Biometrika (XXXX), **XXX**, X, *pp*. 1–3
Printed in Great Britain

Scalable Gaussian Process Regression via Median Posterior Inference for Estimating Multi-Pollutant Mixture Health Effects

BY A.SONABEND
Google Inc, 1600 Amphitheatre Parkway, Mountain View, U.S.A.
asonabend@google.com

J.ZHANG
Department of Statistics, University of California, Davis, One Shields Avenue, Davis, U.S.A.
jiszhang@ucdavis.edu

J.SCHWARTZ
Department of Environmental Health, Harvard University, 677 Huntington Ave, Boston, U.S.A.
jschwrtz@hsph.harvard.edu

B.COULL, AND J.LU
Department of Biostatistics, Harvard University, 677 Huntington Ave, Boston, U.S.A.
bcoull@hsph.harvard.edu junweilu@hsph.harvard.edu

SUMMARY

Humans are never exposed to a single chemical, but rather a wide array of environmental exposures. As a result, a focus of research in the environmental health sciences has been on the quantification of the effects of exposure to a mixture of pollutants. Such research on environmental mixtures yields evidence of effects under more realistic exposure scenarios, leading to regulatory policies that can be more protective of public health. As is now well-documented, quantifying the health effects of environmental mixtures involves addressing several statistical challenges, including complex correlation structures among pollutant levels and potentially complex multivariate expsoure-response relationships between exposure and health. A popular approach to simultaneously address these challenges is a Bayesian semi-parametric Gaussian process regression framework Bobb et al. (2015); Coull et al. (2015). This framework models the exposure-response function with a Gaussian process and performs feature selection to effectively reduce the dimension of the potentially high-dimensional exposure, while accounting for confounders via a linear model. Because the framework was originally motivated by the need to estimate effects in small to moderately sized cohort studies, the algorithms for model fitting do not scale up well in big data settings, such as those encountered when interest focuses on electronic health records or other administrative data. While there are some ad hoc solutions to scale the framework up to big data settings, there is no theoretical guarantee on the inference of the results. In this work, we propose a divide-and-conquer approach in which we split sam-

ples, compute the posterior distribution, and then combine them using the generalized median. Importantly, we provide theoretical guarantees for the convergence of the proposed posterior computation to the posteriors from the original Gaussian process model derived from the full sample. We apply the proposed approach to estimate associations between a mixture of ambient air pollutants and approximately 650,000 birthweights recorded in Massachusetts, USA during 2001-2012. Results suggest that elemental and organic carbon, markers of traffic pollution, as well as total $PM_{2.5}$ mass are negatively associated with birthweight in Massachusetts during this period, while ozone levels and a marker of greenness (i.e. nature of an area) are positively associated with birthweights.

Some key words: Semi-parametric regression, Scalable Bayesian Inference, Median Posterior, Multi-Pollutant Mixtures.

1. INTRODUCTION

Ambient air pollution consists of a heterogeneous mixture of multiple chemical components, with these components being generated by different pollution sources. Therefore, quantification of the health effects of this mixture can yield important evidence on the source-specific health effects of air pollution, which has the potential to provide evidence to support targeted regulations for ambient pollution levels.

As is now well-documented, there are several statistical challenges involved in estimating the health effects of multi-pollutant mixture. First, the relationship between health outcomes and multiple pollutants can be complex, potentially involving non-linear and non-additive effects. Second, pollutant levels can be highly correlated, but only some may impact health. Therefore, models inducing sparsity are often advantageous. Feature engineering, such as basis expansions to allow interaction terms, can lead to high dimensional inference. Alternatively, parametric models can be used, however they require the analyst to impose a functional form, which can yield biased estimates in the likely case that the model is miss-specified.

Finally, large data sets containing information on pollution exposure and population characteristics have become increasingly available. These datasets allow for the estimation of small but significant effects. However, many models used in this context do not scale well with sample size and feature dimension. Efficient methods are needed that can take advantage of massive data and yield results which can be easily interpreted, and which can be computed in a relatively short time.

Several methods address the issues discussed above (Billionnet et al., 2012). A common approach to modelling the complex relationship between pollutants and outcomes is to use flexible models such as random forests which have been shown to be consistent (Scornet et al., 2015), or universal approximators, such as neural networks (Schmidhuber, 2015). These are useful but yield results which are hard to interpret: one cannot report the directionality or magnitude of the feature effect on the outcome. In this context, our interest lies in both prediction as well as interpretation. Another possible way to incorporate flexible multi-pollutant modelling is by clustering pollution-exposure levels and including clusters as covariates in parametric models. This approach essentially stratifies exposure levels which results in important loss of information. It ultimately forces the analyst to adapt the question of interest into one that can be solved by available tools, instead of tackling the relevant questions. A common approach to address the high-dimensionality of multi-pollutants effects is to posit a generalized additive model. This allows one to estimate the association between a health outcome and a single pollutant, which can be repeated for every exposure of interest (Stieb et al., 2012). Flexible modelling such as

quantile regression can be employed to deal with outliers and account for possible differences in associations across the health outcome (Fong et al., 2019b). However, the clear downside is that incorporating multi-pollutant mixtures quickly makes this approach computationally infeasible. Alternatively, generalized linear models can be used to evaluate the associations of interest, with the downside of imposing a functional form (Gaskins et al., 2019). To enforce sparsity on the feature space, variable selection methods such as least absolute shrinkage and selection operator (LASSO) penalty can be used (Tibshirani, 1996), however to use such methods one must specify a parametric model which brings us back to the likely misspecification scenario, in which estimated associations and causal effects may be biased.

A popular approach to simultaneously addressing these issues on small-scale data is the use of semi-parametric Gaussian process model, often referred to as Bayesian kernel machine regression (BKMR)(Bobb et al., 2015; Coull et al., 2015). The pollutants-health outcome relationship is modelled through a Gaussian process, which allows for a flexible functional relationship between the pollutants and the outcome of interest. The model allows for feature selection among the pollutants to discard those with no estimable health effect and to account for high correlation among those with and without a health effect. This framework allows the incorporation of linear effects of baseline covariates, yielding an interpretable model. Even though this framework is frequently employed in the multi-pollutant context, large datasets make it prohibitively slow as it involves Bayesian posterior calculation. There are some ad hoc solutions, which have been shown to work well in practice in certain simulated settings (Bobb et al., 2015). However, there is no theoretical guarantee regarding the inference of the results. Here, we propose a divide-and-conquer approach in which we split samples, compute the posterior distribution, and then combine the smaller samples using the generalized median. This method allows capturing small effects from large datasets in little time. We provide theoretical guarantees for the convergence of the Gaussian process, flexible to different function spaces.

2. METHOD

2·1. *Semi-parametric Regression*

Suppose we observe a sample of n independent, identically distributed (i.i.d.) random vectors $\mathbb{S}_n = \{D_i\}_{i=1}^n$, where $D_i = (Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \sim P_0$ with $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^p$ a vector of possible confounders, and $\boldsymbol{Z}_i \in \mathcal{Z} \subset \mathbb{R}^q$ a vector of exposure to pollution constituents. We will assume health outcome Y has a linear relationship with confounders $\boldsymbol{X}$ and a non-parametric relationship with exposure to pollution $\boldsymbol{Z}$. In particular, for D_i we assume the following semi-parametric relationship:

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^0 + h_0(\boldsymbol{Z}_i) + e_i, \tag{1}$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$, and $h_0 : \mathcal{Z} \mapsto \mathbb{R}$ is an unknown function which we allow to incorporate non-linearity and interaction among the pollutants. We require h_0 to be in an α-Holder space, or to be infinitely differentiable. We formalize this in Section 3.

