ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Bayesian temporal density estimation with autoregressive species sampling models

Youngin Jo [a], Seongil Jo [b], Yung-Seop Lee [c,*], Jaeyong Lee [d]

[a] Kakao corporation, Seongnam 13494, Republic of Korea
[b] Department of Statistics (Institute of Applied Statistics), Chonbuk National University, Jeonju 54896, Republic of Korea
[c] Department of Statistics, Dongguk University-Seoul, Seoul 04620, Republic of Korea
[d] Department of Statistics, Seoul National University, Seoul 08826, Republic of Korea

ARTICLE INFO

Article history:
Received 4 January 2018
Accepted 14 February 2018
Available online xxxx

MSC:
primary 62C10
secondary 62G07

Keywords:
Autoregressive species sampling models
Dependent random probability measures
Mixture models
Temporal structured data

ABSTRACT

We propose a novel Bayesian nonparametric (BNP) model, which is built on a class of species sampling models, for estimating density functions of temporal data. In particular, we introduce species sampling mixture models with temporal dependence. To accommodate temporal dependence, we define dependent species sampling models by modeling random support points and weights through an autoregressive model, and then we construct the mixture models based on the collection of these dependent species sampling models. We propose an algorithm to generate posterior samples and present simulation studies to compare the performance of the proposed models with competitors that are based on Dirichlet process mixture models. We apply our method to the estimation of densities for the price of apartment in Seoul, the closing price in Korea Composite Stock Price Index (KOSPI), and climate variables (daily maximum temperature and precipitation) of around the Korean peninsula.

1. Introduction

In the time-dependent data analysis, the analysts attempt to understand the dependency structure of the observed data in the chronicle order and to predict the future based on the observations in the past, and the distributional assumption is crucial in this process. But the distributional assumptions are often not adequate, because commonly used parametric assumptions poorly approximate the distribution (Rodríguez & Ter Horst, 2008), and thus the subsequent inference becomes inefficient. In such cases, Bayesian nonparametric (BNP) models can be good alternatives. Particularly, dependent BNP models can provide flexible approaches without rigid distributional assumptions for dealing with time dependency structure. See, for examples, Caron, Davy, and Doucet (2007); Caron, Davy, Doucet, Duflos, and Vanheeghe (2008) and Ren, Dunson, and Carin (2008).

Most of BNP models have been developed by using the Dirichlet process (DP) of Ferguson (1973) and its variation. See, e.g., Escobar and West (1995) and Ishwaran and James (2001, 2002). Dependency structure, in particular, has been accommodated by modeling the random support points and/or weights of the DP since MacEachern's dependent DP (DDP) was proposed (MacEachern, 1999, 2000). The examples include the order-based DDP (Griffin & Steel, 2006), the spatial DP (Duan, Guindani, & Gelfand, 2007; Gelfand, Kottas, & MacEachern, 2005), the kernel stick-breaking process (Dunson & Park, 2008), the local DP (Chung & Dunson, 2011), and a dependent Griffiths–Engen–McCloskey (GEM) distribution defined

* Corresponding author.
E-mail address: yung@dongguk.edu (Y.-S. Lee).

https://doi.org/10.1016/j.jkss.2018.02.002
1226-3192/

by transforming a Gaussian process (Arbel, Mengersen, & Rousseau, 2016). See, e.g., Müller and Rodriguez (2013) and Müller, Quintana, Jara, and Hanson (2015) for a discussion.

Recently, BNP models, which are based on the DDP, for estimating density functions of time-dependent data have been proposed. Rodríguez and Ter Horst (2008) proposed the dynamic DDP that models the random support points using Gaussian dynamic linear models, Griffin and Steel (2011) and Gutiérrez, Mena, and Ruggiero (2016) introduced autoregressive structured stick-breaking processes, Rodríguez and Dunson (2011) presented the probit stick-breaking process with latent Markov random fields that is defined by autoregressive process, and Nieto-Barajas, Muller, Ji, Lu, and Mills (2012) developed the time-series DDP that achieve the time dependence by introducing binomial random variables. For more examples, see, Mena and Ruggiero (2016) and references therein.

