Research Aims

Overview

There is growing interest in developing a simple, intuitive air quality index that simultaneously accounts for the health effects of multiple air pollutants (Dominici et al. 2010; Stieb et al. 2008; Bopp et al. 2018). Health effects of air pollution depend on the composition of pollutants in the air, not simply the levels of a single pollutant (Dominici et al. 2010). An air quality index that reflects this understanding should account for the levels and relative contributions of each air pollutant in the ambient air. In this proposal, we will improve statistical methods for conducting inference on the health effects of simultaneous exposure to multiple environmental pollutants, with a focus on quantifying short-term effects of poor air quality on health outcomes at the population level.

The *constrained groupwise additive index model* (cGAIM), introduced by Xia and Tong (2006), is a vehicle for providing a multi-pollutant health index. Writing λ_{it} for the risk of a particular outcome for an individual i on day t, the cGAIM is

$$\lambda_{it} = \exp\left[X_{it}^T\beta + s(\alpha^T Z_{it}) + f_1(W_{1it}) + ... + f_K(W_{Kit})\right].$$

The β parameters are the fixed effects of the potentially time-varying linear covariates X_{it}, and $f_1, \ldots, f_K$ are smooth functions that account for potential confounding variables W_{kit} ($k = 1 \ldots K$). The distinguishing feature of cGAIM is the smooth function s whose argument is a linear combination of covariates Z_t, which might be PM 2.5 and Ozone. The α parameter is a vector of weights on the entries of Z_t, and gives their relative contributions. The smooth functions s and the f_k might be composed of spline functions or Gaussian processes such as random walks.

While cGaim has until now been used with daily case counts as the response variable, we will employ case crossover models, which have seen increased attention in the air pollution literature (Wei et al. 2019; Stringer et al. 2020). These models define one or more *control days* for each case, for example the same day of the week on the previous two weeks, and use a partial likelihood for the probability the event occurs on the case day rather than the control days. The advantage of case crossover models is any risk factors which vary slowly or not at all, or are the same on the case and control days, are automatically adjusted for. The challenge introduced by case crossover models is the likelihood depends on non-linear combinations of the latent variables.

Estimating α is the main statistical challenge with the cGAIM, which Masselot et al. (2020) accomplish with frequentist inference methods that use sequential quadratic programming. We will develop a Bayesian methodology for inference with the cGAIM — the bcGAIM — which will fully quantify the uncertainty around α and propogate the uncertainty into inference on s. While there is consistent evidence that air pollution increases daily incidence of adverse health outcomes, we expect that information contained in data on the effects of specific combinations of pollutants at different lags is weak. The bcGAIM will identify both what we can and cannot infer from historical health and pollution data, a feature which will become increasingly important as the dimensionality of α and Z_{it} increase.

Objectives

This proposal has three main research objectives. The first is to develop an air quality index for Canadian cities that accounts for the combined effects of multiple air pollutants. This index will be developed in collaboration with Health Canada and INSQ with the intention of being used in a public warning system. The second is to investigate how exposure to different mixtures of pollutants affects daily COVID-19 mortality. The third is to undertake epidemiological studies involving exposures to environmental pollutants in areas where the Centre for Global Health Research or other project collaborators have suitable data (Canada, India, and the United States).

Consider the first objective – building an air quality index. There is evidence that health effects estimated by single pollutant models may be caused by correlated pollutants omitted from the model. For example, Franklin and Schwartz (2008) found that the effect of ozone on non-accidental mortality was "substantially reduced" after adjusting for particle sulfate and Liu et al. (2019) found significant differences in the percentage change of all-cause mortality attributable to $PM_{2.5}$ and PM_{10} after adjusting for NO_2 or SO_2. Furthermore, there is evidence that some health outcomes are nonlinearly related to health outcomes levels (Feng et al. 2016). Our three research objectives require non-linear/semi-parametric dose-response curves and a combined-effect exposure model. Our bcGAIM model will meet the requirements necessary to fulfill these objectives due to its being able to estimate weights using data and allow for nonlinear relationships between pollutants and health outcomes. It is also applicable to both Poisson time series and case crossover models.

The bcGAIM model parameters must be interpretable. Unsupervised methods such as principle components analysis and clustering are difficult to interpret (Davalos et al. 2017). A popular nonparametric method is Bayesian Kernel Machine Regression (BKMR), which models an exposure-response surface via a kernel function (Bobb et al. 2015). Using a hierarchical Bayesian variable selection method, it can select one pollutant from a group of correlated ones, and is interpreted by visualizing cross-sections of a potentially high-dimensional exposure-response surface. The bcGAIM will provide similar flexibility to the BKMR, while being able to meet the communication needs of inter-disciplinary research teams.

