Estimating Perinatal Critical Windows of Susceptibility to Environmental Mixtures via Structured Bayesian Regression Tree Pairs

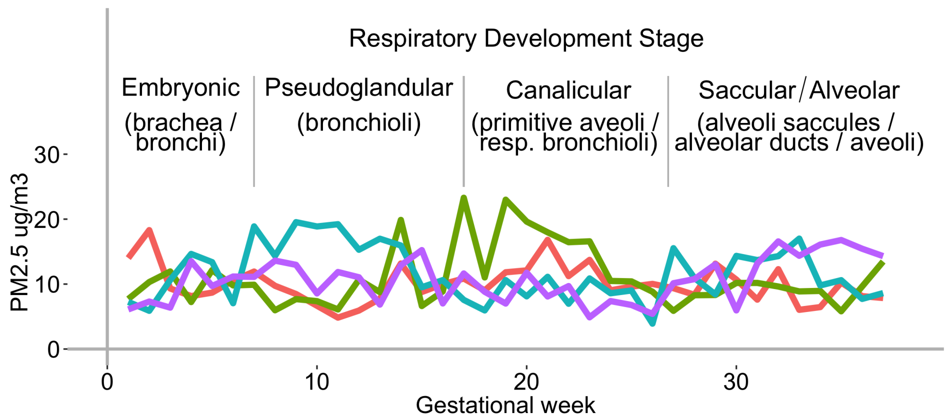
Daniel S. Mork Harvard T. H. Chan School of Public Health
Ander Wilson Colorado State University

ENAR 2022

Critical Windows of Susceptibility

Definition

A period in time during which an exposure can alter phenotype.

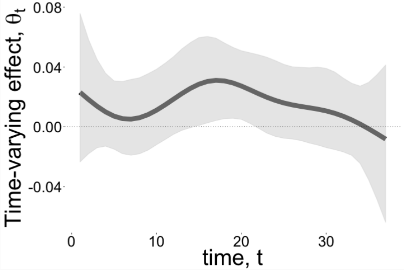


Distributed Lag Model (DLM)

$$y_i = \sum_{t=1}^{T} x_{it}\theta_t + z_i'\gamma + \varepsilon_i$$

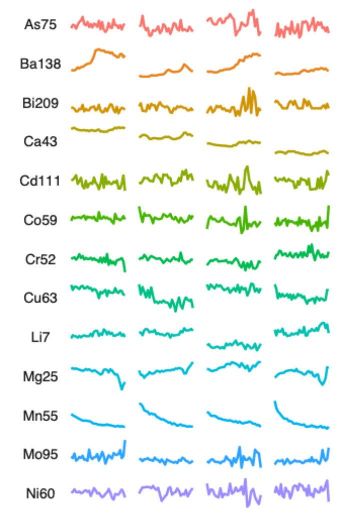
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)'$ *constrained* to vary smoothly in time (e.g. spline, Gaussian process, ...)
    - adds stability to the model
    - conforms with biological hypothesis that exposure at proximal time points are likely to have similar effects

DLM analysis of $PM_{2.5}$ and asthma among boys in the ACCESS cohort.



[1]Figure source: Wilson et al. (2017a) *Biostatistics*.

Critical Windows with Mixtures



Challenges of Mixtures Assessed at Longitudinally

- High dimensional exposure space
- High correlation between mixture components
- High autocorrelation within each component
- Nonlinear associations
- Interactions between components including time-sensitive interactions (e.g. priming)

Limitations of DLM

- Tendency to over-smooth the distributed lag function
- Lack of DLM methods for mixtures
- This talk: How to use Bayesian additive regression trees (BART) to better estimate a DLM and extend DLM to mixtures

Bayesian Additive Regression Trees (BART)

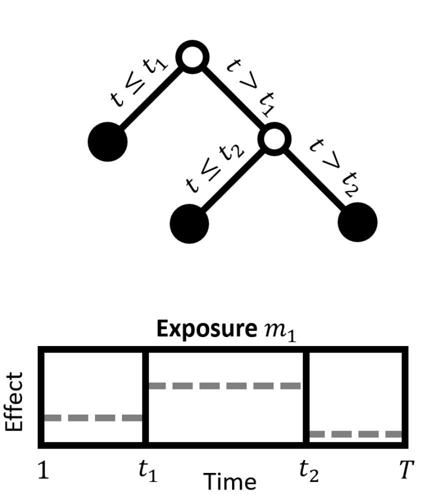
$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i$$

