Applied Mathematics and Computation 339 (2018) 105–112

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Empirical likelihood based inference for generalized additive partial linear models

Zhuoxi Yu[a,c,*], Kai Yang[b], Milan Parmar[a,c]

[a] School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, China
[b] School of Mathematics and Statistics, Changchun University of Technology, Changchun, China
[c] Jilin Province Key Laboratory of Fintech, Changchun, China

ARTICLE INFO

MSC:
62G08
62G20

Keywords:
Generalized Additive partial linear models
Empirical likelihood
Quasi-likelihood equation
χ^2 distribution
Confidence region

ABSTRACT

Empirical-likelihood based inference for the parameters in generalized additive partial linear models (GAPLM) is investigated. With the use of the polynomial spline smoothing for estimation of nonparametric functions, an estimated empirical likelihood ratio statistic based on the quasi-likelihood equation is proposed. We show that the resulting statistic is asymptotically standard chi-squared distributed and the confidence regions for the parametric components are constructed. Some simulations are conducted to illustrate the proposed methods.

1. Introduction

The generalized additive partial linear model (GAPLM) [3] is a realistic, parsimonious candidate when one believes that the relationship between the dependent variable and some of the covariates has a parametric form, while the relationship between the dependent variable and the remaining covariates may not be linear. GAPLM enjoys the simplicity of the generalized linear model (GLM) and the flexibility of the generalized additive model (GAM), because it combines both parametric and nonparametric components.

Let Y be the response variable, $\mathbf{X} = (X_1, \dots, X_{d_1})^{\mathrm{T}} \in R^{d_1}$ and $\mathbf{Z} = (Z_1, \dots, Z_{d_2})^{\mathrm{T}} \in R^{d_2}$ be the covariates. We assume the conditional density of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ belongs to the exponential family

$$f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z}) = \exp[y\xi(\mathbf{x}, \mathbf{z}) - \mathcal{B}\{\xi(\mathbf{x}, \mathbf{z})\} + \mathcal{C}(y)], \tag{1.1}$$

where $\mathcal{B}$ and $\mathcal{C}$ are known functions, ξ is the so-called natural parameter in parametric generalized linear models (GLM), is related to the unknown mean response by

$$\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \mathcal{B}'\{\xi(\mathbf{x}, \mathbf{z})\}.$$

In parametric GLM, the mean function μ is defined via a known link function g by $g\{\mu(\mathbf{x}, \mathbf{z})\} = \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parametric vectors to be estimated. In GAPLM, $g(\mu)$ is modeled as additive partial linear function

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \sum_{k=1}^{d_1} \eta_k(x_k) + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1.2}$$

* Corresponding author at: Jilin University of Finance and Economics, Changchun, China.
E-mail address: 109026@jlufe.edu.cn (Z. Yu).

https://doi.org/10.1016/j.amc.2018.06.050

where $\boldsymbol{\beta}$ is a d_2-dimensional regression parameter, $\{\eta_k\}_{k=1}^{d_1}$ are unknown and smooth functions and $E\{\eta_k(X_k)\} = 0$ for $1 \le k \le d_1$ for identifiability. Furthermore, without loss of generality, we assume the distribution of X_k, $1 \le k \le d_1$ is supported on [0,1].

If the conditional variance function $Var(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \sigma^2 V\{\mu(\mathbf{x}, \mathbf{z})\}$ for some known positive function V, then estimation of the mean can be achieved by replacing the conditional loglikelihood function $\log\{f_{Y|\mathbf{X}, \mathbf{Z}}(y|\mathbf{x}, \mathbf{z})\}$ in (1.1) by a quasi-likelihood function $Q(m, y)$, which satisfies

$$\frac{\partial}{\partial m} Q(m, y) = \frac{y - m}{V(m)}.$$

Wang et al. [10] provided a simple method of estimating $\boldsymbol{\beta}$ and $\{\eta_k\}_{k=1}^{d_1}$ in model (1.2) based on a quasi-likelihood procedure [7] with polynomial spline. However, it is theoretically complicated to prove the convergence results of the maximum likelihood estimates for the nonparametric functions and the asymptotic normality of the estimators for the parametric components. In this paper, we extend the empirical likelihood method to the GAPLM with polynomial spline smoothing for the nonparametric parts. Based on an efficient one-step procedure of maximizing the quasi-likelihood function, an estimated empirical likelihood ratio statistic for the parameter $\boldsymbol{\beta}$ in the GAPLM is defined, and the confidence regions for the parametric components can be constructed accordingly.

