Modeling Linear Combinations of Multiple Pollutants

Name and affiliations of lead investigators

Name: Professor Patrick Brown
Affiliations: Centre for Global Health Research, St. Michael's Hospital
Department of Statistical Sciences, University of Toronto

Name: Professor Fateh Chebana
Affiliation: Institut national de la recherche scientifique (INRS)

Name: Professor Cindy Feng
Affiliation: School of Epidemiology and Public Health, University of Ottawa

List of proposed callaborators, titles, and affiliations

Name: Kamal Rai
Title: Post-Doctoral Fellow (?)
Affiliations: Centre for Global Health Research, St. Michael's Hospital
Department of Statistical Sciences, University of Toronto

Name: Hwashin Shin
Title: Adjunct Associate Professor (?)
Affiliations: Environmental Health Science and Research Bureau, Health Canada
Department of Mathematics and Statistics, Queen's University

Name: Pierre Masselot
Title: Post-Doctoral Fellow (?)
Affiliation: London School of Hygiene & Tropical Medicine

List of potential partner organizations (optional)

Centre for Global Health Research, St. Michael's Hospital
Institut National de Santé Publique du Québec (?)
Health Canada

Research Aims

The Model

We propose to develop a statistical methodology for modeling the *combined effect* of multiple covariates with non-linear relationships to the outcome of interest. For a response distribution D from the exponential family with a d-dimensional parameter $\theta = (\theta_1, \dots, \theta_d)$ and invertible (canonical) link function g, the general form of the proposed model is,

$$\begin{aligned} Y_t|\theta_t &= D(\theta), \\ g(\theta) &= X\beta + \sum_{i=1}^{I} s_i(\alpha_1\gamma_1 + \dots + \alpha_{n_i}\gamma_{n_i}) + \sum_{k=1}^{K} f_k(\eta_{k,*}) \end{aligned}$$

In the above, X is a regression matrix and β a vector of fixed effects. The second sum consists of K smoothing functions $f_1, \dots, f_K$ applied to K potential confounders $\eta_{k,*}$ (where $*$ indexes possible values of $\eta_{k,*}$). Our interest lies in the first sum, which fits smooth functions to convex combinations of covariates. This is a flexible and extensible model that allows for nonlinear interactions between an outcome of interest Y_t and covariates $(\alpha_1, \dots, \alpha_{n_i})_{i=1}^{I}$ via smoothing functions $\{f_i\}$.

To explore the flexibility of this model, we compare it to a standard regression model and one with a more general functional specification of s_i. To ease the notation in the notation, let us assume there are no confounders and that are only considering a convex combination of two covariates γ_1 and γ_2. Then, there three models we are considering have link function,

1. $g(\theta) = X\beta + \alpha_1\gamma_1 + \alpha_2\gamma_2$ Standard link function
2. $g(\theta) = X\beta + s_1(\alpha_1\gamma_1 + \alpha_2\gamma_2)$ Proposed link function
3. $g(\theta) = X\beta + s_1^*(\gamma_1, \gamma_2)$ 2D link function

Here, it easy to see that the complexity of the second model, proposed in this paper, is between that of a standard regression model and one that uses a 2-dimensional smoother. The proposed model fits a convex combination of γ_1 and γ_2 using a 1-dimensional smoothing function, which has two major advantages over using a 2-dimensional smoother. It's first advantage is that it is less computationally demanding, as it does *not* suffer from the curse of dimensionality. The convex combination $\alpha_1\gamma_1 + \alpha_2\gamma_2$ is always 1-dimensional, and remains so regardless of how many terms are the sum. This constrasts to s_1^*, whose dimensionality scales with its number of inputs. It's second advantage is that it is more interpretable. The *weights* α_1 and α_2 reflect the relative contribution of γ_1 and γ_2 (respectively) and are thus readily interpretable. In comparison, $s^*(\gamma_1, \gamma_2)$ has no such weights and is perhaps best interpreted via (two-dimensional) visualizations.

