应用概率统计 第 35 卷
第 2 期 2019 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2019, Vol. 35, No. 2, pp. 126-140
doi: 10.3969/j.issn.1001-4268.2019.02.002

Marginal Empirical Likelihood Independence Screening in Sparse Ultrahigh Dimensional Additive Models *

ZHANG Junying[1,2] ZHANG Riquan[1,3⋆] WANG Hang[2] LU Zhiping[1]

([1] *School of Statistics, East China Normal University, Shanghai, 200062, China*)

([2] *Department of Mathematics, Taiyuan University of Technology, Taiyuan, 030024, China*)

([3] *Department of Mathematics, Shanxi Datong University, Datong, 037009, China*)

Abstract: The additive model is a more flexible nonparametric statistical model which allows a data-analytic transform of the covariates. When the number of covariates is big and grows exponentially with the sample size the urgent issue is to reduce dimensionality from high to a moderate scale. In this paper, we propose and investigate marginal empirical likelihood screening methods in ultra-high dimensional additive models. The proposed nonparametric screening method selects variables by ranking a measure of the marginal empirical likelihood ratio evaluated at zero to differentiate contributions of each covariate given to a response variable. We show that, under some mild technical conditions, the proposed marginal empirical likelihood screening methods have a sure screening property and the extent to which the dimensionality can be reduced is also explicitly quantified. We also propose a data-driven thresholding and an iterative marginal empirical likelihood methods to enhance the finite sample performance for fitting sparse additive models. Simulation results and real data analysis demonstrate the proposed methods work competitively and performs better than competitive methods in error of a heteroscedastic case.

Keywords: marginal empirical likelihood screening; nonparametric regression model; variable selection; dimensionality reduction

2010 Mathematics Subject Classification: Primary 62G05; Secondary 62E20

Citation: ZHANG J Y, ZHANG R Q, WANG H, et al. Marginal empirical likelihood independence screening in sparse ultrahigh dimensional additive models [J]. Chinese J Appl Probab Statist, 2019, 35(2): 126–140.

§1. Introduction

In current practical problems, high-dimensional data are more frequently seen in finance, biomedical sciences, geological studies and many more areas. Statistical methods

*The research was supported in part by the National Natural Science Foundation of China (Grant Nos. 11171112; 11201190), the Doctoral Fund of Ministry of Education of China (Grant No. 20130076110004) and the 111 Project of China (Grant No. B14019).

⋆Corresponding author, E-mail: zhangriquan@163.com.

Received May 15, 2017. Revised April 16, 2018.

for high dimensional data analysis play more important roles to deal with large volume of data containing considerably many features. The general cases can be considered as the number of variables p may be larger than the number of observations n. We often assume $\ln p = O(n^{\iota})$ for some $\iota \in (0, 1/2)$ that can be seen in [1–3] for overviews. Identifying relevant features becomes a fundamental objective of statistical analysis with high dimensional data.

To selection variables more effectively, statisticians proposed and investigated different screening methods to eliminate uncorrelated variables. Sure independent screening procedure arrives, for example, Fan and Lv [4] for linear model, Fan and Song [5] for generalized linear models, Fan et al. [6] for nonparametric additive models, He et al. [7] for model-free nonparametric quantile regression, respectively. Fan and Lv [4] and Fan and Song [5] screened variables by ranking the absolute values of marginal estimates of model coefficients, and Fan et al. [6] performed screening by ranking integrated squared marginal nonparametric curve estimates. Fan and Song [5] also carried out independence screening by examining the magnitudes of the likelihood ratios. He et al. [7] screened variables by marginal nonparametric quantile curve estimates. More development about feature screening are in [8–12] and so on.

In this paper we consider the additive models, as $Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon$, introduced by Stone [13]. In the screening framework based marginal empirical likelihood [14, 15], our technical shares some similarity with that [12] in which the empirical likelihood screening is discussed for linear model and the the generalized linear models. The empirical likelihood [14, 15] is demonstrated effectively with less restrictive distributional assumptions for statistical inferences by incorporating the moment constraints into the classical likelihood-based framework; see [16, 17] and reference therein. The statistical literature contains recently numerous procedures of the empirical likelihood approach to deal with high-dimensional data; see [18–22]. The empirical likelihood approach, however, encounters built-in challenge when data dimensionality is high. We refer to [12] in which the marginal empirical likelihood was introduce to screen variables in linear model with ultra-high dimensional data and we propose the marginal empirical likelihood screening method for additive model. we apply B-spline bases to appropriate the marginal effects of additive components and select the important variables by ranking a measure of the marginal empirical likelihood ratio evaluated at zero. Simulation results and real data analysis demonstrate the proposed methods work competitively and performs better than competitive methods in error of a heteroscedastic case.

The remainder of the article is organized as follows. In Section 2, we introduce the nonparametric marginal empirical likelihood for additive model. In this section we describe the marginal empirical likelihood methodology and present the theoretical properties for nonparametric independent screening. Section 3 gives the algorithm about an iterative sure screening procedure. Numerical examples and real data analysis are given in Section 4. We relegate the proofs to Section 5.

