**pre-processing**

Prior to modeling, we have pre-processed the time-dependent observed variables (for 121 persons, we have 17 independent observed variables, 1 discrete observed outcome variable. 117 persons were used for our analysis.) so that it won't run into errors due to missing values. After this, we have run a Confirmatory Factor Analysis (CFA) in order to extract 7 latent factors (which will be denoted as $X_{it}$) out of 17 observed variables.

**GMM (to estimate latent regime indicator $S_{it}$)**

Then, we cluster the persons based on the intra-individual variables using Gausian Mixture Modeling (GMM). GMM is a soft clustering method in which we express the given dataset as a combination of $K$ probability distributions where $K$ denotes the pre-specified number of clusters. The resulting model is described by the ratios of each cluster $\pi_k$, each cluster mean $\mu_k$ and covariance matrix $\Sigma_k$ which each multivariate normal distribution is parameterized by. Given the information, we can obtain the probability of $i$-th person belonging to each of the $K$ clusters ($Pr[S_i = k|X_i]$). What we want for our analysis is to collect those $K$ probabilities. Note that knowing only $K-1$ of them suffices: in two clusters case ($K = 2$), the probability of belonging to the second cluster can be readily computed given $Pr[S_i = 1|X_i]$ as $(1 - Pr[S_i = 1|X_i])$. For each available time point, we apply the GMM separately ($K = 2$). Given $N$ persons with $T$ measurement occasions, this will give the $N \times T$ matrix $P[S_{it} = 1|X_{it}]$. We use this results as the building block to apply the Extended Kalman Filter (EKF).

**EKF (to estimate latent variable $\eta^s_{it|t}$)**

Taking advantage of the cluster memberships obtained by the GMM, we estimate the state-dependent intra-individual latent variables that are related to the corresponding intra-individual observed variable of interest. The EKF procedure consists of computing the following quantities:

$$\eta^s_{it|t-1} = \alpha_s + \beta_s \eta_{i,t-1|t-1} + \gamma_s X_{it} \tag{1}$$

$$P^s_{it|t-1} = \beta_s^2 P^s_{i,t-1|t-1} \tag{2}$$

$$v_{it} = y_{1it} - \sum_{s\in\{1,2\}} \text{expit}(d_s + \Lambda_s \eta^s_{it|t-1} + A_s X_{it}) \cdot Pr[S_{it} = s|X_{it}] \tag{3}$$

$$F_{it} = \sum_{s\in\{1,2\}} \{\Lambda_s^2 P^s_{it|t-1} + R_s\} \cdot Pr[S_{it} = s|X_{it}] \tag{4}$$

$$\eta^s_{it|t} = \eta^s_{it|t-1} + K^s_{it} v_{it} \tag{5}$$

$$P^s_{it|t} = P^s_{it|t-1} - K^s_{it} \Lambda_s P^s_{it|t-1} \tag{6}$$

where $K^s_{it} = P^s_{it|t-1}\Lambda_s F^{-1}_{it}$ is called the Kalman gain function and *expit* represents the logistic function. The EKF summarized in Equation (1-6) works recursively from time 1 to $T$ and $i = 1, \ldots, n$ until $\eta^s_{it|t}$ and $P^s_{it|t}$ have been computed for all time points and people. The latent variable score at $t = 0$ is assumed to be distributed as $\eta^s_0 \sim MVN(\eta^s_{0|0}, P^s_{0|0})$.

**Parameter estimation**

**Results**

Comparing the time-dependent discrete observed outcome variable $y_{1it}$ and the estimated discrete latent variable $Pr[S_{it}|X_{it}]$, we see that the GMM managed to capture some cluster membership. Out of 117 people which we obtained after pre-processing, 43 people actually dropped out. Our $S_{it}$ shows a large False Positive rate i.e., incorrectly predicting (intention to) drop out when the person actually did not drop out. Nevertheless, the $S_{it}$ shows a large True Negative rate i.e., correctly predicting (intention to) not drop out when the person actually did not drop out in the end.

**To do**

The model has not yet encoded parameter optimization with respect to the Extended Kalman Filter. For this, we would need a way to explore a set of parameter values for $(\alpha_s, \beta_s, \gamma_s, d_s, \Lambda_s, A_s)$. In our case, however, this amounts to optimizing the 36 dimensional parameter space. Therefore, we would need a "smart" algorithm that overcomes the curse of dimensionality while moving around the parameter space. In the frequentist approach, we need to be able to compute the gradient in the parameter space.

Also, the model has the time-dependent discrete latent classes (regimes) which are temporally independently estimated based on the GMM. Although the estimated $S_{it}$ indicated a good sign of fit e.g., a large True Negative rate, it would be better to add some temporal dependency such as the Markov model. Such an extension may be easily done by taking a weighted average of the clustering results based on $X_{it}$ i.e., $E[S_{it}|X_{it}]$ and the previous one $E[S_{i.t-1}|X_{i,t-1}]$.

Further, the time-dependent factor $X_{it}$ was estimated using the CFA model. Although it is a confirmatory model, we should carefully look at the goodness of fit to verify the fitted model.

Finally, what I worked so far had only considered the intra-individual variables, so it has to be extended to accommodate the inter-individual variables. However, I think this should be easy.