Name:	Rebecca Ferrell and Tyler McCormick* (presenter)
Affiliation:	University of Washington
Email:	rferrell@uw.edu, tylermc@uw.edu
Presentation type (select one):	Poster

Bayesian Feature Selection for Interpretable Prediction in Observational Medical Data

Rebecca L. Ferrell[1], Tyler H. McCormick, PhD[2]
[1]Department of Statistics, University of Washington, Seattle, WA; [2]Departments of Statistics and Sociology, University of Washington, Seattle, WA

Abstract

Observational medical databases are typically not immediately "regression ready". Analysts must make various choices in order to derive a design matrix of covariates from the source data. We describe these feature creation decisions in a longitudinal context in which we have granular medical events by date with subject identifiers. In these observational health databases, feature creation often involves coding for characteristics present during a designated baseline period through discretization of the temporal element of the records, e.g. coarsening patient timelines over a specified period into a feature to capture prior disease history. Our proposed approach to create meaningful covariates in these applications is to develop a set of generic concepts and compile a set of potential definitions for each, which may vary in hierarchical or temporal resolution. These correlated "competing definitions" for the same concept are treated as mutually exclusive to retain interpretability in regression models using structured prior distributions and stochastic search variable selection (SSVS). We develop this framework, present simulation results demonstrating its model selection recovery performance, and illustrate its use in selecting parsimonious interpretable logistic regression models for prediction.

Introduction

For fitting statistical models such as logistic regression, data from longitudinal observational medical databases requires further processing. Typically records must be flattened or collapsed within a subject in order to define a design matrix $\mathbf{X}$ containing covariates, a process which necessitates analyst decision-making in situations where information about appropriate time scales or code sets may not always exist. Further, it is often desired for the features to have simple definitions for model interpretation, e.g. for use in rapid risk assessment based on information that would be available to a clinician during the course of routine care. Our motivating example is in creating "lookback" covariates, that is, binary features indicating presence or absence of certain diagnoses or exposures using a particular coding resolution within a specific time window prior to an index date. The analyst may be uncertain as to the appropriate time scale to consider for each such covariate, as some diagnoses or exposures may only have predictive value when occurring shortly before the index date, while others may be informative over a longer time scale. One possible approach is to formulate a large set of these lookback features and then fit an L1-penalized regression model to select a sparse set of definitions. However, this does not guarantee an appealing solution in terms of model interpretation and multicollinearity: for example, if features defined as to "exposed to Drug X in previous 6 months" and "exposed to Drug X in previous 12 months" are both selected, it is difficult to interpret the effect of use of Drug X in previous 12 months conditional on holding exposure to Drug X in the previous 6 months constant as any subject taking Drug X in the more recent period must also have taken in the longer period. Thus, it is desirable to impose more structure on the model to only permit one of these closely related exposure definitions to be active.

Model

We propose a hierarchical model for multiphase inference[1] in which the analyst formulates groups of "competing variables". A toy example of a group of competing variables for a covariate corresponding to a diabetes diagnosis is shown in Figure 1. In this example, the analyst is uncertain which code set to use and how long of a lookback window prior to the study index date to check for the presence of these diagnosis codes. The analyst also allows the possibility that any of these definitions of diabetes may not be useful in predicting the outcome, and hence may all be excluded. This is encoded by placing a multinomial prior with *n=1* over nine indicator variables for inclusion of the listed possibilities and embedding this within a Bayesian stochastic search variable selection (SSVS, "spike-and-slab") logistic regression model[2,3,4]. We then obtain a posterior distribution over the competing variables and can identify which definitions, if any, have high posterior probability and merit inclusion in a single predictive model.

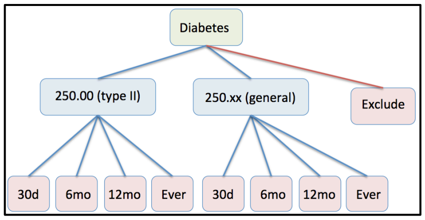

Figure 1. Example of competing definitions for a covariate indicating presence of diabetes at baseline.

Simulation study and model evaluation

We demonstrate the utility of this model in predicting a binary outcome in simulated observational medical data with 10,000 subjects and 100 disease/exposure definition groups, each group containing 5 competing definitions using different lookback windows. Simulation results are shown in Figure 2. Models selected using the competing variables framework are much sparser, do not include multiple definitions for the same feature, and have posterior mean coefficient estimates closer to true values than lasso or unregularized logistic regression.

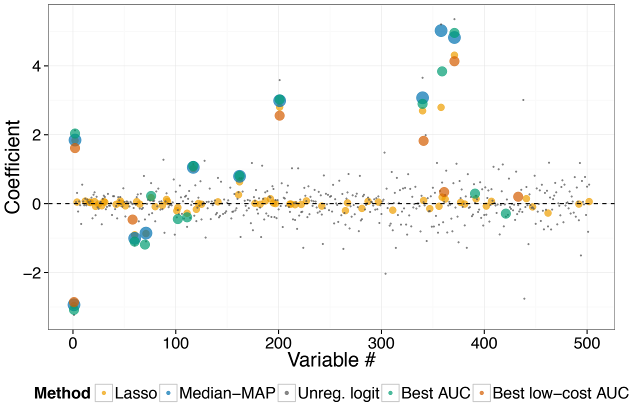

Figure 2. Coefficient estimates in simulation study comparing several models selected using the competing variables framework (Median-MAP, Best AUC, Best low-cost AUC) with a lasso model selected by cross-validation and an unregularized logistic regression model.

Conclusion

Our Bayesian hierarchical competing variable framework accommodates model uncertainty arising in the form of covariate definitions in a transparent and easily interpretable way.

References

1. Blocker AW, Meng XL. The potential and perils of preprocessing: building new foundations. Bernoulli 2013;19(4):1176-1211.
2. George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993;88(423):881-9.
3. Chipman, H. Bayesian variable selection with related predictors. Can J Stat. 1996;24(1):17-36.
4. Farcomeni A. Bayesian constrained variable selection. Stat Sinica. 2010;20:1043-62.