Spatial and Spatio-temporal Epidemiology 000 (2016) 1–11

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Modelling collinear and spatially correlated data

Silvia Liverani[a,b,d,*], Aurore Lavigne[c], Marta Blangiardo[d]

[a] Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, UK
[b] Medical Research Centre Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK
[c] Université Lille 3, UFR MIME, Domaine universitaire du Pont de Bois, BP 60149 59653 Villeneuve d'ascq Cedex, France
[d] MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, 2 Norfolk Place, London W2 8PG, UK

ARTICLE INFO

Article history:
Received 19 November 2015
Revised 23 February 2016
Accepted 5 April 2016
Available online xxx

Keywords:
Profile regression
Bayesian clustering
Spatial modelling
Collinearity
Index of multiple deprivation
Pollution

ABSTRACT

In this work we present a statistical approach to distinguish and interpret the complex relationship between several predictors and a response variable at the small area level, in the presence of (i) high correlation between the predictors and (ii) spatial correlation for the response.

Covariates which are highly correlated create collinearity problems when used in a standard multiple regression model. Many methods have been proposed in the literature to address this issue. A very common approach is to create an index which aggregates all the highly correlated variables of interest. For example, it is well known that there is a relationship between social deprivation measured through the Multiple Deprivation Index (IMD) and air pollution; this index is then used as a confounder in assessing the effect of air pollution on health outcomes (e.g. respiratory hospital admissions or mortality). However it would be more informative to look specifically at each domain of the IMD and at its relationship with air pollution to better understand its role as a confounder in the epidemiological analyses.

In this paper we illustrate how the complex relationships between the domains of IMD and air pollution can be deconstructed and analysed using profile regression, a Bayesian non-parametric model for clustering responses and covariates simultaneously. Moreover, we include an intrinsic spatial conditional autoregressive (ICAR) term to account for the spatial correlation of the response variable.

1. Introduction

In many statistical applications a common challenge arises when trying to assess meaningful relationships between explanatory variables and outcomes through regression models, due to the potential collinearity of the explanatory variables. This issue is well known in epidemiological or social studies, for instance where questionnaires or surveys collect information on a large number of potential risk factors for particular end points; in this context a simplistic approach consists in examining each variable in turn to avoid the instability in the estimates due to the collinearity, making it impossible to judge the more realistic complex relationship involving several risk factors at the same time. A different approach combines all the

* Corresponding author.
E-mail addresses: silvia.liverani@brunel.ac.uk (S. Liverani), aurore.lavigne@univ-lille3.fr (A. Lavigne), m.blangiardo@imperial.ac.uk (M. Blangiardo).

http://dx.doi.org/10.1016/j.sste.2016.04.003

relevant variables into summary scores or indexes and assesses the relationship of these with the outcome of interest, which is free from the collinearity issue, but loses information on the single variables included in the summary.

Recently, Dirichlet process mixture models have been used as an alternative to regression models (Bigelow and Dunson, 2009; Dunson et al., 2008). In this paper we focus on the model known as profile regression and proposed by Molitor et al. (2010). Profile regression is a Bayesian non–parametric method which assesses the link between potentially collinear variables and a response through cluster membership. This allows to formally take into account the correlation between the variables without the need to create a summary score, giving more flexibility to the inferential process. Profile regression has been used on several applications in environmental and social epidemiology and the R package PReMiuM (Liverani et al., 2015) makes it readily available to any applied researcher. For instance (Molitor et al., 2010) considered the National Survey of Children's Health and in particular investigated a large number of health and social related variables on mental health of children age 6–17, while (Papathomas et al., 2011) focussed on profiles of exposure to environmental carcinogens and lung cancer in the EPIC European cohort. Profile regression has also been used in environmental epidemiology (Pirani et al., 2015), for studying risk functions associated with multi-dimensional exposure profiles (Hastie et al., 2013; Molitor et al., 2014) as well as for looking for gene–gene interactions (Papathomas et al., 2012).