We propose to develop a fully Bayesian implementation of this model using Stan, a widely used software package that allows for rapid iteration and development of Bayesian models. As Bayesians, we will use random walks as our smoothing function; their equivalent to smoothing splines is detailed in [@, @].

ADD: Non-MCMC implementations?

Research Questions

Overview

Recent years have seen an increased interest in modeling the joint effect of two or more pollutants in health outcomes (Dominici et al. 2010; Billionnet et al. 2012; Davalos et al. 2017); Bayesian approaches can be seen in (Blangiardo et al. 2019; Bobb, Dominici, and Peng 2013; Huang, Lee, and Scott 2018).

1. What is the combined effect of multiple pollutants on various daily mortality outcomes?
2. What is the relationship between daily COVID-19 mortality and air pollution?

Research Question #1

The workhorse of the air pollution literature is the one-pollutant Poisson regression model that accounts for confounders using fixed effects and smooth functions, typically a natural cubic splines (Samet et al. 2000; Dominici et al. 2002; Laden et al. 2006). Let the *average* λ_t. Then, a typical model is,

$$
\begin{aligned}
Y_t|\lambda_t &= \text{Poisson}(\lambda_t), \\
\log(\lambda_t) &= X\beta + \gamma_1 P_{1,t} + \sum_{k=1}^{K} f_k(\eta_{k,*}).
\end{aligned}
$$

Here, Y_t is the health outcome of interest, such as respiratory mortality or morbidity. The design matrix X typically contains day-of-the-week effects and seasonal terms, and the $f_i(\eta_{k,*})$ are smooth functions of potential confounders such as time and temperature. A more flexible version of this model is to fit $P_{1,t}$ to its own smooth function s_i. More generally, we could fit N pollutants using smooth functions. A third alternative is to model multiple pollutants by applying smooth functions to linear combinations of pollutants.

These are same three scenarios as considered above. *Why move beyond the linear model*?

Using linear combinations of pollutants will allow the model to estimate their *combined* effect in a natural and informative way. For instance, this weighted sum naturally captures synergistic effects that can occurs when more than one pollutant is present at higher levels (Xia and Tong 2006). The levels of different pollutants are also highly correlated (**ADD**: Reference); modeling sums as nonlinear functions will help alleviate the statistical concerns introduced by this high correlation (**ADD**: Reference).

Comparisons to other Multi-Pollutant Models

Recent years have seen more interest in modeling the health impact of multiple pollutants. The review article by Davalos et al. (2017) identified various statistical approaches used to accounting for multiple pollutants: additive main effects, effect measure modification, dimension reduction, and nonparametric methods. Additive main effects models are difficult to extend to multiple pollutants (Dominici et al. 2010), while nonparameteric methods are generally less interpretable than parametric ones. A principle goal of air pollution modeling is to build evidence regarding levels that cause health effects. Given the multidisciplinary nature of the research teams and potential stakeholders regarding this research, the models *must* be easily interpetable to meet this goal.

A number of dimension reduction methods are identified in Davalos et al. (2017). Unsupervised methods include principle components analysis (PCA) and k-means clustering. Unfortunately, the coefficients obtained from using PCA are often difficult to interpret, and clustering is also difficult to interpret. Some papers using supervised methods consider weighted sums of pollutant concentrations. The authors of Pachon et al. (2012) specify the weights from data rather than estimating them, while the authors of Roberts and Martin (2006) consider model a *mixture* of pollutants linearly; this corresponds to assuming that $s_1(\alpha_1 P_{1,t} + \alpha_2 P_{2,t}) = \beta_m(\alpha_1 P_{1,t} + \alpha_2 P_{2,t})$ is a linear function.

