VARIABLE SELECTION IN NONPARAMETRIC ADDITIVE MODELS

by

Jian Huang
University of Iowa

Joel L. Horowitz
Northwestern University

Fengrong Wei
University of Iowa

INTRODUCTION

- We consider the nonparametric additive model

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

where Y is scalar and X is a continuous p-vector.

  - $\{Y_i, X_i : i = 1,...,n\}$ is a random sample of (Y, X).
  - X_{ij} is the i'th observation of the j'th component of X.
  - The functions f_j and constant μ are unknown.
  - The ε_i's are unobserved iid random variables with mean 0 and variance $\sigma^2 < \infty$.

- We suppose that some of the f_j's are zero.
  - We want to distinguish between the non-zero and zero f_j's and estimate the non-zero f_j's.

- We allow p to be much larger than n.

- We describe a penalized method that correctly selects the non-zero f_j's with high probability.

BACKGROUND

- There has been much work on penalization methods for variable selection and estimation in high-dimensional settings.
  - LASSO: Tibshirani (1996); Knight and Fu (2001); Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Zou (2006); Meinshausen and Yu (2008); Bunea, Tsybakov and Wegkamp (2006); Huang Ma, and Zhang (2008); van de Geer (2008); Zhang and Huang (2008).
  - Bridge: Huang, Horowitz, and Ma (2008)
  - SCAD: Fan and Li (2001), Fan and Peng (2004)
  - ENet: Zou and Hastie (2006)
  - Minimum concave penalty: Zhang (2007)
- Most work on penalized methods for high-dimensional models is for linear models.
  - There is often little justification for assuming linearity or any other finite-dimensional parametric family of models.
  - Later, an empirical example from development economics will illustrate usefulness of nonlinear models.

BACKGROUND (cont.)

- Lin and Zhang (2006) proposed the COSSO method for model selection and estimation in a class of nonparametric models that includes the additive model.
  - They assumed a fixed dimension p as $n \to \infty$
  - They showed that with a tensor product design, the COSSO selects components correctly with high probability.
- There is also a large literature on nonparametric estimation of additive models. E.g., Stone (1985, 1986); Mammen, Linton and Nielsen (1999); Horowitz and Mammen (2004).
  - Some estimators have no curse of dimensionality asymptotically and are oracle efficient and pointwise asymptotically normal when p is fixed as $n \to \infty$.
  - Oracle efficiency means that the estimator of each f_j has the same asymptotic distribution that it would have if the other f_j's were known.
  - None of these methods is concerned with model selection and none treats the case of $p > n$.

WHAT THIS PAPER DOES

- Carries out model selection for a nonparametric additive model in which
  - The number of non-zero f_j's is fixed.
  - The dimension of X and the number of f_j's that are zero may exceed the sample size.
- The method allows the dimension of X to be $\exp[o(n^{2d/(2d+1)})]$, where d measures the smoothness of the f_j's.
- X is assumed to be continuously distributed.
  - In other respects, the assumptions about X are very mild.

OUTLINE OF THE METHOD

- Approximate the f_j's with linear combination of B-splines.
- Model selection consists of setting all of the coefficients of some f_j's equal to zero.
  - Coefficients selected to be zero or not in groups.
- We use the group LASSO to do this.
- Model selection is carried out in two steps.
  - In the first step, the group LASSO is used to reduce the dimension of the model.
    - This step is not model-selection consistent because it selects too many non-zero f_j's.
  - In the second step, the adaptive group LASSO is used to make a final selection of f_j's.
  - The second-step estimator is model-selection consistent.
- Following model selection, existing methods can be used to obtain oracle-efficient, asymptotically normal estimators of the non-zero f_j's.

OUTLINE OF REMAINDER OF TALK

- Details of the estimation method
- Asymptotic properties including conditions for model-selection consistency
- Monte Carlo experiments illustrating finite-sample performance
- A real data example

THE ESTIMATION METHOD

- Assume that $Ef_j(X_j) = 0$ for each $j = 1, ..., p$.
  - This is a location normalization that is needed for identification
- Assume that each X_j takes values in [0,1].
- Let $\{\phi_k : k = 1, 2, ...\}$ be a B-spline basis for smooth functions on [0,1].
- Let $\|a\|_2 = \left(\sum_{j=1}^{m} a_j^2 \right)^{1/2}$ denote ℓ_2 norm of any vector $a \in \mathbb{R}^m$.
- Approximate each f_j by

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} \phi_k(x)$$

where the β_{jk}'s are constant coefficients and $m_n \to \infty$ at a suitable rate as $n \to \infty$.