2·2. *Prior Specification*

To perform inference on h_0, we will use a re-scaled Gaussian process prior (Williams & Rasmussen, 2019). In particular, we will use a squared exponential process equipped with an inverse Gamma bandwidth. That is, we will use prior

$$h_0(\boldsymbol{Z}) \sim \mathcal{N}(0, \boldsymbol{K}),$$

where $\text{Cov}[\boldsymbol{Z}, \boldsymbol{Z}'] = K(\boldsymbol{Z}, \boldsymbol{Z}'; \rho) = \exp\left\{-\frac{1}{\rho^{2q}}\|\boldsymbol{Z} - \boldsymbol{Z}'\|_2^2\right\}$, and ρ is a Gamma distributed random variable. We choose this kernel as it is flexible enough to approximate smooth functions, more so when the bandwidth parameter ρ can be estimated from the data.

Alternatively, we will augment the Gaussian kernel to allow for sparse solutions on the number of pollutants which contribute to the outcome (Bobb et al., 2015). Let the augmented co-variance function be $\text{Cov}[\boldsymbol{Z}, \boldsymbol{Z}'] = K(\boldsymbol{Z}, \boldsymbol{Z}'; \boldsymbol{r}) = \exp\{-\sum_{j=1}^{q} r_j (Z_j - Z_j')^2\}$. To select pollutants we assume a "slab-and-spike" prior on the selection variables $r_j \sim g_{r|\delta}$, with

$$g_{r|\delta}(r, \delta) = \delta f_1(r) + (1 - \delta) f_0,\ \delta \sim \text{Bernoulli}(\pi),$$

where f_1 has support on $\mathbb{R}^+$ and f_0 is the point mass at 0. The random variables δ_j can then be read as the indicators for whether pollutant Z_j is relevant to the health outcome. Their relevance is measured by the mean $\mathbb{E}[\delta_j] = \pi_j$. Finally, for simplicity we will assume an improper prior on the linear component: $\boldsymbol{\beta} \sim 1$. This linear component will capture the effects of confounders. We further use a Gamma prior distribution for the error term variance σ^2.

2·3. *Estimation*

Let $\boldsymbol{h} \equiv (h_0(\boldsymbol{Z}_1), \ldots, h_0(\boldsymbol{Z}_n))^\top$, Liu et al. (2007) have shown that model (1) can be expressed as

$$Y_i \sim \mathcal{N}(h_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^T \boldsymbol{\beta}^0, \sigma^2),\ \boldsymbol{h} \sim \mathcal{N}(0, \tau \boldsymbol{K}).$$

This will allow us to simplify our inference procedure and split the problem into tractable posterior estimation (Bobb et al., 2015) for each component of interest. In particular, we can use Gibbs steps to sample the conditionals for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{h}$ analytically, also letting $\lambda = \frac{\tau}{\sigma^2}$, we use a Metropolis-Hastings step, the full set of posteriors is given in equation (2).

$$\begin{aligned}
\boldsymbol{\beta} \mid \sigma^2, \lambda, \boldsymbol{r}, \boldsymbol{Y} &\sim N\left(\boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{X}^\top \boldsymbol{V}_{\lambda,\boldsymbol{Z},\boldsymbol{r}}^{-1} \boldsymbol{Y}, \sigma^2 \boldsymbol{V}_{\boldsymbol{\beta}}\right), \\
\sigma^2 \mid \boldsymbol{\beta}, \lambda, \boldsymbol{r}, \boldsymbol{Y} &\sim \text{Gamma}\left(\alpha_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} WSS_{\boldsymbol{\beta},\lambda,\boldsymbol{r}}\right), \\
\boldsymbol{h} \mid \boldsymbol{\beta}, \sigma^2, \lambda, \boldsymbol{r}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z} &\sim N\left(\lambda \boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{r}} \boldsymbol{V}_{\lambda,\boldsymbol{Z},\boldsymbol{r}}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \sigma^2 \lambda \boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{r}} \boldsymbol{V}_{\lambda,\boldsymbol{Z},\boldsymbol{r}}^{-1}\right), \\
f(\lambda \mid \boldsymbol{\beta}, \boldsymbol{r}, \delta, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) &\propto \left|\boldsymbol{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} WSS_{\boldsymbol{\beta},\lambda,\boldsymbol{r}}\right\} \text{Gamma}(\lambda \mid a_\lambda, b_\lambda).
\end{aligned} \tag{2}$$

where $\boldsymbol{V}_{\lambda,\boldsymbol{Z},\boldsymbol{r}} = \mathbf{I}_n + \lambda \boldsymbol{K}_{\boldsymbol{Z},\boldsymbol{r}}$, $\boldsymbol{V}_{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1} \boldsymbol{X})^{-1}$, $WSS_{\boldsymbol{\beta},\lambda,\mathbf{r}} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}_{\lambda,\mathbf{Z},\mathbf{r}}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$.

To perform inference for functions of interest in (2), we will use Markov Chain Monte Carlo (MCMC) techniques. Furthermore, even though function $\boldsymbol{h}$ has a closed form posterior, large samples will require large matrix inversions. As sample size increases, posterior sampling becomes increasingly challenging. This is particularly true for sampling $\boldsymbol{h}$, as the posterior Gaussian process is n-dimensional.

With an infinite number of samples, MCMC is known to converge to the true posterior. However, in practice the number of samples for the burn-in state required from the true posterior significantly increases with dimension.

This problem is substantially worse when sampling from the Gaussian process posterior as the dimension is the sample size. The computational cost for each iteration is $\mathcal{O}(n^3)$, since to sample from the posterior of $\boldsymbol{h}$ we need to compute an inverse of an $n \times n$ kernel matrix indicated in (2). This renders the method prohibitively slow for real applications on large data sets. On the

other hand given the epidemiological context, these are precisely the data sets needed as they can actually shed light on the small effects of pollutants on health outcomes. This predicament motivates the development of a fast version of inference for (2), and particularly for $\boldsymbol{h}$.

2·4. *Fast Inference on Posteriors via Sub-sampling*

In order to make posterior sampling computationally feasible, we propose a sample splitting technique which is guaranteed to satisfy the needed theoretical properties. Our approach consists of computing multiple *noisy* versions of the posteriors we are interested in, and using the median of these as a proxy for the full data posterior.

First, we randomly split the entire data set into several disjoint subsets with roughly equal sample size. Let $\mathbb{S}_{k=1}^{K}$ denote a random partition of $\mathbb{S}_n$ into K disjoint subsets of size $n_k = n/K$ with index sets $\{\mathcal{I}_k\}_{k=1}^{K}$. Then for each subset $\mathbb{S}_k$, we run a modified version of the estimation approach described in Section 2·3 using sub-sampling sketching matrices $S_k \in \mathbb{R}^{n \times n_k}$. This will yield K posterior distributions for each parameter and function in (2).

