Journal of Multivariate Analysis 123 (2014) 304–310

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric Bayesian information criterion for model selection in ultra-high dimensional additive models

Heng Lian

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 19 January 2012
Available online 12 October 2013

AMS subject classifications:
62G20
62G05

Keywords:
Bayesian information criterion (BIC)
Selection consistency
Sparsity
Ultra-high dimensional models
Variable selection

ABSTRACT

For linear models with a diverging number of parameters, it has recently been shown that modified versions of Bayesian information criterion (BIC) can identify the true model consistently. However, in many cases there is little justification that the effects of the covariates are actually linear. Thus a semiparametric model, such as the additive model studied here, is a viable alternative. We demonstrate that theoretical results on the consistency of the BIC-type criterion can be extended to this more challenging situation, with dimension diverging exponentially fast with sample size. Besides, the assumptions on the distribution of the noises are relaxed in our theoretical studies. These efforts significantly enlarge the applicability of the criterion to a more general class of models.

1. Introduction

With rapid increases in the production of large dimensional data by modern technology, more and more studies have focused on variable selection problems where the goal is to identify the few relevant predictors among a large collection of predictors, which might even outnumber the sample size due to the constraint of experimental costs. For example, in microarray experiments investigating the genetic mechanisms of a certain disease, thousands of genes are assayed all at once while the number of samples is constrained by the cost of arrays as well as by the rarity of the disease in the population.

In linear models with a fixed dimension, performances of various criteria for variable selection are well known [11], including AIC [1], BIC [10], C_p [8], etc. In particular, BIC was shown to be consistent in variable selection. More recently, penalization approaches to variable selection have drawn increasing attention due to their stability and computational attractiveness [13,19,5,21,17]. Following this trend, Wang et al. [14] has shown that BIC computed along the solution path of the penalized estimator is also selection consistent.

Nevertheless, these traditional criteria are too liberal for regression problems with high dimensional covariates, in that they tend to incorporate many spurious covariates in the model selected. On the positive side, modifications of BIC by using a statistically motivated larger penalty term can successfully address this problem, make the criterion provably consistent, and exhibit satisfactory performance in real applications [15,2]. Despite these efforts, the works mentioned above, particularly the theoretical investigations, entirely focused on parametric linear models with Gaussian noise, while in many applications there is little a priori justification that the covariates actually have such simple linear effects on the responses.

The additive model introduced by Stone [12] represents a more flexible class of semiparametric models that allow a general transformation of each covariate to enter as an additive component. This raises an interesting question: is there an

E-mail address: henglian@ntu.edu.sg.

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.09.015

appropriately modified BIC-type criterion that can consistently identify the nonzero components in this class of semiparametric models? Although a similar question has been answered in an affirmative way in [16] for fixed-dimensional varying-coefficient models, it remains a conjecture for high dimensional semiparametric problems. We note that Huang et al. [7] has used the modified BIC-type criterion in selecting the tuning parameter in the group LASSO penalty for additive models, but they did not demonstrate the theoretical properties of such a criterion. Compared to parametric models, the approximation errors for the component functions pose additional challenges to the analysis.

In this paper, we will investigate the theoretical properties of BIC-type criterion in additive models with the number of components p growing much faster than sample size n. To be more specific, we assume $\log p = o(n^{2d/(2d+1)})$, where d characterizes the smoothness (roughly the number of derivatives) of the component functions. Following the existing literature, we say the problem has an ultra-high dimensionality. On the other hand, the number of truly nonzero components is assumed to be fixed and does not diverge with sample size. In stark contrast to our result, Chen and Chen [2] showed the BIC consistency by imposing this assumption of fixed number of nonzero components with growing number of parameters $p = O(n^\kappa)$ for $\kappa > 0$. On the other hand, Wang et al. [15] does not impose this condition at the cost of more slowly increasing number of parameters $p = O(n^\kappa)$ for $\kappa < 1$. The assumption of a fixed number of nonzero components enables us to impose only simple and natural assumptions that are easy to interpret. In particular, we will use Lemma 3 in [7] which was only shown for fixed dimensions, whose extension to diverging dimensions seems to be totally nontrivial. Besides, although we acknowledge that it might be restrictive to assume that all components have the same smoothness, it would be hard, if not impossible, to satisfactorily deal with the more general case. Finally, it is worth noting that we relax the Gaussian noise assumption used in [2,15] to sub-Gaussian noises. The Gaussian assumption was key to make the theoretical analysis tractable in those studies (see for example (B.3) in [15]). With sub-Gaussian noises, we need to resort to study the tail probability of some quadratic forms involving sub-Gaussian random variables.

2. Bayesian information criterion for unpenalized polynomial spline estimators

Consider a regression problem with observations (Y_i, X_i), $i = 1, \ldots, n$ that are independent and identically distributed (i.i.d.) as (Y, X), where Y is a scalar response and $X = (X_1, \ldots, X_p)^T$ contains p covariates. Substantial progress has been made on linear regression when p is large, with or without penalty. Since fitting fully nonparametric models is infeasible for large dimensions, an elegant solution to relax the strong linearity assumption, known as the additive model [12,6], was proposed to avoid this difficulty, which is specified by

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i, \tag{1}$$

where μ is the intercept, f_j are unknown univariate component functions and ϵ_i are i.i.d. mean zero noises.

Without loss of generality, we assume that the distribution of X_j is supported on $[0, 1]$ and also impose the condition $Ef_j(X_j) = 0$ for identifiability. We use polynomial splines to approximate the components. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1})$, $k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots although a data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q-1$ and globally $q-2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + q$. More specifically, the B-spline basis functions of order 1 for a knot sequence are

$$B_k^{(1)}(t) := \begin{cases} 1, & \text{if } t_k \le t < t_{k+1} \\ 0, & \text{otherwise.} \end{cases}$$

Starting from this, higher-order B-splines can be defined recursively by

$$B_k^{(m)} := \omega_{k,m} B_k^{(m-1)} + (1 - \omega_{k+1,m}) B_{k+1}^{(m-1)}$$

with

$$\omega_{k,m} := \begin{cases} \dfrac{t - t_k}{t_{k+m-1} - t_k} & \text{if } t_k \ne t_{k+m-1} \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that the second order B-spline basis functions are piecewise linear, third order B-spline basis functions are piecewise quadratic, etc. More details on B-splines can be found in the popular monograph [3]. B-splines of order four (also called cubic splines) are the most frequently used ones in the literature. We get rid of the superscript (q) for the basis functions for simplicity of notation.

Because of the centering constraint $Ef_j(X_j) = 0$, we instead focus on the subspace of spline functions $S_j^0 := \{s : s = \sum_{k=1}^{K} b_{jk} B_{jk}(x), \sum_{i=1}^{n} s(X_{ij}) = 0\}$ with basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^{n} B_k(X_{ij})/n, k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). Using spline expansions, we can approximate the