- Proposed by by Chipman, George, McCulloch (1998, *JASA* & 2010, *AOAS*)
- Estimate a general mean function
- State of the art predictive performance
- Allows for coherent Bayesian inference

Treed Distributed Lag Model (TDLM)

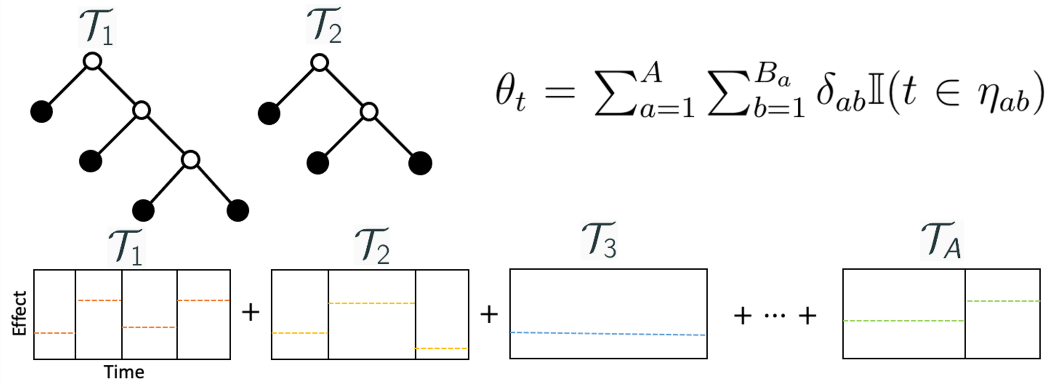
$$y_i = \sum_{t=1}^{T} x_{it}\theta_t + \mathbf{z}_i'\gamma + \varepsilon_i$$

- Apply BART to time ($t = 1, \ldots, T$) to define structure in the lag function $\theta_1, \ldots, \theta_T$
- Constant effect of exposure in each terminal node or time segment

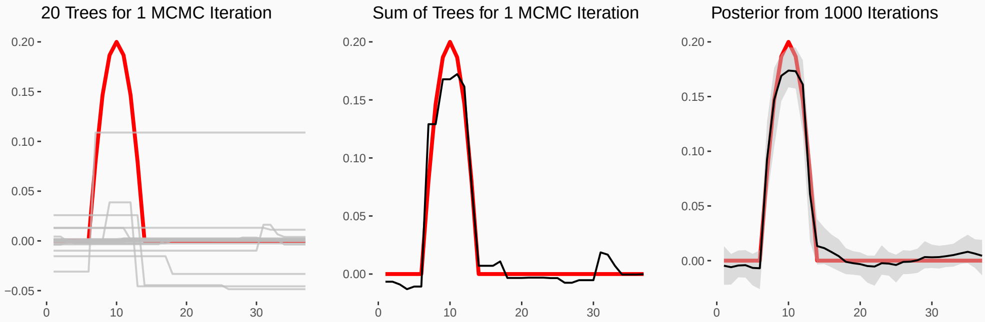


TDLM: Ensemble of Trees

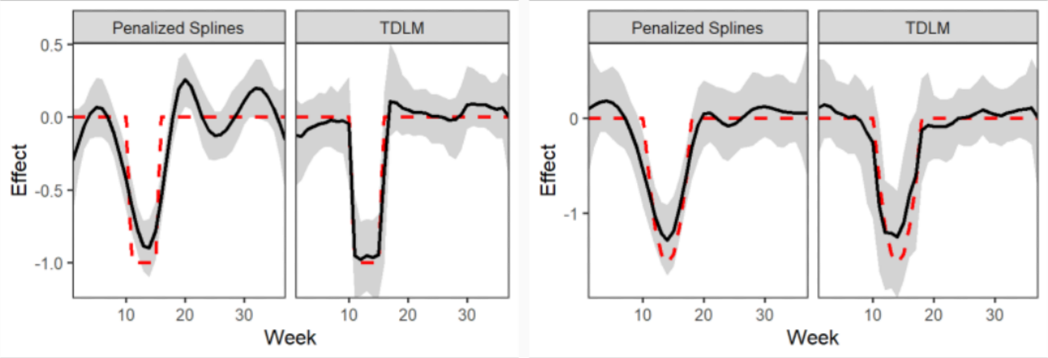
- Use ensemble of A trees
- Adds robustness and can approximate smooth distributed lag functions
- η_{ab} and δ_{ab} is the terminal node and effect for node b on tree a



