Computational Statistics and Data Analysis 68 (2013) 251–261

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian analysis of generalized partially linear single-index models

Wai-Yin Poon [a], Hai-Bin Wang [b,*]

[a] Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong, China
[b] School of Mathematical Sciences, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 19 February 2012
Received in revised form 30 June 2013
Accepted 10 July 2013
Available online 17 July 2013

Keywords:
Free-knot spline
Generalized linear model
Gibbs sampler
Overdispersion
Reversible jump Markov chain Monte Carlo
Single-index model

ABSTRACT

We extend generalized partially linear single-index models by incorporating a random residual effect into the nonlinear predictor so that the new models can accommodate data with overdispersion. Based on the free-knot spline techniques, we develop a fully Bayesian method to analyze the proposed models. To make the models spatially adaptive, we further treat the number and positions of spline knots as random variables. As random residual effects are introduced, many of the completely conditional posteriors become standard distributions, which greatly facilitates sampling. We illustrate the proposed models and estimation method with a simulation study and an analysis of a recreational trip data set.

1. Introduction

Let $Y = (y_1, \ldots, y_n)^T$ be a vector of the independent random observations, where y_i follows an exponential family distribution with density

$$f(y_i|\eta_i, \phi) = \exp\left\{\phi^{-1}[y_i\zeta_i - b(\zeta_i)] + c(y_i, \phi)\right\}, \quad i = 1, \ldots, n, \tag{1}$$

where ϕ is a scalar parameter, and $b(\cdot)$ and $c(\cdot)$ are two specified functions. Generalized linear models (GLMs) (McCullagh and Nelder, 1989), including linear regression models, probit regression models, Poisson regression models, among others, relate the natural parameter ζ_i to a linear combination of predictor vectors $x_i = (x_{i1}, \ldots, x_{ip})^T$ and $z_i = (z_{i1}, \ldots, z_{iq})^T$ by

$$h^{-1}(\zeta_i) = \eta_i = x_i^T\alpha + z_i^T\beta, \quad i = 1, \ldots, n, \tag{2}$$

where $h(\cdot)$ is a known monotone univariate function that is called the link function, and $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ and $\beta = (\beta_1, \ldots, \beta_q)^T$ are unknown parameter vectors.

GLMs provide a unified family of models in regression analysis. However, the assumption that the effects of all the explanatory variables in (2) are linear is often questionable in practice. Relaxing this assumption leads to generalized non-/semi-parametric models, such as generalized additive models (Hastie and Tibshirani, 1990), generalized partially linear

* Corresponding author. Tel.: +86 13696903263; fax: +86 5922580608.
E-mail address: whb@xmu.edu.cn (H.-B. Wang).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.018

single-index models (GPLSIMs) (Carroll et al., 1997), and so on. Specifically, the GPLSIM is given by

$$h^{-1}(\zeta_i) = \eta_i = g(x_i^T\alpha) + z_i^T\beta, \quad i = 1, \ldots, n, \tag{3}$$

where $g(\cdot)$ is an unknown univariate function.

We consider a generalization of GPLSIM with random residual effect. The new model is defined jointly by (1) and

$$h^{-1}(\zeta_i) = \eta_i = g(x_i^T\alpha) + z_i^T\beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

where ε_i's are independent and identically distributed normal random variables with mean 0 and variance σ^2. Note that unlike η_i in (3), which is a function of observable explanatory variables, η_i in (4) is a latent random variable. To ensure identifiability, we assume as usual that the index vector is such that $\|\alpha\| = 1$ with $\alpha_1 > 0$, where $\|\cdot\|$ is the Euclidean norm.

The use of random residual effect ε_i in (4), similar to that in GLMs with random effects (Zeger and Karim, 1991; Breslow and Clayton, 1993; Clayton, 1996; Sun et al., 2000), enables the modeling of unexplained sources of variation in the data, such as explanatory variables that are not recorded and/or the overdispersion that is often encountered in binomial and Poisson data. Another advantage of using a residual component lies in statistical computation, where conditional on latent variables $\eta = (\eta_1, \ldots, \eta_n)^T$, model (4) is independent of Y and becomes a partially linear single-index model. This makes it easier to draw inferences about the model parameters by means of a data augmentation algorithm (Tanner and Wong, 1987).

