cemmap
centre for microdata methods and practice

Variable selection and estimation in high-dimensional models

Joel Horowitz

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP35/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

VARIABLE SELECTION AND ESTIMATION IN HIGH-DIMENSIONAL MODELS

by

Joel L. Horowitz
Department of Economics
Northwestern University
Evanston, IL 60201 USA

October 2014

Abstract

Models with high-dimensional covariates arise frequently in economics and other fields. Often, only a few covariates have important effects on the dependent variable. When this happens, the model is said to be sparse. In applications, however, it is not known which covariates are important and which are not. This paper reviews methods for discriminating between important and unimportant covariates with particular attention given to methods that discriminate correctly with probability approaching 1 as the sample size increases. Methods are available for a wide variety of linear, nonlinear, semiparametric, and nonparametric models. The performance of some of these methods in finite samples is illustrated through Monte Carlo simulations and an empirical example.

This article is based on my State of the Art Lecture at the 2014 meeting of the Canadian Economics Association in Vancouver, BC. I thank Russell Davidson for inviting the lecture.

VARIABLE SELECTION AND ESTIMATION IN HIGH-MODELS

1. Introduction

This paper is about estimating the model

(1.1) $Y_i = f(X_{i1},...,X_{ip}) + \varepsilon_i; \quad i = 1,...,n; \quad j = 1,...,p$,

where n is the sample size, Y_i is the i'th observation of the dependent variable Y, X_{ij} is the i'th observation of the j'th component of a $p \times 1$ vector of explanatory variables X, and the ε_i's are independently and identically distributed random variables that satisfy $E(\varepsilon_i) = 0$ or Quantile$(\varepsilon_i) = 0$. The number of explanatory variables, p, may be larger than the sample size, n. It is assumed that a few components of the vector X have effects on Y that are "large" in a sense that will be defined. The rest of the components of X have effects that are small though not necessarily zero. Suppose the components of X that have large effects on Y are denoted by the vector X_{A_0}. The objective of the analysis is to determine which components of X belong in X_{A_0} (variable selection) and to estimate $E(Y \mid X_{A_0})$ or Quantile$(Y \mid X_{A_0})$. In most of this paper, f is a linear function and $E(\varepsilon_i) = 0$. Other versions of model (1.1) are discussed briefly.

There is a large statistics literature on high-dimensional variable selection and estimation. This literature is cited throughout the discussion in this paper. In a typical application in statistics, one wants to learn which genes out of thousands in a species are associated with a disease, but data are available for only 100 or so individuals in the species. In this application, Y_i is a measure of the intensity of the disease in individual i and X_{ij} is a measure of the activity level of gene j in individual i. There is no hope of discriminating between genes that are and are not associated with the disease if the number of associated genes p exceeds the sample size n. However, it is usually believed that the number of genes associated with a disease is small. In particular, it is much smaller than the size of the sample. Most genes have little or no influence on the disease. Models in which only a few components of X have important influences on Y are called sparse. In a sparse model, it is possible, using methods that are described in this paper, to discriminate between components of X that have important effects on Y and components of X that have little or no influence on Y.

High-dimensional problems also arise in economics. For example, survey data sets such as the National Longitudinal Survey of Youth (NLSY) may contain hundreds or thousands of variables that arguably affect productivity and, therefore, wages. Depending on how the data are stratified, there may be more potentially relevant explanatory variables than observations in a wage equation. However, only a few variables such as education and years of labor force experience are thought to have large effects on

wages. Thus, a wage equation is sparse, although one may not know which variables should be classified as unimportant. A second example is a study by Sala-i-Martin (1997), who carried out 2 million regressions in attempt to determine which of 59 potential explanatory variables should be included in a linear growth model. In an earlier attempt to identify variables relevant to a growth model, Sala-i-Martin (1996) carried out 4 million regressions. These examples illustrate the need in economics and applied econometrics for a systematic way to decide which variables should be in a model.

This paper reviews and explains methods that are used to estimate high-dimensional models. The discussion is informal and avoids technical details. Key results are presented and explained in as intuitive a way as possible. Detailed technical arguments and proofs are available in references that are cited throughout the paper.

