POLITECNICO
MILANO 1863

Bayesian Statistics, 2022-23
Prof. Alessandra Gugleilmi, Tutor Argiento

ANALYSING THE HEALTH EFFECTS OF SIMULTANEOUS EXPOSURE TO PHYSICAL AND CHEMICAL PROPERTIES OF AIRBORNE PARTICLES

Summary of the proposed paper
14th October 2022

Pietro Bogani, 10622041, pietro.bogani@mail.polimi.it
Paolo Botta, 10612869, paolo.botta@mail.polimi.it
Silvia Caresana, 10630163, silvia.caresana@mail.polimi.it
Romeo Carrara, 10616603, romeo.carrara@mail.polimi.it
Gabriele Corbo, 10629702, gabriele.corbo@mail.polimi.it
Luca Mainini, 10576440, luca1.mainini@mail.polimi.it

Abstract

This paper summarize the article "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles"[1] for the course of Bayesian Statistics in the Msc of Mathematical Engineering at Politecnico di Milano.

1 Introduction

In this paper, we present a Bayesian approach that can tackle the problem of estimation of how simultaneous exposure to air particles affects the risk of adverse health responsewithin the framework of time series analysis. More recently, the evidence derived from studies of long- and short-term exposure has been judged sufficient to infer causality for fine particles. Air pollution exists, however, as a heterogeneous mix of different compounds. To gain better insight into the features of air pollution mixtures and their effect, there is a consequent need to explore new statistical methods able to integrate standard methodological tools for a better understanding of these complex systems. Previous temporal clustering analyses have been successfully applied in air pollution exposure assessment, involving mainly heuristic methods such as agglomerative hierarchical clustering ecc... Despite the increasing popularity of these methods, they have some well known drawbacks. First, they do not allow an assessment of the statistical properties of the solutions provided, for example they do not provide an assessment of clustering uncertainties. Moreover, because these methods are based on similarity/dissimilarity measures between objects that are essentially described in terms of distance (e.g., Euclidean distance), they require that the time series of each pollutant has exactly the same dimensionality (i.e., they do not allow the inclusion of records which have missing data). This can represent a limitation when applied to air pollution monitoring data. Mixture models have been proposed as an alternative to heuristic clustering techniques. Generally, model-based clustering methods are based on the idea that the data follow a finite mixture of probability distributions such that each component distribution represents a cluster. A long-standing issue that finite mixture models share with many traditional clustering methods (e.g., k-means), is the a priori determination of the number of clusters. The authors adopt a Bayesian nonparametric modelling approach, where the number of mixture components is not fixed in advance, but is determined by the model and the data. These models can be implemented using a Dirichlet process. The support of the DP is restricted to discrete distributions and this results in a clustering effect that avoids the selection of a pre-defined number of clusters. In this paper we propose an approach within the Bayesian paradigm to analyse the impact of multiple particle metrics on daily mortality, using the DP mixture model. Specifically, we provide a model that addresses, in a one-step procedure, both dimensionality reduction and regression. The model, known as profile regression, performs a Bayesian clustering of the covariates by identifying exposure profiles and, simultaneously, links these to a response variable in non-parametric form (even though the model continues to be parametric within clusters). In this paper we extend this technique to analyse time series data, accounting for their typical features like trends, seasonality and temporal components through smooth functions.

The resulting probabilistic solution groups time points with similar multipollutant and response profiles.

2 Analysis

2.1 Description of the data

Results from an epidemiological time series study examining the effect of different metrics of particulate collected in London, on cardiorespiratory hospital admission and mortality using univariate log-linear Poisson models. We selected a subset of exposure data for the period January 2002 to December 2005. Information about a lot of chimical elemnts in the PM_10 particle that i did not find relevant to summarize. Typically, time series studies of mortality and morbidity control for long-term trends, seasonality, and time-varying factors, including climatology, which can potentially confound the association between an adverse health effect and polluted air. In our model, calendar time and temperature were considered as confounding variables and assumed to potentially influence the response variable via smooth functions. We transformed the original measurements as:

$$z_{t,p} = \frac{(x_{t,p} - Median(x_p))}{(Median(|x_{t,p} - Median(x_p)|))}$$

The estimated regression coefficients were obtained fitting separate univariate log-linear Poisson models.