METHOD (cont.)

- We choose the β_{jk}'s to solve the penalized least-squares problem

$$\underset{b}{\text{minimize}}: \sum_{i=1}^{n}\left[Y_i - \mu - \sum_{j=1}^{p}\sum_{k=1}^{m_n} b_{jk}\phi_k(X_{ij})\right]^2$$

$$+ \lambda_n \sum_{j=1}^{p} w_{nj} \left\| b_j \right\|_2$$

subject to

$$\sum_{i=1}^{n}\sum_{k=1}^{m_n} b_{jk}\phi_k(X_{ij}) = 0; \quad j = 1,..., p,$$

where λ_n is a penalty parameter and the w_{nj}'s are weights

- The w_{nj}'s are all 1 in the first estimation step and are estimates of $\left\| \beta_j \right\|_2^{-1}$ in the second.

- The constraints are empirical analogs of the location normalization.

CENTERING

- The constrained optimization problem can be turned into an unconstrained one by centering the basis functions.

- Define

$$\psi_{jk}(x) = \phi_k(x) - n^{-1}\sum_{i=1}^{n}\phi_k(X_{ij}),$$

$$Z_{ij} = [\psi_{j1}(X_{ij}),...,\psi_{jm_n}(X_{ij})]'$$

and

$$\tilde{Y} = Y - n^{-1}\sum_{i=1}^{n} Y_i$$
$$= Y - \bar{Y}.$$

- With the centered basis functions, the optimization problem becomes

$$\underset{b}{\text{minimize}} : \left\|\tilde{Y} - Z'b\right\|_2^2 + \lambda_n \sum_{j=1}^{p} w_{nj} \left\|b_j\right\|_2 .$$

- The estimator of μ is $\hat{\mu} = \bar{Y}$.

THE TWO-STEP PROCEDURE

- Model selection and estimation takes place in two steps.
  - The first step (group LASSO) reduces the number of variables to a fixed number that is independent of n but is larger than the number of non-zero f_j's.
  - The second step (adaptive group LASSO) further reduces the number of selected f_j's and is model-selection consistent.
- STEP 1: Compute the group LASSO estimator by solving

$$\underset{b}{\text{minimize}} : \left\| \tilde{Y} - Z'b \right\|_2^2 + \lambda_{n1} \sum_{j=1}^{p} \left\| b_j \right\|_2 .$$

  - Denote the resulting estimates of the vector β by $\tilde{\beta}_n = \tilde{\beta}_n(\lambda_{n1})$.
- The Step 1 estimate of f_j is:

$$\tilde{f}_{nj}(x) = \sum_{k=1}^{m_n} \tilde{\beta}_{jk} \psi_k(x); \;\; 1 \le j \le p .$$

THE TWO-STEP PROCEDURE (cont.)

- STEP 2: Use the results of Step 1 to define weights

$$w_{nj} = \begin{cases} \left\| \tilde{\beta}_{nj} \right\|_2^{-1} \text{ if } \left\| \tilde{\beta}_{nj} \right\|_2 > 0 \\ \infty \text{ otherwise} \end{cases}$$

  - The adaptive group LASSO estimator solves

$$\underset{b}{\text{minimize}} : \left\| \tilde{Y} - Z'b \right\|_2^2 + \lambda_{n2} \sum_{j=1}^{p} w_{nj} \left\| b_j \right\|_2 ,$$

    where we define $0 \times \infty = 0$.

  - Denote the vector of estimators by $\hat{\beta}_n = \hat{\beta}_n(\lambda_{n2})$

- The resulting adaptive group LASSO estimators of μ and f_j are

$$\hat{\mu} = \bar{Y}$$

and

$$\hat{f}_{nj}(x) = \sum_{k=1}^{m_n} \hat{\beta}_{jk} \psi_k(x); \;\; 1 \leq j \leq p .$$

NOTATION FOR ASSUMPTIONS AND RESULTS

- Let k be a non-negative integer and $\alpha \in (0,1]$ be such that $d = k + \alpha > 0.5$.

