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 fully Bayesisan implementation of the *groupwise additive index model* (GAIM) (Masselot et al. 2020). Sequential quadratic programming is used in Masselot et al. (2020) to optimize the GAIM in the frequentist setting, but there is no discussion of a Bayesian implementation. Putting priors aside, for response distribution D from the exponential family with a d-dimensional parameter $\theta = (\theta_1, ..., \theta_d)$ and canonical link function g, the GAIM is,

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

In the above, X is a regression matrix and β a vector of fixed effects. The second sum consists of K smoothing functions $f_1, ..., f_K$ applied to K potential confounders $\eta_k = (\eta_1, ..., \eta_{n_k})$. The novelty lies in the first sum, which fits smooth functions s_i to a weighted sum of potentially time-varying covariates.

The GAIM is a flexible and extensible statistical model with a wide variety of potential applications, as the researcher can choose the number I of smooth functions s_i and the number of covariates n_i in each s_i. For example, the researcher may specify a model with one smooth function s_1 and two covariates, where $\gamma_{1,t}$ is the observed levels of the pollutant ozone (O_3) and $\gamma_{2,t}$ is the observed levels of particulate matter less than 2.5 μg ($PM_{2.5}$). In this case, the 2 weights give the *relative* contribution of O_3 and $PM_{2.5}$ to Y_t, the health outcome of interest. If the researcher added 1-day lagged observations of O_3 and $PM_{2.5}$, the 4 weights would give the *relative* contribution of the 0-day and 1-day lagged O_3 and $PM_{2.5}$ levels. If the researcher instead added two more pollutants, nitrogen dioxide (NO_2) and particulate matter less than 10 μg (PM_{10}), the 4 weights would give the *relative* contribution of each pollutant.

To explore the benefits the GAIM provides over other possible model specifications, we next compare it to a standard regression model and one where the s_i have more general form. To ease notation, let us assume there are no confounders and that we are only considering a linear combination of two terms, $\gamma_{1,t}$ and $\gamma_{2,t}$. Then, these three models have link functions,

$$
\begin{aligned}
g(\theta) &= X\beta + \alpha_1\gamma_1 + \alpha_2\gamma_2 && (1) \\
g(\theta) &= X\beta + s_1(\alpha_1\gamma_1 + \alpha_2\gamma_2) && (2) \\
g(\theta) &= X\beta + s_1^*(\gamma_1, \gamma_2) && (3)
\end{aligned}
$$

The model with link function (1) is a regression model with two fixed effects, α_1 and α_2. The GAIM has link function (2) with weights α_1 and α_2, while link function (3) specifies s_1^* as a 2-dimensional smoothing function. The GAIM is more flexible than a regression model, as the inclusion of s_1 allows it to capture nonlinearities in the relationship between $\gamma_{1,t}$, $\gamma_{2,t}$ and the outcome Y_t. Compared to the 2-dimensional smoothing function, the GAIM is less computationally demanding – it does *not* suffer from the curse of dimensionality. The linear combination $\alpha_1\gamma_{1,t} + \alpha_2\gamma_{2,t}$ is always 1-dimensional, and remains so regardless of how many terms are in its sum. This constrasts with s_1^*, whose dimensionality scales with its number of arguments. The second advantage of the GAIM is that it is more interpretable, as the *weights* α_1 and

α_2 reflect the *relative* contribution of $\gamma_{1,t}$ and $\gamma_{2,t}$, respectively. In comparison, $s_1^*(\gamma_1, \gamma_2)$ has no such weights; indeed, s_1^* may not have *any* directly interpretable parameters.

We will develop a fully Bayesian implementation of the GAIM in Stan, a widely used statistical modeling language that allows for the rapid iteration and development of Bayesian models with inference performed using Hamiltonian Monte Carlo (Carpenter et al. 2017). No such Bayesian implementation currently exists; developing a robust implementation and releasing the software would be a significant contribution towards having accessible and interpretable dimension-reducing models. Developing a robust Bayesian implementation includes, among other things, identifying suitable priors for the weights α and suitable priors for imposing a variety of shape constraints, such as monotonicity and convexity, on the s_i. Kimeldorf and Wahba (1970) describes a natural correspondence between random walks and smoothing splines, such that random walks are the natural smoothng function for Bayesians. The second goal is to develop non-MCMC estimation methods, and to characterize the performance of this implementation under a variety of different different response distributions.

