ARTICLE IN PRESS

Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Sure independence screening in ultrahigh dimensional generalized additive models

Guangren Yang [a], Weixin Yao [b], Sijia Xiang [c,*]

[a] Department of Statistics, School of Economics, Jinan University, Guangzhou, 510632, PR China
[b] Department of Statistics, University of California, Riverside, CA 92521, USA
[c] School of Data Sciences, Zhejiang University of Finance & Economics, Hangzhou, Zhejiang, 310018, PR China

ARTICLE INFO

Article history:
Received 22 October 2017
Received in revised form 10 April 2018
Accepted 10 April 2018
Available online xxxx

Keywords:
Generalized additive models
Feature screening
Sure independence screening
Ultrahigh dimensional data
Variable selection

ABSTRACT

Generalized additive models (GAMs) have gained popularity by addressing the curse of dimensionality in multivariate nonparametric regressions with non-Gaussian responses, including continuous, binary or count data. In this paper, we propose a fast and efficient feature screening method for GAMs with ultrahigh dimensional covariates. We provide some theoretical justifications for our screening method and establish the sure screening property. We further examine the finite sample performance of the proposed screening procedure and compare it with some existing methods via Monte Carlo simulations. Three real data examples are used to illustrate the effectiveness of the new method.

1. Introduction

Generalized linear models (GLMs) have been well studied in the literature, and variable selection via penalized likelihood has been developed for GLMs with high-dimensional covariates. With modern data gathering devices and vast data storage space, ultrahigh-dimensional data have been collected in various research areas such as proteomics studies, finance, tumor classification and biomedical imaging. Direct applications of variable selection based on penalized likelihood may not perform well for ultrahigh dimensional data due to the algorithmic stability, computational cost and statistical accuracy. Fan and Lv (2008) proposed a sure independence screening (SIS) procedure for linear models using Pearson correlation coefficient as the marginal utility and further established the sure screening property of their procedure under Gaussian linear model framework. Fan and Song (2010) extended SIS to GLMs with NP-dimensionality. Hall and Miller (2012) proposed a feature screening procedure for transformation of linear models using generalized correlation. Fan et al. (2009) proposed a SIS procedure for generalized linear models based on marginal likelihood estimates. Xia et al. (2016a, b) considered the variable screening with dichotomous response data under generalized varying coefficient models. More details about these marginal feature screening procedures can be found in the recent review paper on feature screening by Liu et al. (2015).

In many applications, it may be too restrictive to assume the effect of all covariates be captured by a simple linear form. Unlike GLMs, generalized additive models (GAMs) introduced by Hastie and Tibshirani (1986, 1990) allow for greater flexibility by modeling the linear predictor as a sum of nonparametric functions of each covariate. GAMs include GLMs as special cases and have less bias. In addition, after fitting a GAM, we can also test whether a GLM is well-specified. GAMs have gained popularity by addressing the curse of dimensionality in multivariate nonparametric regressions with non-Gaussian responses, including continuous, binary or count data. In this paper, we propose a fast and efficient feature

* Corresponding author.
E-mail addresses: tygr@jnu.edu.cn (G. Yang), weixin.yao@ucr.edu (W. Yao), fxbxsj@live.cn (S. Xiang).

https://doi.org/10.1016/j.jspi.2018.04.004
0378-3758/

screening procedure for GAMs with ultrahigh dimensional covariates. We provide some theoretical justifications for our screening method and establish the sure screening property. A greedy version of the proposed feature screening method is also provided to further improve the performance. We examine the finite sample performance of the proposed screening procedure and compare it with some existing methods via Monte Carlo simulations. We further illustrate the effectiveness of the proposed procedure by applications of three real data examples.

Many penalized methods have been proposed to select significant components for additive models when the response is continuous. Cui et al. (2013) proposed method that is a combination of penalized regression spline approximation and group variable selection in additive models. Huang et al. (2010) proposed a two-step approach to select and estimate the nonzero components simultaneously in additive models when p is fixed. It used the group Lasso in the first stage and the adaptive group Lasso in the second stage. Meier et al. (2009) considered the problem of estimating a high-dimensional additive model when $p \gg n$. Ravikumar et al. (2009) studied a new class of methods for high dimensional non-parametric regressions and classifications, namely the sparse additive models. Xue (2009) considered a penalized polynomial spline method for additive models when p is fixed. The method can select significant components and estimate non-parametric additive function components simultaneously. More recently, Fan et al. (2011) suggested several closely related variable screening procedures in sparse ultrahigh dimensional additive models. However, as far as we know, none of the above methods can be directly applied to binary or count data for ultrahigh dimensional setting. Our paper tries to fills this gap and makes additive models much more applicable.