TDLM: Illustrative Example



TDLM: Illustrative Example

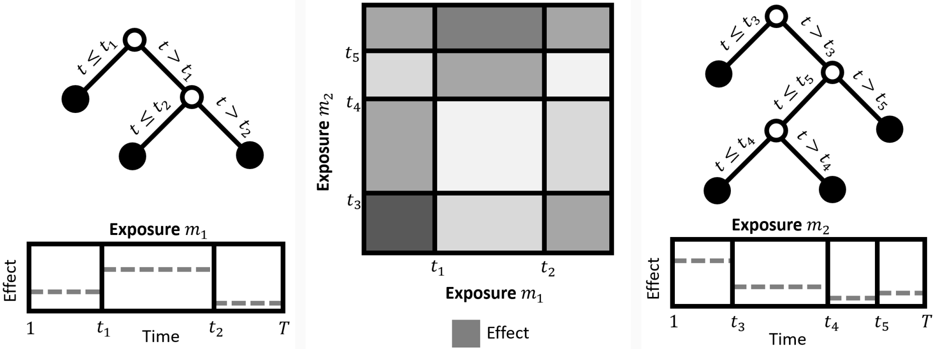


Distributed Lag Mixture Model (DLMM)

$$y_i = \sum_{m=1}^{M}\sum_{t=1}^{T} x_{imt}\theta_{mt} + \sum_{m_1=1}^{M}\sum_{m_2=m_1}^{M}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} x_{im_1t_1}x_{im_2t_2}\theta_{m_1m_2t_1t_2} + \mathbf{z}_i'\gamma + \varepsilon_i$$

- θ_{mt} is the main effect of exposure m $(m = 1, \ldots, M)$ at time t
- $\theta_{m_1m_2t_1t_2}$ is the interaction among exposures m_1 at time t_1 and m_2 at time t_2
- Includes time-sensitive interactions
- Includes quadratic main effects if we include self interactions
- $MT + \binom{M+1}{2}T^2$ parameters (20,720 in our analysis with $M = 5$ and $T = 37$)

Treed Distributed Lag Mixture Model (TDLMM)



- Structured regression tree pairs add structure to the θ's
- Tree pairs define the main effect and pairwise interaction for two exposures (or a self interaction / quadratic)

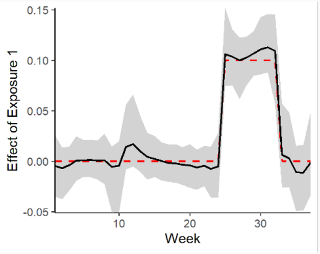
Tree Pairs & Exposure Selection

- Prior on the exposure that each tree is applied to

$$\begin{aligned} S_{aj} &= m \qquad \text{if tree } j \text{ in pair } a \text{ is applied to exposure } m \\ S_{aj}|\mathcal{E} &\sim \text{Categorical}(\mathcal{E}) \\ \mathcal{E} &\sim \text{Dirichlet}(\kappa, \ldots, \kappa) \end{aligned}$$

- New tree proposal update: switch exposure
- If no tree uses exposure m, that exposure is selected out of the model
- Enforces hierarchical variable selection

TDLM Simulation (single pollutant)

