Estimating interactions between exposures using Gaussian processes and strong heredity

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

PhD student, Duke University[†]

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

Motivation

It is well known that humans are exposed to a complex **mixture of different chemicals**, but very little is known about if and how these exposures interact to impact binary and continuous health outcomes. Moreover:

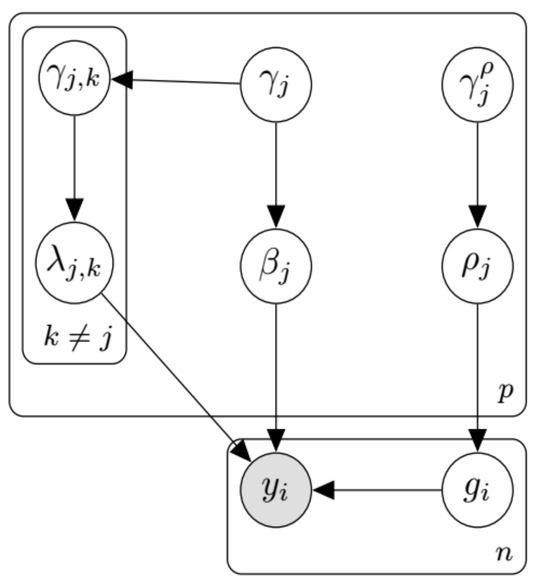
- The main focus in toxicology and epidemiology literatures has been on examining the health effects of chemicals one at a time.
- The models that consider health effect of multiple chemicals have limited interpretability, essentialy providing a black box.

We propose a model based on **Gaussian processes** and **strong heredity structure** in order to disentangle the linear main effect and interactions, in particular:

- Using spike and slab priors, we allow for variable selection for the main effects as well as for interactions and non-linear effects.
- Through a simple projection, we attain identifiability between the GP random effects and the linear terms

Model

$$
\begin{aligned}
&Logit[pr(y_i = 1|x_i, z_i)] = \eta_i \\
&\eta_i = x_i^T\beta + \sum_{j=1}^{p}\sum_{k\neq j}\lambda_{ij}x_{ik}x_{il} + g_i^* \\
&g^* = (I - P)g \\
&(g_1, ..., g_n) \sim GP(0, K) \\
&K_{i,j} = \sigma^2 exp(-x_i^T D x_j) \\
&D = diag(-log(\rho_1), ..., -log(\rho_p))
\end{aligned}
$$



- P is the projection matrix on the column space of the matrix containing the main effects X_j and the statistical interactions X_jX_k. This projection is done in order to have identifiability between the random effects distributed as a Gaussian Process and the linear part of the model.
- We impose **strong heredity condition** as done in [5] between the main effects and the interactions, i.e. we allow the presence of interactions only when both the main effects are present. Mathematically:

$$
\begin{aligned}
&\gamma_{j,k}|\gamma_j = \gamma_k = 1 \sim F(\gamma_{l,j}) \\
&\gamma_{j,k}|(\gamma_j = \gamma_k = 1)^C \sim \delta_0
\end{aligned}
$$

This makes the model invariant to affine transformations of the regressors, which would not happen if we only assumed the **weak heredity condition**.

- We choose spike and slab priors for β, λ and ρ. In particular the prior for β is a mixture of normals as in [3]. On the other hand the spike of λ is a Dirac delta sitting at zero, so that when few main effects are present the computations are easier thanks to the **strong heredity condition**. Finally, the prior for ρ_j is $\pi(\rho_j|\gamma_j^\rho) = \gamma_j^\rho \mathbb{1}_{(\rho_j \in (0,1))} + (1 - \gamma_j^\rho)\delta_1(\rho_j)$ as in [8].
- When we have a large sample size, we employed the reduced-dimensional approach in [2] and [4].
- In the logistic regression, we use Polya-Gamma data augmentation [7] in order to have joint updates of the parameters that have normal priors (or mixture of normals). For the probit model, we use the data augmentation strategy in [1].