The second objective is to build a COVID-19 mortality model. The relationship between daily COVID-19 deaths and air pollution levels has recently become an active area of research. For instance, Wu et al. (2020) apply a zero-inflated negative binomial to U.S. data, where the zero-inflation accounts for counties with no COVID-19 deaths. They find that a 1 μg increase in long-term exposure to ambient $PM_{2.5}$ increases the COVID-19 mortality rate by 15%. Here, it is important to the bcGAIM is a general modeling framework. For the COVID-19 mortality model, we will use COVID-19 mortality as the response and the included variables and confounders may be different than the multi-pollutant model used to build the air quality index. As well as exploring daily variations in the case fatality rate (and its relation to air quality), the COVID-19 model will be adapted to consider long-term exposures and COVID-19 incidence rates.

The third objective is to pursue additional epidemiological applications of the bcGAIM. **TODO**: Add.

Methods

The bcGAIM will make four methodological advancements for modeling health effects of mixtures of exposures. These are:

1. extending the cGAIM to higher dimensional problems;
2. fully exploring the parameter space to identify all plausible values for α;
3. developing priors for shape-constrained Bayesian inference on s; and
4. using case crossover models in place of the Poisson response variable.

Regarding the first methodological innovation, the cGAIM uses an iterative two-step optimization scheme. In the first step α is updated using a quadratic program, while in the second s is updated using the methodology from Pya and Wood (2015). Any linear constraint can be placed on α. For example, it can be constrained to have non-negative components that sum to one, so that it is a vector of weights. Once estimated, these weights give the relative contribution of each component to the outcome of interest. The cGAIM can also constrain the shape of s by requiring it to, for example, be monotonic or convex.

The cGAIM considers both constraints and groupwise additive index terms, while much of the existing literature only considers groupwise additive terms. For example, Hardle et al. (1993) focus on a single index and minimizes a least-squares criteria where a trimmed version of a leave-one-out Nadaraya-Watson estimator of s is used to jointly choose the bandwidth parameter and estimate α. For several indices, T. Wang et al. (2015) minimize a least-squares criteria via a two-stage estimation procedure. They derive large-sample properties of this least-squares estimator, and propose a penalized least-squares estimator for sparse high-dimensional settings. A few other papers propose alternative objective function similar to least-squares but none of these papers consider constrained estimation (Li et al. 2010; Guo et al. 2015; K. Wang and Lu Lin 2017).

One paper that considers constraints is Xia and Tong (2006), where the authors constrain s to be monotonic and α to be non-decreasing. Another is Fawzi et al. (2016), where the authors constrain α to be non-negative and sum to one but do not constrain s. In comparison, the cGAIM allows for any linear constraint on α and different shape constraints on s including monotonicity, convexity, and concavity (Masselot et al. 2020). Finally, while there are R packages, such as `scam` and `cgam` that facilitate shape-constrained inference, they do not estimate α (Pya and Wood 2015; Liao and Meyer 2019). The cGAIM considers shape constrained inference of s while estimating α under a variety of possible constraints. The bcGAIM will also allow users to specify a variety of constraints on α and s simultaneously, and report posterior distributions that communicate estimation uncertainty for both.

For the second extension, the bcGAIM It will initially be implemented in Stan, a statistical modeling language that facilitates iterative model development (Carpenter et al. 2017). For the multi-pollutant model, doing so will allow us to extend the bcGAIM to additional pollutants, additional lags for pollutants, and additional smooth functions s. We expect that α will not always be well identified, and the results will be sensitive to model assumptions and prior distributions. A major task in this component of the research will be to find reparametrizations and multivariable prior distributions that enable prior elicitation from subject-area specialists. After bcGAIM is implemented for a three-dimensional α (with covariates O_3, $PM_{2.5}$, and NO_2 at two day lags), additional time lags will be added with the resulting α being 9-12 dimensional. The computational and methodological challenges at this stage are expected to be significant, and parallelizing the algorithm on cloud platforms will be used to dramatically increase the number of candidate values of α considered.

