Supervised Bayesian Latent Class Models for High-Dimensional Data

S DeSantis[1], EA Houseman[2,3], B Coull[3], C Nutt[4], RA Betensky[3,4]

[1]Medical University of South Carolina, Charleston, SC, [2]University of Massachusetts, Lowell, [3]Harvard School of Public Health, Boston, MA, [4]Massachusetts General Hospital, Boston, MA

High grade gliomas, the most common primary brain tumors in adults, are diagnosed using histopathological techniques. However, these diagnostic categories are highly heterogeneous and do not always correlate well with survival. In an attempt to refine these diagnoses, several immunohistochemical measurements were made of YKL-40, a gene previously shown to be differentially expressed between diagnostic groups. We propose two penalized supervised latent class models for high-dimensional binary data, which are fit using Bayesian MCMC techniques. Penalization and model selection are incorporated in this setting by including prior distributions on the unknown parameters. In simulations, these new methods provide parameter estimates under conditions in which standard supervised latent class models break down, and outperform a two-stage approach to variable selection and model estimation in a variety of settings. Resulting latent classes correlate well with survival. We apply our new methodologies to the glioma study, for which identifiable 3-class parameter estimates cannot be obtained without penalization. With penalization, the resulting latent classes not only correlate well with clinical tumor grade but potentially offer additional information on survival prognosis that is not captured by clinical diagnosis alone. In addition, the inclusion of YKL-40 features increases the precision of survival estimates. Fitting models with and without YKL-40 further highlights patients who have GBM diagnosis but appear to have better prognosis than the average GBM patient.