Statistics and Its Interface Volume 8 (2015) 123–124

Editorial: Special Issue on Modern Bayesian Statistics (Part II)

Part II of the special issue on modern Bayesian statistics features 11 innovative and original research articles. These articles cover a wide spectrum of applications of Bayesian statistical methods arising from cancer studies, design of clinical trials, epidemiology, genomic studies, HIV studies, and survey research.

Estimating the size of the sub-populations whose behavior puts them at higher risk of contracting and transmitting HIV is important for assessing overall HIV prevalence and designing effective interventions. **Bao, Raftery, and Reddy** developed a Bayesian hierarchical model for estimating the sizes of local and national HIV key affected populations. Their proposed approach allows for the incorporation of multiple commonly used data sources, including mapping data, surveys, interventions, capture-recapture data, estimates or guesstimates from organizations, and expert opinion. They then applied the proposed approach to data used to estimate the number of males who injected drugs in Bangladesh. Complex diseases, such as cancer, result from genetic and environmental factors. Single-nucleotide polymorphisms (SNPs) are commonly used genetic markers, each explaining a small amount to the overall risk of disease. By combing pathway-based approaches with multiple SNP analyses of a specified region of interest and incorporating information on SNP-to-SNP associations via network priors that model the linkage disequilibrium between SNPs, **Stingo, Swartz, and Vannucci** developed a novel Bayesian modeling framework to identify molecular biomarkers for disease prediction. To carry out posterior inference, they developed a stochastic search method that identifies significant biomarkers for disease prediction. They then demonstrated the ability of the proposed methodology to detect relevant genes and associated SNPs in a lung cancer dataset.

Cross-validation is a widely-used method for estimating out-of-sample prediction error and comparison of statistical models. However, for multilevel data (as well as other dependent structures such as time series, spatial, and network data), there are several challenges in the use of cross-validation for estimating out-of-sample prediction error and model selection. In particular, in multilevel models, the observed loss function for data-level cross-validation can be so close to flat that the cross-validation estimates of prediction errors under candidate models can be outweighed by random fluctuations. Using a hierarchical model fit to large survey data with a battery of questions, **Wang and Gelman** demonstrated that even though cross-validation might give good estimates of pointwise out-of-sample prediction error, it is not always a sensitive instrument for model comparison. An innovative aspect of their analysis is that they evaluated separately on 71 different survey responses, taking each in turn as the outcome in a comparison of regression models, which allowed them to construct a relatively large corpus of data out of a single survey. The Approximate Bayesian Computation (ABC) methodology is a way to handle models for which the likelihood function may be intractable or even unavailable and/or too costly to evaluate. By eliminating the nuisance parameters from a complex statistical model in order to produce a likelihood function depending on the quantity of interest only, **Grazian and Liseo** approximated the integrated likelihood by the ratio of kernel estimators of the marginal posterior and prior for the quantity of interest given a proper prior for the entire vector parameter. They used several case studies to demonstrate the proposed methodology.

Detection of bacterial species has improved dramatically in recent years due to the development of next generation sequencing. **Clarke, Valdes, Dobra, and Clarke** developed a new Bayesian framework for strain detection from bacterial metagenomic samples generated by next generation sequencing and assessment of strain dependence. Their method uses posterior marginal probabilities to detect specific bacterial strains, and quantifies the dependence between pairs of strains by comparing the joint probability of detection to the product of the marginal probabilities of detection. One attractive feature of the proposed method is that it is scalable to large genomic data. They demonstrated their proposed method on two metagenomic samples from the Human Microbiome Project (HMP) Data Analysis and Coordination Center (DACC). Binary time series data are often encountered in the behavioral study. One of the critical issues in modeling binary response data is the choice of links. To address this important issue, **Abanto-Valle, Dey, and Jiang** developed state space models with the generalized extreme value and symmetric power logit links for binary time series data. They demonstrated the proposed method to estimate the effects of deep brain stimulation (DBS) on attention of a macaque monkey performing a reaction-time task. Continuous density estimation has received abundant attention in the Bayesian nonparametrics literature. However, there is limited theory on multivariate mixed scale density estimation. In this regard, **Canale and Dunson** considered a general framework to jointly model continuous, count and categorical variables under a nonparametric