§2. Marginal Empirical Likelihood of Nonparametric Additive Model

2.1 Models and Notation

Let $Y_i \in R^1$ $(i = 1, 2, \ldots, n)$ be the response from the ith subject, $\boldsymbol{x}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\tau \in R^p$ be the associated p-dimensional predictor, X_j is the jth covariate and its n independent observation denoted as $(X_{1j}, X_{2j}, \ldots, X_{nj})^\tau$. We assume that

$$Y_i = \sum_{j=1}^{p} m_j(X_{ij}) + \varepsilon_i, \tag{1}$$

where $m_j(\cdot)$ is a general unknown smooth function, ε_i is the random error with mean 0. Let $\mathscr{M}_* = \{j : \mathsf{E}[m_j(X_j)]^2 > 0\}$ be the true sparse model and nonsparsity size is $s = |\mathscr{M}_*|$. We allow p to grow with n and denote it by p_n.

Without loss of generality, we assume that each X_{ij} takes values on the interval $[0, 1]$. Let $\mathscr{S}$ be the space of the functions defined in condition A1 in Section 2.3 and $0 = t_0 < t_1 < \cdots < t_k = 1$ be a partition of the interval. Using the t_i as knots, we construct $N = k + l$ normalized B-spline basis functions which form a basis of order $l + 1$. We denote these basis functions as vector $B(t) = \{B_1(t), B_2(t), \ldots, B_N(t)\}^\tau$, where $\|B(t)\|_\infty \leqslant 1$ and $\|\cdot\|_\infty$ denotes the sup norm. Assume that $f_j(t) \in \mathscr{S}$. Then $f_j(t)$ can be well approximated by a linear combination of the basis functions $B^\tau(t)\beta_j$, for some $\beta_j \in R^N$. Denote nonparametric marginal projections as $\{f_j(X_{ij})\}_{j=1}^p$ that is approximated as $\{f_{nj}(X_{ij})\}_{j=1}^p$, a.e., $f_{nj}(X_{ij}) = B^\tau(X_{ij})\beta_j$.

2.2 Marginal Empirical Likelihood Methodology

We firstly consider the marginal moment condition of the least squares estimator:

$$\mathsf{E}\{B_k(X_j)[Y_i - B^\tau(X_j)\beta_j]\} = 0 \qquad (k = 1, 2, \ldots, N,\ j = 1, 2, \ldots, p). \tag{2}$$

From (2) we can note that $\mathsf{E}[B_k(X_j)B^{\mathsf{T}}(X_j)\beta_j] = \mathsf{E}[B_k(X_j)Y_i]$. Let $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$ be collected independent data, $g_{ijk}(\beta) = B_k(X_{ij})[Y_i - B^{\mathsf{T}}(X_{ij})\beta]$ $(k = 1, 2, \ldots, N,\ j = 1, 2, \ldots, p)$. We define the following marginal empirical likelihood:

$$L_{jk}(\beta_{jk}) = \sup\Big\{ \prod_{i=1}^n \omega_i : \omega_i \geqslant 0,\ \sum_{i=1}^n \omega_i = 1,\ \sum_{i=1}^n \omega_i g_{ijk}(\beta_{jk}) = 0\Big\} \tag{3}$$

for $j = 1, 2, \ldots, p,\ k = 1, 2, \ldots, N$. For any given β such that 0 in the convex hull of $\{B_k(X_{ij})Y_i\}_{i=1}^n$, the marginal empirical likelihood ratio is defined as

$$\ell_{jk}(\beta_{jk}) = -2\ln[L_{jk}(\beta_{jk})] - 2n\ln n = 2\sum_{i=1}^n \ln[1 + \lambda g_{ijk}(\beta_{jk})], \tag{4}$$

where λ is the Lagrange multiplier satisfying

$$0 = \sum_{i=1}^n \frac{g_{ijk}(\beta)}{1 + \lambda g_{ijk}(\beta)}. \tag{5}$$

To obtain the objective of dimensional reducing we select active variable X_j if marginal signal $|f_j(X_j)|$ is more larger than some a given positive constant, i.e., $|B^{\mathsf{T}}(X_{ij})\beta|$ is large enough, where $|a|$ denotes the absolute value of a. We know that $|B^{\mathsf{T}}(X_{ij})\beta|$ may be small when β_k $(k = 1, 2, \ldots, N)$ are not all 0. However, to differentiate parametric marginal signal by $\mathsf{E}[B_k(X_j)Y_i]$, it is only needed $\beta_j = 0$. We test the null hypothesis $H_0 : \beta_j = 0$ so that $\ell_{jk}(0)$ has a very clear practical interpretation for the marginal empirical likelihood ratio (4) with $g_{ijk}(\beta_{jk}) = B_k(X_{ij})[Y_i - B^{\mathsf{T}}(X_{ij})\beta_{jk}]$. So we can use $\ell_{jk}(0)$ as a device for feature screening.

Let $\ell_j(0) = \max\{\ell_{j1}(0), \ell_{j2}(0), \ldots, \ell_{jN}(0)\}$. We select the variable as

$$\widehat{\mathscr{M}}_{\gamma_n} = \{1 \leqslant j \leqslant p : \ell_j(0) \geqslant \gamma_n\},$$

where the threshold level γ_n is predefined.

2.3 Main Results

To achieve the theoretical basis of the sure screening, we impose the following assumption:

A1 The nonparametric marginal projections $\{f_j\}_{j=1}^p \in \mathscr{S}$. $\mathscr{S}$ is the collection of a class of functions defined on $[0, 1]$, whose rth derivative f_j^r exists and is Lipschitz of order α: $|f_j^r(s) - f_j^r(t)| \leqslant K|s - t|^\alpha$ for some positive constant K, where $s, t \in [0, 1]$ and $\alpha \in (0, 1]$ such that $d = r + \alpha > 0.5$.