In Sun et al. (2013), the authors make quantitative comparisons of five different methods. These include

Research Question #2

The relationship between daily coronavirus deaths and air pollution levels has become an active area of research in recent months. For instance, Wu et al. (2020) finds that a 1 μg increase in long-term exposure to ambient $\text{PM}_{2.5}$ increases the coronavirus death rate by 15%. Additional studies that examine this relationship include (Conticini, Frediani, and Caro 2020; Sciomer et al. 2020; Setti et al. 2020).

However, much work remains to be done. For instance, non-COVID-19 daily mortality data is generally not yet available, such that we do not have an accurate meaasure of *excess* deaths attributable to COVID-19. These excess deaths could be attributable to under-reported COVID-19 case and death counts (due to limited testing), restricted access to care for patients with other health conditions, or potential reporting delays. Moreover, cumulative COVID-19 mortality will likely continue to rise for some time, making the question of excess deaths due to COVID-19 best suited for an ongoing inquiry that help inform public health responses.

Anticipated roles of trainees (students and post-doctoral fellows)

Kamal will develop the Bayesian implementation of the multiple pollutant models in Stan. This includes exploring determining appropriate prior distributions for the weights α and developing visualizations that communicate model results in a simple and interpretable way. He will be responsible for producing paper(s) summarizing the results of this model when run on Canadian air pollution and mortality data. To facilitate team communication and cohesion, he will also split time between Toronto (at the Centre for Global Health Research) and Ottawa (at the Universtiy of Ottawa), and use the proximity to Quebec to visit the collaborators there.

The University of Toronto PhD student will compare the results from the Bayesian random walk models to those obtained from non-MCMC methods. For example, these could include frequentist methods that fit (natural cubic) splines or Bayesian inference using R-INLA (Rue, Martino, and Chopin 2009). The University of Laval/University of Ottawa PhD student will compare the results from the Stan implementation to those obtained by a case-crossover model.

Plans for dissemination and communication

The results and findings of this multiple pollutant inquiry will be shared with Health Canada and the Institut National de Santé Publique du Québec. The lead investigators have a track record of publishing their research results in statistical and epidemiolgical journals, and aim to publish the results of this project in high-impact journals. They (or the trainees) will also attend appropriate conferences to present the work while it in progress.

Suggested reviewers

(Possible) CVs

- Patrick, Fateh, Hwashin, Meredith, Cindy

Preliminary budget description

The CANSSI Collaborative Research Team (CRT) grant is for $180,000 over 3 years. We propose the budget:

1. $30,000/year to support post-doctoral funding; matched by the Centre for Global Health Research.
2. $12,000/year to support a Laval University or University of Ottawa PhD student.
3. $12,000/year to support a University of Toronto PhD student.
4. $6,000/year to support travel to/from the cities of the lead investigators – Toronto, Ottawa, and Quebec – and annual team meetings held around the Statistical Society of Canada conference.

Other funding (optional)

The Centre for Global Health Research at St. Michael's Hospital will contribute $30,000 to post-doctoral funding. Additional expenses (travel, conference registration, etc.) related to disseminating results will be paid for by the lead investigators.

References

Billionnet, Cécile, Duane Sherrill, Isabella Annesi-Maesano, and others. 2012. “Estimating the Health Effects of Exposure to Multi-Pollutant Mixture.” *Annals of Epidemiology* 22 (2). Elsevier: 126–41.

Blangiardo, Marta, Monica Pirani, Lauren Kanapka, Anna Hansell, and Gary Fuller. 2019. “A Hierarchical Modelling Approach to Assess Multi Pollutant Effects in Time-Series Studies.” *PloS One* 14 (3). Public Library of Science.

Bobb, Jennifer F, Francesca Dominici, and Roger D Peng. 2013. “Reduced Hierarchical Models with Application to Estimating Health Effects of Simultaneous Exposure to Multiple Pollutants.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62 (3). Wiley Online Library: 451–72.

Conticini, Edoardo, Bruno Frediani, and Dario Caro. 2020. “Can Atmospheric Pollution Be Considered a Co-Factor in Extremely High Level of Sars-Cov-2 Lethality in Northern Italy?” *Environmental Pollution*. Elsevier, 114465.

