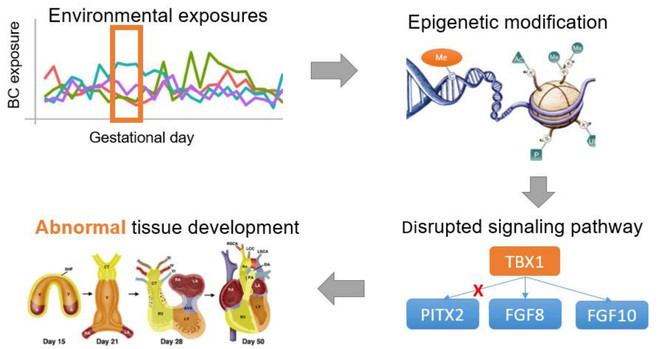
Identifying Epigenetic Regions Exhibiting Critical Windows of Susceptibility to Air Pollution

Michele Zemplenyi[1], Mark J. Meyer[2], and Brent A. Coull[1]

[1]Harvard University [2]Georgetown University

Michele Zemplenyi, Doctoral student
Harvard University Dept. of Biostatistics
mzemplenyi@g.harvard.edu

Motivation

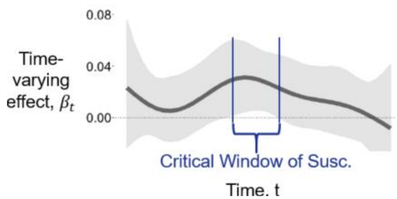


We want to pinpoint genomic regions and time periods during which epigenetic changes in those regions occur in order to understand the biological mechanisms by which air pollution may lead to adverse health outcomes.

Methodology Background

- Our goal is to identify critical windows of susceptibility: time periods during which there is an increased association between an exposure and a future outcome of interest.
- A distributed lag model (DLM) can estimate the time-varying relationship between an exposure and outcome.

$$y_i = \alpha + \sum_{t \in T} \beta_t \, x_{it} + \epsilon_i$$



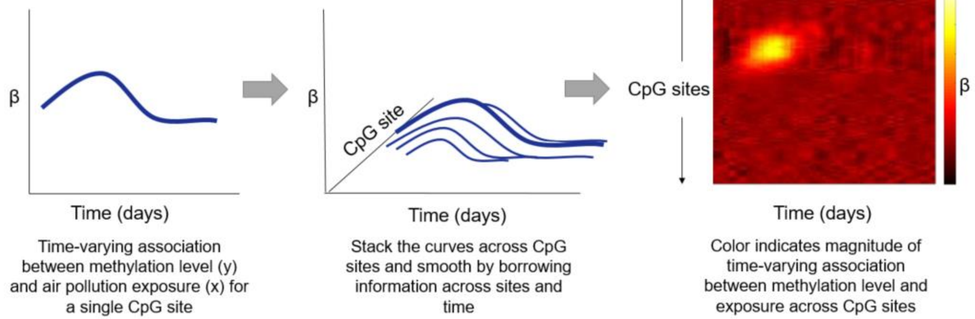
- If we let time get increasingly dense, we have the continuous analog of the DLM:

$$y_i = \alpha + \int_{t \in T} \beta(t) \, x_i(t) \, dt \, + \epsilon_i$$

Function-on-Function Regression

- While we could run the DLM for every methylation (CpG) site separately, modeling methylation profiles as functions rather than independent sites allows information to be borrowed across neighboring sites and samples, giving greater power to detect differentially methylated regions.
- We are interested in the relationship between two functions:
  1) y(s): methylation level as a function of CpG site position, s, in the genome.
  2) x(t): maternal air pollution exposure on day t of gestation.

$$y_i(s) = \alpha(s) + \int_{t \in T} \beta(t,s) \, x_i(t) \, dt \, + E_i(s), \qquad E_i(s) \sim GP(0, \Sigma)$$

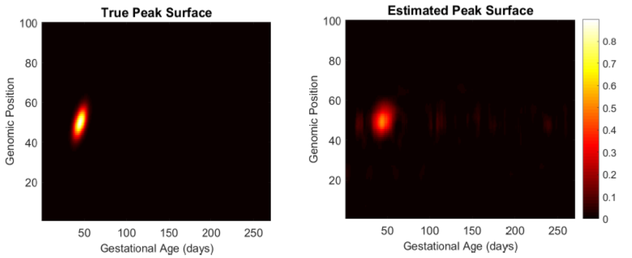


Fitting the Function-on Function Model

1. Transform both the air pollution and methylation profiles from the original data space into the wavelet space using the Discrete Wavelet Transform (DWT).
   - Wavelet regression is useful for modeling data with local features (e.g. spikes).
   - The DWT tends to concentrate signal on a small subset of wavelet coefficients.
   - We can regularize the original functions by shrinking the small wavelet coefficients towards zero.
2. Use MCMC to obtain posterior samples of the coefficient surface β* in the wavelet space.
   - Place spike-and-slab priors on the coefficients where d_0 is a point-mass at zero.

   $\beta_p^* \sim \gamma_p \, N(0, \tau_p) + (1 - \gamma_p) d_0 \qquad \gamma_p \sim Bern(\pi_p)$
   - γ_p indicates whether the wavelet coefficient represents signal or noise.
3. Transform posterior samples of β* back to the data space for estimation and inference.

Results



- We simulate a multivariate normal peak surface centered at the 50th probe and 45th exposure day with a signal-to-noise ratio of 0.5.
- We compare the function-on-function (FF) model and DLM for a positive control site (s=50, strong signal) and negative control site (s=80, zero signal).

Method	Control	MSE	Bias^2	Var.	Rel. Eff.
FF	Pos	8.51	7.46	1.05	2.10
DLM	Pos	9.28	7.08	2.20	1
FF	Neg	0.21	0.00	0.21	7.81
DLM	Neg	1.64	0.00	1.64	1

Conclusions

- The ability to borrow information across neighboring methylation sites provides efficiency gains to the function-on-function model relative to the DLM approach that models probes independently.
- The function-on-function regression framework is a powerful method for capturing complex associations between variables that may vary spatially and temporally.

Selected References

[1] Meyer, Mark J et al. (2015). "Bayesian function-on-function regression for multilevel functional data." Biometrics 71, pp.565-574.
[2] Morris, Jeffrey et al. (2006). "Wavelet-based functional mixed models." Journal of the Royal Statistical Society: Series B 68(2), p.179-199.