prior, which is induced through rounding latent variables having an unknown density with respect to Lebesgue measure. For the proposed class of priors, they provided sufficient conditions for large support, strong consistency and rates of posterior contraction. They further applied the proposed procedure via a rounded multivariate nonparametric mixture of Gaussians to the crime and communities data.

HIV RNA viral load measures are often subjected to some upper and lower detection limits depending on the quantification assays. Thus, the responses are either left or right censored. Linear/nonlinear mixed-effects models, with slight modifications to accommodate censoring, are routinely used to analyze this type of data. Usually, the inference procedures are based on normality (or elliptical distribution) assumptions for the random terms. However, those analyses might not provide robust inference when the distribution assumptions are questionable. Alternatively, quantile regression can characterize the entire conditional distribution of the outcome variable, and is more robust to outliers and misspecification of the error distribution. To model longitudinal data in the presence of censored responses, **Lachos, Chen, Abanto-Valle, and Azevedo** developed Bayesian quantile regression via a hierarchical mixed-effects model under the assumption that the error term follows an asymmetric Laplace distribution. They illustrated the proposed procedure with two HIV AIDS studies on viral loads that were initially analyzed using the typical normal (censored) mean regression mixed-effects models. To address several issues that arise in designing a phase I/II trial, **Guo, Zang, and Yuan** proposed a Bayesian phase I/II design that jointly models efficacy, toxicity, and dropout as time-to-event data. Correlations among the three time-to-event outcomes are taken into account by a shared frailty. Their joint model strategy accounts for the informative dropouts and has an additional advantage of accommodating a high accrual rate without suspending patient enrollment when toxicity or efficacy outcomes require a long follow-up. Their simulation studies empirically showed that the proposed design has more desirable operating characteristics with a high probability of selecting the target dose and assigning the most patients to the target dose.

Masked data such as competing risk data are becoming more prominent in reliability studies, medical diagnostic studies and biological systems. For competing risk data, each failure time is associated to a known cause of failure, whereas for masked data, the causes for a failure may be unknown (masked) for a group of subjects. Under the exponential model for masked data, **Xu, Tang, and Sun** showed that the parameters are nonidentifiable under a general masking probability assumption and established the connection between the maximum likelihood estimates and the posterior estimates under the symmetric assumption. They further derived the Jeffreys prior and the reference prior under the symmetric assumption and examined the propriety of the posteriors under the Jeffreys prior and the reference prior. The finite mixture regression is a method to account for heterogeneity in relationship between the response variable and the predictor variables. The variable selection issue within each component in the finite mixture regression has not been fully studied in the literature from a Bayesian perspective. Thus, **Chen and Ye** proposed an approach by embedding variable selection into the data augmentation method that iteratively updates estimation in two steps: estimate parameters for each component and determine the latent membership of each observation. Componentwise variable selection is realized by imposing special priors or procedures designed for parsimony in the first step. They further illustrated how two popular variable selection techniques can be embedded in the proposed approach: g-prior and Stochastic Search Variable Selection. They applied the proposed approach to analyze the data from bioinformatics.

We would like to thank all the authors for contributing their fine pieces of research work to both Part I and Part II of this special issue. We would also like to thank Aileen McElroy and Ina Talandienė for their editorial assistance. Without their support, this special issue would not be possible. Finally, we sincerely hope that this special issue would further promote developments, innovations, and applications of Bayesian statistics.

Ming-Hui Chen (Guest Editor), University of Connecticut
Heping Zhang (Editor-in-Chief), Yale University