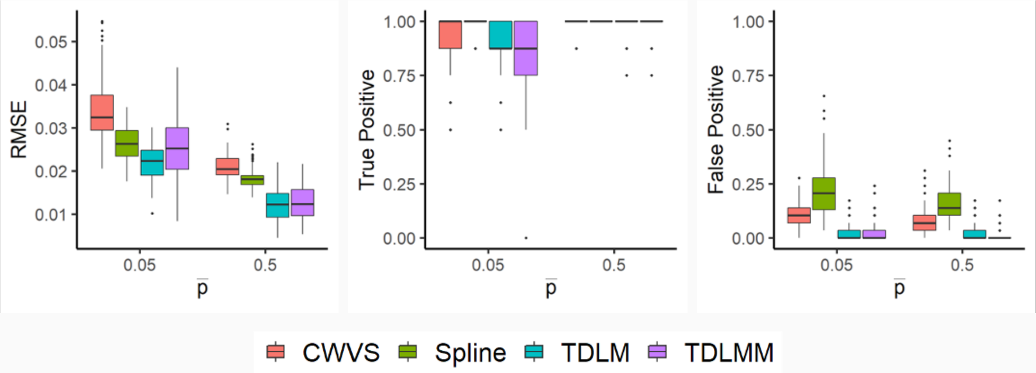


- Scenario 1: Binary outcome, single exposure
- $n = 5000$, two different average probabilities of success (0.05, 0.5)
- Randomly placed, eight-week critical window
- Real Colorado exposure data for $PM_{2.5}$
- Compare:
  - TDLM with a single exposure
  - Penalized cubic regression splines[1]
  - Critical window variable selection (CWVS)[2]
  - TDLMM with four additional exposures in mixture model (NO_2, SO_2, CO, temperature)

[1] Gasparrini et al. (2017) *Biometrics*
[2] Warren et al. (2020) *Biostatistics*

TDLM Simulation (single pollutant)

- Better distributed lag function estimation
- More accurate critical window detection
- Minimal penalty for using TDLMM when only one exposure has a true effect



TDLMM Simulation (mixture with five components)

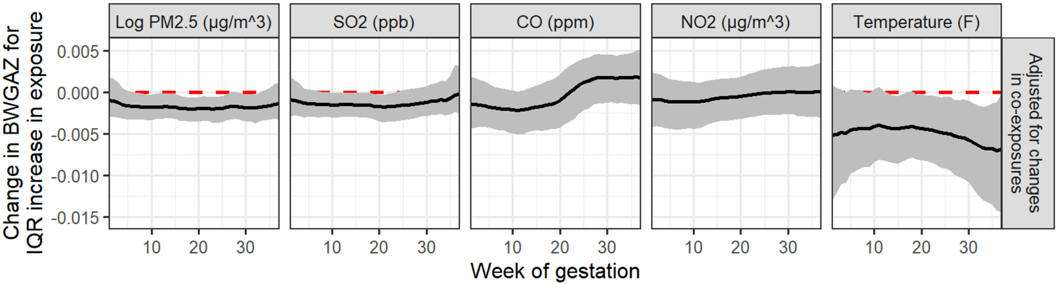
- Second simulation from a mixture with time-sensitive interactions
- Gaussian model
- Overall good performance
    - acceptable RMSE
    - proper 95% interval coverage
    - high precision identifying windows
    - high rate of selecting correct exposures and lower rate of selecting incorrect exposures

Analysis of Colorado Administrative Birth Cohort

- 195,701 full term (37 weeks) births
- Outcome: birth weight z-score (BWGAZ), adjusted for sex, gestational age
- Five exposures assessed weekly during gestation: $PM_{2.5}$, NO_2, SO_2, CO, temperature
- Controlled for: maternal age, weight, income, education, smoking, prenatal care, race, Hispanic, county, elevation, year and month of conception

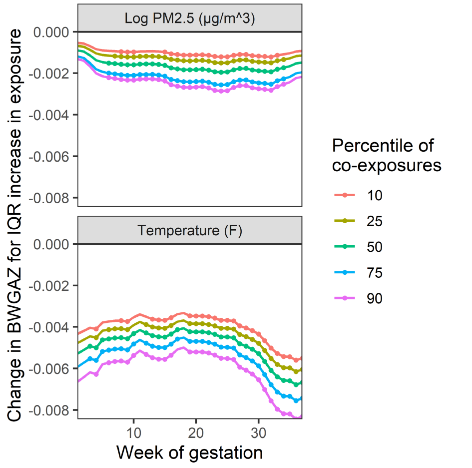
Main Effects

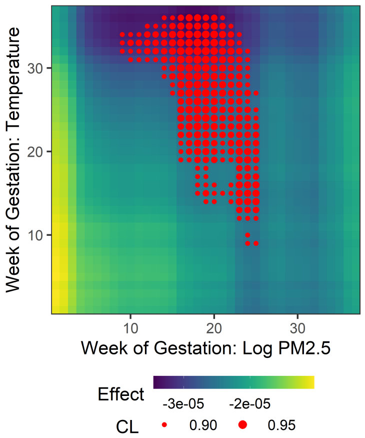
- Many "main effects"
- Here: IQR change of one exposure and the expected corresponding change in the co-exposures



$$E\left[Y\middle|\widetilde{x}_t = E\left\{x_t\middle|x_{mt} = x_{m(0.75)}\right\}, \widetilde{x}_{[t]} = \overline{x}, z = z_0\right]$$
$$- E\left[Y\middle|\widetilde{x}_t = E\left\{x_t\middle|x_{mt} = x_{m(0.25)}\right\}, \widetilde{x}_{[t]} = \overline{x}, z = z_0\right]$$

