Bayesian nonlinear regression for the air pollution effects on daily clinic visits in small areas of Taiwan

Atanu Biswas
Indian Statistical Institute, Kolkata
atanu@isical.ac.in

Joint work with Jean-François Angers and Jing-Shiang Hwang
Outline

- Air pollution and health effects
- Spatiotemporal data
- Two-stage Bayesian hierarchical models
- Bayesian nonlinear regression models
- Preliminary results
- Discussions
Epidemiological studies

- Short-term studies for *acute effects*
  - Ecological studies: (Time series analysis)
    - Examine the associations between day-to-day changes in air pollution levels and routinely measured health outcomes such as emergency room visit, hospital admission and mortality *in an area*
  - Most commonly used approach
  - Usually not possible to infer causality
Problems of epidemiological studies

- Two common features of ecological studies
  1. Mainly carried out in a single large city
  2. Aggregated data in the large area are used to represent population exposures.
- Misclassification is often compounded.
Solution: Small area design

- Cluster clinics around a monitoring station to create relatively homogeneous area of size about 20 km².
  - **Population at risk** of each area is estimated by the service coverage of all clinics in that area.
  - **Population exposure** is represented by measurements from the monitoring station.
  - **Health outcome** is the daily clinic visits for minor lower respiratory illness in a designated area, which are not sparse.

The data

- The study communities: 50 townships and city districts across Taiwan
  - Include rural, urban and industrial areas
  - Population densities range from 250 to 28,000 persons/km²
  - The distribution of these ambient air quality monitoring stations covers the whole island (96% of the total population)
Study communities include 50 townships and city districts where ambient air monitoring stations of Taiwan Air Quality Monitoring Network (TAQMN) are situated.
Taiwan Ambient Air Quality Network
The data

- Environmental variables from EPA
  - Daily average for NO$_2$, SO$_2$ and PM$_{10}$
  - Daily maximum O$_3$ and maximum 8-hour running average for CO
  - Daily maximum temperature, average dew point, wind direction, wind speed, precipitation.
  - We include clinic visits due to respiratory diseases as outcomes, which are acute bronchitis, acute bronchiolitis, and pneumonia.
Statistical analyses

- Estimate population at risk for each area and convert daily clinic visit counts to daily rates.

- Phase I: Use linear models to model temporal patterns in order to obtain estimated acute pollution-health effect for each area.

- Phase II: Use Bayesian hierarchical models to combine the estimated pollution-health effects across the 50 communities.
Result summary
Alternative analysis

- Consider a regression spline model for analyzing the data for each small area (He and Shi, 1998)
- Assume a conjugate prior and obtain the Bayesian predictive values
- Choose the appropriate model using Bayesian model averaging
- Use Occam’s Razor principle (Madigan and Raftery, 1994)
Notation for data

- Let $y_{jt}$ be the daily clinic visit rate for the $j^{th}$ area and the $t^{th}$ day, $t=1,2,\ldots,T$
- Let $x_{clt}$ be the $l^{th}$ available covariate for the $c^{th}$ area and the $t^{th}$ day
- Let $w_{ht}$ be the $h^{th}$ available covariate for the $t^{th}$ day and for all areas, $h=1,2,\ldots,H$
- Even though there are three groups (children, adult and elderly), we only consider the 3rd gr.
Frequentist model

- **Regression spline (of order P) Model $M_{cl}$**
  
The model for the $j^{th}$ area using the $l^{th}$ covariate at $c^{th}$ area $X_{cl}$ and other covariate $W_h$
  
$$v_{jt} = \log(y_{jt}) = \sum_{p=0}^{P} \beta_{clp} u_{clt}^p + \sum_{k=1}^{K} \eta_{clk} (u_{clt} - z_{clk})_+^p + \sum_{h=1}^{H} \gamma_h W_{ht} + \varepsilon_{jt}$$

$t = 1, 2, \ldots, T$, and $\varepsilon_{jt} \sim N(0, \sigma^2)$

$u_{clt} = (x_{clt} - \min_t \{x_{clt}\}) / (\max_t \{x_{clt}\} - \min_t \{x_{clt}\})$

$z_{clk}$ are knots to be chosen.
Frequentist model

\[ \theta_{cl} = (\beta_{cl0}, \ldots, \beta_{clP}, \ldots, \eta_{cl1}, \ldots, \eta_{clK}, \gamma_1, \ldots, \gamma_H)' \]

\[ X_{cl} = \begin{pmatrix} 1 & u_{cl1} & \cdots & (u_{cl1} - z_{cl1})_+ & \cdots & (u_{cl1} - z_{clK})_+ & W_{11} & \cdots & W_{H1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 1 & u_{clT} & \cdots & (u_{clT} - z_{cl1})_+ & \cdots & (u_{clT} - z_{clK})_+ & W_{1T} & \cdots & W_{HT} \end{pmatrix} \]

