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 (Bopp *et al.*, 2018; Dominici *et al.*, 2010; Stieb *et al.*, 2008). 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) is introduced by Masselot *et al.* (2020), who employs frequentist inference methods that use sequential quadratic programming. For a response distribution D, parameter $\theta = (\theta_1, \ldots, \theta_d)$, link function g, and constraints as below, the cGAIM is,

$$\begin{aligned} &Y_t|\lambda_t \overset{i.i.d.}{\sim} D(\lambda_t, \tau), \\ &g(\lambda_t) = X^T\beta + s(\alpha^T Z_t) + f_1(W_{1t}) + \ldots + f_K(W_{Kt}). \end{aligned}$$

Here Y_t is the outcome of interest, β are the fixed effects of the potentially time-varying linear covariates X, and $f_1, \ldots, f_K$ are smoothing functions that account for potential confounding variables $W_{k,t}$ $(k = 1, \ldots, K)$. The distinguishing feature of cGAIM is s, a smooth function fit to a linear combination of covariates Z_t. α is a vector of weights on the entries of Z_t, and gives the relative contribution the components of Z_t. The smooth function s is specified as a member of a flexible parametric family of functions, such as cubic splines or random walks. Estimating α is the main statistical challenge with the cGAIM. We will develop a Bayesian methodology for inference with the cGAIM – the bcGAIM.

The bcGAIM provides two main statistical benefits compared to the cGAIM. The first is quantification of the uncertainty for α; the cGAIM does not provide confidence intervals for α. Consider the multi-pollutant model, where Z_t is a matrix of daily pollutant levels. Pollutants tend to be positively correlated, and it is possible that the posterior of α is multi-modal. If this holds, multiple different values of α will be equally predictive of health outcomes. It is also possible that α is not multi-modal and the importance of each pollutant can be estimated with some certainty. Either result would be a significant contribution research on the health effects of multiple pollutant. The second advantage is that the bcGAIM will be able to accommodate high-dimensional α, allowing the multi-pollutant model to be extended to Z_t that contain pollution levels at different time lags. The joint posterior distribution of this high-dimensional α will provide information on how quickly poor air quality affects different mortality and morbidity outcomes. For example, it could show if different combinations of lagged pollutants are equally predictive for the outcome of interest, or if they are not.

We will also develop methods for fitting a case crossover models using the bcGAIM, which offers advantages over the log-linear Poisson model. The case crossover has seen increased attention in the air pollution literature (Stringer *et al.*, 2020; Wei *et al.*, 2019). It can be viewed as a proportional hazards survival model, where each individual is in a separate strata and ‘control days’ are chosen to have the same baseline hazard as the event day. Additional details are given in the Methods section. Finally, note that the bcGAIM has many applications beyond building an air quality index, some of which are detailed below.

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, 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 (Guo *et al.*, 2015; Li *et al.*, 2010; Wang and 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 (Liao and Meyer, 2019; Pya and Wood, 2015). 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 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). 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$^{\text{th}}$ 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

Agrell, C. (2019) Gaussian processes with linear operator inequality constraints. *arXiv preprint arXiv:1901.03134*.

Bobb, J. F., Valeri, L., et al. (2015) Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, **16**, 493–508.

Bopp, S., Richarz, A., et al. (2018) Something from nothing: Ensuring the safety of chemical mixtures. *Ensuring the safety of chemical mixtures, Publications Office of the European Union, EUR*, **29258**.

Bürkner, P. and Charpentier, E. (2020) Modelling monotonic effects of ordinal predictors in bayesian regression models. *British Journal of Mathematical and Statistical Psychology.* Wiley Online Library.

Carpenter, B., Gelman, A., et al. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Davalos, A. D., Luben, T. J., et al. (2017) Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Annals of Epidemiology*, **27**, 145–153.

Dominici, F., Peng, R. D., et al. (2010) Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology (Cambridge, Mass.)*, **21**, 187.

Fawzi, A., Fiot, J., et al. (2016) Structured dimensionality reduction for additive model regression. *IEEE Transactions on Knowledge and Data Engineering*, **28**, 1589–1601.

