Computational Statistics and Data Analysis 117 (2018) 162–181

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Dual-semiparametric regression using weighted Dirichlet process mixture

Peng Sun [a], Inyoung Kim [a,*], Ki-Ahm Lee [b,c]

[a] Department of Statistics, Virginia Polytechnic Institute and State University, USA
[b] Department of Mathematical Sciences, Seoul National University, Republic of Korea
[c] Korea Institute for Advanced Study, Republic of Korea

ARTICLE INFO

Article history:
Received 16 April 2016
Received in revised form 4 July 2017
Accepted 10 August 2017
Available online 31 August 2017

Keywords:
Additive model
Bayes factor
Cubic splines
Dual-semiparametric regression
Generalized pólya urn
Gibbs sampling
Metropolis–Hastings
Nonparametric Bayesian model
Ordinal data
Semiparametric regression
Weighted Dirichlet process

ABSTRACT

An efficient and flexible Bayesian approach is proposed for a dual-semiparametric regression model that models mean function semiparametrically and estimates the distribution of the error term nonparametrically. Using a weighted Dirichlet process mixture (WDPM), a Bayesian approach has been developed on the assumption that the distributions of the response variables are unknown. The WDPM approach is especially useful for real applications that have heterogeneous error distributions or come from a mixture of distributions. In the mean function, the unknown functions are estimated using natural cubic smoothing splines. For the error terms, several different WDPMs are proposed using different weights that depend on the distances between the covariates. Their marginal likelihoods are derived, and the computation of marginal likelihood for WDPM is provided. Efficient Markov chain Monte Carlo (MCMC) algorithms are also provided. The Bayesian approaches based on different WDPMs are compared with the parametric error model and the Dirichlet process mixture (DPM) error model in terms of the Bayes factor using a simulation study, suggesting better performance of the Bayesian approach based on WDPM. The advantage of the proposed Bayesian approach is also demonstrated using the credit rating data.

1. Introduction

Semiparametric regression (Ruppert et al., 2003) has been widely used in many fields, including economics (Chib and Greenberg, 2010), finance (Hannah et al., 2011; Jensen and Maheu, 2013), environmetrics (Mahmoud et al., 2016) and biostatistics (Kim et al., 2003; Kim and Cohen, 2004; Kim et al., 2011; Pang and Zhao, 2012; Ortega Villa et al., 2017), in which some covariates have a known relationship with a response while others do not. It mixes the parametric and nonparametric parts together in the regression. The nonparametric part can be estimated using cubic smoothing splines (Durrleman and Simon, 1989; Green and Silverman, 1994; Pagan and Ullah, 1999; Marsh and Cormier, 2002; Li and Racine, 2006; Chib and Greenberg, 2010). Although semiparametric regression has the flexibility of modeling both parametric and nonparametric parts, parametric distributions often impose strong assumptions about the distribution of the unobserved error or the distribution of the underlying latent variable. The error distribution is often assumed to be parametric, as with normal distribution and t-distribution, when the outcomes are continuous. In ordinal models for categorical outcomes, the model is almost always specified with logit or probit links. Chu and Ghahramani (2005) proposed Gaussian processes (GP) for ordinal regression and introduced a generalization of the probit function.

* Correspondence to: Department of Statistics, Virginia Tech., Blacksburg, VA 24061, USA
E-mail address: inyoungk@vt.edu (I. Kim).

http://dx.doi.org/10.1016/j.csda.2017.08.005
0167-9473/© 2017 Elsevier B.V. All rights reserved.

Parametric assumptions of error distributions are often not satisfied in real applications that have heterogeneous error distributions or come from a mixture of a heterogeneous distribution. As a result, the model may easily turn out to be misspecified, which thus influences the statistical inference. Therefore, we adopt a nonparametric Bayesian approach to allow the semiparametric regression to be more flexible. Our nonparametric Bayesian approach can apply to both continuous and ordinal responses. Because we are interested in the functional relationship between the covariate and response variable, in this paper, we model each covariate using the natural cubic smoothing splines.

First, let us consider that we have $p + ns$ covariates and a continuous response variable y. The first p covariates can be parametrically modeled with unknown parameters $\boldsymbol{\beta}_p$, and the other ns covariates $w_1, \ldots, w_{ns}$ are nonparametrically modeled with $g_\zeta(\cdot)$, $\zeta = 1, \ldots, ns$. Our semiparametric model can be written as:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_p + g_1(w_{i1}) + g_2(w_{i2}) + \cdots + g_{np}(w_{i,np}) + \varepsilon_i \quad i = 1, \ldots, n, \tag{1.1}$$

where $\mathbf{x}_i'\boldsymbol{\beta}_p$ is the parametric function, $g_\zeta(\cdot)\,(\zeta \leq ns)$ are unknown functions that are estimated using the cubic smoothing splines by adopting the basis illustrated in Lancaster and Šalkauskas (1986) and Wood (2006), and ε is the error term whose distribution shape needs to be captured.