The Model \( M_{cl} \)

\[ V_j = X_{cl} \theta_{cl} + \varepsilon \]

\[ V_j = (v_{j1}, \ldots, v_{jT})' \quad \varepsilon \sim N_T(0, \sigma^2 I_{P+K+H+1}) \]
Prior distribution

- Conjugate prior

\[
\theta_{cl} \mid \sigma^2 \sim N_{P+K+H+1}(\theta_{cl0}, \frac{\sigma^2}{n_0}(X'_{cl}X_{cl})^{-1})
\]

\[
\sigma^2 \sim \text{Inv}\Gamma(\alpha/2, \gamma/2)
\]

- Parameter values

\[
\theta_{cl0} = (T^{-1}\sum_{t=1}^{T} \nu_{jt}, 0, \ldots, 0),
\]

\[
\alpha = \gamma = 0, \ n_0 = 0.1
\]
Posterior and marginal

- Lindley and Smith, 1972

\[ \theta_{cl} \mid V_{j}, \sigma^2 \sim N_{P+K+H+1}(\theta_{cl}^{(LS)} - \frac{n_0}{n_0 + 1}(\theta_{cl}^{(LS)} - \theta_{cl0}), \frac{\sigma^2}{n_0 + 1}(X_{cl}'X_{cl})^{-1}) \]

\[ \sigma^2 \mid V_{j} \sim \Gamma(\frac{T + \alpha}{2}, \frac{\gamma_{*}}{2}) \]

\[ \theta_{cl}^{(LS)} \sim T_{P+K+H+1}(T + \alpha - P - K - H - 1, \theta_{cl0}, \frac{S + \gamma}{T + \alpha - P - K - H - 1}) A_{*} \]

where

\[ \theta_{cl}^{(LS)} = (X_{cl}'X_{cl})^{-1}X_{cl}'V_{j} \]

\[ \gamma_{*} = S + \gamma + (n_0 + 1)(\theta_{cl}^{(LS)} - \theta_{cl0})'X_{cl}'X_{cl}(\theta_{cl}^{(LS)} - \theta_{cl0}) \]

\[ S = (V_{j} - X_{cl}\theta_{cl}^{(LS)})'(V_{j} - X_{cl}\theta_{cl}^{(LS)}) \]

\[ A_{*} = \frac{n_0 + 1}{n_0}(X_{cl}'X_{cl})^{-1} \]
Bayesian estimator

- Bayesian estimator of $\theta_{cl}$

$$\hat{\theta}_{cl} = \theta_{cl}^{(LS)} - \frac{n_0}{n_0 + 1} (\theta_{cl}^{(LS)} - \theta_{cl0})$$

- Bayesian predictor for $V_j$ using model $M_{cl}$

$$\hat{V}_{jcl} = X_{cl} \hat{\theta}_{cl}$$
Several methods in the literature to select “optimal model”

However, the standard criteria like BIC or AIC are too conservative and they often lead to model using only one covariate.
Model averaging

- Let \( m_{cl}(V_j) \) denote the marginal density.
- Occam’s Razor principle (Madigan and Raftery, 1994)
- Consider models in the set

\[
A_d = \left\{ M_{cl} \mid \max_{ab} \frac{m_{ab}(V_j)}{m_{cl}(V_j)} \leq d \right\}
\]

- Let

\[
B_d = \{(c, l) \mid M_{cl} \in A_d \}
\]
Model averaging

- Using Bayesian averaging, the predictor for $V_j$ is given by

$$
\hat{V}_{(d)j} = \sum_{(c, l) \in B_d} p_{c, l} \hat{V}_{jcl}
$$

where

$$
p_{c, l} = \frac{m_{cl}(V_j)}{\sum_{(a, b) \in B_d} m_{ab}(V_j)}
$$
The relative ratios of predicated clinic visit rates with measured pollutant concentrations at Site 14 to the predicated clinic visit rate with minimum concentration.
The relative ratios of predicated clinic visit rates with measured pollutant concentrations at Site 42 to the predicated clinic visit rate with minimum concentration.
The relative ratios of predicated clinic visit rates with measured pollutant concentrations at Site 55 to the predicated clinic visit rate with minimum concentration.
The relative ratios of predicated clinic visit rates with measured pollutant concentrations at Site 60 to the predicated clinic visit rate with minimum concentration.
Identified covariates associated with clinic visits of the elderly
Discussions

- NO$_2$ was associated with clinic visit rates of respiratory illness of the elderly in most of areas in Taiwan.
- PM$_{10}$ and O$_3$ identified in central and southern Taiwan
- CO and SO$_2$
- Temperature in northern Taiwan
THANK YOU