Research Questions

Overview

We will demonstrate the computational and interpretation benefits of the GAIM in the context of the health effects of multiple air pollutants. Recent years have seen 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). We will consider two research questions in the course of our inquiry, namely,

1. What is the combined effect of multiple pollutants on various daily mortality outcomes?, and
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 log-linear Poisson regression model. This model accounts for confounders using fixed effects and smooth functions, such as the natural cubic spline (Samet et al. 2000; Dominici et al. 2002; Liu et al. 2019). Let the *average* rate an outcome Y_t occurs on day t be denoted by λ_t. Then, a typical one-pollutant 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 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. An extension of this model would fit $P_{1,t}$ to its own smooth function s_1. More generally, we could fit N pollutants to N smooth functions, $s_1, \ldots, s_N$. A third alternative, which we propose here, is to model N pollutants using the GAIM model.

We have previously discussed the computational and interpretability advantages of the GAIM. For one, people are exposed to the mixture of *all* the pollutants present in their area. When extending the one pollutant model to account for this, it is important to note that common pollutants are highly correlated (Huang, Brown, and Shin 2020). When fitting a single pollutant model, health effects that are attributed to this one pollutant may very well be caused by another correlated pollutant(s). Thus, the one-pollutant model should be extended in a way that recognizes that the natural target for inference is the weighted sum of *all* pollutants. The GAIM is particularly well-suited to estimating the effects of mixture components, as the smooth functions s_i gives the total effect of the mixture and the weights α give the relative contribution of each component. Using a nonlinear function of the weighted sum also has advantages. For example, it allows the model to capture the synergistic effects that can occur when more than one pollutant is present at higher levels, a result previously seen in Xia and Tong (2006).

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 multi-disciplinary 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.

Returning to non-parameteric methods, Davalos et al. (2017) discusses two major non-parametric methods that have been applied to the multiple pollutant problem: Classification and Regression Trees (CART) and Bayesian Kernel Machine Regression (BKMR). CART bins pollution data into observations with similar health outcomes, while BKMR allows for estimation and variable selection. BKMR was introduced in Bobb et al. (2015), with software and applications discussed in Bobb et al. (2018).

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 $PM_{2.5}$ increases the coronavirus death rate by 15%. Additional studies that examine this relationship include Conticini, Frediani, and Caro (2020), Sciomer et al. (2020), and 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.

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.

Bobb, Jennifer F, Birgit Claus Henn, Linda Valeri, and Brent A Coull. 2018. "Statistical Software for Analyzing the Health Effects of Multiple Concurrent Exposures via Bayesian Kernel Machine Regression." *Environmental Health* 17 (1). BioMed Central: 1–10.

Bobb, Jennifer F, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. 2015. "Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures." *Biostatistics* 16 (3). Oxford University Press: 493–508.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1).

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, Brown Partrick, and Hwashin H. Shin. 2020. "Daily Mortality/Morbidity and Air Quality: Using Multivariate Time Series with Seasonally Varying Covariances." *Submitted*.

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.

Kimeldorf, George S, and Grace Wahba. 1970. "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines." *The Annals of Mathematical Statistics* 41 (2). JSTOR: 495–502.

Liu, Cong, Renjie Chen, Francesco Sera, Ana M Vicedo-Cabrera, Yuming Guo, Shilu Tong, Micheline SZS Coelho, et al. 2019. "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities." *New England Journal of Medicine* 381 (8): 705–15.

Masselot, Pierre, Fateh Chebana, Céline Campagna, Éric Lavigne, Taha B.M.J. Ouarda, and Pierre Gosselin. 2020. "Constrained Groupwise Additive Index Models." *Submitted*.

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.

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.

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.

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.

Suggested reviewers

Patrick might know?

(Possible) CVs

- Patrick, Fateh, Hwashin, Meredith, Cindy