Bayesian Factor Analysis for Inference on Interactions (FIN)

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 inference on interactions among chemical exposures impacting human health outcomes.

We propose a **latent factor joint model**, which includes shared factors in both the predictor and response components while assuming conditional independence.

- By including a quadratic regression in the latent variables in the response component, we induce characterize main effects and interactions.
- FIN can accommodate higher order interactions and multivariate outcomes.

Model

- y_i denotes a continuous health response for individual i, and $X_i = (x_{i1}, \cdots, x_{ip})^T$ denotes a vector of exposure measurements.
- Ω is a $k \times k$ symmetric matrix inducing a quadratic latent variable regression that characterizes interactions among the latent variables.
- Λ is the $p \times k$ factor loadings matrix.

$$y_i = \eta_i^T \omega + \eta_i^T \Omega \eta_i + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N(0, \sigma^2),$$
$$X_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Psi),$$
$$\eta_i \sim N_k(0, I),$$

- We can show that the induced regression of X_i on y_i from model (1) is indeed a **quadratic regression**.

$$E(y_i|X_i) = tr(\Omega V) + (\omega^T A)X_i + X_i^T (A^T \Omega A) X_i$$

where $V = (\Lambda^T \Psi^{-1} \Lambda + I)^{-1}$, $A = V \Lambda^T \Psi^{-1}$.

- We can include **covariates** as follows:

$$y_i = \eta_i^T \omega + \eta_i^T \Omega \eta_i + Z_i^T \alpha + \eta_i^T \Delta Z_i + \epsilon_{y,i}$$
$$\mathbb{E}(\eta_i^T \Delta Z_i | X_i, Z_i) = \mathbb{E}(\eta_i^T | X_i) \Delta Z_i = X_i^T (A^T \Delta) Z_i,$$

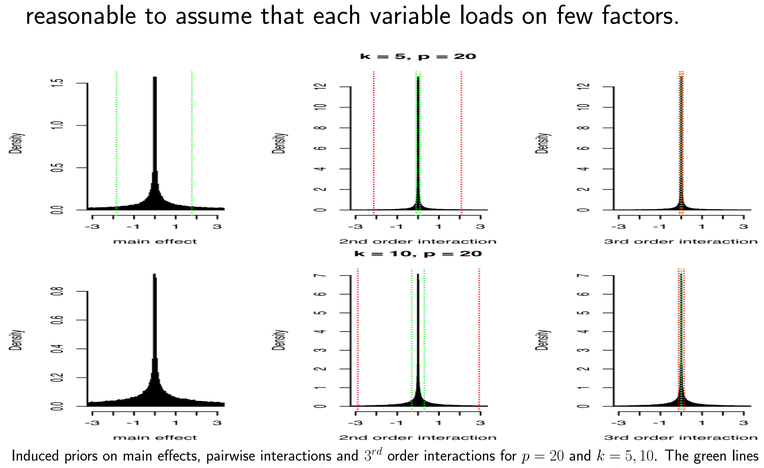
where $(A^T \Delta)$ is a $p \times q$ matrix of pairwise interactions between exposures and covariates.

Prior for Λ

For Λ, we choose the **Dirichlet-Laplace** (DL) prior for each row, corresponding to:

$$\lambda_{j,h} | \phi_{jh}, \tau_j \sim DE(\phi_{jh} \tau_j) \qquad h = 1, \cdots, k$$
$$\phi_j \sim Dir(a, \cdots, a) \qquad \tau_j \sim Gamma(ka, 1/2),$$

- DE refers to the zero mean double-exponential or Laplace distribution.
- k is an upper bound on the number of factors, as the prior allows effective deletion of redundant factors through setting all elements of columns of Λ close to zero.
- The DL prior induces near sparsity row-wise in the matrix Λ, as it is reasonable to assume that each variable loads on few factors.



Induced priors on main effects, pairwise interactions and 3^{rd} order interactions for $p = 20$ and $k = 5, 10$. The green lines corresponds to 0.25 and 0.75 quartiles and the red lines to the 0.05 and 0.95.

Higher Order Interactions

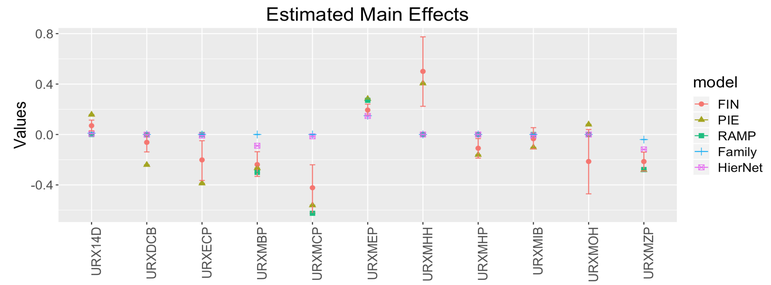
FIN can be generalized to allow for higher order interactions. In particular, let us define a polynomial regression in the latent variables:

$$E(y_i|\eta_i) = \sum_{h=1}^{k} \omega_h^{(1)} \eta_{ih} + \sum_{h=1}^{k} \omega_h^{(2)} \eta_{ih}^2 + \cdots + \sum_{h=1}^{k} \omega_h^{(Q)} \eta_{ih}^Q,$$

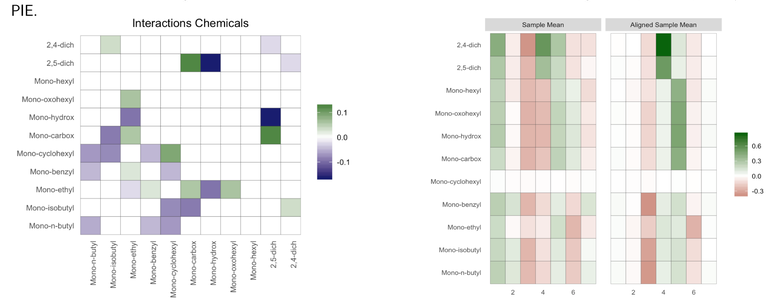
- We do not include interactions between the factors, so that the number of parameters to be estimated is Qk.
- We obtain estimates for the interaction coefficients up to the Q^{th} order with above model.

Environmental Epidemiology Analysis

- The goal of our analysis is to assess the effect of nine phthalate metabolites and two dichlorophenol pesticides on body mass index. There is a growing health concern for the association of phthalates and dichlorophenol pesticides with obesity.
- The data are taken from the National Health and Nutrition Examination Survey (NHANES) for the years 2009 and 2011.
- We select a subsample of 2239 individuals.
- We also include in the analysis cholesterol, creatinine, race, sex, education and age.



Estimated main effects using FIN with 95% credible intervals and estimated coefficients using RAMP, hierNet, Family and PIE.



Left, posterior mean of the matrix of chemicals interactions. The white boxes indicates that the 95% credible interval does not contain zero. Right, posterior mean of the matrix Λ of factor loadings before and after applying the Clustalign algorithm.

- The signs of the coefficients are to be consistent across different methods.
- The matrix of factor loadings reflects the correlation structure of the chemicals. In fact, the chemicals appear to be grouped in three families: the environmental chemicals and two groups of phthalates.
- We did not identify any significant interactions between the chemical exposures and covariates.