The rest of the paper is organized as follows. In Section 2, we propose polynomial spline estimator via a quasi-likelihood approach, the estimated empirical likelihood ratio of the parametric part is defined and the main results are also given. In Section 3, we provide examples based on simulated data, the converge accuracy of the proposed empirical likelihood method is investigated. The proofs of the main results are collected in Section 4.

2. Methodology and main results

Let $(Y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \ldots, n,$ be independent copies of $(Y, \mathbf{X}, \mathbf{Z})$. To avoid confusion, let $\eta_0 = \sum_{k=1}^{d_1} \eta_{0k}(x_k)$ and $\boldsymbol{\beta}_0$ be the true parameter values, respectively. We use polynomial splines to approximate the nonparametric components. We introduce a knot sequence with J interior knots:

$$\tau_{-r+1} = \cdots = \tau_{-1} = \tau_0 = 0 < \tau_1 < \cdots < \tau_J < 1 = \tau_{J+1} = \cdots = \tau_{J+r},$$

where $J \equiv J_n$ increases when sample size n increases. We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order r is a function whose restriction to each subinterval is a polynomial spline of degree $r-1$ and globally $r-2$ times continuously differentiable on [0,1]. The collection of spline with a fixed sequence of knots has a B-spline basis $\{B_1(x), \ldots, B_{J_n+r}(x)\}$. Because of the centering constraint $E\{\eta_k(X_k)\} = 0$, let $N_n = J_n + r - 1$, we adopt the normalized B-spline space $\mathcal{S}_n^0$ introduced in Ref. [11] with the following normalized basis

$$B_{j,k}(x_k) = \sqrt{N_n}\left\{b_{j+1,k}(x_k) - \frac{E(b_{j+1,k})}{E(b_{1,k})} b_{1,k}(x_k)\right\}, 1 \le j \le N_n, 1 \le k \le d_1,$$

for the kth covariate x_k, where $\{b_{j,k}(x_k), j = 1, \ldots, J_n + r, k = 1, \ldots, d_1\}$ be the B-spline basis functions of order r. Let $\mathbf{B}(\mathbf{x}) = \{B_{j,k}(x_k), j = 1, \ldots, N_n, k = 1, \ldots, d_1\}^T$ and $\mathbf{B}_i = \mathbf{B}(\mathbf{X_i})$, and $\boldsymbol{\gamma} = \{\gamma_{j,k}, j = 1, \ldots, N_n, k = 1, \ldots, d_1\}^T$ is the spline coefficient vector, write $\eta(\mathbf{X}_i) = \mathbf{B}_i^T \boldsymbol{\gamma}$. The two-step estimation procedure for η_0 and $\boldsymbol{\beta}_0$ is as follows:

Step 1. Find $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ by maximizing the quasi-likelihood function:

$$(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = \arg\max_{\boldsymbol{\gamma}, \boldsymbol{\beta}} n^{-1} \sum_{i=1}^{n} Q[g^{-1}\{\boldsymbol{B}_i^T \boldsymbol{\gamma} + \mathbf{Z}_i^T \boldsymbol{\beta}\}, Y_i].$$

Then the spline estimator of η_0 is $\hat{\eta}(\mathbf{x}) = \hat{\boldsymbol{\gamma}}^T \mathbf{B}(\mathbf{x})$, and the centered spline estimators of component functions are

$$\hat{\eta}_k(x_k) = \sum_{j=2}^{N_n} \hat{\gamma}_{j,k} B_{j,k}(x_k) - n^{-1} \sum_{i=1}^{n} \sum_{j=2}^{N_n} \hat{\gamma}_{j,k} B_{j,k}(X_{ik}), \qquad k = 1, \ldots, d_1.$$

Step 2. Update $\hat{\boldsymbol{\beta}}$ by maximizing the following function:

$$n^{-1} \sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}(\mathbf{X}_i) + \mathbf{Z}_i^T \boldsymbol{\beta}\}, Y_i], \tag{2.1}$$

with respect to $\boldsymbol{\beta}$.

Step 3. Obtain the final estimator $\hat{\eta}(\mathbf{x})$ by maximizing the quasi-likelihood function:

$$n^{-1} \sum_{i=1}^{n} Q[g^{-1}\{\boldsymbol{B}_i^T \boldsymbol{\gamma} + \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}\}, Y_i],$$