Computational Statistics and Data Analysis 95 (2016) 161–175

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A time dependent Bayesian nonparametric model for air quality analysis

Luis Gutiérrez [a,*], Ramsés H. Mena [b], Matteo Ruggiero [c]

[a] Escuela de Salud Pública, Fac. Medicina, Universidad de Chile, Chile
[b] IIMAS-UNAM, Mexico
[c] University of Torino & Collegio Carlo Alberto, Italy

ARTICLE INFO

Article history:
Received 6 July 2014
Received in revised form 14 September 2015
Accepted 1 October 2015
Available online 22 October 2015

Keywords:
Dirichlet process
Density estimation
Dependent process
Stick-breaking construction
Particulate matter

ABSTRACT

Air quality monitoring is based on pollutants concentration levels, typically recorded in metropolitan areas. These exhibit spatial and temporal dependence as well as seasonality trends, and their analysis demands flexible and robust statistical models. Here we propose to model the measurements of particulate matter, composed by atmospheric carcinogenic agents, by means of a Bayesian nonparametric dynamic model which accommodates the dependence structures present in the data and allows for fast and efficient posterior computation. Lead by the need to infer the probability of threshold crossing at arbitrary time points, crucial in contingency decision making, we apply the model to the time-varying density estimation for a $PM_{2.5}$ dataset collected in Santiago, Chile, and analyze various other quantities of interest derived from the estimate.

1. Introduction

Human exposure to high levels of hazardous air pollutants has long been known to have adverse health effects. In October 2013, the International Agency for Research on Cancer, the specialized cancer agency of the World Health Organization, has announced that outdoor air pollution has been formally classified in Group 1, meaning that there is sufficient evidence that the agent is carcinogenic to humans (see IARC, 2013). Urban air pollution is estimated to cause worldwide 9% of lung cancer deaths, 5% of cardiopulmonary deaths, 1% of respiratory infection deaths, and to increase the risk of bladder cancer. Air pollutants associated with health risks include sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), tropospheric ozone (O_3) and particulate matter (PM). The latter is constituted by solid and liquid particles emitted by combustion engines and households heating or formed as residual from other products such as vehicles tyres, brakes and road pavement among other sources. Particulate matter was evaluated separately by the International Agency for Research on Cancer and also classified as carcinogenic to humans. Of particular concern, due to their harmful character, are the levels of PM_{10} and $PM_{2.5}$, that is particles with diameter smaller than 10 and 2.5 micrometers respectively, which can reach the lungs through inhalation and even get to the inner organs, where they settle and become cause of serious health problems, including death (Préndez, 1993; Dockery et al., 1992).

The awareness of the hazard caused by air pollution, and PM in particular, generates an emerging concern in overpopulated metropolitan areas around the world. As a response, environmental authorities have set forth a series of actions to reduce and control the levels of pollutants as well as to set policies to enact alerts propitiously. One of the

* Corresponding author.
E-mail address: luisgutierrez@med.uchile.cl (L. Gutiérrez).

http://dx.doi.org/10.1016/j.csda.2015.10.002
0167-9473/

main actions by environmental authorities was the establishment of monitoring networks that collect information about concentrations of different pollutants. These sources of information generate spatially and temporally dependent data, which present asymmetries, heavy tails and sometimes multi-modality. For these reasons, robust and flexible statistical models are needed to accommodate this kind of phenomena with time-varying distributions. Indeed, statistical models constitute a key aspect when elaborating policies to decrease pollution levels (WHO, 2011), to set appropriate thresholds for emission contingencies and similar actions.

