_{it} = 1, \ldots, S$ denotes the categorical (unobserved) latent state of subject i at occasion t. Thus, $\mathbf{z}_i = (z_{i1}, \ldots, z_iT)'$ is the state sequence of subject i through the latent space over time. A first-order Markov chain for

the latent variable assumes that the latent state at occasion $t + 1$ depends only upon the latent state at occasion t; that is, for any $i, p(z_{it} \mid z_{i1}, \ldots, z_{iT}) = p(z_{it} \mid z_{it-1})$. Accordingly, the basic hidden Markov model contains the following three sets of parameters:

1. $\boldsymbol{\tau} = (\tau_{rs})$, the $S \times S$ matrix of time-homogeneous transition probabilities between latent states, where for any $i, \tau_{rs} = p(z_{it} = s \mid z_{it-1} = r), t = 2, \ldots, T$

2. $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1S})'$, the $S \times 1$ vector of marginal probabilities of the states at occasion 1 $\alpha_{1s} = p(z_{i1} = s)$. The vector $\boldsymbol{\alpha}_t = (\alpha_{t1}, \ldots, \alpha_{tS})', t = 2, \ldots, T$, can be recursively derived by the formula $\boldsymbol{\alpha}'_t = \boldsymbol{\alpha}'_{t-1}\boldsymbol{\tau}$

3. $\boldsymbol{\pi} = (\pi_{jsk})$, the $J \times S \times K_j$ array of state-conditional response probabilities, where k is the index of outcome category, $k = 1, \ldots, K_j$, and for any $i, t, \pi_{jsk} = \Pr(y_{itj} = k \mid z_{it} = s)$, where $\sum_{k=1}^{K_j} \pi_{jsk} = 1$

Assuming conditional independence between responses given the latent state, the marginal probability of a response pattern $\mathbf{y}_i$ is given by

$$p(\mathbf{y}_i) = \sum_{\mathcal{Z}} p(\mathbf{z}_i) \prod_{t=1}^{T} p(\mathbf{y}_{it} \mid z_{it})$$

with the summation over the entire latent state space $\mathcal{Z}$ of size S^T, and

$$p(\mathbf{z}_i) = \alpha_{1Z_{i1}} \prod_{t=2}^{T} \tau_{z_{it-1}z_{it}}$$

Furthermore, we write out the likelihood equation using the local independence assumption given a latent state:

$$p(\mathbf{y}_{it} \mid z_{it} = s) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \pi_{jsk}^{\delta(y_{itj},k)}$$

where $\delta(y_{itj}, k) = 1$ if $y_{itj} = k$, and 0 otherwise.

3.2 Mixed effects Hidden Markov model

Altman [2007] proposed a generalized mixed effects linear model for each individual transition within a hidden Markov model (HMM):

$$\text{logit} \quad p(z_{it} = s \mid z_{it-1} = r) = x'_i \gamma_{rs} + w'_i \theta_i$$

where z_{it} denotes the latent state for individual i at time t, s is a vector of regression coefficients for covariate x_i for transitioning from state r at time $t - 1$ (state of origination) to state s at time t (state of destination), and w_i and θ are the random-effects design vector and random-effects vector, respectively.

3.3 Model selection

The appropriate number of hidden states was determined by selecting the model according to the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

References

B. Alafchi, S. Yazdi-Ravandi, R. Najafi-Vosough, A. Ghaleiha, and M. Sadeghifar. Forecasting new cases of bipolar disorder using poisson hidden markov model. *International Clinical Neuroscience Journal*, 5(1):7, 2018.

R. M. Altman. Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477): 201–210, 2007.

V. P. Cornet and R. J. Holden. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics*, 77:120–132, 2018.

E. Dogan, C. Sander, X. Wagner, U. Hegerl, and E. Kohls. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? systematic review. *Journal of medical Internet research*, 19(7):e262, 2017.

R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

E. A. Holmes, M. Bonsall, S. Hales, H. Mitchell, F. Renner, S. Blackwell, P. Watson, G. Goodwin, and M. Di Simplicio. Applications of time-series analysis to mood fluctuations in bipolar disorder to promote treatment innovation: a case series. *Translational psychiatry*, 6(1):e720–e720, 2016.

E. Ip, Q. Zhang, J. Rejeski, T. Harris, and S. Kritchevsky. Partially ordered mixed hidden markov model for the disablement process of older adults. *Journal of the American Statistical Association*, 108(502):370–384, 2013.

A. Kolliakou, I. Bakolis, D. Chandran, L. Derczynski, N. Werbeloff, D. P. Osborn, K. Bontcheva, and R. Stewart. Mental health-related conversations on social media and crisis episodes: a time-series regression analysis. *Scientific reports*, 10(1):1–7, 2020.

I. L. MacDonald and W. Zucchini. *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press, 1997.

A. Maruotti and R. Rocci. A mixed non-homogeneous hidden markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine*, 31(9):871–886, 2012.

P. J. Moore, M. A. Little, P. E. McSharry, J. R. Geddes, and G. M. Goodwin. Forecasting depression in bipolar disorder. *IEEE Transactions on Biomedical Engineering*, 59(10): 2801–2807, 2012.

B. Mor, S. Garhwal, and A. Kumar. A systematic review of hidden markov models and their applications. *ARCHIVES OF COMPUTATIONAL METHODS IN ENGINEERING*, 2020.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7 (1):1–11, 2017.

J. Torous, P. Staples, and J.-P. Onnela. Realizing the potential of mobile mental health: new methods for new data in psychiatry. *Current psychiatry reports*, 17(8):61, 2015.

J. Torous, P. Staples, I. Barnett, L. R. Sandoval, M. Keshavan, and J.-P. Onnela. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ digital medicine*, 1(1):1–9, 2018.

S. Tuarob, C. S. Tucker, S. Kumara, C. L. Giles, A. L. Pincus, D. E. Conroy, and N. Ram. How are you feeling?: A personalized methodology for predicting mental states from temporally observable physical and behavioral information. *Journal of biomedical informatics*, 68:1–19, 2017.

N. Vangeepuram, V. Mayer, K. Fei, E. Hanlen-Rosado, C. Andrade, S. Wright, and C. Horowitz. Smartphone ownership and perspectives on health apps among a vulnerable population in east harlem, new york. *Mhealth*, 4, 2018.

Q. Zhang, A. Snow Jones, F. Rijmen, and E. H. Ip. Multivariate discrete hidden markov models for domain-based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics*, 19(3):746–765, 2010.

W. Zucchini, I. L. MacDonald, and R. Langrock. *Hidden Markov models for time series: an introduction using R*. CRC press, 2017.