In its present formulation, profile regression has only been used for studies based on cohorts or surveys where information on the predictors/outcomes is available on each individual; in this paper we extend the method to fit small area studies, commonly used in epidemiological surveillance (see for instance Elliott and Wartenberg, 2004) or in studies where the interest lies on the spatial variability of an outcome (Barcelo et al., 2009) or on cluster detection (Abellan et al., 2008; Li et al., 2012). In this types of studies information is available at the area level rather than at the individual level and space is used as a proxy for any unmeasured variable; the common assumption is that areas which are close to each other are more similar than those further apart, suggesting that an additional source of correlation, namely *spatial correlation* needs to be accommodated in the models. We incorporate it in the model through a conditional autoregressive structure (Besag et al., 1991) based on a neighbourhood definition, thus assuming that conditional on the neighbourhood structure, two areas are independent from each other if they do not share boundaries. We apply the *spatial profile regression* to the problem of environmental and social inequalities in London, jointly modelling social deprivation and air pollution to highlight the presence of environmental justice.

The paper is structured as follows. In Section 2 we present the motivating example for our methodological development of the spatial profile regression, introducing the context of social and environmental inequalities and how they are related; we also describe the available data. In Section 3 we provide a brief summary of the profile regression and present how to extend it to include spatial correlation. In Section 4 we illustrate how the model works on evaluating the relationship between social deprivation and air pollution. Section 5 presents some discussion points and ideas for future work.

2. Example: social deprivation and air pollution in London

The scientific literature reports mixed evidence on the link between socio-economic status and air pollution. Recent studies indicated that air pollution tends to influence most deprived groups, suggesting that people with lower socio-economic status are more likely to live in a more hazardous and polluted living environment, accidentally or deliberately (Blowers and Leroy, 1994; Brown, 1995; Morello-Frosch et al., 2002; O'Neill et al., 2003). In particular, ecological studies using small areas such as neighbourhoods, census tracts and post codes, report this association, while studies carried out at a lower spatial resolution (e.g. region, country), thus characterised by more aggregate measurements of socio-economic characteristics, showed either non-existent or negative associations (Davidson and Anderton, 2000; Laurent et al., 2007), presumably due to the large within-area variability not taken into account, or even an inverse association, with higher exposures in less deprived groups (Perlin et al., 1995). In the UK several studies reported positive or non-linear correlation between environmental pollution and the deprivation index at both small area level and country level. However the results varied depending on the selection of environmental hazards and scale of analysis (Briggs et al., 2008), calling for some more research on the topic.

Understanding environmental and social inequalities is a key issue as growing health disparities appear between people with socially disadvantaged and privileged social classes, which can translate into increased mortality or morbidity for the low socio-economic groups across a wide range of diseases (Benach et al., 2001; Brulle and Pellow, 2006), including lung cancer (Pope et al., 2011), cardiovascular events (Peters et al., 2004; Tonne et al., 2007), and childhood respiratory diseases (Morgenstern et al., 2007).

In addition, the exposure of air pollution can lead to negative health outcomes acutely or chronically (Chen et al., 2008). Previous studies reported possible mechanisms to explain how environmental exposures result in greater health impact among socially disadvantaged groups, who may have increased susceptibility to the effect of these exposures because of limited access to health care and psychosocial stress; underlying health conditions such as cardiovascular diseases and respiratory diseases that increase susceptibility to the effect of these exposure may also vary between deprived and privileged populations (Morello-Frosch and Jesdale, 2006; O'Neill et al., 2003). These environmental exposure inequalities are increasingly considered as a potential determinant of health disparities (Morello-Frosch and Jesdale, 2006). In addition, it has been suggested that the disparities grow in more deprived areas as health improves faster in high socio-economic groups (Higgs et al., 1998; Leyland et al., 2007).

Although individual determinants (such as smoking) or individual risk responses (such as closing windows to avoid

Please cite this article as: S. Liverani et al., Modelling collinear and spatially correlated data, Spatial and Spatio-temporal Epidemiology (2016), http://dx.doi.org/10.1016/j.sste.2016.04.003