A2 Random variables Y has bounded variance and $\min\limits_{j \in \mathscr{M}_*} \mathsf{E}|\mathsf{E}(Y \mid X_j)| \geqslant c_1 N^{1/2} n^{-\kappa}$ for some $0 < \kappa < d/(2d + 1)$ and $c_1 > 0$.

A3 There exist positive constant c_3 such that $N^{-d-1/2} = c_3 n^{-\kappa}$.

A4 There exist positive constants T_1, T_2, ι_1, ι_2 such that

$$\begin{aligned} &\mathsf{P}\{|X_j| > z\} \leqslant T_1 \exp(-T_2 z^{\iota_1}) \qquad \text{for } u > 0 \ (j = 1, 2, \ldots, p), \\ &\mathsf{P}\{|Y| > z\} \leqslant T_1 \exp(-T_2 z^{\iota_2}) \qquad \text{for } u > 0 \ (j = 1, 2, \ldots, p). \end{aligned}$$

From [12], $\ell(\beta_j)$ was showed to tend towards infinity in probability with $n \to \infty$, and $\ell(\beta_j)$ don't exceed some number depending n in probability, under some additions. We now establish the general result for the distribution of empirical likelihood ratio.

Theorem 1 Under assumptions A1 − A4, there exists a positive constant C_1 depending only on T_1, T_2, ι_1 and ι_2 appeared in additions A4 such that

$$\begin{aligned} &\max_{j \in \mathscr{M}_*} \mathsf{P}\{\ell_j(0) < c_5^2 N n^{2\tau}\} \\ &\leqslant \begin{cases} \exp(-C_1 N n^{(1-\kappa) \wedge [(1-\kappa-\tau)\iota/2]}), & \text{if } (1-\kappa)(1-\delta) < 1; \\ \exp(-C_1 N n^{[(1-\kappa/2)/(1+\delta)] \wedge [(1-\kappa-\tau)(\iota/2)]}), & \text{if } (1-\kappa)(1+2\delta) \geqslant 1, \end{cases} \end{aligned}$$

where $\iota = \iota_1 \iota_2 / (\iota_1 + \iota_2)$ and $\delta = \max\{2/\iota - 1, 0\}$.

Theorem 2 Under assumptions A1 − A4, there exists a positive constant C_1 depending only on T_1, T_2, ι_1 and ι_2 appeared in additions A4 such that, for any $\tau \in (0, (1-\kappa)/2)$ and $\gamma_n = c_5^2 N n^{2\tau}$,

$$\begin{aligned} &\mathsf{P}\{\mathscr{M}_* \subset \widehat{\mathscr{M}_{\gamma_n}}\} \\ &\geqslant \begin{cases} 1 - s\exp(-C_1 N n^{(1-\kappa) \wedge [(1-\kappa-\tau)\iota/2]}), & \text{if } (1-\kappa)(1-\delta) < 1; \\ 1 - s\exp(-C_1 N n^{[(1-\kappa/2)/(1+\delta)] \wedge [(1-\kappa-\tau)(\iota/2)]}), & \text{if } (1-\kappa)(1+2\delta) \geqslant 1, \end{cases} \end{aligned}$$

where $\iota = \iota_1 \iota_2 / (\iota_1 + \iota_2)$ and $\delta = \max\{2/\iota - 1, 0\}$.

We know that the results of Theorem 1 and Theorem 2 holds under the conditions $N = O(n^{2\kappa/(1+2d)})$ and

$$\ln p = \begin{cases} o(N n^{(1-\kappa) \wedge [(1-\kappa-\tau)\iota/2]}), & \text{if } (1-\kappa)(1-\delta) < 1; \\ o(N n^{[(1-\kappa/2)/(1+\delta)] \wedge [(1-\kappa-\tau)(\iota/2)]}), & \text{if } (1-\kappa)(1+2\delta) \geqslant 1. \end{cases}$$

§3. Conditional Permutation Iterative Screening Method

To tackle the problem of correlated explanatory variables, we propose the following random permutation iterative screening method.

Step 1: Compute $\ell_{jk}(0)$ for $k = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, p$ by (4) and (5). Select the top K explanatory variables by ranking their marginal empirical likelihood $\ell_j(0)$, where $\ell_j(0) = \max\limits_{k\in\{1,2,\ldots,N\}} \{\ell_{jk}(0)\}$. We select $K = 1$ or the number of less than the number of prespecified covariables. Denote the index set of selected variables as $\mathscr{M}_0$.