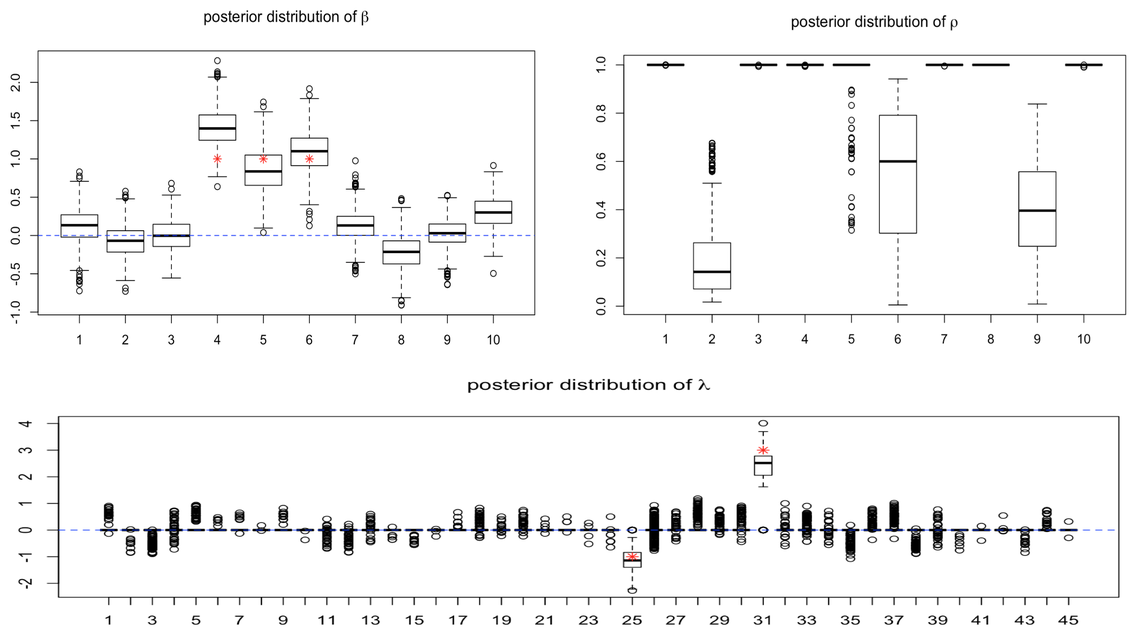
Simulation Study

For this simulated example, I consider a logistic regression with 300 data points. The true model is:

$$
\begin{aligned}
&Logit[pr(y_i = 1|x_i, z_i)] = \eta_i \\
&\eta_i = X_4 + X_5 + X_6 - X_4X_5 + 3X_5X_6 - sin(3X_2) + cos(2X_9)
\end{aligned}
$$

The pictures below show the posterior distribution of β, ρ and λ, notice that:

- The model correctly selects and estimates the true coefficients for the main effects and the interactions.
- Thanks to the **strong heredity condition**, the coefficients of the interactions that are not included in the model are set to zero throughout most of the MCMC run.
- The model correctly excludes the presence of non-linear effects for 7 out of 8 covariates.



Application to mixture of chemicals

The data is taken from of the *National Collaborative Perinatal Project (NCPP)* and was used in [6]. Women have been enrolled during pregnancy and then the kids are followed in order to collect both pregnancy and childhood development outcomes. In this analysis the response variable is *premature birth*. The covariates include:

- **Chemicals**: *DDE and PCBs*.
- **Demographic variables**: *race*, *age*, *smoking status*, *socio- economic index*, *heigth* and *BMI before pregnancy*.
- **Lipids**: *triglycerides*.

The **results** were consistent with the analysis in [6]. In particular:

- The model selects the variables *DDE*, *PCBs*, *triglycerides*, *race* and *BMI before pregnancy*.
- There is evidence of interactions between *DDE* and *Pcb* and between *Pcb* and *triglycerides*.
- There is no strong evidence for the presence of non-linear effects.

References

1 J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association.*, 88(422):66979, 1993.

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 Guan, Y. and Haran, M. (2018) A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models, *to appear in the Journal of Computational and Graphical Statistics*

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

6 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

7 Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association.*. 2013; 108:13391349.

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