As important dimension-reduction tools in multivariate nonparametric regression, single-index-type models have been substantially discussed in the literature: see, for instance, Stoker (1986), Härdle and Stoker (1989), Li (1991), Härdle et al. (1993), Ichimura (1993), Horowitz and Härdle (1996), Carroll et al. (1997), Xia et al. (1999, 2002), Naik and Tsai (2001), Yu and Ruppert (2002) and Kong and Xia (2007) for classical statistical methods, and Antoniadis et al. (2004), Park et al. (2005), Wang (2009), Choi et al. (2011), and Gramacy and Lian (2012) for Bayesian methods. The main drawback of classical methods is that unstable estimates are often obtained in small samples. Although Bayesian methods can produce more reliable estimates, the existing work has been limited to the continuous response case and, to the best of our knowledge, discrete response data have not been discussed.

Free-knot spline approximation is one of the most widely used approaches to nonparametric curve estimation, especially in the Bayesian setting (Smith and Kohn, 1996; Denison et al., 1998; Biller, 2000; Dimatteo et al., 2001; Holmes and Mallick, 2001, 2003; Lindstrom, 2002). An appealing feature of free-knot splines is that treating the number and the positions of spline knots as random variables can make the models spatially adaptive in a sense that a variable bandwidth or smoothing parameter is automatically determined (Holmes and Mallick, 2001, 2003). This approach is used to analyze the model in (4).

Assume that the unknown curve $g(\cdot)$ is an m-order polynomial spline with k ordered interior knots $\xi = (\xi_1, \ldots, \xi_k)^T$, i.e.,

$$g(u) = \sum_{j=1}^{K} B_j(u)\gamma_j = B^T(u)\gamma, \quad u \in [a, b], \tag{5}$$

where $K = 1+m+k$, a and b are two boundary knots, $B(u) = (B_1(u), \ldots, B_K(u))^T$ is a spline basis vector that is determined by knot ξ, and $\gamma = (\gamma_1, \ldots, \gamma_K)^T$ is a spline coefficient vector. Eq. (4) can then be represented as

$$\eta_i = B^T(x_i^T\alpha)\gamma + z_i^T\beta + \varepsilon_i = \mathcal{B}_i^T(\alpha)\theta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $\mathcal{B}_i^T(\alpha) = (B^T(x_i^T\alpha), z_i^T)^T$ and $\theta = (\gamma^T, \beta^T)^T$. In general, the boundary knots a and b take the minimum and maximum values of $\{x_i^T\alpha, i = 1, \ldots, n\}$, respectively. We denote

$$a_\alpha = \min_{1\leqslant i\leqslant n}\{x_i^T\alpha\} \quad \text{and} \quad b_\alpha = \max_{1\leqslant i\leqslant n}\{x_i^T\alpha\} \tag{7}$$

to emphasize their dependence on the index α. Clearly, given η, α and $B(u)$, (6) is an ordinary linear regression model, which allows us to adopt a conjugate normal-inverse gamma prior for regression coefficient θ and error variance σ^2. An advantage of the conjugate prior is that by integrating θ and σ^2 out of the joint posterior we can obtain a marginal posterior that greatly reduces the dimensionality of the parameter space and thus produces a faster algorithm for estimating the index vector and finding the number and locations of knots.

We propose a fully Bayesian method to analyze GPLSIMs with random residual effects. Section 2 completes the Bayesian framework by specifying the priors on all of the unknown parameters. Section 3 develops a hybrid Gibbs sampler to generate samples from the joint posterior. To speed up the convergence to the target distribution, we treat the marginal posterior of the index vector and the number and positions of knots, which is obtained by integrating over the regression coefficient and the error variance, as a target distribution. Section 4 illustrates the proposed method using simulated and real data examples. Section 5 concludes the paper with a summary.