We proceed as follows: Let $\{p_1, \ldots, p_n\}$ be drawn uniformly from the n-dimensional identity matrix I_n with indexes of columns as in $\mathcal{I}_k$, and define sketching matrix S_k as a matrix with n_k columns of the form $S_{k,i} = \sqrt{K} \cdot p_i$. Using S_k we denote by $\widetilde{\boldsymbol{V}}_k$ and $\widetilde{\boldsymbol{A}}_k$, any vector and matrix transformation respectively as $\widetilde{\boldsymbol{V}}_k = S_k^\top \boldsymbol{V}$, $\widetilde{\boldsymbol{A}}_k = S_k^\top \boldsymbol{A} S_k$. We can then redefine model (1) for sample $\mathbb{S}_k$ as

$$\widetilde{\boldsymbol{Y}}_k \sim N(\widetilde{\boldsymbol{h}}_k + \widetilde{\boldsymbol{X}}^{(k)}\boldsymbol{\beta}, \sigma^2 S_k S_k^\top), \;\; \widetilde{\boldsymbol{h}}_k \sim N(\boldsymbol{0}, \tau\widetilde{\boldsymbol{K}}^{(k)}). \tag{3}$$

We then implement our inference from Section 2·3 to the above by using $\widetilde{\boldsymbol{V}}^{(k)}_{\lambda,\boldsymbol{Z},\boldsymbol{r}}$, $\widetilde{\boldsymbol{V}}^{(k)}_{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}_k^\top (\widetilde{\boldsymbol{V}}^{(k)}_{\lambda,\boldsymbol{Z},\boldsymbol{r}})^{-1} \widetilde{\boldsymbol{X}}_k)^{-1}$, $\widetilde{WSS}^{(k)}_{\boldsymbol{\beta},\lambda,\boldsymbol{r}} = (\widetilde{\boldsymbol{Y}}_k - \widetilde{\boldsymbol{X}}_k\beta)^\top (\widetilde{\boldsymbol{V}}^{(k)}_{\lambda,\boldsymbol{Z},\boldsymbol{r}})^{-1} (\widetilde{\boldsymbol{Y}}_k - \widetilde{\boldsymbol{X}}_k\boldsymbol{\beta})$ in (2) and sample from each of the K posteriors.

For any i.i.d. random vectors $D_1, \ldots, D_n \sim P_0$, let $P_0 \equiv P_{\theta_0}$ be indexed by $\theta_0 \in \Theta$. Bayesian inference usually consists of specifying a prior distribution Π for θ_0, and using sample $\mathbb{S}_n$ to compute a posterior distribution for θ_0 defined as

$$\Pi_n(\theta \mid \mathbb{S}_n) \equiv \frac{\prod_{i=1}^n p_\theta(D_i)\Pi(\theta)}{\int_\Theta \prod_{i=1}^n p_\theta(D_i)\Pi(\theta)}.$$

Note that this definition is general enough that θ can be any parameter in (2) as well as function h_0, in which case prior $\Pi(\theta)$ is a Gaussian process. Thus, we compute $\Pi_k(\theta \mid \mathcal{I}_k)$ $k = 1, \ldots, K$ for each split and each parameter of interest. Naturally, posterior $\Pi_k(\theta \mid \mathcal{I}_k)$ will be a noisy approximation of $\Pi_n(\theta \mid \mathbb{S}_n)$ we combine each Π_k using the geometric median. To formalize this, we first define the geometric median, which is a multi-dimension generalization of the univariate median (Minsker, 2015).

To construct the geometric median posterior, we will use K posteriors $\Pi_k(\theta \mid \mathcal{I}_k) \in \mathcal{F}$ where $\mathcal{F} = \{f : \mathcal{A} \mapsto \mathbb{R}, f \in \mathbb{H}, \|f\|_{\mathbb{H}} = \sqrt{\langle f, f\rangle} \leq 1\}$, then define metric $\|\cdot\|_{\mathcal{F}}$ as

$$\|G_1 - G_2\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{A}} f(x) d(G_1 - G_2)(x) \right|,$$

and define the geometric median $\overline{\Pi}_n$ as

$$\overline{\Pi}_n = \underset{\Pi \in \mathcal{F}}{\operatorname{argmin}} \sum_{k=1}^{K} \|\Pi - \Pi_k\|_{\mathcal{F}} \tag{4}$$

As the univariate version, the geometric median $\overline{\Pi}_n$ is robust to outlier observations, meaning that our estimation procedure will preserve consistency on a finite sample within a reasonable number of outliers present. Next, the convergence rate will be in terms of n_k, however this rate improves geometrically with K with respect to the rate at which the K estimators are weakly concentrated around the true parameter (Minsker et al., 2017).

One way to estimate (4) numerically, is through the barycenter of subset posterior distributions in the Wasserstein space of probability measures (Srivastava et al., 2018). We first recall the definition of Wasserstein space. Let (Θ, ρ) be a metric space which is completely separable, and $P(\Theta)$ denote the space of all probability measures on Θ with finite first and second moments. Then the Wasserstein space of order 2 is defined as:

$$P_2(\Theta) := \{\mu \in P(\Theta) : \int_\Theta \rho^2(\theta_0, \theta)\mu(d\theta) < \infty\}, \tag{5}$$

where $\theta_0 \in \Theta$ and the space does not depend on θ_0. In our method, this Wasserstein space of order 2 is defined using Euclidean metric ρ. Now let $\mu, \nu \in P_2(\Theta)$ and $\Pi(\mu, \nu)$ be the set of all probability measures on $\Theta \times \Theta$ with marginals μ, ν,the Wasserstein distance of order 2 between μ and ν is defined as:

$$W_2(\mu, \nu) = (\inf_{\pi \in \pi(\mu,\nu)} \int_{\Theta\times\Theta} \rho^2(x, y)d\pi(x, y))^{\frac{1}{2}}. \tag{6}$$

Suppose probability measures $\Pi_1, \ldots, \Pi_K$ are in $P_2(\Theta)$, and we use Euclidean metric as the metric ρ, then the barycenter of a collection $\Pi_1, \ldots, \Pi_K$ becomes:

$$\overline{\Pi}_n = \operatorname*{argmin}_{\Pi} \sum_{k=1}^{K} W^2(\Pi, \Pi_k), \tag{7}$$

In fact, this barycenter $\overline{\Pi}_n$ is the geometric median in the Wasserstein space. Thus, we can consider finding the posterior geometric median function of the subset posterior probability measure in the Wasserstein space.

As the median function $\overline{\Pi}_n$ is generally analytically intractable, an achievable solution is to estimate $\overline{\Pi}_n$ from samples of subsets posteriors. We can approximate the median function by assuming that subset posterior distributions are empirical measures and their atoms can be simulated from the subset posteriors by a sampler (Srivastava et al., 2015). Let $\{\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{kN}\}$ be N samples of parameters $\boldsymbol{\theta}$ obtained from subset posterior distribution $\Pi_n^{(k)}$. In our method, samples can be directly generated from subsets posteriors by using an MCMC sampler , then we can approximate the $\Pi_n^{(k)}$ by the empirical measure corresponding with $\{\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{kN}\}$, which is defined as:

$$\widehat{\Pi}_n^{(k)}(\cdot) = \sum_{i=1}^{N} \frac{1}{N}\delta_{\boldsymbol{\theta}_{ki}}(\cdot),\ (k = 1, \ldots, K) \tag{8}$$

where $\delta_{\boldsymbol{\theta}_{ki}}(\cdot)$ is the Dirac measure concentrated at $\boldsymbol{\theta}_{ki}$. In order to approximate the subset posterior accurately, we need to make N large enough. Then the empirical probability measure of median function $\widehat{\Pi}_{n,\mathcal{I}}$ can be approximated by estimating the geometric median of the empirical probability measure of subset posteriors. Using samples from subsets posteriors, the empirical

probability measure of the median function is defined as:

$$\widehat{\overline{\Pi}}_{n,\mathcal{I}}(\cdot) = \sum_{k=1}^{K}\sum_{i=1}^{N} a_{ki}\delta_{\boldsymbol{\theta}_{ki}}(\cdot),\ 0 \le a_{ki} \le 1,\ \sum_{k=1}^{K}\sum_{i=1}^{N} a_{ki} = 1 \tag{9}$$

where a_{ki} are unknown weights of atoms. Here the problem of combing subset posteriors to give a problem measure is switched to estimating a_{ki} in (9) for all the atoms across all subset posteriors. Fortunately, a_{ki} can be estimated by solving the linear program in (7) with posterior distributions restricted to atom forms in (8) and (9). There are several different algorithms, such as the entropy-smoothed sub-gradient Sinkhorn algorithm developed by Cuturi & Doucet (2014) and non-smooth optimization algorithm developed by Carlier et al. (2015). We use an efficient algorithm developed by Srivastava et al. (2018), which is summarized as Algorithm 1.