- Let $\mathcal{F}$ be the class of functions f on [0,1] whose k'th derivative exists and satisfies

$$\left| f^{(k)}(s) - f^{(k)}(t) \right| \leq C \,|\, s - t \,|^{\alpha};\ C < \infty$$

- Order the additive components so that the first q are non-zero and the remaining $p - q$ are zero.

- Define $\| f \|_2 = \left[\int_0^1 f(x)^2 \, dx \right]^{1/2}$.

ASSUMPTIONS

- (A1) The number of non-zero f_j's, q, is fixed, and $\min_{1 \le j \le q} \left\| f_j \right\|_2 \ge c_f$ for some constant $c_f > 0$

- (A2) The random variables $\{\varepsilon_j\}$ are IID with means of 0, finite variances, and $P(|\varepsilon_i| > z) \le K \exp(-Cx^2)$ for all i, all $z > 0$, and finite constants K and C.

- (A3) $Ef_j(X_j) = 0$ and $f_j \in \mathcal{F}$ for all $j = 1, ..., q$.

- (A4) There are constants C_1 and C_2 such that the density function g_j of X_j satisfies

  $$0 < C_1 \le g_j(x) \le C_2 < \infty$$

  on [0,1] for every $j = 1, ..., p$.

- Under (A3) and (A4), the eigenvalues of the submatrix of $Z'Z/n$ corresponding to non-zero f_j's are bounded away from 0 and ∞ with probability approaching 1 as $n \to \infty$.

ASYMPTOTIC BEHAVIOR OF STEP 1 ESTIMATOR

- Theorem 1: Let (A1)-(A4) hold. Also, let

  $\lambda_{n1} \propto \sqrt{n \log(pm_n)}$,

  $m_n \propto n^{1/(2d+1)}$,

  and

  $n^{-2d/(2d+1)} \log p \to 0$ as $n \to \infty$.

  Then

  - With probability converging to 1, all the non-zero f_j's are selected.
  - With probability converging to 1, the number of selected f_j's is no more than Mq, where $M > 6$ is a finite constant that does not depend on n
  - $\left\| \tilde{f}_{nj} - f_j \right\|_2^2 = O_p[n^{-2d/(2d+1)} \log(pm_n)]$ for f_j in the set of selected additive components.

- The number of f_j's that are zero can be $\exp[o(n^{d/(2d+1)})]$ -- e.g., $\exp[o(n^{4/5})]$ if the f_j's are twice continuously differentiable.

ASYMPTOTIC BEHAVIOR OF STEP 2 ESTIMATOR

- Theorem 2: Let (A1)-(A4) hold. Also, assume that $\lambda_{n1} \propto \sqrt{n \log(pm_n)}$, $m_n \propto n^{1/(2d+1)}$, and $n^{-2d/(2d+1)} \log p \to 0$ as $n \to \infty$. Assume, further, that $\lambda_{n2} \geq O(n^{1/2})$,

$$\frac{\lambda_{n2}}{n^{(8d+3)/(8d+4)}}, \quad \frac{n^{1/(4d+2)} \log(pm_n)}{\lambda_{n2}} = o(1).$$

Then

- The Step 2 estimator consistently selects the non-zero f_j's. That is, with probability approaching 1 as $n \to \infty$

$$\left\| \hat{f}_{nj} \right\|_2 > 0 \quad \text{for all} \quad j = 1, \ldots, q$$

and

$$\left\| \hat{f}_{nj} \right\|_2 = 0 \quad \text{for all} \quad j = q+1, \ldots, p.$$

- In addition, $\sum_{j=1}^{q} \left\| \hat{f}_{nj} - f_j \right\|_2^2 = O_p(n^{-2d/(2d+1)})$.