The rest of this paper is organized as follows. In Section 2, we introduce a new feature screening procedure for ultrahigh-dimensional GAMs, and study its theoretical properties. Effective algorithms for the procedure are presented in Section 3. In Section 4, we use Monte Carlo studies and three real data examples to demonstrate the finite sample performance of the new procedure. Some discussion and conclusion remarks are given in Section 5. Technical proofs are deferred to the supplement file.

2. Feature screening procedure for GAM

Let Y be the response, and $\mathbf{x} = (X_1, \ldots, X_p)$ be p-dimensional covariates. We postulate a generalized additive model between Y and $\mathbf{x}$ as follows. Given $\mathbf{x}$, the density function of Y is in the form of

$$f(y; \theta) = \exp\{\theta y - b(\theta) + c(y)\}$$

with respect to a σ-finite measure ν, where $b(\cdot)$ and $c(\cdot)$ are some known functions. Parameter θ is called the natural parameter and the set $\Theta = \{\theta : \int f(y; \theta) d\nu < \infty\}$ is the natural parameter space. We assume that $\theta \in \bar{\Theta}$ with $\bar{\Theta}$ denoting a compact subspace of Θ and $b(\cdot)$ is twice continuously differentiable. Under this model, $E(Y|bx) = b'(\theta) \triangleq \mu$ and $Var(Y|bx) = b''(\theta) \triangleq \nu(\mu)$ with the primes being derivatives. Denote by $\mu(\mathbf{x}) = E(Y|bx)$ the regression. The parameter θ is connected to $\mathbf{x}$ through a pre-specified link function $g(\cdot)$, and so the generalized additive model (GAM) assumes that

$$\eta(\mathbf{x}) \triangleq g\{\mu(\mathbf{x})\} = \beta_0 + \sum_{j=1}^{p} f_j(\mathbf{x}_j), \tag{2.1}$$

where β_0 is a constant and $(f_1(\cdot), \ldots, f_p(\cdot))$ are unspecified smooth regression coefficient functions. To ensure unique identification of $f_j(\cdot)$'s, we assume that $Ef_j(X_j) = 0$, $1 \le j \le p$. The canonical link $g(\cdot) = b'^{-1}(\cdot)$ leads model (2.1) to a particularly simple form as $\theta = \beta_0 + \sum_{j=1}^{p} f_j(\mathbf{x}_j)$. The classical generalized linear model with canonical link covers the normal linear regression, the logistic regression, and the Poisson regression, and so on.

Suppose that $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ constitutes an independent and identically distributed sample. Conditional on $\mathbf{x}_i$, the quasi-likelihood (McCullagh and Nelder, 1989) of the collected data $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is

$$\ell_n\{\eta(\mathbf{x}), \mathbf{y}\} = \frac{1}{n} \sum_{i=1}^{n} \ell(\eta(\mathbf{x}_i), Y_i), \tag{2.2}$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, $\ell(\eta(\mathbf{x}_i), Y_i) = Q\left[g^{-1}\left\{\beta_0 + \sum_{j=1}^{p} f_j(X_{ij})\right\}; Y_i\right]$, and $Q(\mu, y) = \int_y^{\nu} \frac{y-t}{\nu(t)} dt$. A cubic B-spline parameterization method is applied to each nonparametric regression coefficient function $f_j(\cdot)$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \ge 1$ and $\{\psi_{jk}, k = 1, \ldots, d_{n_j}\}$ denote a normalized B-spline basis with $\|\psi_{jk}\|_\infty \le 1$ and $d_n = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the sup norm (Stone, 1982, 1985). For any $f_{nj} \in \mathcal{S}_n$, it is assumed that

$$f_{nj}(x) = \sum_{k=1}^{d_{n_j}} \beta_{jk} \psi_{jk}(x) = \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\psi}_j(x), \quad j = 1, \ldots, p, \tag{2.3}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_{n_j}}$. Here we allow d_{n_j} to increase as n increases, and be different for different j's since different coefficient functions may have different smoothness. Under some conditions, the nonparametric coefficient functions $\{f_j(\cdot)\}_{j=1}^{p}$ can be well approximated by functions in $\mathcal{S}_n$, i.e., $\eta(\mathbf{x}) \approx \sum_{j=0}^{p} \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\psi}_j(X_{ij}) \widehat{=} \tilde{\eta}(\mathbf{x})$, where $\boldsymbol{\beta}_0 = (\beta_0, 0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{\psi}_0(\cdot) = \boldsymbol{\psi}_0 = (1, \ldots, 1)^{\mathrm{T}}$.

Please cite this article in press as: Yang G., et al., Sure independence screening in ultrahigh dimensional generalized additive models. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.04.004.