

Research Statement - Howard Bondell

My research interests are varied, but all connected by the fact that the motivation behind them stems from applications to real-world problems. I have collaborated with, and/or consulted for, researchers in various fields, including medicine, biology, nursing, pharmacy, psychology, political science, oceanography, and sociology. The truly interdisciplinary nature of statistical methodology and the interplay between theory and application are the driving forces of my research work. Due to my consulting work, I have received acknowledgements in at least 5 papers in scientific journals and co-authorship on an additional paper.

My dissertation research is in the area of robust estimation for logistic regression. I have also worked with researchers at the National Institutes of Health (NIH) on estimation procedures for pooled biological specimens, and I have recently become interested in the area of functional data analysis. I will now briefly describe my work in these three areas, but as mentioned previously, much of my future research will likely come from new applications, particularly in collaborative work with scientists in other fields.

Robust Logistic Regression

Logistic regression is widely used in all areas of research throughout the natural and social sciences, and beyond. However, it is well known that Maximum Likelihood Estimation can be greatly affected by atypical observations. While working on a consulting project with social science data, I noticed that the estimation of the logistic regression coefficients was heavily dependent on a few influential observations, as most likelihood-based methods tend to be. I decided that this deserved further investigation and discovered that several robust techniques had already been proposed in the literature and implemented in standard software packages. However, the standard robust approaches currently implemented are mainly ones that are borrowed from linear regression, downweighting in terms of influence, or leverage, typically by using elliptical contours from some notion of “center”, commonly known as “bounded influence approaches”.

When considered via the case-control viewpoint (conditioning on the binary variable), it is clear that the marginal distribution of the covariates is a mixture of two groups, so the standard elliptical-based approaches are not appropriate. In the extreme case, if one group is much smaller than the other, and their centers were sufficiently far apart, all observations in the smaller group would be systematically considered outliers and receive little or no weight. This realization, which to my knowledge has not appeared in the literature, then became the key underlying principle in my dissertation research.

In my dissertation, I have developed a new minimum distance approach to logistic regression after identifying the case-control formulation with a two-sample semi-parametric biased sampling problem. Based on this equivalence, two empirical distributions are constructed, one fully nonparametric and one semi-parametric. A measure of discrepancy between these two distributions is minimized to obtain parameter estimates whose properties are determined by the choice of discrepancy measure. These estimators can be highly efficient if the model is true, yet robust to atypical observations. This approach remains valid even if the sampling was not done via the case-control setup. The initial paper introducing this idea is currently under review, but also available as a technical report. A follow-up paper is in progress.

The proposed minimum distance approach also yields a natural family of goodness-of-fit tests. Future work is needed to determine the properties of these tests, as well as to further examine the choices of discrepancy measures and the resulting estimates. Extending the idea to multiclass and ordinal logistic regression is a natural next step. It is also worth investigating the use of this minimum distance idea for estimation and testing goodness-of-fit in other semi-parametric models such as proportional odds, proportional hazards, and other techniques used

regularly in fields such as survival analysis.

In addition to the minimum distance approach, I am currently working on an alternate version of the bounded influence approach by considering each group individually, as opposed to downweighting in terms of a single distribution. This is being done in two ways. The first is to directly downweight in terms of leverage (position in covariate space) for each group separately, from their own center. This is an alternative to the class commonly known as Mallows-type estimates, although it is no longer strictly in the Mallows class. The second is a modification of the influence function to contaminate each conditional distribution individually, as opposed to the standard definition of contaminating the joint distribution. This form of contamination makes more sense in the case-control setting, as the resulting contaminated distribution maintains the same proportions of cases and controls as the original, which under case-control sampling is typically fixed beforehand. This new approach to “bounded influence” heuristically centers the score function separately for each group before measuring influence. These modifications seem extremely promising, both in theory and simulation, and further work is warranted to fully develop the theory and then implement these approaches in standard software packages.

Inference for Pooled Data

Some continuous biomarkers can be highly costly to measure, but easy to acquire multiple specimens. A number of medical researchers have proposed to physically pool, or mix, equal amounts of small samples into one sample and then analyze the result. Assuming that the measures are concentrations, for example, the measurement for the pooled specimen is now the average of the individual measurements. However, it is desired to conduct inference on the distribution of the biomarker itself. The original application was to compare the distribution of the biomarker in a control group with that of a diseased group via an ROC curve analysis.

The general problem is then how to recreate the desired attributes of the original distribution based solely on observing a random sample of averages (or, equivalently, sums). The initial approach is to assume a Normal (or Gamma) distribution for the biomarker, so that the sums are Normal (or Gamma) whose parameters are easily related to the original parameters. An alternative parametric form may be assumed for the biomarker instead. In this case, the sum no longer has a simple distribution. However, the first few central moments of a sum is related to the first few central moments of the original data so that a method of moments approach can be used to estimate the parameters of the original distribution. I am a co-author with researchers at the NIH on two papers written in this area, one accepted for publication, the other under review.

Further work in this area, with possible continued collaboration, is to extend via a fully nonparametric approach. The approach could use the empirical moments and an Edgeworth expansion, or use the empirical characteristic function and the simple relationship between the characteristic function of the original random variables and that of its sum, and then the inversion formula. Once an estimate of the original density is recovered, one can proceed with, for example, estimating the area under the ROC curve, which is just $P(Y > X)$, where Y is a random measurement from the diseased group and X is a random measurement from the control group, which can then be estimated by using the available density estimate for each group.

Functional Data Analysis

Functional data can arise in a variety of ways. The common theme is that the true underlying random variable is actually a function, not a finite dimensional vector. For example, data collected at discrete time points for an individual is often modeled as a vector of correlated

random variables, commonly known as longitudinal data. However, the nature of the true process for each individual is not a finite vector of measurements, there is an underlying continuous function that is only measured at discrete time points. The continuity, and some smoothness assumptions, allow for modeling and analysis techniques that may supplement and complement standard techniques of multivariate and longitudinal-type data analysis.

Functional data analysis is a relatively new and growing research area and I am currently investigating some of the literature in this field to fully formulate possible topics for additional future work.