For the third innovation, a major task is to develop Gaussian process priors, such as random walks, for shape-constrained Bayesian inference on s. In addition to having desirable statistical properties, the prior should be simple and interpretable so that it can be elicited from subject-area experts. One approach to achieving this is a nested model approach. Consider a prior $\pi(\phi)$ on s that encourages monotonicity. Viewing the bcGAIM with s monotonic as nested within the bcGAIM with s unconstrained, we want ϕ to control how strongly s is encouraged towards monotonicity. Moreover, how strongly $\pi(\phi)$ encourages monotonicity should be easy to communicate visually. This will facilitate prior elicitation and improve our ability to communicate modeling results.

Priors can have subtle negative effects on the posterior, which cam be difficult to discern in hierarchical models and/or high dimensional settings. For example, a truncated multivariate normal (tMVN) prior can induce monotonicity if placed on the coefficients of a basis expansion of s (Maatouk and Bay 2017). However, a tMVN prior subject to linear constraints places negligible mass in near-flat regions of s in high-dimensional settings. This is remedied in Zhou et al. (2020), who introduce a scale parameter on the coordinates of the tMVN, and use the half-Cauchy distribution as a shrinkage prior on these parameters. We will perform iterative development of our priors, conducting simulation studies to verify that they do not introduce undesirable side effects.

There is a vast literature on Bayesian shape-constrained inference for Gaussian processes. The distribution of a constrained Gaussian process is no longer a Gaussian process. However, the derivative of a Gaussian process is. Riihimäki and Vehtari (2010) use this to enforce monotonicity under a data augmentation scheme where derivatives are required to be positive at the virtual locations. Agrell (2019) and X. Wang and Berger (2016) find that a relatively small number of virtual observations are needed to to ensure the shape constraint holds globally with high probability. However, we have found that the effect of air pollution can substantially deviate from monotonicity (Rai et al. 2020). Also, our air pollution data sets have over 6,000 daily observations per region, and adding more virtual observations may not be computationally feasible. Therefore, data augmentation is not optimal for this project.

Another approach is to approximate the Gaussian process with a basis expansion and constrain the coefficients of that expansion, but it can be difficult to relate the priors of these coefficients to the shape of s (López-Lopera et al. 2018; Maatouk and Bay 2017). L. Lin and Dunson (2014) introduce a method that projects unconstrained Gaussian processes onto a shape-constrained space. This approach has two limitations. It cannot conduct inference on covariance parameters as those posterior distributions are affected by the projection, and the projection often produces non-smooth sample paths (which reduces interpretability) (Golchi et al. 2015). Both limitations make it undesirable for this project. Lenk and Choi (2017) assume the q^{th} derivative of s are squares of Gaussian processes, where $q = 1$ for monotonicity and $q = 2$ for convexity. They place priors on the coefficients of a Karhunen-Loeve expansion, which are not particularly interpretable. Many basis expansions have been proposed

– Zhou et al. (2020) list Bernstein polynomials, regression splines, penalized spines, cumulative distribution functions, and restricted splines – but priors on these coefficients are also not particularly interpretable. Finally, Shively et al. (2009) uses a mixture of constrained normals $N^*(0, c\sigma^2\Sigma)$ as the prior on the coefficients of a spline regression to encourage monotonicity. However, this prior can be difficult to interpret – the constrained normal N^* can be hard to explain as the dimension of Σ increases, and the scale parameter c has to be tuned by the user

Finally, consider an approach similar in spirit to our own. Bürkner and Charpentier (2020) propose a Bayesian model to estimate ordinal predictors with monotonic effects. They employ a simplex parameter ζ to model normalized differences between categories, and a scale parameter b. The prior on b expresses prior knowledge on the average differences between adjacent categories, while the prior on ζ expresses prior knowledge on individual differences between adjacent categories. The authors suggest an $N(0, \sigma)$ prior on b and a Dirichlet(α) prior on ζ. Then, σ and α would express how heavily average and individual differences between adjacent categories are penalized. Not only are ζ and b interpretable, but so are the prior parameters σ and α. The bcGAIM seeks to achieve this ease of interpretation of its parameters and priors. This will encourage adoption of the bcGAIM in other research areas, which is one of the goals of this project.

For the fourth methodological extension, we will develop non-MCMC inference methods similar in spirit to INLA (Rue et al. 2009). The Latent Gaussian approximation in INLA separates the parameter space into covariance parameters θ and linear predictors $\eta = (\beta, \theta, f)$, and considers $\pi(\eta|Y, \theta)$, $\pi(\theta|Y)$, and $\pi(\eta|Y) = \int \pi(\eta|Y, \theta)\pi(\theta|Y)d\theta$ (the last one numerically). INLA performs approximate inference on θ by estimating $\phi(\theta|Y, \phi)$ with a normal distribution with mean θ^* and variance Σ^*. If the likelihood is log-concave and Gaussian priors are used, $\pi(\theta|Y, \phi)$ is unimodal and is well-approximated by the Laplace approximation. In Margossian et al. (2020), the authors estimate $\pi(\theta|Y, \phi)$ with the Laplace approximation and $\pi(\theta|T)$ with Hamiltonian Monte Carlo. They find that this performs well for their examples, both of which have log-concave likelihoods.