Feng, C., Li, J., et al. (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. *Environmental Health*, **15**, 17.

Franklin, M. and Schwartz, J. (2008) The impact of secondary particles on the association between ambient ozone and mortality. *Environmental Health Perspectives*, **116**, 453–458.

Golchi, S., Bingham, D. R., et al. (2015) Monotone emulation of computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, **3**, 370–392. SIAM.

Guo, Z., Li, L., et al. (2015) Groupwise dimension reduction via envelope method. *Journal of the American Statistical Association*, **110**, 1515–1527. Taylor & Francis.

Hardle, W., Hall, P., et al. (1993) Optimal smoothing in single-index models. *The Annals of Statistics*,, 157–178. JSTOR.

Lenk, P. J. and Choi, T. (2017) Bayesian analysis of shape-restricted functions using gaussian process priors. *Statistica Sinica*,, 43–69. JSTOR.

Li, L., Li, B., et al. (2010) Groupwise dimension reduction. *Journal of the American Statistical Association*, **105**, 1188–1201. Taylor & Francis.

Liao, X. and Meyer, M. C. (2019) Cgam: An r package for the constrained generalized additive model. *Journal of Statistical Software*, **85**, 1–24.

Lin, L. and Dunson, D. B. (2014) Bayesian monotone regression using gaussian process projection. *Biometrika*, **101**, 303–317. Oxford University Press.

Liu, C., Chen, R., et al. (2019) Ambient particulate air pollution and daily mortality in 652 cities. *NEJM*, **381**, 705–715.

López-Lopera, A. F., Bachoc, F., et al. (2018) Finite-dimensional gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, **6**, 1224–1255. SIAM.

Maatouk, H. and Bay, X. (2017) Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, **49**, 557–582.

Margossian, C. C., Vehtari, A., et al. (2020) Hamiltonian monte carlo using an adjoint-differentiated laplace approximation. *arXiv preprint arXiv:2004. 12550.*

Masselot, P., Chebana, F., et al. (2020) Constrained groupwise additive index models. *Submitted.*

Pya, N. and Wood, S. N. (2015) Shape constrained additive models. *Statistics and Computing*, **25**, 543–559.

Rai, K., Brown, P. E., et al. (2020) Trend detection paper. *Submitted.*

Riihimäki, J. and Vehtari, A. (2010) Gaussian processes with monotonicity information. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 645–652.

Rue, H., Martino, S., et al. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *JRSS: Series B*, **71**, 319–392.

Shively, T. S., Sager, T. W., et al. (2009) A bayesian approach to non-parametric monotone function estimation. *JRSS: Series B*, **71**, 159–175. Wiley Online Library.

Stieb, D. M., Burnett, R. T., et al. (2008) A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. *Journal of the Air & Waste Management Association*, **58**, 435–450. Taylor & Francis.

Stringer, A., Brown, P., et al. (2020) Approximate bayesian inference for case-crossover models. *Biometrics*. Wiley Online Library.

Wang, K. and Lin, L. (2017) Robust and efficient direction identification for groupwise additive multiple-index models and its applications. *Test*, **26**, 22–45. Springer.

Wang, T., Zhang, J., et al. (2015) Estimation of a groupwise additive multiple-index model and its applications. *Statistica Sinica*,, 551–566. JSTOR.

Wang, X. and Berger, J. O. (2016) Estimating shape constrained functions using gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, **4**, 1–25. SIAM.

Wei, Y., Wang, Y., et al. (2019) Short term exposure to fine particulate matter and hospital admission risks and costs in the medicare population: Time stratified, case crossover study. *BMJ*, **367**.

Wu, X., Nethery, R. C., et al. (2020) Exposure to air pollution and covid-19 mortality in the United States. *medRxiv*.

Xia, Y. and Tong, H. (2006) Cumulative effects of air pollution on public health. *Statistics in Medicine*, **25**, 3548–3559.

Zhou, S., Ray, P., et al. (2020) On truncated multivariate normal priors in constrained parameter spaces. *arXiv preprint arXiv:2001.09391.*