The spatial and temporal nature of pollutant measurements rise many statistical questions. One of the most important is a reliable assessment of the probability that a given pollutant concentration is or will be above a given threshold; such information is then used by the environmental authorities in order to call for environmental alerts and initiate policies for decreasing the pollutants levels. Various methods have been used in order to provide answers to statistical questions in air pollution research. These include extreme value theory (Roberts, 1979a,b; Horowitz, 1980; Smith, 1989; Davison and Smith, 1990), multivariate analysis (Guardani et al., 2003), neural networks (Comrie, 1997; Guardani et al., 1999; Pérez et al., 2000; Ordieres et al., 2005), Poisson models (Raftery, 1989; Achcar et al., 2008, 2010) and time series and spatial statistics (Draghicescu and Ignaccolo, 2009) among others. However, most of these either rely on parametric assumptions, such as symmetry or uni-modality on the pollutant level distribution or ignore the temporal dependence inherent to pollution data. The data structure resulting from air pollutant measurements cannot be robustly captured by time-dependent parametric models as the temporal dimension can change structural features of the involved distributions, such as the number of modes, the shape of tails and so on.

These type of datasets instead require models with enough generality to avoid unnecessary prior constraints in the estimation, flexibility to account for particular structures in the data, and relative computational simplicity to avoid excessive algorithmic effort in the presence of multivariate data.

In this paper we develop a flexible, nonparametric model, suitable to analyze multivariate air pollution data which exhibit asymmetries, multi-modality and spatio-temporal dependence. We provide illustrations of the proposed model with an air quality analysis of the city of Santiago, Chile, where pollution contingencies have recently been enforced as a consequence of long periods during which the daily standard threshold of 50 $\mu g/m^3$ of $PM_{2.5}$ has been surpassed. The guiding question will be to estimate the probability that a given pollutant concentration is above a given threshold δ_0 at time t. However, our aim will be rather general, in that the object of inference will be the entire shape of the time dependent data generating distribution. From the time-varying density estimate, other quantities of interest can be derived, such as the mean functional or the probability of exceeding an arbitrary threshold. We emphasize that, unlike other approaches such as extreme value theory and Poisson models, the model and the estimation procedure are by construction invariant to the choice of the threshold, which can be determined *ex post*.

Specifically, we propose to model the pollutant levels through a simple time measure-valued Markov process. This will be a nonparametric mixture of parametric kernels, whose mixing measure is a stochastic process that induces the temporal dependence. The use of a multivariate kernel enables the model to capture the spatial dependence among the observations, and the temporal dependence structure built in the process allows to infer and reproduce that present in the data. The nonparametric approach avoids unrealistic constraints on the shape of the distributions involved, guaranteeing full flexibility in the estimation procedure and the ability to capture features such as asymmetries and multimodality. The relative simplicity of the mechanism which induces the temporal dependence in the mixture makes our proposal particularly appealing for these type of datasets. Unlike similar approaches based on dependent random probability measures, the proposal finds a good trade-off between generality and ease of implementation, in that the simplicity of the induced temporal dependence, together with usual techniques for dealing with the infinite dimensionality of the model, enables to design a fast and efficient algorithm for the posterior computation. Furthermore, the model allows to range along all degrees between fully correlated and uncorrelated adjacent probability measures in the collection, thus enabling the researcher to calibrate the use of the procedure for different frameworks.

The paper is organized as follows. Section 2 presents the methodology we introduce for the analysis. After briefly reviewing some general background notions about Bayesian nonparametric density estimation, we develop the model, which falls into the class of dependent Dirichlet process mixtures. We discuss its properties and outline the strategy for posterior computation. In Section 3, we illustrate the performance of the proposed model with a simulated dataset, comparing it also with a spline regression alternative. In Section 4, we apply the model to the air quality analysis on a $PM_{2.5}$ dataset collected in Santiago, Chile. The results include estimation of the time-varying density for a four-dimensional spatially correlated time series for the $PM_{2.5}$ levels recorded in different monitoring stations in the metropolitan area, together with other quantities of interest which are derived from the estimate. Furthermore, although these are not explicitly enforced in the model formulation, a study of the seasonality trends for one monitoring location and the probability of exceeding an arbitrary threshold in single days of the year are also derived as a byproduct.

2. The model

After collecting some considerations on the density estimation problem from a Bayesian standpoint, we present a model for studying the air pollution data with temporal and spatial dependence. The proposed model falls in the realm of Bayesian nonparametric dependent models, and is tailored to multivariate data which exhibit this type of dependence. Since we aim at the entire shape of the time-varying distribution, from which other quantities of interest can be derived, such type of data