Step 2: For data $\{(X_j, Y),\ j \in \mathscr{M}_0\}$, apply B-spline to estimate $m_j(X_j)$ and get $\widehat{m}_j(X_j)$. Conddition on $\mathscr{M}_0$, the partial residual is

$$Y^* = Y - \sum_{j\in\mathscr{M}_0} \widehat{m}_j(X_j).$$

Compute $\ell_j^*(0)$ using $\{(X_j, Y^*),\ j \in \mathscr{M}_0^c\}$. We determine the threshold value by applying random permutation on the partial residual Y^*, which yields Y_Q^*, where $Y_Q^* = \{Y_{Q1}^*, Y_{Q2}^*, \ldots, Y_{Qn}^*\}^\tau$ and $\{Q1, Q2, \ldots, Qn\}$ is a permutation of $\{1, 2, \ldots, n\}$. Compute $\ell_j^{Q*}(0)$ for $\{(X_j, Y_Q^*),\ j \in \mathscr{M}_0^c\}$. Let γ_q^* be the qth-ranked magnitude of $\{\ell_j^{Q*}(0),\ j \in \mathscr{M}_0^c\}$. Then, the active variable set of variables is chosen as

$$\mathscr{A}_1 = \{j : \ell_j^*(0) \geqslant \gamma_q^*,\ j \in \mathscr{M}_0^c\} \cup \mathscr{M}_0.$$

In our numerical studies, $q = 1$.

Step 3: Apply penalized empirical likelihood [21] to explanatory variables in $\mathscr{A}_1$ to select variables and get $\mathscr{M}_1$.

Step 4: Repeat Step 2 and Step 3 k steps until $\mathscr{A}_k = \mathscr{A}_{k+1}$ or $\mathscr{A}_{k+1}$ reaches a pre-specified number.

§4. Numerical Results

In this section, we study the performance of our proposed methods on the simulated data and in a real data analysis. We appropriate the marginal function by cubic B-spline, $N = 7$. We compare our proposed nonparametric empirical likelihood-based screening procedure (denoted by NEL-SIS) and corresponding iterative procedure (denoted by NEL-ISIS), with the screening methods proposed in [6] (denoted by NLS-SIS and NLS-ISIS) and [7] (denoted by QaSIS) for nonparametric additive models, respectively. For the method QaSIS [7], we consider two case that the quantile $\alpha = 0.5$ and $\alpha = 0.75$. We set $n = 100$ and $p = 1,000$ and 200 repetition for all the examples. We get the number of the final model with the respective SCAD penalized variables and screen to a much smaller number d of explanatory variables. For each example, we report the number that each important explanatory variables is selected in the final model for 200 repetitions and the average number of unimportant explanatory variables being selected.

Let

$$h_1(x) = x, \qquad h_2(x) = (2x-1)^2, \qquad h_3(x) = \frac{\sin(2\pi x)}{2-\sin(2\pi x)},$$

and

$$h_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3.$$

Example 3 The response is generated as $Y = 5h_1(X_1) + 3h_2(X_2) + 4h_3(X_3) + 6h_4(X_4) + \sqrt{1.74}\varepsilon$, with ε being independent of explanatory variables. We consider ε as three different distribution $\mathrm{N}(0,1)$, $\mathrm{N}(0,2^2)$, t_4. The covariates are generated as follows:

$$X_j = \frac{U_j + tU_{p+1}}{1+t}, \qquad j = 1, 2, \ldots, p,$$

where $\{U_i\}_{i=1}^{p+1}$ are i.i.d. uniform random variables on $[0,1]$. When $t=0$, X_i is uncorrelated with X_j, $i \neq j$ as $t=1$ the pairwise correlation of covariates is 0.5.

Example 4 The example has 12 important variables with different coefficients

$$\begin{aligned} Y = {} & h_1(X_1) + h_2(X_2) + h_3(X_3) + h_4(X_4) \\ & + 1.5h_1(X_5) + 1.5h_2(X_6) + 1.5h_3(X_7) + 1.5h_4(X_8) \\ & + 2h_1(X_9) + 2h_2(X_{10}) + 2h_3(X_{11}) + 2h_4(X_{12}) + \sqrt{0.518}\varepsilon. \end{aligned}$$

The covariates generated and the error ε are as in Example 3.

We set random samples of size $n = 100$ and $d = [n/(2\ln n)] = 10$ in Example 3 and $d = [n/(\ln n)] = 21$ in Example 4, where $[a]$ denotes the largest integer that is less than or equal to a. The results of Example 3 and Example 4 are reported in Table 1 and Table 2, where we report the number of repetitions of the important explanatory variables selected. We report their average number of repetitions of unimportant explanatory variables. It shows that the proposed empirical likelihood-based screening methods perform very competitively compared to the screening method proposed by [6] and similarly by iterative algorithm, respectively. When true important variables are fewer, the selection effects gets better.

Example 5 We use the example of [23]. The data is generated as $Y = 2Z_1 + 2Z_2 + 2Z_3 - 3\sqrt{2}Z_4 + \varepsilon$, where ε is simulated as three different distribution $\mathrm{N}(0,1)$, $\mathrm{N}(0,2^2)$, t_4. The covariates $Z_1, Z_2, \ldots, Z_p$ are jointly Gaussian, marginally $\mathrm{N}(0,1)$, and with $\mathrm{corr}(Z_j, Z_4) = 1/\sqrt{2}$ for all $j \neq 4$ and $\mathrm{corr}(Z_i, Z_j) = 1/2$ if $i \neq j$, $i \neq 4$, $j \neq 4$. Note Z_4 is independent of Y, even though it is the most important variable in the joint model. In Example 5 we set random samples of size $n = 100$ and $d = [n/(2\ln n)] = 10$. The results are reported in Table