Algorithm 1. Fast Posterior Inference Via Sub-sampling.

REQUIRE Observed sample $\mathbb{S}_n = \{D_i\}_{i=1}^n$, subset number K, parameter sample size N
Randomly partition $\mathbb{S}_n$ into K subsets $\mathbb{S}_{k=1}^K$ with size n_k
For $\mathbb{S}_k \in \mathbb{S}_{k=1}^K$ do
1. Get index set $\mathcal{I}_k$ for $\mathbb{S}_k$
2. Get sub-sampling sketching matrix S_k
3. Run MCMC sampling on modified model described as (2) and (3) to generating parameter samples $\{\boldsymbol{\theta}_{k1}, \ldots, \boldsymbol{\theta}_{kN}\}$
Solve the linear program in (7) with (8)-(9)
RETURN empirical approximation posterior median function $\widehat{\overline{\Pi}}_{n,\mathcal{I}}$

Algorithm 1 provides a sample splitting approach to decrease computational complexity for posterior inference. Sampling $\boldsymbol{\beta}$, σ^2, λ and $\boldsymbol{h}$ requires computing K, and inverting $V_{\lambda,\mathbf{Z},\mathbf{r}}$, this translates into $O(n^2q)$, $O(n^3)$ operations respectively per iteration. For λ we also need to compute $|V_{\lambda,\mathbf{Z},\mathbf{r}}|$ which is $O(n^3)$. Bobb et al. (2015) recommend using at least 10^4 iterations, which translates into $O(10^4 n^3)$ operations (assuming $q << n$). There is a clear trade-off between a large number of splits K which decreases computational complexity, and using the whole sample $K = 1$ which yields better inference. For example, choosing $K = n^{1/2}$ yields a computational complexity of $O(10^4 n^{3/2})$ for Algorithm 1. Next in Section 3 we discuss the posterior convergence rate and its dependence on the number of splits in detail.

3. Theoretical Results

In this section we go over the assumptions needed for our theoretical results, state our main theorem and discuss its implications. Our results focus on estimation of h_0 as this is the main function of interest and our main contribution. We first assume that the confounder and pollution exposure space $\mathcal{X}$, $\mathcal{Z}$ respectively are compact bounded sets. This is easily satisfied in practice.

Next we define two function spaces. Letting $\alpha > 0$, we define $C^\alpha[0,1]^q$ to be the Holder space of smooth functions $h : [0,1]^q \mapsto \mathbb{R}$ which have uniformly bounded derivatives up to $\lfloor\alpha\rfloor$, and the highest partial derivatives are Lipschitz order $\alpha - \lfloor\alpha\rfloor$. More precisely, we define the vector of q integers as $\boldsymbol{v} = (v_1, \ldots, v_q)$ and

$$D^v g(\boldsymbol{z}) = \frac{\partial^{(\sum_i v_i)} g(\boldsymbol{z})}{\partial z_1^{v_1}, \ldots, \partial z_q^{v_q}}.$$

Then for function h we define

$$\|h\|_\alpha = \max_{\sum_i v_i \leq \lfloor \alpha \rfloor} \sup_{D^v g(\boldsymbol{z})} + \max_{\sum_i v_i = \lfloor \alpha \rfloor} \sup \frac{|D^v g(\boldsymbol{z}) - D^v g(\boldsymbol{z}')|}{\|\boldsymbol{z} - \boldsymbol{z}'\|^{\alpha - \lfloor \alpha \rfloor}}, \text{ for} \boldsymbol{z} \neq \boldsymbol{z}'.$$

With the above we say that

$$C^\alpha[0,1]^q = \left\{ h : [0,1]^q \mapsto \mathbb{R} \middle| \|h\|_\alpha < M \right\}$$

(Vaart, 1996).

Note that $C^\alpha[0,1]^q$ might be too large of a space as it is highly flexible in terms of differentiability restrictions. In light of this, if we only consider smooth functions, we introduce the following space.

Let the Fourier transform be $\widehat{h}(\lambda) = 1/(2\pi)^q \int e^{i(\lambda,t)} h(t) dt$ and define

$$\mathcal{A}^{\gamma,r}(\mathbb{R}^q) = \left\{ h : \mathbb{R}^q \mapsto \mathbb{R} \middle| \int e^{\gamma \|\lambda\|^r} |\widehat{h}(\lambda)|^2 d\lambda < \infty \right\}.$$

Set $\mathcal{A}^{\gamma,r}(\mathbb{R}^q)$ contains infinitely differentiable functions which are increasingly smooth as γ or r increase (van der Vaart & van Zanten, 2009).

THEOREM 1. *Let $\delta_0(h)$ be the Dirac measure concentrated at h_0. For any $\delta \in (0,1)$, there exists a constant C_1 such that if we choose the number of splits $K \leq C_1 \log 1/\delta$, then with a probability of at least $1-\delta$, we have*

$$\|\widehat{\Pi}_{n,g} - \delta_0(h_0)\|_{\mathcal{F}} \leq \begin{cases} C_2 (n/\delta)^{-\left(\frac{\alpha}{2\alpha+q}\right)} \left(\log(n/\delta)\right)^{\left(\frac{4\alpha+q}{4\alpha+2q}\right)} & \text{if } h_0 \in \mathcal{C}^\alpha[0,1]^q, \\ C_3 (n/\delta)^{-\frac{1}{2}} \left(\log(n/\delta)\right)^{(q+1+q/(2r))} & \text{if } h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q) \text{ and } r < 2, \\ C_3 (n/\delta)^{-\frac{1}{2}} \left(\log(n/\delta)\right)^{(q+1)} & \text{if } h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q) \text{ and } r \geq 2, \end{cases}$$

where C_2, C_3 are sufficiently large constants.

The proof follows from results on convergence of the posterior median and scaled squared exponential Gaussian process properties. We defer the proof to Appendix A. The rate in Theorem 1 is achieved for all levels of regularity α simultaneously. If $h_0 \in \mathcal{C}^\alpha[0,1]^q$, then the adaptive rate is $\widetilde{\mathcal{O}}\left((n/\delta)^{-(\alpha/(2\alpha+q))}\right)$, however further assuming h_0 is infinitely differentiable, then $h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q)$ and we recover the usual $\widetilde{\mathcal{O}}\left(n^{-1/2}\right)$ rate. Intuitively, understanding α as the number of derivatives of h_0, this $n^{-1/2}$ rate is obtained letting $\alpha \to \infty$. Theorem 1 sheds light into the trade-off between choosing the optimal number of splits K: large K negatively impacts the statistical rate as it slows down convergence, however it helps with respect to computation complexity. Finally, dimension q affects the rate on a logarithmic scale if h_0 is infinitely differentiable; in the case that $h_0 \in \mathcal{C}^\alpha[0,1]^q$ then q has a larger effect in the rate. This trade-off is further illustrated in Section 4.

4. SIMULATION RESULTS

To study our method's empirical performance in a finite sample we evaluated it in several simulation settings. The simulated data is generated with the following procedure. We generated a data set with n observations, $\mathbb{S}_n = \{D_i\}_{i=1}^n$, $D_i = (y_i, x_i, \boldsymbol{z}_i)$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^\top$ is the profile for observation i with q mixture components. x_i is a confounder of the mixture profile generated by $x_i \sim N(3\cos(z_{i1}), 2)$. The outcomes were generated by $y_i \sim N(\beta x_i + h_0(\boldsymbol{z}_i), \sigma^2)$. We considered $q = 4$ total number of mixture components, and each exposure vector $\{\boldsymbol{z}_i\}_{i=1}^n$

was obtained by sampling each component $\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{i4}$ from the standard normal distribution $N(0, 1)$.

