Functional clustering methods for extracting features from EHR biomarker history data

Jason Roy

Rutgers School of Public Health
E-mail: jason.roy@rutgers.edu

Abstract: In many modern applications, there is interest in predicting subject-specific functions of a variable over time. For example, we might want to know patient specific trends in a biomarker over time. Modeling is needed If there is measurement error in the variable, or if gaps between data collection times is too wide. We propose a novel semiparametric model for the joint distribution of a continuous longitudinal outcome and the baseline covariates using an enriched Dirichlet process (EDP) prior. This joint model decomposes into subject-specific linear mixed models for the outcome given the covariates and simple marginals for the covariates. The nonparametric EDP prior is placed on the regression and spline coefficients, the error variance, and the parameters governing the predictor space. We predict the outcome at unobserved time points for subjects with data at other time points as well as for completely new subjects with covariates only. We find improved prediction over mixed models with Dirichlet process (DP) priors when there are a large number of covariates. Our method is demonstrated with electronic health records consisting of initiators of second generation antipsychotic medications, which are known to increase the risk of diabetes. We use our model to predict laboratory values indicative of diabetes for each individual and assess incidence of suspected diabetes from the predicted dataset. Our model also serves as a functional clustering algorithm in which subjects are clustered into groups with similar longitudinal trajectories of the outcome over time.