Table 1 Model selection results for Example 3

ε		Method	X_1	X_2	X_3	X_4	Unimportant explanatory variables
$\mathrm{N}(0,1)$	$t=0$	NLS-NIS	196	196	199	198	1.516013
		NEL-NIS	192	190	189	182	1.533261
		NLS-INIS	200	200	200	200	0.987642
		NEL-INIS	200	200	200	200	0.723519
	$t=1$	NLS-NIS	196	195	197	200	1.530386
		NEL-NIS	196	186	188	183	1.542543
		NLS-INIS	200	200	200	200	1.504923
		NEL-INIS	200	200	200	200	1.097082
$\mathrm{N}(0,2^2)$	$t=0$	NLS-NIS	195	197	199	196	1.516123
		NEL-NIS	192	185	183	185	1.520312
		NLS-INIS	200	200	200	200	1.502164
		NEL-INIS	200	200	200	200	1.089732
	$t=1$	NLS-NIS	195	196	195	198	1.518326
		NEL-NIS	190	192	189	183	1.550329
		NLS-INIS	200	200	200	200	1.512904
		NEL-INIS	200	200	200	200	1.102114
t_4	$t=0$	NLS-NIS	194	198	198	199	1.517438
		NEL-NIS	190	186	183	182	1.551326
		NLS-INIS	200	200	200	200	1.503214
		NEL-INIS	200	200	200	200	1.087921
	$t=1$	NLS-NIS	194	199	198	196	1.518458
		NEL-NIS	190	186	182	185	1.548921
		NLS-INIS	200	200	200	200	1.504116
		NEL-INIS	200	200	200	200	1.095410

3. We note from Table 3 that the proposed empirical likelihood-based screening methods is challenged by the important explanatory variable Z_4 but the corresponding iterative screening can easily pick it up and perform competitively compared to the screening method proposed by [6].

Example 6 We investigate the performance of the nonparametric empirical likelihood-based screening procedure under a heteroscedastic example. The data is generalized as following: $Y = c[h_1(X_1) - h_2(X_2) + h_3(X_3)] + \varepsilon/(X_1^2 + X_2^2 + X_3^2)$ with independent

Table 2 Model selection results for Example 4

ε		Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	Unimportant variables
N(0,1)	$t=0$	NLS-NIS	192	191	187	187	191	188	182	180	185	184	180	180	1.553213
		NEL-NIS	189	183	180	182	184	180	178	177	175	175	172	171	1.586431
		NLS-INIS	200	198	198	195	198	195	193	192	195	195	194	192	1.132041
		NEL-INIS	196	192	192	189	193	190	185	186	192	190	186	186	1.042106
	$t=1$	NLS-NIS	192	190	187	188	189	186	183	180	184	183	178	182	1.622905
		NEL-NIS	192	186	185	183	185	184	180	178	180	175	178	180	1.638423
		NLS-INIS	199	197	197	195	198	195	197	199	196	194	194	192	1.102865
		NEL-INIS	199	196	195	195	195	193	195	197	195	191	195	195	1.08632
$N(0,2^2)$	$t=0$	NLS-NIS	192	190	190	189	192	187	183	179	186	183	180	181	1.553185
		NEL-NIS	188	183	178	181	183	176	176	177	175	174	171	170	1.553890
		NLS-INIS	199	199	198	196	198	195	194	193	194	195	193	193	1.102232
		NEL-INIS	194	193	193	189	191	189	185	185	190	190	184	183	1.002365
	$t=1$	NLS-NIS	192	189	187	188	189	186	184	181	185	184	177	182	1.622915
		NEL-NIS	190	185	185	182	185	184	180	176	180	173	176	180	1.627295
		NLS-INIS	200	196	197	195	198	195	193	193	194	195	193	193	1.102439
		NEL-INIS	199	196	194	194	195	192	193	196	195	191	192	192	1.08465
t_4	$t=0$	NLS-NIS	192	191	190	189	192	187	182	180	186	183	181	181	1.553169
		NEL-NIS	190	185	180	183	183	180	178	178	176	175	174	174	1.550321
		NLS-INIS	200	199	197	197	198	195	194	193	194	195	193	193	1.102264
		NEL-INIS	194	193	195	195	196	193	194	192	194	195	192	193	1.002342
	$t=1$	NLS-NIS	193	189	188	187	188	186	185	181	186	184	178	181	1.622890
		NEL-NIS	190	184	184	182	183	183	178	178	179	173	175	175	1.627287
		NLS-INIS	200	196	196	196	197	196	193	192	194	194	193	193	1.102428
		NEL-INIS	199	195	194	194	195	190	192	194	194	190	191	192	1.08273

$\varepsilon \sim \mathrm{N}(0,1)$, $X_j \sim \mathrm{N}(0,1)$ $(j=1,2,3)$, $\mathrm{Cov}(X_i, X_j) = 0$ $(i,j=1,2,3)$ for $i \neq j$ and $c > 0$ controls the signal level. In Example 6 we set random samples of size $n = 100$ and $d = [n/(2\ln n)] = 10$. The results are reported in Table 4. It shows that the proposed empirical likelihood-based screening methods perform better compared to the screening method proposed by [6], which is affected by the heteroscedasticity.