Temperature-$PM_{2.5}$ Interaction





Summary

- We can add structure to BART to get interpretable estimates of DLMs
- Allows for identifying critical windows
- Tree-pairs allows for mixtures
- Overall good finite sample properties
- Available for linear and logistic regression (zero inflated count data coming soon)
- Similar approach for heterogeneity (Mork et al. 2022, ArXiv:2109.13763)
- Treed distributed lag nonlinear model also available (Mork and Wilson 2021, *Biostatistics*)
- R code available: `github.com/danielmork/dlmtree`

Thank You

anderwilson.github.io

ander.wilson@colostate.edu

@ander_wilson

Mork, D., Wilson, A. (In press). Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs. *Biometrics*. https://arxiv.org/abs/2102.09071

NIEHS Grants: ES029943, ES028811

USEPA grant: RD-839278

Contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

References i

Chipman, HA, EI George, and RE McCulloch (2010). "BART: Bayesian additive regression trees". *The Annals of Applied Statistics* 4 (1).

Chipman, HA, EI George, and RE McCulloch (1998). "Bayesian CART Model Search". *Journal of the American Statistical Association* 93 (443).

Gasparrini, A, F Scheipl, B Armstrong, and MG Kenward (2017). "A penalized framework for distributed lag non-linear models". *Biometrics* 73 (3).

Mork, D, MA Kioumourtzoglou, M Weisskopf, BA Coull, and A Wilson (2021). "Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution". arXiv: 2109.13763.

Mork, D and A Wilson (2021a). "Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs". *Biometrics*.

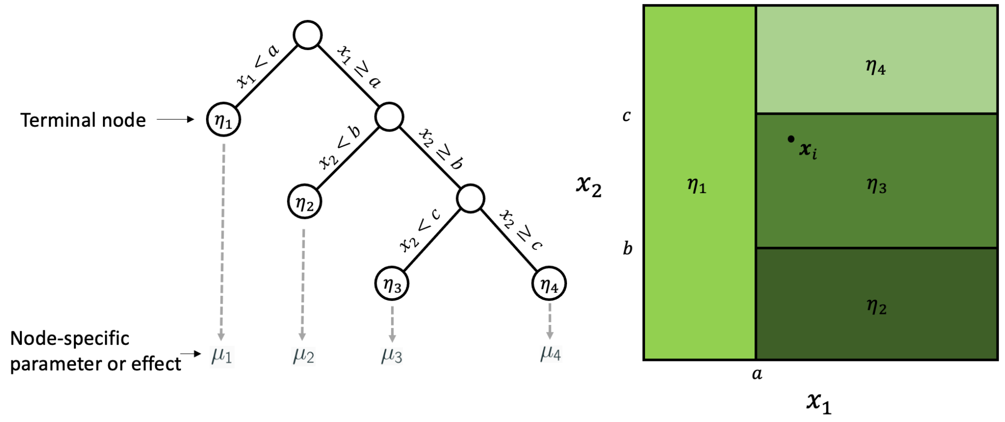
Mork, D and A Wilson (2021b). "Treed distributed lag nonlinear models". *Biostatistics*.

Warren, JL, W Kong, TJ Luben, and HH Chang (2020). "Critical window variable selection: estimating the impact of air pollution on very preterm birth". *Biostatistics* 21 (4).