Davalos, Angel D, Thomas J Luben, Amy H Herring, and Jason D Sacks. 2017. “Current Approaches Used in Epidemiologic Studies to Examine Short-Term Multipollutant Air Pollution Exposures.” *Annals of Epidemiology* 27 (2). Elsevier: 145–53.

Dominici, Francesca, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. 2002. “On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health.” *American Journal of Epidemiology* 156 (3). Oxford University Press: 193–203.

Dominici, Francesca, Roger D Peng, Christopher D Barr, and Michelle L Bell. 2010. “Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-Pollutant Approach.” *Epidemiology (Cambridge, Mass.)* 21 (2). NIH Public Access: 187.

Huang, Guowen, Duncan Lee, and E Marian Scott. 2018. “Multivariate Space-Time Modelling of Multiple Air Pollutants and Their Health Effects Accounting for Exposure Uncertainty.” *Statistics in Medicine* 37 (7). Wiley Online Library: 1134–48.

Laden, Francine, Joel Schwartz, Frank E Speizer, and Douglas W Dockery. 2006. “Reduction in Fine Particulate Air Pollution and Mortality: Extended Follow-up of the Harvard Six Cities Study.” *American Journal of Respiratory and Critical Care Medicine* 173 (6). American Thoracic Society: 667–72.

Pachon, Jorge E, Sivaraman Balachandran, Yongtao Hu, James A Mulholland, Lyndsey A Darrow, Jeremy A Sarnat, Paige E Tolbert, and Armistead G Russell. 2012. “Development of Outcome-Based, Multipollutant Mobile Source Indicators.” *Journal of the Air & Waste Management Association* 62 (4). Taylor & Francis: 431–42.

Roberts, Steven, and Michael A Martin. 2006. “Investigating the Mixture of Air Pollutants Associated with Adverse Health Outcomes.” *Atmospheric Environment* 40 (5). Elsevier: 984–91.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2). Wiley Online Library: 319–92.

Samet, Jonathan M, Francesca Dominici, Frank C Curriero, Ivan Coursac, and Scott L Zeger. 2000. “Fine Particulate Air Pollution and Mortality in 20 Us Cities, 1987–1994.” *New England Journal of Medicine* 343

(24). Mass Medical Soc: 1742–9.

Sciomer, Susanna, Federica Moscucci, Damiano Magrı‘, Roberto Badagliacca, Gianfranco Piccirillo, and Piergiuseppe Agostoni. 2020. "SARS-Cov-2 Spread in Northern Italy: What About the Pollution Role?" *Environmental Monitoring and Assessment* 192. Springer: 1–3.

Setti, Leonardo, Fabrizio Passarini, Gianluigi De Gennaro, Pierluigi Barbieri, Alberto Pallavicini, Maurizio Ruscio, Prisco Piscitelli, Annamaria Colao, and Alessandro Miani. 2020. "Searching for Sars-Cov-2 on Particulate Matter: A Possible Early Indicator of Covid-19 Epidemic Recurrence." Multidisciplinary Digital Publishing Institute.

Sun, Zhichao, Yebin Tao, Shi Li, Kelly K Ferguson, John D Meeker, Sung Kyun Park, Stuart A Batterman, and Bhramar Mukherjee. 2013. "Statistical Strategies for Constructing Health Risk Models with Multiple Pollutants and Their Interactions: Possible Choices and Comparisons." *Environmental Health* 12 (1). Springer: 85.

Wu, Xiao, Rachel C Nethery, Benjamin M Sabath, Danielle Braun, and Francesca Dominici. 2020. "Exposure to Air Pollution and Covid-19 Mortality in the United States." *medRxiv*. Cold Spring Harbor Laboratory Press.

Xia, Yingcun, and Howell Tong. 2006. "Cumulative Effects of Air Pollution on Public Health." *Statistics in Medicine* 25 (20). Wiley Online Library: 3548–59.