2.2 Profile regression model for time series of multiple particles and health events

Denote by $t = 1, \ldots, T$ a series of temporal points. Let the data consist of realizations of a response data vector $y = (y_1, \ldots, y_T)$, a set of (normalised) covariates (i.e., predictors) $z_{t,p}, p = 1, \ldots, P$, and a collection. of confounding factors $u_{t,h}, h = 1, \ldots, H$. In our study, y_t denotes the count number of deaths for respiratory diseases on day t, $z_t = (z_{t,1}, \ldots, z_{t,P})'$ represents a daily covariate profile of air particles, and $u_t = (u_{t,1} \ldots, u_{t,H})'$ is a B-spline basis matrix for natural cubic splines of calendar time and temperature.

We assumed a joint probability model for the data, which takes the following form:

$$p\left(y_t, z_t \mid \theta, u_t\right) = \sum_{k=1}^{\infty} w_k p\left(y_t \mid \theta_k, \Theta_0, u_t\right) p\left(z_t \mid \theta_k, \theta_0\right)$$

where w_k are the mixture probabilities satisfying $\sum_{k=1}^{\infty} w_k = 1$ almost surely and indicating the probability of belonging to the k th component. Θ denotes the collection of model parameters, that includes component specific parameters θ_k and global parameters θ_0, that is, $\theta = (\theta_k, \theta_0)$. The inference for such mixture models can be simplified by introducing latent variables that indicate the group memberships of objects (i.e., the cluster to which day t belongs to). We define these latent group labels as: $g = (g_1, \ldots, g_T)$, such that $p\left(g_t = k\right) = w_k$. Thus, g_t is chosen using a multinomial distribution parameterised by the mixing probabilities, $g_t \mid w \sim \text{Multinomial}(w)$.

Rather than specifying a parametric distribution for the mixture probabilities, w_k, we modelled them as unknown quantities to be estimated by the data. Specifically, we assumed that w_k are generated using a stick-breaking representation of the DP. The mixture probabilities break the stick into a potentially infinite number of pieces, such that $\sum_{k=1}^{\infty} w_k = 1$. The first mixture probability is equal to V_1, i.e., $w_1 = V_1$, where $V_1 \sim \text{Beta}(1, \alpha)$ and for $k \geq 2$ the k-th mixture probabilities are given by $V_k \prod_{i=1}^{k-1} (1 - V_i)$. We used a Gamma distribution to specify prior uncertainty for the precision parameter of DP, namely $\alpha - \text{Gamma}(a, b)$, where $a = 2$ and $b = 1$ are the shape and the inverse-scale (rate) parameter respectively.

We assumed a multivariate normal distribution for the P covariates:

$$p\left(z_t \mid \theta_k, \theta_0\right) = (2\pi)^{-\frac{p}{2}} \left|\Sigma_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(z_t - m_k\right)' \Sigma_k^{-1} \left(z_t - m_k\right)\right\}$$

where $m_k = (m_{k,1}, \ldots, m_{k,P})$ is the mean vector for component k (i.e., location parameters), and Σ_k is the $P \times P$ symmetric positivedefinite variance-covariance matrix. We specified hyperpriors for m_k and Σ_k similar to Molitor et al. (2011), adopting an empirical Bayesian approach. We assumed a normal distribution for the location parameters, that is, $m_k \sim \text{N}\left(m_0, \Sigma_0\right)$ (with m_0 equal to the empirical mean of each covariate, and Σ_0 having a diagonal structure with elements equal to the square of empirical range of each covariate). We specified a Wishart distribution for the precision matrix $Q_k = \Sigma_k^{-1}$ (i.e., inverse variance-covariance matrix), that is, $Q_k \sim W(\Phi, \nu)$, where Φ is a symmetric (non-singular) matrix parameter (set equal to the inverse of the empirical variance multiplied by $1/P$) and v is the degrees of freedom parameter (set equal to P). The response was modelled as a Poisson:

$$p\left(y_t \mid \theta_k, \theta_0, u_t\right) = \frac{\lambda_t^{y_t}}{y_t!} \exp\left(-\lambda_t\right)$$

where

$$\lambda_t = E_t \exp\left(\mu_t\right)$$

and

$$\mu_t = \mu_k + \sum_{h=1}^{H} f_h\left(u_{t,h}, df_h\right) + \varepsilon_t$$

assuming ε_t be normal distributed with zero mean and variance σ_ε^2. Here μ_t is the mean response for day t and E_t is the expected offset given by the average number of deaths for respiratory diseases in the full period in study. The parameter of interest is μ_k, which represents the log relative risk for the outcome of interest associated with the k th cluster, were each cluster includes days with similar multipollutant profile.