Most of the dependent random probability models are based on DP or DDP which generates weights from stick-breaking process. It is well known that stick-breaking weights are stochastically ordered in a descending way and their expectations decrease at a geometrical rate. That is, the sequence of stick-breaking weights decreases rapidly and values of the weights are very small except a few. This property is adequate to data with small number of large cluster, but may not be to data sets with more complex structures. To overcome this limitation of stick-breaking weights of DP, more general BNP models have been developed recently. These models are based on the species sampling model (SSM) which is introduced by Pitman (1996). For example, Airoldi, Costa, Bassetti, Leisen, and Guindani (2014) and Bassetti, Leisen, Airoldi, and Guindani (2015) introduced the Beta-GOS process, which is a special case of a generalized Ottawa sequence (GOS), with weights derived from the product of independent Beta random variables and Jo, Lee, Müller, Quintana, and Trippa (2017) developed CAR SSMs with normalized weights based on conditional autoregressive (CAR) models.

In this paper, we extend the idea of Jo et al. (2017) to the BNP model for the time series data. The proposed model can be viewed as an extension of CAR SSMs Jo et al. (2017) in two directions. First, the proposed model is for the time series data. Second, the proposed model has stochastic atoms as well as stochastic weights, while CAR SSMs have only stochastic weights. In particular, we propose a dependent SSM, which assumes that the weights and/or atoms of SSM vary in time, as a prior for the distribution of the data with time dependency structure. The time dependency is accommodated through the autoregressive (AR) models. The main advantages of the proposed model are two-folds. First, since it is based on the SSM, its cluster structure is more flexible than the DP based models. A SSM is a discrete random probability measure (RPM) with the independent and identically distributed (i.i.d.) random support points and the weights defined by an arbitrary probability model. It enables us to define a flexible and intuitive RPM easily (see, e.g., Jo et al. (2017) and Lee, Quintana, Müller, and Trippa (2013)). Second, instead of complicated order-based dependence structures, the proposed model uses well-understood AR models, which have been widely used for inference of time dependent data because of flexibility and interpretability. Additionally, high order AR processes can approximate any stationary process (see, e.g., Rodríguez and Ter Horst (2008)). Therefore, giving AR models to both provides more flexible density estimation as the random support points and weights of distribution changes according to time.

The structure of this paper is as follows. In Section 2 we review species sampling models and autoregressive model briefly. In Section 3, we describe our autoregressive species sampling model and its posterior computation algorithm. In Sections 4 and 5, we present an extensive simulation study and three real data examples to verify the performance of the proposed autoregressive species sampling model. Some concluding remarks are made in Section 6.

2. Background

2.1. Species sampling model

The *proper* species sampling model, which is a version of the SSM, is a discrete random probability measure defined as

$$G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h},$$

where the atoms (θ_h) are formed by a sequence of random samples from a diffuse probability measure G_0, the weights (p_h) are nonnegative random variables with $\sum_{h=1}^{\infty} p_h = 1$ with probability 1, and the atoms and the weights are independent from each other. The Dirichlet process (Ferguson, 1973, 1974) and Pitman–Yor process (Ishwaran & James, 2001; Pitman, 1996) also follow the same probabilistic structure. See, e.g., Müller and Rodriguez (2013) and Müller et al. (2015) for a review and more examples.

Lee et al. (2013) considered a family of SSMs whose weights consist of normalized positive random variables

$$G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h} \text{ with } p_h = \frac{w_h}{\sum_{h=1}^{\infty} w_h} \text{ for } h = 1, 2, \ldots, \tag{1}$$

where (w_h) is a sequence of positive random variables and (θ_h) is an independent and identically distributed sequence of random samples from G_0, and (w_h) and (θ_h) are independent. They have shown that a sufficient condition for $\sum_{h=1}^{\infty} w_h < \infty$ a.s. is that $\sum_{h=1}^{\infty} E(w_h) < \infty$ and gave an example of (w_h) satisfying the sufficient condition. Let $N(c, d^2)$ and $N(\cdot; c, d^2)$ denote the normal distribution with mean c and variance d^2 and its density. Let

$$w_h = e^{u_h} \text{ and } u_h \overset{ind}{\sim} \mathrm{N}\left\{\log\left(1 - \frac{1}{1 + e^{b-ah}}\right), \tau^2\right\}, \; h \geq 1, a, b, \tau^2 > 0, \tag{2}$$

Please cite this article in press as: Jo, Y., et al., Bayesian temporal density estimation with autoregressive species sampling models. Journal of the Korean Statistical Society (2018), https://doi.org/10.1016/j.jkss.2018.02.002.