COMMENTS

- The 2-step procedure
  - Correctly distinguishes between zero and non-zero f_j's as $n \to \infty$.
  - Allows the number of zero f_j's to be much larger than the sample size.
  - Estimates the non-zero f_j's with the usual one-dimensional L_2 rate of convergence.
  - Does not require irrepresentable or weak orthogonality conditions on the design matrix.
    - This is because (A3)-(A4) imply the sparse Riesz condition of Wei and Huang (2008) with probability approaching 1 as $n \to \infty$.
- The number of non-zero f_j's is small relative to the sample size in the sense that it remains constant as $n \to \infty$.
- Oracle efficient, asymptotically normal estimates of the f_j's can be obtained by applying existing methods after consistent model selection.

MONTE CARLO EXPERIMENTS

- These compare the numerical performances of the group LASSO, adaptive group LASSO and "ordinary" LASSO.
  - The ordinary LASSO minimizes

$$\left\|\tilde{Y} - Z'b\right\|_2^2 + \lambda_n \sum_{j=1}^{p} \sum_{k=1}^{m_n} |\, b_{jk}\,|.$$

  - Ordinary LASSO does not take account of the grouping of the spline components, so it selects among individual components, not the f_j's.

- The model generating the data is

$$Y_i = \mu + \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i \equiv f(X_i) + \varepsilon_i .$$

- The experiments use
  - $n = 100$
  - $q = 4$ (4 non-zero f_j's)
  - $p = 21$ and $p = 100$

MONTE CARLOS (cont.)

- $\varepsilon \sim N(0,1)$.
- The X's are supported on [0,1] and have correlation coefficients of 0.5 within non-zero and zero components.
  - The zero- and non-zero covariates are independent.
- The basis functions are cubic B-splines with 6 evenly space knots.
- Penalty parameters chosen using BIC.
- The functions f are
  - $-7+8X_1-3X_2+10X_3^2-6X_4(X_4-1)$
  - $-5+8X_1^3+10X_2(1-X_2)-10X_3^5-8X_4^2$
  - $-4+4X_1+\cos(2\pi X_2)-8X_3^3$

$$+\sqrt{X_4(1-X_4)}\sin\left[\frac{2\pi(1+2^{-3/5})}{X+2^{-3/5}}\right]$$

RESULTS

$f(x)$	Adaptive Group LASSO	Group LASSO	LASSO
	$n=100,\ p=21$		
1	100	100	100
	80	51	27
2	99	99	100
	77	40	42
3	99	99	97
	60	27	18
	$n=100,\ p=100$		
1	91	91	50
	59	45	11
2	90	90	47
	52	45	11
3	94	95	51
	53	21	12

ILLUSTRATIVE APPLICATION TO DATA ON ECONOMIC GROWTH

- Illustrates usefulness of nonparametric approach to model selection.

- Sala-i-Martin (1997) investigated relation between economic growth and 59 covariates in 99 countries.

- He used the linear model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$$

  where Y_i is the average rate of growth of the GDP in country i from 1960-1992, X_{ij} is the j'th covariate for country i.

- The data describe economic, political, social, and geographical characteristics of countries.

- Sala-i-Martin selected covariates by using a heuristic method requiring 2 million regressions.
  - This procedure selected 22 of the 59 covariates.
  - Sala-i-Martin noted that other investigators have found nonlinear relations in models of economic growth.

APPLICATION (cont.)

- Many of Sala-i-Martin's covariates are binary.
- We modeled the relation between GDP growth and the 21 continuous covariates in the data.
- We estimated models using
  - Adaptive group LASSO
  - Group LASSO
  - Ordinary LASSO
- The adaptive group LASSO and group LASSO both selected the same 13 variables.
  - The ordinary LASSO selected all of the variables.
- The figure shows adaptive group LASSO estimates of some of the f_j's.

