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

First, 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 π_k, each cluster mean μ_k and covariance matrix Σ_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:

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

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

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

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

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

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

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 $\eta^s_0 \sim MVN(\eta^s_{0|0}, P^s_{0|0})$.