Let us translate this reasoning to the bcGAIM, which has link function $g(\lambda_t) = X^t\beta + s(\alpha^T Z_t) + f_1(W_{1,t}) + ... + f_K(W_{K,t})$. Note that conditional on α, $\alpha^T Z_t$ is known. Thus, we can simplify the estimation problem by considering parameters ϕ, θ, and α and estimating $\pi(\eta|Y, \theta, \alpha)$, $\pi(\alpha|Y, \theta)$, $\pi(\theta|Y)$, and $\pi(\eta|Y) = \int \pi(\eta|Y, \theta, \alpha)\pi(\alpha|Y, \theta)\pi(\theta|Y)d\theta d\alpha$ (the last one numerically). The first and third densities in the integrand, $\pi(\eta|Y, \theta, \alpha)$ and $\pi(\theta|Y)$ are well-suited to the Laplace approximation while $\pi(\alpha|Y, \theta)$ can be estimated using HMC. Introducing two Laplace approximations will lessen the computational burden, and enable us to fit a hierarchical bcGAIM model to air pollution data. This will allow us to produce national estimates of air quality while fitting the bcGAIM to over 25 regions across Canada, each with over 6,000 daily observations, an otherwise daunting computational task. Therefore, this non-MCMC inference method will provide significant computational and ease-of-use benefits, and will expand the types of problems and number of users who can use the bcGAIM methodology. To facilitate use by other researchers, all bcGAIM software will be released in an R package.

References

References

Agrell, C. (2019). “Gaussian Processes with Linear Operator Inequality Constraints”. In: *arXiv preprint arXiv:1901.03134*.

Bobb, J. F., L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2015). “Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures”. In: *Biostatistics* 16.3, pp. 493–508.

Bopp, S., A. Richarz, A. Worth, E. Berggren, and M. Whelan (2018). “Something from Nothing: Ensuring the Safety of Chemical Mixtures”. In: *Ensuring the safety of chemical mixtures, Publications Office of the European Union, EUR* 29258.

Bürkner, P.C and E. Charpentier (2020). “Modelling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models”. In: *British Journal of Mathematical and Statistical Psychology*.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). “Stan: A Probabilistic Programming Language”. In: *Journal of Statistical Software* 76.1.

Davalos, A. D., T. J. Luben, A. H. Herring, and J. D. Sacks (2017). “Current Approaches used in Epidemiologic Studies to Examine Short-Term Multipollutant Air Pollution Exposures”. In: *Annals of Epidemiology* 27.2, pp. 145–153.

Dominici, F., R. D. Peng, C. D. Barr, and M. L. Bell (2010). “Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-Pollutant Approach”. In: *Epidemiology (Cambridge, Mass.)* 21.2, p. 187.

Fawzi, A., J.B. Fiot, B. Chen, M. Sinn, and P. Frossard (2016). “Structured Dimensionality Reduction for Additive Model Regression”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.6, pp. 1589–1601.

Feng, C., J. Li, W. Sun, Y. Zhang, and Q. Wang (2016). “Impact of Ambient Fine Particulate Matter (PM 2.5) Exposure on the Risk of Influenza-Like-Illness: A Time-Series Analysis in Beijing, China”. In: *Environmental Health* 15.1, p. 17.

Franklin, M. and J. Schwartz (2008). “The Impact of Secondary Particles on the Association Between Ambient Ozone and Mortality”. In: *Environmental Health Perspectives* 116.4, pp. 453–458.

Golchi, S., D. R. Bingham, H. Chipman, and D. A. Campbell (2015). “Monotone Emulation of Computer Experiments”. In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1, pp. 370–392.

Guo, Zifang, Lexin Li, Wenbin Lu, and Bing Li (2015). “Groupwise Dimension Reduction via Envelope Method”. In: *Journal of the American Statistical Association* 110.512, pp. 1515–1527.

Hardle, W., P. Hall, and H. Ichimura (1993). “Optimal Smoothing in Single-Index Models”. In: *The Annals of Statistics*, pp. 157–178.

Lenk, P. J. and T. Choi (2017). “Bayesian Analysis of Shape-restricted Functions using Gaussian Process Priors”. In: *Statistica Sinica*, pp. 43–69.