We considered the mixture-response function $h_0(\cdot)$ as a non-linear and non-additive function of only (z_{i1}, z_{i2}) with an interaction. In particular, let $\phi(x) = 1/(1 + e^{-x})$, we generated h as

$$h_0(\boldsymbol{z}_i) = 4\phi\left(\frac{5}{6}\left\{z_{i1} + z_{i2} + \frac{1}{2}z_{i1}z_{i2}\right\}\right).$$

We set $\beta^0 = 2$, and $\sigma^2 = 0.5$. We considered the total number of samples $n \in \{512, 1024, 2048, 4096\}$, and the number of splits $K = n^t$, with $t \in \{0, 0.05, 0.1, 0.15, \ldots, 0.7\}$. Note that for $n = 512$, the subset sample size is not enough for preforming the MCMC sampler when $t = 0.7$. Therefore, simulation under the setting wasn't performed. Each simulation setting is replicated 300 times. All computations are performed on a server with 2.6GHz 10 core compute nodes, with 15000MB memory.

Figure 1 and 2 show the method performance for approximating h_0 by the median posterior. To evaluate performance, we ran a linear regression of the estimated $\widehat{h}$ on true h_0, i.e. $h_0(\boldsymbol{z}_i) = \gamma_0 + \gamma_1\widehat{h}(\boldsymbol{z}_i) + \epsilon_i \; i = 1, \ldots, n$ and plot the estimated slope $\widehat{\gamma}_1$, intercept $\widehat{\gamma}_0$ and R^2 while varying number of sample splits K. A good $\widehat{h}(\cdot)$ would yield $\widehat{\gamma}_0 = 0$, $\widehat{\gamma}_1 = 1$, $R^2 = 1$ as $h_0(\boldsymbol{z}) \approx \widehat{h}(\boldsymbol{z})$. As the figure shows, as the number of splits increases with t, inference on h_0 starts to lose precision. This is natural; although the median geometrically improves rate n_k, as splits increase each posterior sample becomes noisier. However, near $t = 1/2$ the median performance for $\widehat{h}$ is close to performance when the entire sample is used ($t = 0$) as measured by $\widehat{\gamma}_0, \widehat{\gamma}_1, R^2$, with significant computation time gains. Figure 2 shows computing time for inference on h_0 through the posterior median. There is a clear trade-off between sampling from a high dimensional Gaussian process posterior of n samples, and a large number of data splits which require almost equivalent computation power to sample. Results suggest that splits with $t \in [1/4, 1/2]$ decrease computational burden significantly. On the other hand Figure 1,2 and theoretical results in Section 3 suggest that $t \leq 1/2$ offers a good approximations to the full-data posterior. Theoretical and empirical results suggest choosing $t \in [1/4, 1/2]$, with $t \to 1/2$ as n increases will optimize the computation-cost vs. precision trade-off.

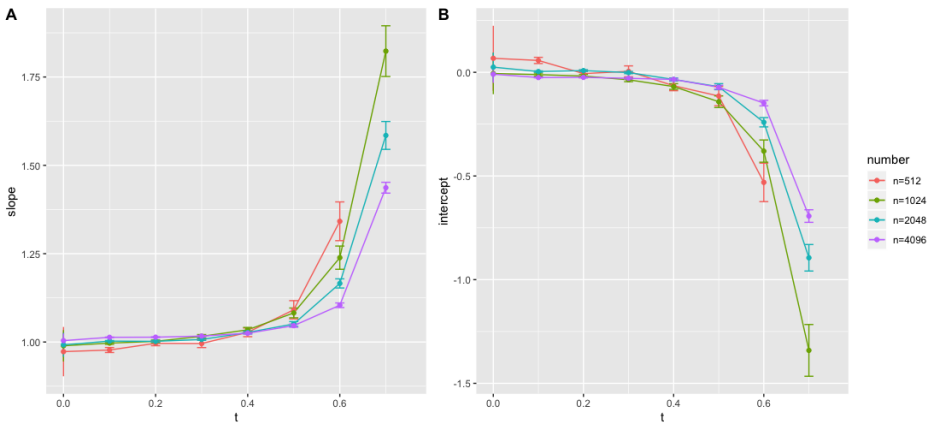

Fig. 1: Regression summary results for $\mathbf{h} = \gamma_0 + \gamma_1\widehat{\mathbf{h}}$ across different sample size n and data set splits. The setting of number of subsets are described above as n^t. We show (A) intercept: $\widehat{\gamma}_0$, (B) slope: $\widehat{\gamma}_1$.

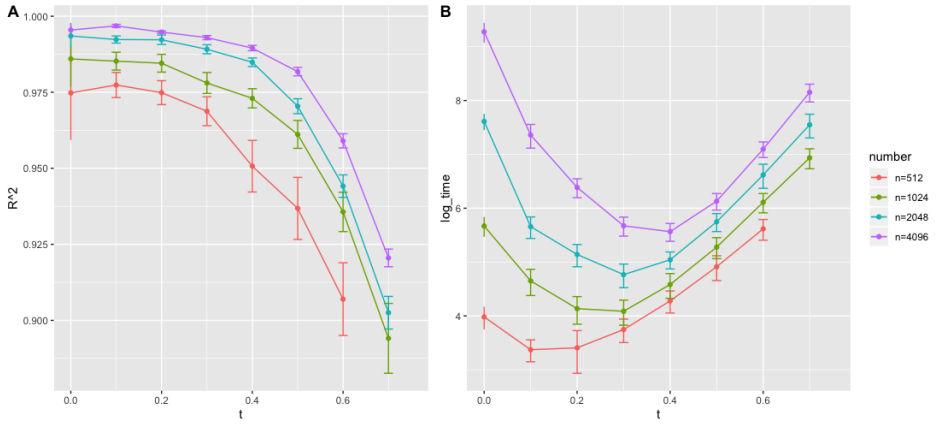

Fig. 2: (A)Regression R^2 for $\mathbf{h} = \gamma_0 + \gamma_1 \widehat{\mathbf{h}}$ and (B) Logarithmic runtime for fast BKMR across different sample size n and data set splits. The setting of number of subsets are described above as n^t.

5. A Toxicology Study of Major Particulate Matter Constituents on Birth-weight in Massachusetts

To further evaluate our method on a real data set, we consider a study of major particulate matter constituents and birth-weight, in which preliminary data from 907,766 newborns in Massachusetts from January 2001 to 31 December 2012 were collected Fong et al. (2019a). After excluding the observation records with missing data, there were $n = 685,857$ observations used for analysis. We treated normalized Ozone , NDVI, $PM_{2.5}$, EC, OC, nitrate, and sulfate as mixture components for non-parametric parts, and other variables such as maternal characteristics, as confounders. We randomly split the sample to $K = 686$ (using $t \approx 1/2$) different splits, each split contains ≈1000 samples. For each split, we ran the MCMC sampler for 1000 iterations after 1 000 burn in iterations, and every fifth sample was kept for further inference, thus we have parameter sample size $N = 200$ for each split. Further details on the confounders included in the analysis can be found in Appendix B.

Figure 3 shows univariate estimated effects for each major constituent and $PM_{2.5}$ on birth weight with other constituents fixed at their median. The figure suggests that for the $PM_{2.5}$, EC, and OC terms, increasing values of the constituents are associated with decreasing values of birth-weight. On the other hand, we have the Ozone, nitrate, and NDVI terms, for which increasing values are associated with increasing birth-weight. Furthermore, it seems there is no association between birth-weight and maternal exposure to sulfate. Among negatively associated constituents, EC and remaining $PM_{2.5}$ constituents have stronger linear negative associations compared to OC. Among positive associations, NDVI and Ozone seem to have a strong linear relationship with birth-weight. However, for nitrate, when its concentration is lower than +1 standard deviation, it is positively associated with birth weight increase, whereas when it is above the mean level over 1 standard deviation, it is negatively associated with birth-weight. This suggests an effect modification type of relationship.