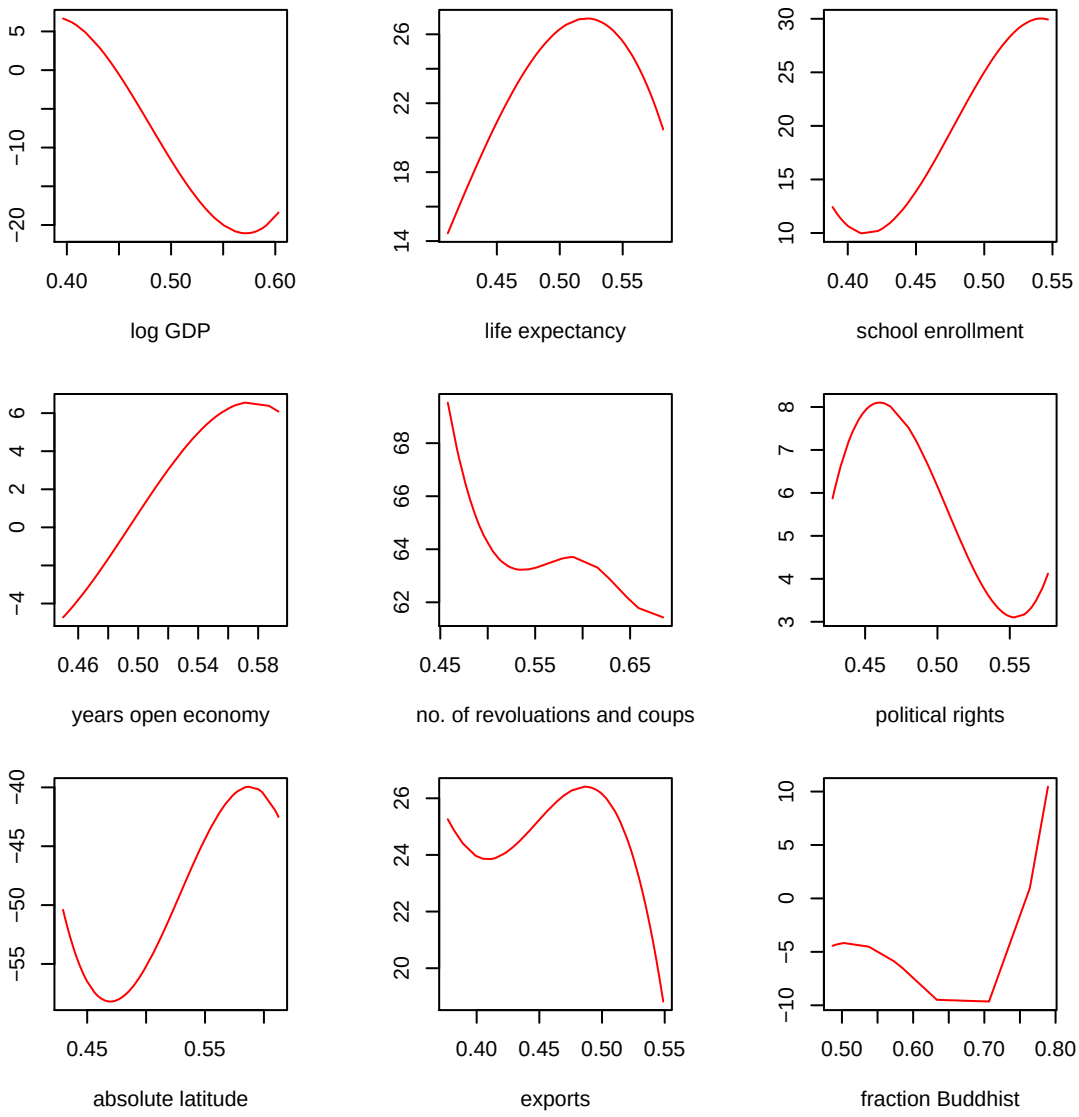

Figure 1: Plots of the estimated nonzero components by the adaptive group Lasso.

CONCLUSIONS

- Most methods for selecting covariates in sparse, high-dimensional models assume that the model in question is linear.

- Often in applications, there is no justification for assuming linearity.

- This paper has investigated the properties of the group LASSO and adaptive group LASSO for sparse, high-dimensional nonparametric additive models.

- We assume that the number of non-zero additive components is fixed and, therefore, “small” compared to the sample size.

- We show that when this condition and other mild assumptions hold, the adaptive group LASSO is model selection consistent even if the total number of covariates is much larger than the sample size.

- An empirical example illustrates the usefulness of allowing the model to be nonlinear.