2.3 Predictions

- We compared two predictive scenarios based on: (i) concentrations of particles measured in 2005, and (ii) concentration of the same particles

measured in 2012, to analyse any changes in respiratory mortality arising from the combined effects of local, city, national and EU policies to manage air pollution in interval of seven years period.

- We partitioned the four years time series, using the data collected in 2002–2004 as training sample and the data in 2005 as validation sample. We predicted the respiratory deaths for the 2005 and we compared the validation predictions with the actual observations.
- We then used the full time series data and computed the posterior predictive distribution of the count of respiratory-related deaths in 2012, and we compared this with the one computed for the year 2005.
- We quantified an average reduction in mortality attributable to the decrement of the ambient air particles analysing the distribution of the percent change between the two years.

Slice sampling methods was run for 70,000 iterations with the first 20,000 discarded as burn-in. Using 1 in 10 thinning, this gave us a total of 5000 draws from the posterior distribution of parameters and predictions. Finally, we performed a sensitivity analysis with respect to changes in the prior for the DP precision parameter, , a hyperparameter that influences the number of clusters (i.e., mixture components).

2.4 Post-processing

At each iteration of the sampler, we recorded a $T \times T$ score matrix with (i, j) th elements set equal to 1 if day i and day j belong to the same cluster and 0 otherwise. The end of this process leads to a probability matrix, S, formed by averaging the score matrices obtained at each iteration, thus element $S_{i,j}$ denotes the probability that day i and j are assigned to the same cluster. A model averaging approach was adopted to evaluate the uncertainty related to the characteristics of the clusters that involved running through the MCMC run, obtaining an average value for the model parameters (effects and cluster related parameters) across all days in a certain cluster.

3 Results

The representative clustering separated the days into three main clusters, which included respectively 1156, 63 and 242 days. Compared to clusters 1 and 3, cluster 2 had larger posterior errors as the number of days included was lower. The risk of mortality for respiratory diseases varied according to these cluster profiles. Cluster 1 was characterised by low posterior estimates for most of the particles (except chloride), and had the lowest risk of mortality when compared to the average mortality in 2002-2005. Cluster 2 was characterised by

low posterior estimates of inorganic anions and secondary particles and higher posteriors for primary emissions and included mainly winter days. Finally, cluster 3 was dominated by secondary aerosol, especially nitrate and sulphate, with high posteriors of non-primary airborne particles, including mainly spring and autumn days. The large decrease in PNC was most likely due to a decrease in the sulphur content of diesel in 2008 which also contributed to decreased sulphate concentrations. Based on the observed number of deaths for respiratory-related diseases which occurred in 2005, we would expect an average reduction in mortality of approximately 270 subjects. The results essentially confirmed the reliability of the three representative clusters obtained in the post-processing. The diagnosis performed setting different starting points in the number of clusters in the initialization of the model, showed the consistency of the results

4 Conclusions

A clear benefit of our model is the simultaneous estimation of the contribution of all pollutants to the mortality risk. First, it is able to address the challenging question of uncertainty in the cluster assignment. In our application we found that the uncertainty associated with the partitioning of the days to clusters was quite low, and this supports the use of the partitioning around medoids method on the posterior dissimilarity matrix to obtain a representative partition. These issues are avoided by instead looking at mixtures. As an illustration of this approach we estimated the changes in health response from changes in pollution concentrations in all 12 exposure variables measured in our data set. Between 2005 and 2012 we predicted a decrease in annual respiratory mortality of 3.51% (95% CI: 0.12%, 5.74%) in London. we found that cluster membership seemed to be an effect modifier in health effects analysis, denoting pollutant mixtures that could be targeted as part of air quality control strategy for health. In population-based time series studies, individual risk factors (age, diet, smoking etc.) are unlikely to be confounders as they do not vary temporally with air pollution over relatively short-term periods

References

[1] Monica Pirani et al. 'Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles'. In: *Environment international* 79 (2015), pp. 56–64.