Figure 4 investigates the bivariate relationship of two major constituents with birth weight, with other constituents fixed at their median levels. The figure suggests different levels of non-linear relationships between constituent concentrations and birth weight. Unlike the pattern of sulfate shown in figure 3, there exists a strong inverted u-shaped relationship between sulfate

and mean birth-weight when nitrate concentration is at around -1 standard deviation. A similar relationship is visible between nitrate and mean birth-weight when sulfate concentration is higher than +0.5 standard deviation. Moreover, the $PM_{2.5}$ shows no association with birth weight when its concentration is lower than 0 standard deviation, with sulfate concentration lower than -1 standard deviation.

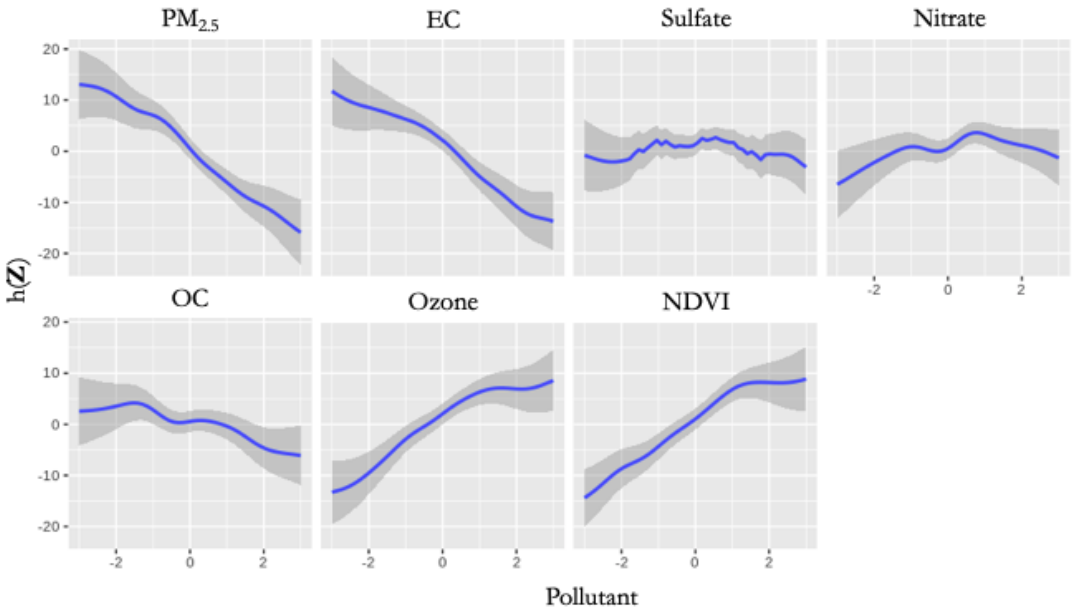

Fig. 3: Univariate estimated effects on birth-weight per standard deviation increase in $PM_{2.5}$, EC, OC, nitrate, sulfate, NDVI, and Ozone. 95% confidence bands of estimates are in gray. All of the other mixture components are fixed to 50th percentile level when investigating single mixture effect on birth-weight. We show h(Z): difference of birth-weight comparing to the mean birth-weight of samples in grams; Pollutant(Z): change of each of the major constituents with the measure of standard deviation of that constituent.

6. Discussion

As industry and governments invest in new technologies that ameliorate their environmental and pollution impact, the need to quantify the effects of pollution on health is prioritized. In parallel, electronic data registries such as the Massachusetts birth weights data set are increasingly common and larger. These rich data sets allow measuring small, highly non-linear effects of pollutant mixtures that impact public health. To the best of our knowledge, we propose the first semi-parametric Gaussian process regression framework that can be used to estimate effects using large datasets. In particular, we model the pollutant-health outcome surface with a Gaussian process that allows for feature selection. Additionally, we use a linear component to incorporate confounder effects. Previous approaches for similar analysis had to either assume a parametric relationship or use a single pollutant per regression to estimate effects of interest (Fong et al., 2019a).

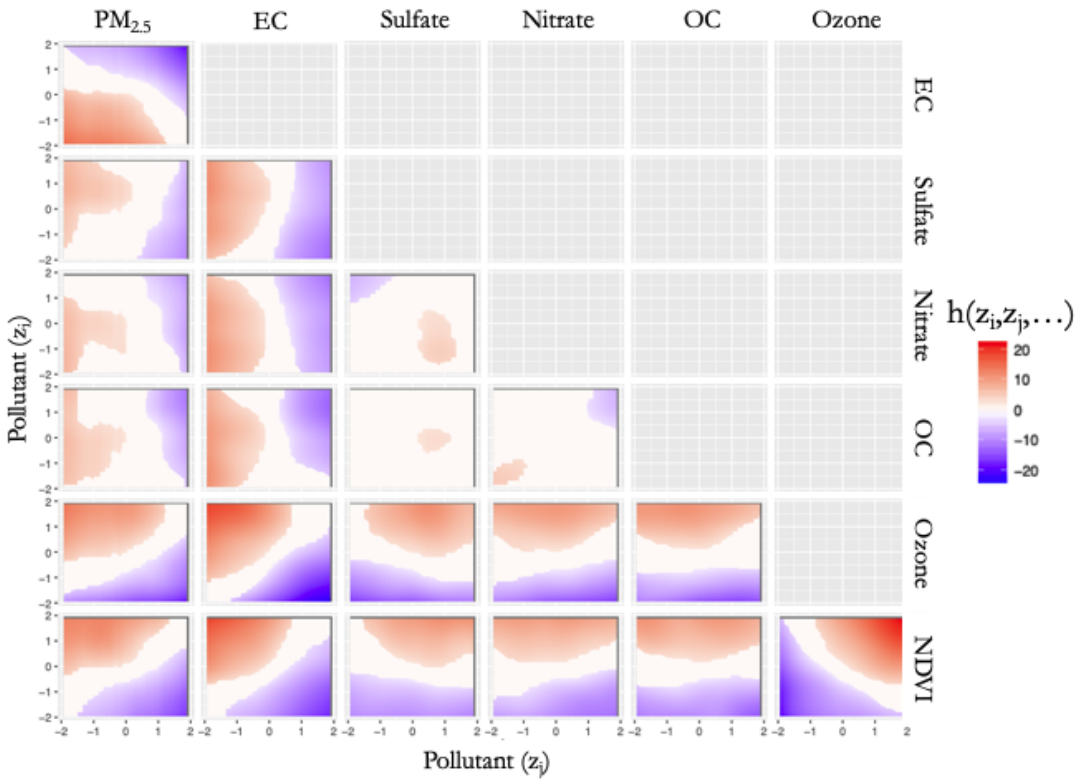

Fig. 4: Bivariate estimated effects on birthweight per standard deviation increase between $PM_{2.5}$, EC, OC, nitrate, sulfate, NDVI, and Ozone. All of the other mixture components are fixed to 50^{th} percentile level when investigating bivariate mixture effects on birthweight. We show h(z_i, z_j): difference of birth-weight compared to the mean birth-weight of samples in grams; Pollutant(z_i) and Pollutant(z_j): change of each of the major constituents with the measure of standard deviation of that constituent.

To ameliorate the computational burden of computing the Bayesian posteriors of the Gaussian process, we propose a divide-and-conquer approach. Our method consists of splitting samples into subsets, computing the posterior distribution for each data split, and then combining the samples using a generalized median based on the Wasserstein distance (Minsker et al., 2017).

