Journal of Statistical Planning and Inference 166 (2015) 24–35

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes

Eduard Belitser [a], Paulo Serra [b,*], Harry van Zanten [c]

[a] *Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*
[b] *Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany*
[c] *Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 27 November 2013
Accepted 14 March 2014
Available online 9 July 2014

Keywords:
Adaptive estimation
Bayesian nonparametric estimation
Contraction rate
Markov chain Monte Carlo
Poisson process
Splines

ABSTRACT

We apply nonparametric Bayesian methods to study the problem of estimating the intensity function of an inhomogeneous Poisson process. To motivate our results we start by analyzing count data coming from a call center which we model as a Poisson process. This analysis is carried out using a certain spline prior. This prior is based on B-spline expansions with free knots, adapted from well-established methods used in regression, for instance. This particular prior is computationally feasible. Theoretically, we derive a new general theorem on contraction rates for posteriors in the setting of intensity function estimation which can be applied not just to this spline prior but also to a large number of other commonly used priors. Practical choices that have to be made in the construction of our concrete spline prior, such as choosing the priors on the number and the locations of the spline knots, are based on these theoretical findings. The results assert that when properly constructed, our approach yields a rate-optimal procedure that automatically adapts to the regularity of the unknown intensity function.

1. Introduction

Poisson processes have a long-standing history and are some of the most widely used processes in statistics to study temporal and spacial count data, in diverse fields such as communication, meteorology, seismology, hydrology, astronomy, biology, medicine, actuary sciences and queueing, among others. In this paper we focus on inhomogeneous Poisson processes on the real line with periodic intensity functions, which are models for count data in settings with a natural periodicity. We obtain asymptotic results as the number of observed periods goes to infinity but our approach is flexible enough to also deliver asymptotic results for estimating an intensity function on a compact in terms of either the number of observed events or in terms of the scale of the intensity function.

Nonparametric Bayesian methods, which are used more and more in many different statistical settings, have so far only been used on a limited scale to analyze such models. From the applied perspective they can be attractive for making inference about intensity functions, for the same reasons they are appealing in other situations. Estimating the intensity essentially requires some form of smoothing of the count data, and a nonparametric Bayesian approach can provide a natural way of achieving this. Using hierarchical priors we can automatically achieve a data-driven selection of the degree of smoothing. Moreover, Bayesian methods provide a way to quantify the uncertainty about the intensity using the spread of the posterior

* Corresponding author. Tel.: +49 15123765379.
E-mail addresses: e.n.belitser@vu.nl (E. Belitser), pdeandr@uni-goettingen.de (P. Serra), hvzanten@uva.nl (H. van Zanten).

http://dx.doi.org/10.1016/j.jspi.2014.03.009
0378-3758/

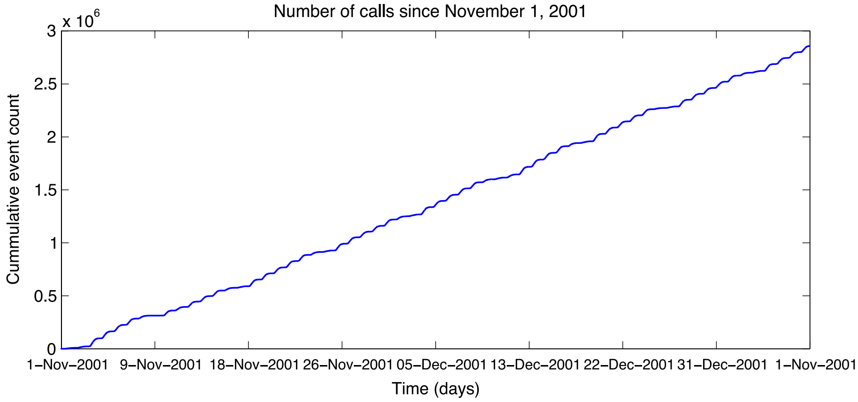

Fig. 1. Number of incoming phone calls between November 1, 2001 and December 31, 2001.

distribution. A typical implementation provides a computational algorithm that can generate a large number of (approximate) draws from the posterior. From this it is usually straightforward to construct numerical credible bands or credible sets.

The relatively small number of papers using nonparametric Bayesian methodology for intensity function smoothing have explored various possible prior distributions on intensities. An early reference is Møller et al. (1998), who consider log-Gaussian priors. Other papers employing Gaussian process priors, combined with suitable link functions, include Adams et al. (2009) and Palacios and Minin (2013). Kottas and Sansó (2007) consider kernel mixtures priors; see also the related paper by DiMatteo et al. (2001), in which count data are analyzed using spline-based priors.

The cited papers show that nonparametric Bayesian inference for inhomogeneous Poisson processes can give satisfactory results in various applications. On the theoretical side however the existing literature provides no performance guarantees in the form of consistency theorems or related results. It is by now well known that nonparametric Bayes methods may suffer from inconsistency, even when seemingly reasonable priors are used (e.g. Diaconis and Freedman, 1986). The purpose of this paper is therefore to propose a Bayesian approach to nonparametric intensity smoothing that is both computationally feasible and at the same time theoretically underpinned by results on consistency and related issues like convergence rates and adaptation to smoothness. Such theoretical results have in the last decade been obtained for various statistical settings, including density estimation, regression, classification, and drift estimation for diffusions (see e.g. Ghosal, 2010 for an overview of some of these results). Until now, intensity estimation for inhomogeneous Poisson processes has remained largely unexplored.

As motivation and starting point for the paper we consider the problem of analyzing count data from a call center. The same type of data were analyzed by frequentist methods in the paper Brown et al. (2005). We revisit the problem using a nonparametric Bayesian method employing a spline-based prior on the unknown intensity function. In addition to a single estimator of the intensity, this method provides credible bounds indicating the degree of uncertainty. In Section 3 we study theoretical properties of our procedure, namely consistency, posterior contraction rates and adaptation to smoothness. The results show that we have set up our procedure in such a way that we obtain consistent, rate-optimal estimation of the intensity and that the method adapts automatically to the unknown smoothness of the intensity curve, up to the level of the order of the splines that are used. Section 4 concludes with some remarks and directions for further research.

2. Analysis of call center data

2.1. Data and statistical model

The approach we propose is motivated by the wish to analyze a dataset consisting of counts of telephone calls arriving at a certain call center. The dataset was obtained from the website of the S.E.E. Center of the Faculty of Industrial Engineering and Management, Technion in Haifa, Israel (http://ie.technion.ac.il/Labs/Serveng/). It consists of counts for calls arriving at a bank's 24 h a day call center in the United States of America. We considered the records for the period from November 1, 2001 until December 31, 2001, covering a total of about 2.8 million incoming phone calls. These events are recorded in 30 s intervals with an average of about 32 calls per minute. The raw data are plotted in Fig. 1.

We model the full count data as the realization of an inhomogeneous Poisson process N with an intensity function λ that is periodic, the period being 24 h; cf. Daley and Vere-Jones (1988). This Poisson assumption is natural and is investigated in some detail in Brown et al. (2005), who could not find significant evidence to the contrary in a similar dataset (same kind of data, but over a different time interval). See also Belitser et al. (2013), who study the periodicity in the data.

This dataset is known to exhibit periodicity on different time scales; (approximate) daily, weekly, monthly and yearly periodicities seem to be present in the data. Different time scales are relevant if one would like to analyze the intensity of the call traffic during, say, the weekends, holidays or specific times of the year. (To analyze the intensity of calls during weekends, for example, a weekly time scale would be appropriate.) By carrying out our estimation procedure under the assumption