From Example 3 – Example 6, we can know our proposed nonparametric empirical likelihood-based screening procedure are better methods under the heteroscedasticity. In

Table 3 Model selection results for Example 5

ε	Method	Z_1	Z_2	Z_3	Z_4	Unimportant explanatory variables
$N(0,1)$	LS-INIS	196	196	195	0	1.523128
	EL-INIS	196	194	191	0	1.527039
	LS-INIS	200	198	198	193	1.103980
	EL-INIS	199	199	198	188	0.853952
$N(0,2^2)$	LS-INIS	196	195	196	0	1.521642
	EL-INIS	195	193	190	0	1.528437
	LS-INIS	200	199	197	193	1.234326
	EL-INIS	200	200	196	186	0.923824
t_4	LS-INIS	196	193	193	0	1.522365
	EL-INIS	198	193	195	0	1.526341
	LS-INIS	200	199	196	193	1.210331
	EL-INIS	200	198	196	182	0.812391

Table 4 Model selection results for Example 6

c	Method	X_1	X_2	X_3	Unimportant explanatory variables
1	LS-INIS	140	132	130	1.234648
	EL-INIS	185	180	176	1.056392
1.5	LS-INIS	169	165	162	1.153876
	EL-INIS	193	190	187	1.082694
2	LS-INIS	180	175	173	1.062794
	EL-INIS	190	185	183	1.010843

symmetric and homoscedasticity, our propose methods are similar to the least squares screening by iterative and our proposed methods have uncomplicated algorithm and more easy compute.

A real data example To illustrate the application of our proposed method we use the dataset reported by [24] analyzed by [25] and [6]. For this dataset, 120 12-week old male rats were selected to harvest tissue from the eyes and subsequent microarray analysis. By microarrays we analyze the RNA from the eyes of these animals contain more than 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multichip averaging method by

[26] to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale.

Following [25] and [6] we were finding the genes that are related to the TRIM32 gene, which was recently found to cause Bardet-Biedl syndrome in [27] and is a genetically heterogeneous disease of multiple organ systems, including the retina. More than 30,000 probe sets are represented on the Rat Genome 230 2.0 Array, but many of these are not expressed in the eye tissue. We use the dataset of the 18,975 probes that are expressed in the eye tissue, were too slowly following [25]. So we used 2,000 probe sets that are expressed in the eye and have the greatest marginal correlation with TRIM32 in the analysis. For the subset of the data ($n = 120$, $p = 2,000$), we apply the proposed empirical likelihood-based screening methods and the method of [6] to model the relationships between the expression of TRIM32 and expression of the 2,000 genes ($p = 2,000$). NLS-SIS selects the following 7 probes: 1371755_at, 1373534_at, 1373944_at, 1376686_at, 1374669_at, 1376747_at, 1377880_at. NEL-SIS selects the following 5 probes: 1389584_at, 1379971_at, 1373944_at, 1376686_at, 1377187_at. NLS-ISIS selects the following 7 probes: 1371755_at, 1373534_at, 1373944_at, 1380033_at, 13782263_at, 1373776_at, 1374106_at. NEL-ISIS selects the following 6 probes: 1385944_at, 1382835_at, 1373944_at, 1376686_at, 1373776_at, 1382263_at.

§5. Proofs

Lemma 7 (Proposition 1 in [12]) Suppose that $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed random variables with $\mathsf{E}(|Z_i|^w) < \infty$ for some $w > 3$. In (4) and (5) let $g_{ijk} = Z_i - z$ for all $i = 1, 2, \ldots, n$ and obtain new $\ell(z)$. If $|z - z_0| = O(n^{-u})$ for some $u \in (1/w, 1/2)$, then

$$\frac{\ell(z)}{n(z - z_0)^2\sigma^{-2}} \xrightarrow{\mathsf{P}} 1 \qquad \text{as } n \to 1,$$

where $z_0 = \mathsf{E}(Z_i)$ and $\sigma^2 = \mathsf{E}(Z_i - z_0)^2$.

Lemma 8 (Proposition 1 in [12]) For independent and identically distributed random variables $U_1, U_2, \ldots, U_n$, suppose that exist three positive constants T_1, T_2 and ι such that $\mathsf{P}\{|U_i| > u\} \leqslant T_1 \exp(-T_2 u^\iota)$ for all $u > 0$. Define $u_0 = \mathsf{E}(U_i)$, $\delta = \max\{2/\iota - 1, 0\}$, $C = 2^{1+\delta}$ and $H = n^{1/2}\sigma/(2K)$, where $\sigma^2 = \mathsf{E}(U_i - u_0)^2$ and $K > \delta$ is a sufficiently large positive constant depending only on T_1, T_2, ι and u_0, then

$$\mathsf{P}\Big\{ \pm \frac{1}{n^{\sigma/2}} \sum_{i=1}^{n} (U_i - u_0) \geqslant x \Big\}$$

$$\leqslant \begin{cases} \exp\Big(-\dfrac{x^2}{4C}\Big), & \text{if } 0 \leqslant x \leqslant (C^{1+\delta}H)^{1/(1+2\delta)}; \\ \exp\Big[-\dfrac{1}{4}(xH)^{1/(1+\delta)}\Big], & \text{if } x \geqslant (C^{1+\delta}H)^{1/(1+2\delta)}, \end{cases}$$

and more then for $C_1 \to \infty$, there exist a positive constant C_2 only depending on T_1, T_2 and ι such that

$$\begin{aligned} &\mathsf{P}\Big\{\ell(0) < \frac{nu_0^2}{C_1^2}\Big\} \\ &\leqslant \begin{cases} \exp\Big(-\dfrac{nu_0^2}{4C\sigma^2}\Big) + \exp(-C_2C_1^{\iota}), & \text{if } n^{1/2}|u_0| \leqslant \sigma(C^{1+\delta}H)^{1/(1+2\sigma)}; \\ \exp\Big[-\dfrac{1}{4}\Big(\dfrac{n|u_0|}{2K}\Big)^{1/(1+\delta)}\Big] + \exp(-C_2C_1^{\iota}), & \text{if } n^{1/2}|u_0| > \sigma(C^{1+\delta}H)^{1/(1+2\delta)}. \end{cases} \end{aligned}$$

