**Title: Multi-view Bayesian matrix factorization for mining large-scale heterogeneous electronic health records**

Electronic health records (EHR) are rich heterogenous collection of patient health information. The broad adoption of EHR systems has provided clinicians and researchers unprecedented opportunities for conducting health informatics research, which promises to provide an unbiased way to characterize patents' disease risks, thereby making actionable clinical recommendations for subsequent follow-ups of precision medicine. However, there are several challenges in modeling EHR data, including noisy irregular text in clinical notes, arbitrary bias in the billing codes, not missing at random (NMAR) lab tests, and heterogeneous data types (e.g., clinical notes, billing codes, lab tests, medications). To address the above challenges, we developed a Bayesian integrative generative model in the ravine of collaborative filtering and latent topic models. Specifically, we propose a multi-view probabilistic matrix factorization framework. In a nutshell, the proposed method factorizes multiple high-dimensional clinical-feature matrices into lower rank (basis) matrices and a common (loading) matrix that spans patients' dimension, which we interpret as the probabilistic disease mixture memberships for each patient. To learn the model parameters, we describe an efficient variational inference algorithm and its online stochastic counterpart.

We demonstrate our method's general utilities using real-world EHR data from MIMIC-III database. By 5-fold cross-validation and prospective imputation, we observe superior imputation accuracy using multiple EHR data categories (except for the target EHR data category) compared to models using individual data categories, suggesting the benefits of borrowing information across data types in otherwise extremely sparse EHR matrices. Qualitative assessment shows that heterogeneous clinical features that tend to co-occur under the same latent topics exhibit meaningful semantics of known diseases under similar epidemiology along with relevant medications and treatment procedures. We then leverage the lower dimensional patient mixture projections to predict prospective mortality of patients in critical conditions using their early admission records 1-6 months in advance. The proposed approach gives state-of-the-art performance compared to existing methods and reveals several distinct and meaningful disease topics related to the prognostic outcomes.