Hierarchical Additive Modeling of Nonlinear Association with Spatial Correlations

Qingzhao Yu, Bin Li, Richard Scribner

Louisiana State University, School of Public Health

Supported by: NIAAA-5 R01 AA013810-03 (PI Richard Scribner)
Outline

- Objective
- Data and Challenges
- Multivariate Additive Regression Trees
- Conditional Autoregressive Model
- Hierarchical Additive Model
- Model Convergence
- Results
The aim is to determine whether a change in the alcohol environment results in a change of assaultive violence rates in the neighborhood.

**Hypothesis:** Neighborhood alcohol outlet density is positively associated with alcohol related crimes, such as the assaultive crime.
In 1992, a civil unrest occurred over a large area of south central Los Angeles. During the civil unrest, a number of alcohol outlets were damaged, so that the alcohol density decreased in that area.

We restrict our analysis to the areas affected by civil unrest thereby controlling for a possible global effect of the unrest on outcomes.

We control for the other factors that are associated with both alcohol density and the related crime rates.
Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years’ data: 1990 to 1999.
Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years’ data: 1990 to 1999.
- Alcohol Environment: data are obtained from the PRC and California ABC.
Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years’ data: 1990 to 1999.
- Alcohol Environment: data are obtained from the PRC and California ABC.
  - outlet license surrender data following the 1992 Civil Unrest: revealing 279 outlets suspended operation following the 1992 unrest, 252 of those outlets were off-sale type outlets.
  - annual count of offsale and onsale alcohol outlets
Data (Cont’d)

- Sociodemographic Data. These data are collected through US Census and LA county department of health services for the years 1990 to 1999. The variables we used include:
  - % Black, % White, % Hispanic, % Asian, % male, % adults and % households in poverty
Sociodemographic Data. These data are collected through US Census and LA county department of health services for the years 1990 to 1999. The variables we used include:

♦ % Black, % White, % Hispanic, % Asian, % male, % adults and % households in poverty

Outcomes. The crime rates are collected from the LAPD annual police reports for the years 1990 to 1999. The data collected include:

♦ Assaultive Violence Rate: homicide, rape, robbery and assault.
♦ Complicated interactions might exist among covariates.
Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.
Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.
- More than 7% of the observations are missing one or more values for some covariates.
Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.
- More than 7% of the observations are missing one or more values for some covariates.
- We should take into account the spatial correlations among observations in adjacent tracts.
MART is a special case of the generic gradient boosting approach.
MART is a special case of the generic gradient boosting approach.

MART approximates the target function $F(x)$ by an additive expansion of trees

$$
\hat{f}(x) = \sum_{m=1}^{M} \nu b_H(x; \gamma_m).
$$
MART is a special case of the generic gradient boosting approach.

MART approximates the target function $F(x)$ by an additive expansion of trees

$$\hat{f}(x) = \sum_{m=1}^{M} \nu b_H(x; \gamma_m).$$

MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.
Why MART?

- MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.
- MART is able to capture complex and/or high order interaction effects.
Why MART?

- MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.
- MART is able to capture complex and/or high order interaction effects.
- MART can handle mixed-type predictors and missing values in covariates.
We use the vector $\{\phi_{T_i, C_i}\}$ to capture spatial autocorrelations among areas $C_i$ at time $T_i$.

Assume that an area $C_i$ is correlated with only the areas that are adjacent to it.

$y$ with a hierarchical structure on its mean function:

$$y_i \sim N(\mu_i, \sigma^2) \text{ and } \mu_i = f(x_i) + \phi_{T_i, C_i};$$  \hspace{1cm} (1)

The hierarchical CAR structure for $\{\phi_{T_i, C_i}\}$ has the form

$$\phi_{T_i, C_i} = j \mid \phi_{T_i, C_i} \neq j \sim N\left(\sum_{k \sim j} \frac{1}{n_j} \phi_{T_i, k}, \frac{1}{n_j \tau_{T_i}}\right).$$
1. Let $\phi_{T_i,C_i}^0 = 0$ where $C_i \in \{1, \ldots, C\}$, $T_i \in \{1, \ldots, T\}$; $q = 0$, $\delta = 1000$, $\mu_{1i} = 0$ and $i = 1, \ldots, n$.