Lemma 9 Under assumption A2, let $Z_{ijk} = [B_k(X_{ij})]Y_i$. Then

$$\mathsf{P}\{|Z_{ijk}| > z\} \leqslant 2M_1 \exp(-M_2 z^{-\iota}).$$

Proof For some $\tau > 0$,

$$\begin{aligned} \mathsf{P}\{|Z_{ijk}| > z\} &= \mathsf{P}\{|B_k(X_{ij})| > z^{\tau},\, |[B_k(X_{ij})]Y_i| > z\} \\ &\quad + \mathsf{P}\{|B_k(X_{ij})| \leqslant z^{\tau},\, |[B_k(X_{ij})]Y_i| > z\} \\ &\leqslant \mathsf{P}\{|B_k(X_{ij})| > z^{\tau}\} + \mathsf{P}\{|Y_i| > z^{1-\tau}\}. \end{aligned} \tag{6}$$

Note that $\mathsf{P}\{|B_k(X_{ij})| > z^{\tau}\} \leqslant M\mathsf{P}\{|X_{ij}| > z^{\tau}\} \leqslant MK_1 \exp(-K_2 z^{\iota_1})$ for some positive constant $M > K_1$.

Hence from A2,

$$\begin{aligned} \mathsf{P}\{|Z_{ijk}| > z\} &\leqslant MK_1 \exp(-K_2 z^{\iota_1}) + K_1 \exp(-K_2 z^{\iota_2}) \\ &\leqslant MK_1 \exp(-K_2 z^{\iota_1}) + MK_1 \exp(-K_2 z^{\iota_2}) \\ &\leqslant 2MK_1 \exp(-K_2 z^{\iota}), \end{aligned}$$

where $\iota = \iota_1\iota_2/(\iota_1 + \iota_2)$, then the result of Lemma 7 holds. □

Proof of the Theorem 1 Without loss of generality, we assume that $B_k(X_{ij})$ for $k = 1, 2, \ldots, p$ are not all 0, a.e., for some k, there exists $\epsilon_1 > 0$ such that

$$\epsilon_1 \leqslant B_k\{X_{ij}\} \leqslant 1. \tag{7}$$

From Theorem 12.7 of [28] that there exists a positive constant c_2 such that

$$|f_j(X_{ij}) - B(X_{ij})^{\mathsf{T}}\beta_j| < c_2 N^{-d}.$$

According to absolute value inequality, then

$$|B(X_{ij})^{\mathsf{T}}\beta_j| > c_2N^{-d} + |f_j(X_{ij})|.$$

From A4, A5 for $j \in \mathscr{M}_*$, we have

$$\begin{aligned}\mathsf{E}|B_k(X_j)B^{\mathsf{T}}(X_j)\beta_j| &> c_2N^{-d} + \mathsf{E}|f_j(X_j)| > c_2N^{-d} + c_1N^{1/2}n^{-\kappa}\\ &> c_4N^{1/2}n^{-\kappa},\end{aligned}$$

where $c_4 = c_1 + c_2c_3$. From above and 7, then

$$\mathsf{E}|B_k(X_j)Y| = \mathsf{E}|B_k(X_j)B^{\mathsf{T}}(X_j)\beta_j| > c_5N^{1/2}n^{-\kappa}, \tag{8}$$

where $c_5 = \epsilon_1c_4$.

We now prove that $\mathsf{E}[B_k(X_j)Y]$ can be bounded by a uniform constant. we denote $\mathsf{E}[B_k(X_j)Y]$ as z_{0j}. We know $\mathsf{E}[B_k(X_j)]^2$ is bounded uniformly [28]. Note that

$$|z_{0j}| \leqslant \{\mathsf{E}[B_k(X_j)]^2\}^{1/2}[\mathsf{E}(Y^2)]^{1/2},$$

then $|z_{0j}|$ is bounded uniformly. According the type (8), it is acquired that $nz_{0j}^2 \geqslant c_5^2Nn^{1-\kappa}$ for $j \in \mathscr{M}_*$. Then by Lemma 8 and Lemma 9, we can acquire the result of Theorem 1. □

Proof of the Theorem 2 We know that

$$\begin{aligned}\mathsf{P}\{\mathscr{M}_* \subsetneqq \widehat{\mathscr{M}_{\gamma_n}}\} &= \mathsf{P}\{\exists\, j \in \mathscr{M}_* \text{ s.t. } \ell_j(0) < c_5^2Nn^{2\tau}\}\\ &\leqslant s\max_{j\in\mathscr{M}_*}\mathsf{P}\{\ell_j(0) < c_5^2Nn^{2\tau}\},\end{aligned}$$

then the result holds. □

References

[1] BÜHLMANN P, VAN DE GEER S. *Statistics for High-Dimensional Data: Methods, Theory and Applications* [M]. Heidelberg: Springer, 2011.

[2] HASTIE T, TIBSHIRANI R, FRIEDMAN J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* [M]. 2nd ed. New York: Springer, 2009.