Li, Lexin, Bing Li, and Li-Xing Zhu (2010). “Groupwise Dimension Reduction”. In: *Journal of the American Statistical Association* 105.491, pp. 1188–1201.

Liao, X. and M. C. Meyer (2019). “cgam: An R Package for the Constrained Generalized Additive Model”. In: *Journal of Statistical Software* 85.9, pp. 1–24.

Lin, L. and D. B. Dunson (2014). “Bayesian Monotone Regression using Gaussian Process Projection”. In: *Biometrika* 101.2, pp. 303–317.

Liu, C., R. Chen, F. Sera, A. M. Vicedo-Cabrera, Y. Guo, S. Tong, M. SZS Coelho, P. HN Saldiva, E. Lavigne, and P. Matus (2019). “Ambient Particulate Air Pollution and Daily Mortality in 652 Cities”. In: *NEJM* 381.8, pp. 705–715.

López-Lopera, Andrés F., F. Bachoc, N. Durrande, and O. Roustant (2018). “Finite-dimensional Gaussian Approximation with Linear Inequality Constraints”. In: *SIAM/ASA Journal on Uncertainty Quantification* 6.3, pp. 1224–1255.

Maatouk, H. and X. Bay (2017). “Gaussian Process Emulators for Computer Experiments with Inequality Constraints”. In: *Mathematical Geosciences* 49.5, pp. 557–582.

Margossian, C. C., A. Vehtari, D. Simpson, and R. Agrawal (2020). “Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation”. In: *arXiv preprint arXiv:2004. 12550.*

Masselot, P., F. Chebana, C. Campagna, E. Lavigne, T. B.M.J. Ouarda, and P. Gosselin (2020). “Constrained Groupwise Additive Index Models”. In: *Submitted.*

Pya, N. and S. N. Wood (2015). “Shape Constrained Additive Models”. In: *Statistics and Computing* 25.3, pp. 543–559.

Rai, K., P. E. Brown, A. Morariu, and H. H. Shin (2020). “Trend Detection Paper”. Submitted.

Riihimäki, J. and A. Vehtari (2010). “Gaussian Processes with Monotonicity Information”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 645–652.

Rue, H., S. Martino, and N. Chopin (2009). “Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations”. In: *JRSS: Series B* 71.2, pp. 319–392.

Shively, T. S., T. W. Sager, and S. G. Walker (2009). “A Bayesian Approach to Non-Parametric Monotone Function Estimation”. In: *JRSS: Series B* 71.1, pp. 159–175.

Stieb, D. M., R. T. Burnett, M. Smith-Doiron, O. Brion, H. H. Shin, and V. Economou (2008). “A New Multipollutant, No-threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses”. In: *Journal of the Air & Waste Management Association* 58.3, pp. 435–450.

Stringer, Alex, Patrick Brown, and Jamie Stafford (2020). “Approximate Bayesian inference for case-crossover models”. In: *Biometrics.*

Wang, Kangning and Lu Lin (2017). “Robust and Efficient Direction Identification for Groupwise Additive Multiple-Index Models and its Applications”. In: *Test* 26.1, pp. 22–45.

Wang, T., J. Zhang, H. Liang, and L. Zhu (2015). “Estimation of a Groupwise Additive Multiple-Index Model and its Applications”. In: *Statistica Sinica*, pp. 551–566.

Wang, X. and J. O. Berger (2016). “Estimating Shape Constrained Functions using Gaussian Processes”. In: *SIAM/ASA Journal on Uncertainty Quantification* 4.1, pp. 1–25.

Wei, Y., Y. Wang, Q. Di, C. Choirat, Y. Wang, P. Koutrakis, A. Zanobetti, F. Dominici, and J. D Schwartz (2019). “Short Term Exposure to Fine Particulate Matter and Hospital Admission Risks and Costs in the Medicare Population: Time Stratified, Case Crossover Study”. In: *BMJ* 367.

Wu, X., R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici (2020). “Exposure to Air Pollution and COVID-19 Mortality in the United States”. In: *medRxiv.*

Xia, Y. and H. Tong (2006). “Cumulative Effects of Air Pollution on Public Health”. In: *Statistics in Medicine* 25.20, pp. 3548–3559.

Zhou, S., P. Ray, D. Pati, and A. Bhattacharya (2020). “On Truncated Multivariate Normal Priors in Constrained Parameter Spaces”. In: *arXiv preprint arXiv:2001.09391*.