Computational Statistics and Data Analysis 68 (2013) 141–154

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Analysis of space–time relational data with application to legislative voting

Esther Salazar [a,*], David B. Dunson [b], Lawrence Carin [a]

[a] Department of Electrical & Computer Engineering, Duke University, United States

[b] Department of Statistical Science, Duke University, United States

ARTICLE INFO

Article history:
Received 21 June 2012
Received in revised form 20 June 2013
Accepted 20 June 2013
Available online 3 July 2013

Keywords:
Bayesian nonparametrics
Gaussian process
Kernel stick breaking process
Probit model
Spatio-temporal process

ABSTRACT

We consider modeling spatio-temporally indexed relational data, motivated by analysis of voting data for the United States House of Representatives over two decades. The data are characterized by incomplete binary matrices, representing votes of legislators on legislation over time. The spatial covariates correspond to the location of a legislator's district, and time corresponds to the year of a vote. We seek to infer latent features associated with legislators and legislation, incorporating spatio-temporal structure. A model of such data must impose a flexible representation of the space–time structure, since the apportionment of House seats and the total number of legislators change over time. There are 435 congressional districts, with one legislator at a time for each district; however, the *total* number of legislators typically changes from year to year, for example due to deaths. A *matrix kernel stick-breaking process* (MKSBP) is proposed, with the model employed within a probit-regression construction. Theoretical properties of the model are discussed and posterior inference is developed using Markov chain Monte Carlo methods. Advantages over benchmark models are shown in terms of vote prediction and treatment of missing data. Marked improvements in results are observed based on leveraging spatial (geographical) information.

1. Introduction

The analysis of data with space–time dependences is of interest in many applications. For example, one may be interested in modeling geospatial demographic data as it evolves with time, or in analyzing time-evolving weather patterns across the globe. In the motivating application of this paper, we consider time-dependent voting patterns of legislators, where the spatial coordinate corresponds to the location of the congressional district and time represents the year of the vote. Specifically, we consider binary vote matrices from the United States House of Representatives, corresponding to the period 1989–2008 (101th–110th US Congress). The data are publicly available from *The Library of Congress* http://thomas.loc.gov and from www.govtrack.us. Each binary matrix, $\boldsymbol{B}_t \in \{0, 1\}^{N_t \times L_t}$, is manifested by mapping all "yes" votes to one and "no" votes to zero. Here, t denotes the time index (year), L_t denotes the number of pieces of legislation in year t, and N_t denotes the number of legislators in year t (the total number of legislators at any time is fixed by law at 435, but the particular people who serve as legislators may change slightly within a given year, with the departure — e.g., death — and arrival of a small subset of legislators). Some relevant features of the voting data are: (i) they consist of a set of incomplete binary matrices, with approximately 5% missing data in a given year; (ii) the number of pieces of legislation, L_t, varies between 444 (102th Congress, 1992) and 1186 (110th Congress, 2007); (iii) the apportionment of House seats changes over time, as a function of changes in the population density (the total number of House districts is constant, but the geographical distribution of

* Corresponding author. Tel.: +1 9194859865.
E-mail address: esther.salazar@duke.edu (E. Salazar).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.018

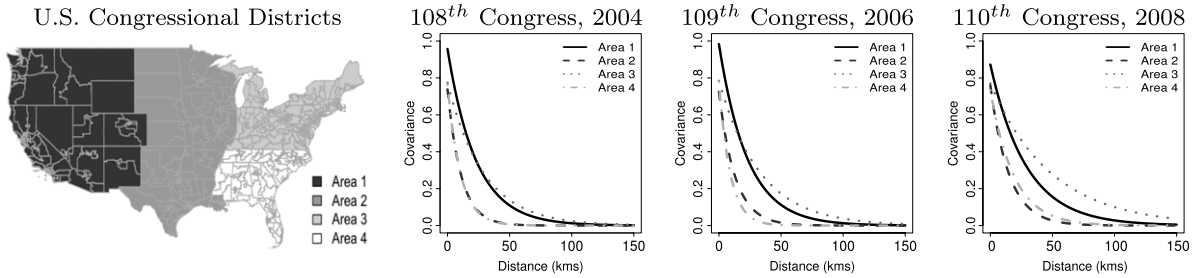

Fig. 1. Left panel: 110th Congressional Districts map divided into four areas, each of them represented by different colors. Last three panels: Estimated exponential spatial covariance as a function of Euclidean distance, for the four areas and for the 108th (2nd Session, 2004), 109th (2nd Session, 2006) and 110th (2nd Session, 2008) US Congress. In the second panel, the covariance function for Areas 2 and 4 is the same.

districts changes with time). Areas that have a low population density have geographically large congressional districts and areas that have a high population density have geographically small congressional districts.

