Identifying main effects and interactions among exposures using Gaussian processes

Federico Ferrari[†] and David B. Dunson[§]

PhD student, Duke University[†]

Arts and Sciences Professor of Statistical Science, Duke University[§]

Motivation

We are motivated by the problem of studying the **joint effect** of different chemical exposures on human health outcomes, which is essentially a nonparametric regression problem.

- The main focus in toxicology and epidemiology literatures has been on examining the health effects of chemicals one at a time.
- Parametric regression techniques provide interpretable estimates but the resulting dose response surface is typically too restrictive.
- Non parametric techniques modeling the joint health effect of multiple chemicals have limited interpretability, essentialy providing a black box.

We **decompose** the regression surface on the health outcome into a linear effect, pairwise interactions and non-linear deviation.

- We address identifiability between the parametric and nonparametric part of the model with a projection approach developed in spatial statistics.
- Using spike and slab priors, we allow for variable selection for the main effects and non-linear effects.
- We reduce the computation by imposing a heredity condition and a dimension reduction approach.

Model

- y_i denotes a continuous health response for individual i, $x_i = (x_{i1}, \cdots, x_{ip})^T$ denotes a vector of exposure measurements and $z_i = (z_{i1}, \cdots, z_{iq})^T$ are covariates.

$$y_i = x_i^T\beta + \sum_{j=1}^{p}\sum_{k>j}\lambda_{jk}x_{ij}x_{ik} + g^* + z_i^T\alpha + \epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2),$$
$$g^* = Pg, \quad g \sim \mathsf{GP}(0, c),$$

- $\beta = (\beta_1, \ldots, \beta_p)^T$ are linear main effects of exposures, $\lambda = \{\lambda_{jk}\}$ are pairwise linear interactions, $g^*(x)$ is a nonparametric deviation, and $\alpha = (\alpha_1, \ldots, \alpha_q)^T$ are coefficients for the covariates.
- $\mathsf{GP}(0, c)$ denotes a Gaussian process (GP) centered at zero with covariance function c controlling the uncertainty and smoothness of the realizations.

Variable Selection

- We choose spike and slab priors for main effects:

$$\beta_k \sim \gamma_k N(0,1) + (1-\gamma_k)\delta_0$$

- We impose the heredity condition for interations. Strong heredity (S) means that an interaction is included in the model only if the main effects are. For weak heredity (W) it sufficies to have one main effect in the model to estimate the interaction of the corresponding variables.

S: $\lambda_{j,k}|\gamma_j = \gamma_k = 1 \sim N(0,1), \quad \lambda_{j,k}|(\gamma_j = \gamma_k = 1)^C \sim \delta_0$
W: $\lambda_{j,k}|(\gamma_j = \gamma_k = 0)^C \sim N(0,1), \quad \lambda_{j,k}|\gamma_j = \gamma_k = 0 \sim \delta_0$

- We use a squared exponential covariance funtion to favor smooth departures from linearity:

$$c(x,x) = \mathrm{cov}\{g(x), g(x)\} = \tau^2 \exp\left\{\sum_{j=1}^{p}\rho_j(x_j - x_j)^2\right\},$$

- We endow the smoothness parameters $\rho_1, \cdots, \rho_p$ with independent spike and slab priors:

$$\rho_k \sim \gamma_k^\rho F_\rho(\cdot) + (1-\gamma_k^\rho)\delta_0$$

where $F_\rho(\cdot)$ is a Gamma distribution with parameters $(1/2, 1/2)$. When $\gamma_k^\rho = 0$, the kth exposure is eliminated from the nonparametric term g in the model.

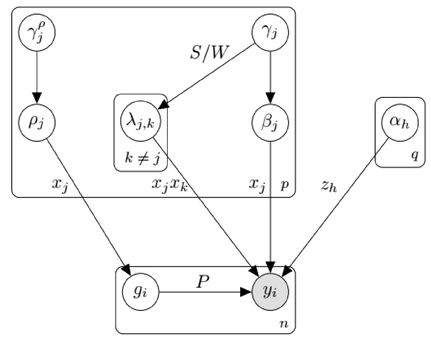
Non Identifiability and Projection

Confounding between the Gaussian Process prior and parametric functions is a known problem in Spatial statistics and occurs when spatially dependent covariates are strongly correlated with spatial random effects. This problem is emphasized when the same features are included in the model both in the linear term and in the nonparametric surface.

$$P_X = X^*(X^{*T}X^*)^{-1}X^{*T}$$
$$g^* = Pg = (I_n - P_X)g$$

- where X_i^* denote the vector containing main effects and interactions for the i^{th} unit. We are projecting the nonlinear effect onto the orthogonal space of the parametric part of the model.

Graphical Model



The arrows between two nodes indicate conditional dependence. Variables that are in the same plate share the same indexes, for example y_i and g_i are such that $i = 1, \cdots, n$

Environmental Epidemiology Analysis

The goal of this analysis is to assess the effect of DDE and PCBs concentrations on pregnancy outcomes, in particular on length of pregnancy. We define the variable *preterm birth* as our outcome and we estimate a Logistic Regression. We use the same data as [6].

$$Logit(P(Y_i = 1)) = \eta_i$$
$$\eta_i = x_i^T\beta + \sum_{j=1}^{p}\sum_{k\neq j}\lambda_{jk}x_{ik}x_{il} + \alpha z_i + g_i^*$$

- The estimated main effects of *DDE* and *Cholesterol* are significantly different from zero. As expected, a higher concentration of *DDE* is linearly associated with an higher probability of a *preterm* delivery.
- The coefficients of pairwise interactions are not significantly different than zero.
- We estimate a significant non linear deviation for the chemical *PCB*, with the smoothness parameter being different than zero in more than 99% of the iterations.

References

1 Bobb, Jennifer F., et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 16.3 (2014): 493-508.

2 Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2012). Efficient gaussian process regression for large datasets. *Biometrika*.

3 Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association.*, 88(423):881889, 1993.

4 Hao Ning, Zhang Hao Helen. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association.* 2014;109(507):12851301.

5 Longnecker, M.P. et al. (2001) Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet* 358, 110114

6 Savitsky, T., Vannucci, M., and Sha, N. (2011). Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Stat. Sci.* 26, 130149