Elizabeth Zhang

Revised Thesis Proposal

Purpose. I plan to explore either the issues of multicollinearity or non-additive interactions in environmental exposure mixture studies, comparing the performance of multiple linear regression and Bayesian Kernel Machine Regression (BKMR).

Motivation. I had originally proposed an idea on multi-ancestry genome-wide association studies, which, while still statistically interesting to me, is less motivating as an application area now, though I still want to retain some relevance toward health disparities in my thesis. This summer, I had the opportunity to contribute to epidemiological research, and it helped me realize that I want to pursue statistical work with a clear public health application going forward. In particular, I am motivated by projects that have implications for policy work around disease prevention. At the moment, a particular field of interest is environmental epidemiology, and so I have shifted the application area of my thesis to environmental health. I will note that I am still interested in learning more about the Bayesian approach, so I have selected a Bayesian methodology of interest in this new application area.

Background on problem. It is well known that exposure to various environmental factors is closely linked to health outcomes. Such exposures are distributed unevenly across space, and, with regards to chemical exposures, low-income communities of color are disproportionately vulnerable to harmful toxicity (Bullard, 1993; Schell et al., 2020). Linking exposure with poor health outcomes can inform important regulations controlling the release of pollutants into the environment. To do this, environmental epidemiological studies have typically focused on linking the effect of single pollutants with health; however, this is not representative of reality, wherein humans are exposed to an assortment of particulate matter, often occurring in similar mixtures due to a common source of origin. As a result, the National Institutes of Environmental Health Services has released a call for renewed attention toward estimating the effects of environmental mixtures (Joubert et al., 2022). Mixture analyses can also have more direct implications for public health interventions, as regulation occurs via controlling the source of pollution, which is responsible for the production of a whole mixture of chemicals with specific combined effects on health (Keil et al., 2020). However, the analysis of exposure mixtures presents unique statistical challenges, including high-dimensionality, multicollinearity, non-additive interactions, and nonlinear effects (Yu et al., 2022). These challenges have motivated the development of a variety of methods focused on assessing the exposure-response relationship while also accounting for the above considerations. See (Yu et al., 2022) for a recent review of methods for mixture exposures.

Methods. I am interested in focusing on Bayesian Kernel Machine Regression (BKMR), which has been specifically developed for analyzing exposure mixtures (Bobb et al., 2015). Publicly available software is available through the *bkmr* package in R (with documentation at (Bobb et al., 2018)). BKMR models the health outcome using a Gaussian kernel function of the exposure variables, with a hierarchical variable selection method and options for visualizing potential three-way interactions (Bobb, 2017). I plan to contrast the performance of this method with a standard multiple linear regression.

I am still deciding whether to focus on multicollinearity or non-additive interactions. In either case, I plan to simulate datasets with different levels of the selected feature, and to assess the performance of BKMR using multiple linear regression as a baseline. For multicollinearity, I would simulate different levels of collinearity along with different sample sizes and conduct a power analysis. For non-additive interactions, there has yet to be a formal approach for quantifying the presence of an interaction, so I would likely use general measures to assess model performance. If time permits, I would also be interested in exploring principal components analysis, an older method used in environmental mixtures.

Potential datasets. I plan to base simulations off a publicly available dataset. The National Health and Nutrition Examination Survey (NHANES) is a commonly used dataset for environmental health applications (National Center for Health Statistics, 2023). I may also explore an application on data available from the All of Us study, an ongoing project that aims to recruit a diverse set of participants across the US, with data regularly updated to a publicly available hub (https://allofus.nih.gov/get-involved/opportunities-researchers).

Sources cited.

Bobb, J. F. (2017, December 15). *Example using the bkmr R package with simulated data from the NIEHS mixtures workshop*. https://jenfb.github.io/bkmr/SimData1.html#1_load_packages_and_download_data

Bobb, J. F., Claus Henn, B., Valeri, L., & Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health*, *17*(1), 67. https://doi.org/10.1186/s12940-018-0413-y

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., & Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, *16*(3), 493–508. https://doi.org/10.1093/biostatistics/kxu058

Bullard, R. D. (1993). *Confronting environmental racism: Voices from the grassroots*. South End Press.

Joubert, B. R., Kioumourtzoglou, M.-A., Chamberlain, T., Chen, H. Y., Gennings, C., Turyk, M. E., Miranda, M. L., Webster, T. F., Ensor, K. B., Dunson, D. B., & Coull, B. A. (2022). Powering Research through Innovative Methods for Mixtures in Epidemiology (PRIME) Program: Novel and Expanded Statistical Methods. *International Journal of Environmental Research and Public Health*, *19*(3), Article 3. https://doi.org/10.3390/ijerph19031378

Keil, A. P., Buckley, J. P., O, 'Brien Katie M., Ferguson, K. K., Zhao, S., & White, A. J. (2020). A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. *Environmental Health Perspectives*, *128*(4), 047004. https://doi.org/10.1289/EHP5838

National Center for Health Statistics. (2023, May 31). *NHANES - About the National Health and Nutrition Examination Survey*. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Schell, C. J., Dyson, K., Fuentes, T. L., Des Roches, S., Harris, N. C., Miller, D. S., Woelfle-Erskine, C. A., & Lambert, M. R. (2020). The ecological and evolutionary consequences of systemic racism in urban environments. *Science*, *369*(6510), eaay4497. https://doi.org/10.1126/science.aay4497

Yu, L., Liu, W., Wang, X., Ye, Z., Tan, Q., Qiu, W., Nie, X., Li, M., Wang, B., & Chen, W. (2022). A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environmental Pollution*, *306*, 119356. https://doi.org/10.1016/j.envpol.2022.119356