BART

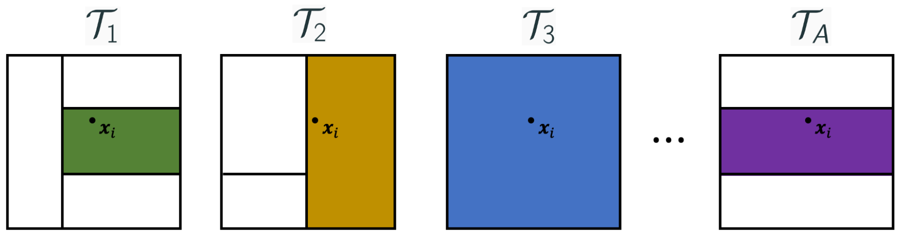
Bayesian Additive Regression Trees (BART)

$$g(\mathbf{x}_i, \mathcal{T}) = \mu_b \qquad \text{if } \mathbf{x}_i \in \eta_b$$

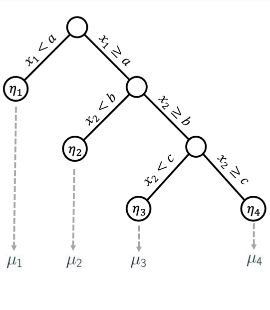


BART

$$f(x_i) = \sum_{a=1}^{A} g(x_i, \mathcal{T}_a)$$



BART Priors

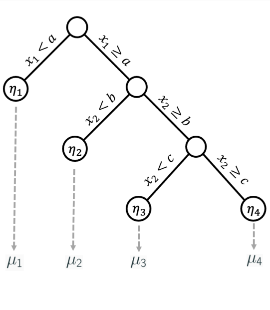


- Implicit prior based on tree generating process
- Three parts:
  - Prior that a node at tree depth d splits
  - Prior on variable that is split at a node (e.g. uniform from all variables)

$$\alpha(1+d)^{-\beta} \qquad \alpha \in (0,1),\ \beta \in [0,\infty)$$

  - Prior on a rule that splits that variable (e.g. uniform breaks in range or uniform of subgroups of categorical variables)
- Independent Gaussian priors on μs

BART Computation



- μs can be integrated out to avoid changing parameter space problem
- Bayesian backfitting updates one tree at a time with Metropolis–Hastings
- Four possible tree-update steps
  - Grow
  - Prune
  - Change splitting rule
  - Swap parent and child node order
- Update other parameters with Gibbs

TDLMM

TDLM Priors

$$
\begin{aligned}
\delta_{ab} | \tau_a^2, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \tau_a^2 \nu^2 \sigma^2) \\
\nu &\sim \mathcal{C}^+(0, 1) \\
\tau_a &\sim \mathcal{C}^+(0, 1) \\
\\
\sigma &\sim \mathcal{C}^+(0, 1) \\
\boldsymbol{\gamma} &\sim \mathcal{MVN}(\mathbf{0}, \sigma^2 c \boldsymbol{I}) \\
\\
\alpha &= 0.95,\ \beta = 2
\end{aligned}
$$

TDLMM Priors

$$
\begin{aligned}
\delta_{ajb} | \mu^2_{S_{aj}}, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \mu^2_{S_{aj}} \nu^2 \sigma^2) \qquad \text{(main effects)} \\
\mu_{S_{aj}} &\sim \mathcal{C}^+(0,1) \\
\zeta_{ab_1 b_2} | \mu^2_{S_{a1} S_{a2}}, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \mu^2_{S_{a1} S_{a2}} \nu^2 \sigma^2) \qquad \text{(interactions terms)} \\
\mu_{S_{a1} S_{a2}} &\sim \mathcal{C}^+(0,1) \\
\nu &\sim \mathcal{C}^+(0,1)
\end{aligned}
$$

$$
\begin{aligned}
\sigma &\sim \mathcal{C}^+(0,1) \\
\boldsymbol{\gamma} &\sim \mathcal{MVN}(\mathbf{0}, \sigma^2 c \boldsymbol{I})
\end{aligned}
$$

$$
\alpha = 0.95, \ \beta = 2
$$

TDLMM Computation

Key modifications to the BART MCMC algorithm:

- Integrate out fixed effect when estimating trees and distributed lag effects
- New proposal step: switch exposure, accepted with Metropolis-Hastings algorithm Simultaneous integration over all distributed lag effects during tree update
- Multivariate draw of tree terminal node and interaction parameters
- Logistic regression method for regression trees using Polya Gamma latent variable (Polson, Scott, Windle, 2013, *JASA*)
- Methods for zero inflated count data coming soon.
- Posterior analysis of tree structures, exposure, and estimates gives distributed lag effects and uncertainty