We tailor the method to incorporate a squared exponential kernel and provide theoretical guarantees for the convergence of the Gaussian process for this choice of kernel. Our convergence results accommodate different assumptions for the underlying space of the true feature-response function. We provide theoretical and empirical results which illustrate a trade-off for the optimal number of splits. As the number of data splits increases, the posterior computation of the small data subsets will be faster; however, these posteriors will be noisy. In other words, there is a tension between computational cost and obtaining precise estimates. We propose using $K = n^{1/2}$ sample splits to efficiently approximate the posterior in a relatively short time. To illustrate the benefit of our method, we analyze the impact of several individual pollution constituents on Massachusetts birth weights using a large dataset. Given our results, we believe this method will be

highly useful in answering similar questions which require high-dimensional inference to analyze the complex relationships underlying environmental health.

REFERENCES

BILLIONNET, C., SHERRILL, D. & ANNESI-MAESANO, I. (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Annals of epidemiology* **22**, 126–141.

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**, 493–508.

CARLIER, G., OBERMAN, A. & OUDET, E. (2015). Numerical methods for matching for teams and wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis* **49**, 1621–1642.

COULL, B. A., BOBB, J. F., WELLENIUS, G. A., KIOUMOURTZOGLOU, M.-A., MITTLEMAN, M. A., KOUTRAKIS, P. & GODLESKI, J. J. (2015). Part 1. statistical learning methods for the effects of multiple air pollution constituents. *Research report - Health Effects Institute* , 5.

CUTURI, M. & DOUCET, A. (2014). Fast computation of wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning* **32**, 685–693.

FONG, K. C., DI, Q., KLOOG, I., LADEN, F., COULL, B. A., KOUTRAKIS, P. & SCHWARTZ, J. D. (2019a). Relative toxicities of major particulate matter constituents on birthweight in massachusetts. *Environmental epidemiology* **3**, e047.

FONG, K. C., KOSHELEVA, A., KLOOG, I., KOUTRAKIS, P., LADEN, F., COULL, B. A. & SCHWARTZ, J. D. (2019b). Fine particulate air pollution and birthweight: Differences in associations along the birthweight distribution. *Epidemiology (Cambridge, Mass.)* **30**, 617–623.

GASKINS, A. J., MÍNGUEZ-ALARCÓN, L., FONG, K. C., ABU AWAD, Y., DI, Q., CHAVARRO, J. E., FORD, J. B., COULL, B. A., SCHWARTZ, J., KLOOG, I., ATTAMAN, J., HAUSER, R. & LADEN, F. (2019). Supplemental folate and the relationship between traffic-related air pollution and livebirth among women undergoing assisted reproduction. *American journal of epidemiology* **188**, 1595–1604.

LIU, D., LIN, X. & GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63**, 1079–1088.

MINSKER, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability* **21**, 2308–2335.

MINSKER, S., SRIVASTAVA, S., LIN, L. & DUNSON, D. (2017). Robust and scalable bayes via a median of subset posterior measures. *Journal Of Machine Learning Research* **18**.

SCHMIDHUBER, J. (2015). Deep learning in neural networks: An overview. *Neural networks* **61**, 85–117.

SCORNET, E., BIAU, G. & VERT, J.-P. (2015). Consistency of random forests. *The Annals of statistics* **43**, 1716–1741.

SRIVASTAVA, S., CEVHER, V., TRAN DINH, Q. & DUNSON, D. B. (2015). Wasp: Scalable bayes via barycenters of subset posteriors. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* **38**, 912 – 920.

SRIVASTAVA, S., LI, C. & DUNSON, D. B. (2018). Scalable bayes via barycenter in wasserstein space. *J. Mach. Learn. Res.* **19**, 312–346.

STIEB, D. M., CHEN, L., ESHOUL, M. & JUDEK, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental research* **117**, 100–111.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Methodological* **58**, 267–288.

VAART, VAN DER, A. & ZANTEN, VAN, J. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics* .

VAART, A. W. (1996). *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer Series in Statistics. New York, NY: Springer New York : Imprint: Springer.

VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics* **37**, 2655–2675.

WILLIAMS, C. K. I. & RASMUSSEN, C. E. (2019). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. The MIT Press.

Supplementary material to

Scalable Gaussian Process Regression Via Median Posterior Inference for Estimating Multi-Pollutant Mixture Health Effects

SUMMARY

This document contains the supplementary material to the paper "Scalable Gaussian Process Regression Via Median Posterior Inference for Estimating Multi-Pollutant Mixture Health Effects".

A. PROOF OF THEOREM 1

The proof for Theorem 1 uses the following results.

Assumption 1. Let A be a random variable with positive support, the distribution of A has a Lebesgue density g such that

$$C_1 a^p \exp\{-D_1 a^{d_1} \log^{d_2} a\} \leq g(a) \leq C_2 a^{d_1} \exp\{-D_2 a^{d_1} \log^{d_2} a\},$$

for every large enough $a > 0$ and constants $C_1, D_1, C_2, D_2 > 0$ and $d_1, d_2 \geq 0$.

Assumption 2. Let H be a Gaussian field, the associated spectral measure μ satisfies

$$\int e^{\delta \|\lambda\|} \mu(\delta\lambda) < \infty,$$

for some $\delta > 0$. We say that H has subexponential tails.

Assumption 3. Let H be a Gaussian field, H possesses a Lebesgue density f such that $a \mapsto f(a\lambda)$ is decreasing on $(0, \infty)$ for every, $\lambda \in \mathbb{R}^q$.

For a random variable A satisfying Assumption 1, let $H^A = \{H_{A\boldsymbol{z}} : \boldsymbol{z} \in [0,1]^q\}$ be a centered rescaled Gaussian process. We consider the Borel measurable map in $C[0,1]^q$, equipped with the uniform norm $\|\cdot\|_\infty$.

THEOREM 2 (THEOREM 3.1 IN (VAN DER VAART & VAN ZANTEN, 2009)). *Let H be a centered homogeneous Gaussian field which satisfies Assumptions* 2, 3*, then there exist Borel measurable subsets B_n of $C[0,1]^q$ and for sufficiently large n and big enough constant C_4*

$$\begin{aligned} \log N\left(\bar{\epsilon}_n, B_n, \|\cdot\|_\infty\right) &\leq n\bar{\epsilon}_n^2, \\ P\left(H^A \notin B_n\right) &\leq e^{-4n\epsilon_n^2}, \\ P\left(\|H^A - h_0\|_\infty < \epsilon_n\right) &\geq e^{-n\epsilon_n^2}, \end{aligned} \tag{10}$$

where

- *if $h_0 \in C^\alpha[0,1]^q$ for $\alpha > 0$ then $\epsilon_n = n^{-\alpha/(2\alpha+q)} (\log n)^{\kappa_1}$, $\bar{\epsilon} = C_4 \epsilon_n (\log n)^{\kappa_2}$, for $\kappa_1 = ((1 + q \vee d_2)/(2 + q/\alpha)$ and $\kappa_2 = (1 + q - d_2)/2$,*
- *if h_0 is the restriction of a function in $\mathcal{A}^{\gamma,r}(\mathbb{R}^q)$ to $[0,1]^q$ with spectral density satisfying $|f(\lambda)| \geq C_3 \exp\{-D_3 \|\lambda\|^\nu\}$ for some constants $C_3, D_3, \nu > 0$, then $\bar{\epsilon}_n = \epsilon_n (\log n)^{(q+1)/2}$ and $\epsilon_n = C_4 n^{-\frac{1}{2}} (\log n)^{(q+1)/2}$ if $r \geq \nu$, and $\epsilon_n = C_4 n^{-\frac{1}{2}} (\log n)^{((q+1)/2 + q/(2r))}$ if $r < 2$.*

THEOREM 3 (THEOREM 2.2 IN (VAN DER VAART & VAN ZANTEN, 2009)). *Let $H = \{H_{\boldsymbol{Z}} : \boldsymbol{z} \in [0,1]^q\}$ be the centered Gaussian process, with covariance function*