In the parametric Bayesian approach, it is typical to assume that the distribution of ε_i has some known form $f_e(\varepsilon_i|\theta_i)$, where θ_i denotes the parameters of the distribution of the error term. The prior distribution of θ_i is also in some known form G_0. However, it is relatively challenging to derive convincing evidence about the distribution of the error. On the other hand, in the nonparametric Bayesian approach, the Dirichlet process, which was formally introduced by Ferguson (1973), enables us to construct the prior of θ in a nonparametric way. If we assume that the prior distribution of θ, say G, is sampled from a Dirichlet process, this Dirichlet process mixture (DPM) model allows more flexibility than does assuming that G belongs to some known family of distributions. Therefore, many existing studies have adopted DPM (Chib and Greenberg, 2010; Zhang et al., 2014; Dunson et al., 2007; Dunson and Stanford, 2005; Ghosal, 2009; Pati and Dunson, 2014; Chae et al., 2016) in various problems such as density estimation, the asymptotic distribution, the outlier problem, and high-dimensional analysis. Ghosal (2009) reviewed the role of the Dirichlet process (DP) and related prior distributions in nonparametric Bayesian inference. Ghosal (2009) also discussed various properties of the Dirichlet process and provided the asymptotic properties of posterior distributions. In addition, Pati and Dunson (2014) proposed robust Bayesian nonparametric regression which automatically downweights outliers and influential observations using the varying residual density based on probit stick-breaking (PSB) scale mixtures and symmetrized PSB (sPSB) location-scale mixtures. Recently, Chae et al. (2016) proposed Bayesian sparse linear regression with unknown symmetric error. They studied full Bayesian procedures for sparse linear regression when errors have a symmetric but otherwise unknown distribution. The unknown error distribution is endowed with a symmetrized Dirichlet process mixture of Gaussian.

Although DPM has flexibility, it does not take into account the covariates' information into the prior of θ. DPM assumes that all of the θ_is follow the same prior distribution G. Therefore, DPM is not flexible enough when the error distributions are a mixture of heterogeneous distributions. To overcome this limitation, we can use the weighted Dirichlet process mixture (WDPM). WDPM can be viewed as the dependent Dirichlet process (DDP), which extended DP by allowing cluster parameters to vary with some covariates (MacEachern, unpublished). The idea of WDPM was proposed by Zellner (1986) and applied by Dunson et al. (2007). The concept of WDPM comes from the idea that one can add the information provided by covariates into the construction of prior distributions. Such a concept relaxes the constraint that all of the observations share the same prior for the error-term parameter. Instead of a single prior, there are multiple candidate priors. The observations with similar predictor values (covariates) are more likely to share the same prior, which is one of the available candidate priors. Therefore, we can see that WDPM allows a higher degree of heterogeneity for the observations compared to DPM or the parametric Bayesian model.

In summary, we have shown that, in the parametric Bayesian model, we have only one prior for all of the θ_i's, and such a prior must belong to some known distribution family. In DPM, there is only one prior as well, although it is not necessarily some known distribution. However, when the observations do not share the same prior distribution for θ, DPM is not appropriate. WDPM can be an efficient alternative approach when a model based on a single prior-distribution assumption fails to produce an adequate degree of accuracy.

In this paper, we incorporate WDPM in semiparametric regressions and propose the dual-semiparametric model because we adopt semiparametric regression for the mean part and nonparametric Bayesian approaches for the error terms. We also propose several WDPM models using different weight functions. We derive their marginal likelihood for statistical inference. Efficient Markov chain Monte Carlo (MCMC) algorithms are provided. To the best of our knowledge, our approach is the first to incorporate WDPM in semiparametric regression, propose efficient weights, and provide the marginal likelihood derivation. Our weight functions in WDPM are compared with the weight function proposed by Dunson et al. (2007). We further compare our WDPM models with parametric and DPM models in terms of the Bayes factor using both a simulation study and real data and suggest some outperformances of our approach.

This article is organized as follows. In Section 2, we briefly review the basic idea of WDPM and propose several weight functions for WDPM. In Section 3, we introduce the potential reason why WDPM can produce a better marginal likelihood based on the Pólya urn for WDPM. In Section 4, we explain how to incorporate WDPM into the semiparametric regression. In Section 5.1, the posterior computation is illustrated. In Section 5.2, the marginal likelihood computation is explained so that it can be used for statistical inference. In Section 6, we conduct a simulation study to understand the performance of our approach. In Section 7, we apply our method to the credit-rating data illustrated in Verbeek (2008). Finally, Section 8 provides our concluding remarks.