2. If $\delta < \Delta$, go to step 3, otherwise $q = q + 1$ and

(a) Let $z_i = y_i - \phi_{T_i,C_i}^{[q-1]}$. Fit MART $f^{[q]}(x)$ on $z$ and the covariates $x$.

(b) Let $e_i = y_i - f^{[q]}(x_i)$, calculate the Moran’s I of $e_i$ within time slots 1 to $T$. Let $S$ be the collection of time slots in which the spatial correlation test show a p-value smaller than 0.01 (this indicates a strong spatial correlation in $e_i$).
(c) If $S$ is empty, let $\phi_{T_i, C_i}^{[q]} = 0$ and go to step 3; otherwise for the observations $i \in \{ i : T_i \in S \}$, let the $f(x_i)$ in Equation (1) be $f^{[q]}(x_i)$ and calculate the generalized MLEs of $\hat{\phi}_{T_i, C_i}$. Let $\phi_{T_i \in S, C_i}^{[q]} = \hat{\phi}_{T_i, C_i}$ and $\phi_{T_i \notin S, C_i}^{[q]} = 0$.

(d) Let $\mu_{0i} = \mu_{1i}$, $\mu_{1i} = f^{[q]}(x_i) + \phi_{T_i, C_i}^{[q]}$ and let $\delta = \frac{\sum_{i=1}^{n}(\mu_{1i} - \mu_{0i})^2}{\sum_{i=1}^{n} \mu_{1i}^2}$, go back to 2.

3. Output the results from step $q$. 

The Two-Stage Iteration Algorithm (Cont’d)
Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
- Let $\Delta = 10^{-7}$ in our analysis.
Figure 1: Relative variable importance in MART-fitted models.
Figure 2: *Partial dependence plots.*
## Spatial Correlation

<table>
<thead>
<tr>
<th>Year</th>
<th>Origin</th>
<th>P-value</th>
<th>Res1</th>
<th>P-value</th>
<th>Res2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>0.46</td>
<td>&lt; 0.0001</td>
<td>0.13</td>
<td>1.23e-04</td>
<td>-0.09</td>
<td>0.017</td>
</tr>
<tr>
<td>1998</td>
<td>0.52</td>
<td>&lt; 0.0001</td>
<td>0.29</td>
<td>0.00e-00</td>
<td>-0.04</td>
<td>0.277</td>
</tr>
<tr>
<td>1997</td>
<td>0.50</td>
<td>&lt; 0.0001</td>
<td>0.25</td>
<td>1.46e-12</td>
<td>-0.02</td>
<td>0.688</td>
</tr>
<tr>
<td>1996</td>
<td>0.29</td>
<td>&lt; 0.0001</td>
<td>0.05</td>
<td>4.78e-02</td>
<td>0.05</td>
<td>0.048</td>
</tr>
<tr>
<td>1995</td>
<td>0.49</td>
<td>&lt; 0.0001</td>
<td>0.23</td>
<td>2.35e-11</td>
<td>-0.01</td>
<td>0.484</td>
</tr>
<tr>
<td>1994</td>
<td>0.49</td>
<td>&lt; 0.0001</td>
<td>0.24</td>
<td>4.98e-12</td>
<td>-0.04</td>
<td>0.335</td>
</tr>
<tr>
<td>1993</td>
<td>0.52</td>
<td>&lt; 0.0001</td>
<td>0.21</td>
<td>4.83e-09</td>
<td>-0.01</td>
<td>0.825</td>
</tr>
<tr>
<td>1992</td>
<td>0.30</td>
<td>&lt; 0.0001</td>
<td>0.14</td>
<td>1.77e-07</td>
<td>-0.16</td>
<td>0.000</td>
</tr>
<tr>
<td>1991</td>
<td>0.18</td>
<td>&lt; 0.0001</td>
<td>0.02</td>
<td>4.43e-01</td>
<td>0.02</td>
<td>0.443</td>
</tr>
<tr>
<td>1990</td>
<td>0.44</td>
<td>&lt; 0.0001</td>
<td>0.16</td>
<td>1.26e-06</td>
<td>-0.11</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Maps of Log Assault Rates

values for loghatrate98

values for logmartrate98

values for spatialerror98

values for randomerror98
Questions