[3] FAN J Q, LV J C. A selective overview of variable selection in high dimensional feature space [J]. *Statist Sinica*, 2010, **20(1)**: 101–148.

[4] FAN J Q, LV J C. Sure independence screening for ultrahigh dimensional feature space (with discussion) [J]. *J R Stat Soc Ser B Stat Methodol*, 2008, **70(5)**: 849–911.

[5] FAN J Q, SONG R. Sure independence screening in generalized linear models with NP-dimensionality [J]. *Ann Statist*, 2010, **38(6)**: 3567–3604.

[6] FAN J Q, FENG Y, SONG R. Nonparametric independence screening in sparse ultra-high-dimensional additive models [J]. *J Amer Statist Assoc*, 2011, **106(494)**: 544–557.

[7] HE X M, WANG L, HONG H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data [J]. *Ann Statist*, 2013, **41(1)**: 342–369.

[8] WANG H. Factor profiled sure independence screening [J]. *Biometrika*, 2012, **99(1)**: 15–28.

[9] XUE L Z, ZOU H. Sure independence screening and compressed random sensing [J]. *Biometrika*, 2011, **98(2)**: 371–380.

[10] ZHU L P, LI L X, LI R Z, et al. Model-free feature screening for ultrahigh-dimensional data [J]. *J Amer Statist Assoc*, 2011, **106(496)**: 1464–1475.

[11] LI R Z, ZHONG W, ZHU L P. Feature screening via distance correlation learning [J]. *J Amer Statist Assoc*, 2012, **107(499)**: 1129–1139.

[12] CHANG J Y, TANG C Y, WU Y C. Marginal empirical likelihood and sure independence feature screening [J]. *Ann Statist*, 2013, **41(4)**: 2123–2148.

[13] STONE C J. Additive regression and other nonparametric models [J]. *Ann Statist*, 1985, **13(2)**: 689–705.

[14] OWEN A B. Empirical likelihood ratio confidence intervals for a single functional [J]. *Biometrika*, 1988, **75(2)**: 237–249.

[15] OWEN A B. *Empirical Likelihood* [M]. New York: Chapman and Hall/CRC, 2001.

[16] QIN J, LAWLESS J. Empirical likelihood and general estimating equations [J]. *Ann Statist*, 1994, **22(1)**: 300–325.

[17] NEWEY W K, SMITH R J. Higher order properties of GMM and generalized empirical likelihood estimators [J]. *Econometrica*, 2004, **72(1)**: 219–255.

[18] HJORT N L, MCKEAGUE I W, VAN KEILEGOM I. Extending the scope of empirical likelihood [J]. *Ann Statist*, 2009, **37(3)**: 1079–1111.

[19] CHEN S X, PENG L, QIN Y L. Effects of data dimension on empirical likelihood [J]. *Biometrika*, 2009, **96(3)**: 711–722.

[20] TANG C Y, LENG C L. Penalized high-dimensional empirical likelihood [J]. *Biometrika*, 2010, **97(4)**: 905–919.

[21] LENG C L, TANG C Y. Penalized empirical likelihood and growing dimensional general estimating equations [J]. *Biometrika*, 2012, **99(3)**: 703–716.

[22] CHANG J Y, CHEN S X, CHEN X H. High dimensional generalized empirical likelihood for moment restrictions with dependent data [J]. *J Econometrics*, 2015, **185(1)**: 283–304.

[23] FAN J Q, SAMWORTH R, WU Y C. Ultrahigh dimensional feature selection: beyond the linear model [J]. *J Mach Learn Res*, 2009, **10**: 2013–2038.

[24] LIN Y, ZHANG H H. Component selection and smoothing in multivariate nonparametric regression [J]. *Ann Statist*, 2006, **34(5)**: 2272–2297.

[25] HUANG J, HOROWITZ J L, WEI F R. Variable selection in nonparametric additive models [J]. *Ann Statist*, 2010, **38(4)**: 2282–2313.

[26] IRIZARRY R A, HOBBS B, COLLIN F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data [J]. *Biostatistics*, 2003, **4(2)**: 249–264.

[27] CHIANG A P, BECK J S, YEN H J, et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11) [J]. *Proc Natl Acad Sci U S A*, 2006, **103(16)**: 6287–6292.

[28] SCHUMAKER L L. *Spline Functions: Basic Theory* [M]. New York: Wiley, 1981.

超高维数据边际经验似然独立筛选方法

张俊英[1,2] 张日权[1,3] 王 航[2] 陆智萍[1]

([1]华东师范大学统计学院, 上海, 200062; [2]太原理工大学数学系, 太原, 030024)

([3]山西大同大学数学系, 大同, 037009)

摘 要: 可加模型通过协变量函数对响应变量起作用, 是更加灵活的非参统计模型. 当协变量个数大于样本数且以指数阶增大时, 将维数降到经典方法可解决的范围是统计学家急需解决的问题. 本文研究了超高维数据可加模型的变量筛选问题, 提出了边际经验似然变量筛选方法. 该方法通过排列在 0 点的边际经验似然率选择变量. 我们证明了选择变量集以概率 1 渐进包含真实变量集; 提出了迭代边际经验似然变量筛选方法. 数据模拟和实数据分析验证了所提方法的可行性.

关键词: 边际经验似然筛选; 非参回归模型; 变量选择; 维数缩减

中图分类号: O212.7