In the political science literature, a number of frequentist statistical models for roll call data analysis have been proposed (Poole and Rosenthal, 1997; Heckman and Snyder, 1997; Jackman, 2001). Later, Clinton et al. (2004) developed a Bayesian procedure for analysis of voting data via a one-factor probit model and, more recently, Wang et al. (2010) proposed a related model for the joint analysis of time-evolving voting matrices. Both approaches do not take advantage of the spatial information inherent in the data. In this paper, we seek to infer latent features associated with legislators and legislation through the analysis of spatio-temporal structure.

Considering the features of the data, a flexible spatio-temporal correlation structure must be imposed to identify nonstationary structures, i.e. covariances that change over space and time. In order to examine and illustrate spatial nonstationarity, we divided the US House district map in four big contiguous areas. Fig. 1, the left panel, shows a division on the 110th Congressional district map; the same division was considered for the other sessions. Across the time period considered, Areas 1, 2, 3 and 4 have on average 95, 106, 154 and 77 districts, respectively. Note that Areas 3 and 4 (eastern part of the country) have a high population density and we can expect higher spatial correlation, particularly in the northeastern part (Area 3). For each area we fitted the voting data using a modeling framework similar to that in Clinton et al. (2004) but incorporating spatial structure into the unique factor associated to legislators. Specifically, we consider a spatial Gaussian process with exponential correlation function given by $\rho(d) = \exp\{-d/\phi\}$ and variance τ (d is the Euclidean distance between the center of two congressional districts). Fig. 1, last three panels, shows the estimated posterior covariance as a function of Euclidean distance (in km), for the four areas and for three different sessions (corresponding to the years 2004, 2006 and 2008). Note that the covariances are different across time and for each area, suggesting a nonstationary spatio-temporal specification for the data.

Motivated by the wide class of interesting applications, spatial and spatio-temporal interactions have been studied in many ways, in a parametric and nonparametric manner (Duan et al., 2007; Figueiredo et al., 2006; Gelfand et al., 2007, 2005b; Griffin and Steel, 2006; Reich and Fuentes, 2007; Rodriguez et al., 2010). One of the most common strategies is to include spatial and temporal dependences separately through underlying features (or factors) (Lopes et al., 2008; Luttinen and Ilin, 2009). However, in many applications the assumption of separability of the total correlation structure may be an over-simplification. Spatial (and also temporal) features are often modeled via Gaussian processes (GPs). However, assuming a common GP for the spatial/temporal region of interest may be too restrictive for some phenomena for which the correlation length is nonstationary. This nonstationary characteristic is of relevance for our motivating example, as the spatial correlation length is expected to be smaller in regions of high population density (*e.g.*, the northeastern part of the United States), as compared to extended regions of modest population (the midwest portion of the United States) for which the correlation length is expected to be larger.

There is a rich literature on nonstationary spatio-temporal models. Some approaches based on dynamic and hierarchical modeling are Wikle et al. (1998), Stroud and Müller (2001) and Gelfand et al. (2005a). Other ways to remove the stationary assumption include spatially varying kernel convolution as given by Higdon (1998) and Higdon et al. (1999), as well as convolving a fixed kernel over independent stationary processes with different covariance parameters (Fuentes, 2001). These approaches (in general within the setting of GPs) are fully parametric with predefined kernels empirically chosen from a pool of parametric kernel functions. In addition, the spatial deformation approach of Sampson and Guttorp (1992) is another alternative to achieve nonstationarity. This method is based on mapping the original input space to a new conceptual deformed space in which the process is assumed stationary. A Bayesian version of this idea can be found in Damian et al. (2001) and Schmidt and O'Hagan (2003). However, a limitation of the deformation approach is the need for replication to obtain a sample covariance matrix.

In contrast with the previous mentioned approaches, we are interested in developing a nonparametric model for spatio-temporal processes, flexible enough to explain the complex structure of the data where the resulting process is nonstationary in space and time. We seek development of a model that can efficiently borrow information across space and time simultaneously, while allowing variable correlation lengths in both dimensions. So motivated, we develop a new *matrix*