Research Statement

Xiaofei Wang
Department of Statistics
Yale University, New Haven, CT 06510
email: xiaofei.wang@yale.edu

December 29, 2013

I enjoy research that is motivated by real-world problems. While at Yale, my three focuses in research are Bayesian change point analysis, statistical computing, and interdisciplinary research. My primary area of research is on the topic of Bayesian change point analysis, which has many applications in econometrics, environmental sciences, and biological sciences. In the age of big data, statistical computing is an important topic, helping to harness the power of cloud computing for analyzing larger and larger datasets. Finally, I enjoy collaborating with researchers in other fields of study; I believe that interdisciplinary research keeps statisticians grounded. In the following, I describe each of my research areas.

1 Bayesian Change Point Analysis

Inspired by the New Haven, Connecticut residential property data[1], my foray into change point analysis was motivated by the question of how to best model real estate values. A naive approach might entail fitting a linear model predicting housing values in a particular year using predicting variables of square footage, the number of bathrooms, the number of bedrooms, and lot size. Such an approach proves unsatisfactory; a simple plot of the residuals on a map with longitude on the x-axis and latitude on the y-axis would show large clusters of positive residuals intermingled with large clusters of negative residuals. These clusters are perhaps explained by an unobserved neighborhood structure. If we define neighborhoods as a group of spatially-near observations that follow the same distribution, then neighborhood detection is equivalent to change point detection.

[1] gave a Bayesian method for the simple change point problem, where a sequential set of observations $y_1, \ldots, y_n$ are assumed to follow normal distributions $N(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$ and θ_i are assumed to be constant within contiguous blocks of a partition of the series. I provided four extensions of this methodology in [6]. The first extension addresses multivariate serial data, where each $\boldsymbol{y_i}$ in the series is k-dimensional and is assumed to follow a $N_k(\boldsymbol{\theta_i}, \sigma^2\boldsymbol{I})$ distribution. The second extension considers sequential observations $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, where the response y_i might be explained by predicting variables $\boldsymbol{x_i}$ via a linear model. For all indices i of observations within a given block S, we model $y_i \sim N(\widetilde{\boldsymbol{x}}_{\boldsymbol{iS}}\boldsymbol{\beta_S}, \sigma^2)$, where $\widetilde{\boldsymbol{x}}_{\boldsymbol{iS}}$ is the row vector of the predicting variables for the i-th observation centered about their means over block S. The third and fourth extensions generalize the previous two extensions by assuming the same types of observed data (multivariate and regression data) lie on a general graph structure.

The generalized Bayesian change point methodology that allows fitting linear models within blocks can be applied to the New Haven real estate problem described in the beginning of this

[1] http://data.visionappraisal.com/newhavenct/

section. In [6], I used a minimum spanning tree based on house latitudes and longitudes. Using this graph structure, my method yielded posterior estimates of coefficients for each house, accounting for the unobserved neighborhood structure.

Future work in this area might include broader generalizations of the underlying models, such as considering a more general covariance structure Σ as opposed to $\sigma^2 \boldsymbol{I}$ in the current multivariate extensions.

2 Statistical Computing

Given the growing availability and affordability of cloud computing resources, conducting data analysis efficiently is easier than ever before. [2] and [3] provide a roadmap for `R` users to harness the power of the Amazon Elastic Compute Cloud for performing parallel processing across multiple machines. A good setup has low latency between machines and can make use of shared volumes, which are helpful when the dataset used during computation is large enough such that both 1) passing data constantly between machines and 2) keeping separate copies of the dataset on each machine are undesirable.

3 Interdisciplinary Research

John W. Tukey once said that a statistician gets "to play in everyone's backyard." I believe that every statistician should play in someone's backyard occasionally in order to stay abreast of the types of real world problems that demand statistical solutions. During my stay at Yale University, I had the opportunity to work on several projects with researchers in other departments. Two of these projects led to papers.

I had the pleasure of working with Professor Stephen Stearns on a genome-wide association study (GWAS) project. In [5], we examined genetic evidence of the tradeoff between lifetime reproductive success and lifespan using the Framingham Heart Study dataset. We found several single nucleotide polymorphisms that were associated with the relationship between lifetime reproductive success and lifespan.

I also had an opportunity to work on a project studying the effectiveness of medication on eye pressure [4]. As a retrospective study, we were unable to make statements regarding causality; we were able to discuss the absence or presence of an association between the variables of interest.

4 Future Work

I will continue to work on the three areas of research mentioned above, with particular emphasis on Bayesian change point analysis. There are many classes of problems, such as those in finance, that require more general assumptions than those used in the existing models mentioned above. As such, one of my priorities is to work on generalizations of these models.

References

[1] D. Barry and J. A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):pp. 309–319, 1993.

[2] J. W. Emerson and X. Wang. Advances in big data and high-performance computing. In *Proceedings of the 2013 IASC Satellite for the ISI WSC and the 8th IASC-ARS Conference*, pages 79–84, August 2013.

[3] J. W. Emerson and X. Wang. Amazon EC2, big data and high-performance computing. *Statistical Computing and Graphics Newsletter*, 23(1), July 2013.

[4] J. Oatts, X. Wang, N. Patel, K. Kaplowitz, and N. Loewen. Effect of alpha-2-agonist premedication on intraocular pressure after selective laser trabeculoplasty. *British Journal of Ophthalmology*, submitted.

[5] X. Wang, S. G. Byars, and S. C. Stearns. Genetic links between post-reproductive lifespan and family size in Framingham. *Evolution, Medicine, and Public Health*, 2013(1):241–253, 2013.

[6] X. Wang and J. W. Emerson. Extensions of Bayesian change point analysis. *Journal of the American Statistical Association*, to be submitted.