$\mathbb{E}[H_{\boldsymbol{Z}} H_{\boldsymbol{Z}'}] = \exp\{-\|\boldsymbol{Z} - \boldsymbol{Z}'\|_n^2\}$. *Also let* σ *be Gamma-distributed random variable. We consider* $H = \{H_{\sigma \boldsymbol{Z}} : \boldsymbol{z} \in [0,1]^q\}$ *as a prior distribution for* h_0*. Then for every large enough* M*,*

$$\Pi_{h,\sigma}\left(\|h - h_0\|_n + |\sigma - \sigma_0| \geq M\epsilon | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\right) \to 0 \text{ as } n \to \infty,$$

where

$$\epsilon = \begin{cases} n^{-\left(\frac{\alpha}{2\alpha+q}\right)} (\log n)^{\left(\frac{4\alpha+2}{4\alpha+2q}\right)} & \text{if } h_0 \in \mathcal{C}^\alpha[0,1]^q, \\ n^{-\frac{1}{2}} (\log n)^{(q+1+q/(2r))} & \text{if } h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q) \text{ and } r < 2, \\ n^{-\frac{1}{2}} (\log n)^{(q+1)} & \text{if } h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^d) \text{ and } r \geq 2. \end{cases}$$

THEOREM 4 (THEOREM 7 IN (MINSKER ET AL., 2017)). *Let* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_k} \sim P_0$ *be an* $i.i.d$ *sample, and assume that* $\epsilon_k > 0$ $\Theta_k \subset \Theta$ *are such that for some constant* $\widetilde{C}_1 > 0$

$$\begin{aligned} &\log M(\epsilon_k, \Theta_k, \rho) \leq n_k \epsilon_k^2, \\ &\Pi(\Theta \backslash \Theta_k) \leq \exp\{-n_k \epsilon_k^2 (\widetilde{C}_1 + 4)\}, \\ &\Pi\left(\theta : -P_0\left(\log \frac{p_\theta}{p_0}\right) \leq \epsilon_k^2, P_0\left(\log \frac{p_\theta}{p_0}\right)^2 \leq \epsilon_k^2\right) \geq \exp\{-\widetilde{C}_1 n_k \epsilon_k^2\}. \end{aligned} \tag{11}$$

Then there exists constants $R(\widetilde{C}_1)$ *and* $\widetilde{C}_2$ *such that*

$$P\left(d_{W_{1,\rho}}\left(\delta_0, \Pi_k(\cdot \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K)\right) \geq R\epsilon_k + e^{-\widetilde{C}_2 n_k \epsilon_k^2}\right) \leq \frac{1}{n_k \epsilon_k^2} + 4e^{-\widetilde{C}_2 n_k \epsilon_k^2}.$$

COROLLARY 1 (COROLLARY 8 IN (MINSKER ET AL., 2017)). *Let* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \sim P_0$ *be an* $i.i.d.$ *sample, and let* $\widehat{\Pi}_{n,g}$ *be defined as in xxx. Under conditions* 11*, if* ϵ_k *satisfies* $1/(n_k \epsilon_k^2) + 4e^{-(1+\widetilde{C}_2/2) n_k \epsilon_k^2/2} < \frac{1}{7}$*, then*

$$\mathbb{P}\left(\left\|\delta_0 - \widehat{\Pi}_{n,g}\right\|_{\mathcal{F}} \geq 1.52\left(R\epsilon_k + e^{-\widetilde{C}_2 n_k \epsilon_k^2}\right)\right) < 1.27^{-K}$$

Proof of Theorem 1. If A has a Gamma distribution, then Assumption 1 is satisfied with $q = 0$. Additionally, as H is squared exponential Gaussian process, it is a density relative to the Lebesgue measure given by

$$\lambda \mapsto \frac{1}{2^q \pi^{q/2}} \exp\{-\|\lambda\|^2/4\}$$

which has sub-exponential tails (see (van der Vaart & van Zanten, 2009)). Therefore, by Theorem 2 conditions (10) are satisfied for H^A with $\epsilon_k = n_k^{-\alpha/(2\alpha+q)} (\log n_k)^{(4\alpha+q)/(4\alpha+2q)}$ if $h_0 \in \mathcal{C}[0,1]^q$ and if $h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q)$, then

$$\epsilon_k = \begin{cases} n_k^{-\frac{1}{2}} (\log n_k)^{(q+1+q/(2r))} & \text{if } r < 2, \\ n_k^{-\frac{1}{2}} (\log n_k)^{(q+1)} & \text{if } r \geq 2. \end{cases}$$

Note that (10) map one-to-one to Conditions in (11) (see (Vaart & Zanten, 2008)), thus by Theorem 4 with $\epsilon_k > 0$ defined as above we have

$$P\left(d_{W_{1,\rho}}\left(\delta_0, \Pi_k(\cdot \mid \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K)\right) \geq R\epsilon_k + e^{-\widetilde{C}_2 k \epsilon_k^2}\right) \leq \frac{1}{n_k \epsilon_k^2} + 4e^{-\widetilde{C}_2 n_k \epsilon_k^2}. \tag{12}$$

note that whether $h_0 \in \mathcal{C}^\alpha[0,1]^q$ or $h_0 \in \mathcal{A}^{\gamma,r}(\mathbb{R}^q)$ we can choose $k(n)$ such that $1/(n_k\epsilon_k^2) + 4e^{-(1+\widetilde{C}_2/2)n_k\epsilon_k^2/2} < \frac{1}{7}$. For example any $k \leq n^{1/2}\log n$ would work well. Therefore, for any $\delta > 0$, and fixed $k(n)$ using Corollary 1 there is an $\epsilon_k(\delta)$ with a large enough n such that

$$\mathbb{P}\left(\left\|\delta_0 - \widehat{\Pi}_{n,g}\right\|_{\mathcal{F}} \geq 1.52\left(R\epsilon_k + e^{-\widetilde{C}_2 n_k\epsilon_k^2}\right)\right) < \delta.$$

B. Details on Application to Boston Birth Weight Data

Each record consists of the outcome of interest which is the birth-weight of the newborn, confounders such as maternal age (years), maternal race (white, black, Asian, American Indian, other), maternal marital status (married, not married), maternal smoking during or before pregnancy (yes, no), maternal education (highest level of education attained: less than high school, high school, some college, college, advanced degree beyond college), parity (first-born, not first-born), maternal diabetes (yes, no), gestational diabetes (yes, no), maternal chronic high blood pressure (yes, no), maternal high blood pressure during pregnancy (yes, no), Kessner index of adequacy of prenatal care (adequate, intermediate, inadequate, no prenatal care), mode of delivery (vaginal, forceps, vacuum, first cesarean birth, repeat cesarean birth, vaginal birth after cesarean birth), clinical gestational age (weeks), year of birth (one of 2001–2012), season of birth (spring, summer, autumn, winter), date of birth, newborn sex (male, female), Ozone concentration, Normalized Difference Vegetation Index (NDVI), Medicaid-supported prenatal care (yes, no). Finally pollution exposure measures are concentration of $PM_{2.5}$ and four major chemical constituents of it: elemental carbon (EC), organic carbon (OC), nitrate, and sulfate. After excluding the observation records with missing data, the final sample with size equal to $n = 685,857$ is used for our model illustration. We treated normalized Ozone , NDVI, $PM_{2.5}$, EC, OC, nitrate, and sulfate as mixture components for non-parametric parts, and other variables as covariates. For the date of birth within one year, in order to control the temporal effect on birth weight, we implement a cosine transformation on it, with birth date on January 1^{st} has highest positive effect on birth weight, and June 15^{th} has lowest negative effect on the birth weight. The model used is (1). In our main analysis, we scaled the estimated effects per a standard deviation increase per each pollutant, which is more representative of a real world scenario than mass scaling.