Section 2 presents basic concepts and definitions that are used throughout the subsequent discussion. Section 3 discusses the linear model in detail. Nonlinear and nonparametric models are discussed in Section 4. Section 5 presents some Monte Carlo results and an empirical example. Section 6 presents conclusions.

2. Basic Concepts and Definitions

This section presents concepts and definitions that are used in discussing methods for high-dimensional models. The concepts and definitions are presented first in the setting of a linear mean-regression model. The linear model has received the most attention in the literature, and methods for it are highly developed. Extensions to nonlinear models are presented later in the paper.

When f in model (1.1) is a linear function, the model becomes

$$(2.1) \qquad Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i; \quad i = 1,...,n,$$

where $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) < \infty$, and the β_j's are constant coefficients that must be estimated. Assume without loss of generality that Y_i and the covariates X_{ij} are centered and scaled so that $n^{-1}\sum_{i=1}^{n} Y_i = 0$, $n^{-1}\sum_{i=1}^{n} X_{ij} = 0$, and $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$ for each $j = 1,..., p$. Because of the centering, there is no intercept term in model (2.1). The scaling makes it possible to define "large" and "small" β_j's unambiguously. Without the scaling, each β_j could be made to have any desired magnitude by choosing the scale of X_{ij} $(i = 1,...,n)$ appropriately.

Most known properties of variable selection and estimation methods for high-dimensional models are asymptotic as $n \to \infty$. If p is fixed, then $p > n$ is not possible as $n \to \infty$. Similarly, if the

coefficients β_j in (2.1) are fixed, then coefficients whose magnitudes are small compared to random sampling error but not zero are not possible as $n \to \infty$. To enable asymptotic approximations to be used while allowing the possibility that $p > n$, p is allowed to increase as n increases. Thus, $p = p(n)$. Similarly, the coefficients, β_j may approach zero as $n \to \infty$. The possible dependence of p and the β_j's on n is a mathematical device to enable the use of asymptotic approximations. The values of p and the β_j's in the sampled population do not depend on n.

Suppose for the moment that the non-zero β_j's are bounded away from 0. Then the objectives of variable selection and estimation in (2.1) are to discriminate between coefficients that are non-zero and zero and to estimate the non-zero coefficients. A model selection procedure is called model-selection consistent if it discriminates correctly between β_j's that are zero and non-zero with probability approaching 1 as $n \to \infty$. An estimation procedure is called oracle efficient if the estimated non-zero β_j's have the same asymptotic distribution that they would have if the variables with coefficients of zero were known *a priori*, dropped from model (2.1), and ordinary least squares (OLS) were used to estimate the non-zero β_j's.

Now suppose that some β_j's can be small but not zero, whereas others are large. Let A_S denote the indices j of β_j's that are small or zero and $A_0 = \bar{A}_S$ denote the indices of β_j's that are large. The precise definitions of small and large vary, depending on the variable selection and estimation method. Roughly speaking, however, the small β_j's satisfy $\sum_{j \in A_S} |\beta_j| = o(n^{-1/2})$ as $n \to \infty$. The large β_j's satisfy $n^{1/2} |\beta_j| \to \infty$ as $n \to \infty$. Thus, the small β_j's are smaller in magnitude than the random sampling errors of their estimates, and the large β_j's are larger in magnitude than random sampling error. Let q denote the number of large β_j's, and suppose that q remains fixed as $n \to \infty$. Then using the algebra of ordinary least squares, it can be shown that the large β_j's can be estimated with a smaller mean-square error if the covariates with small β_j's are omitted from model (2.1) (or, equivalently, the small β_j's are set equal to zero *a priori*). Accordingly, a model selection method is said to be model-selection consistent if it discriminates correctly between large and small β_j's (that is between A_0 and A_S) with probability approaching 1 as $n \to \infty$. An estimation method is said to be oracle efficient if the estimates of the large β_j's have the same asymptotic distribution that they would have if the small β_j's

were known, the covariates X_{ij} ($j \in A_S$) were dropped from (2.1), and the large coefficients were estimated by OLS applied to the resulting reduced version of (2.1).

If $p > n$, then model (2.1) cannot be estimated by OLS. OLS estimation is possible if $p < n$, but OLS estimates of the β_j's cannot be zero. Even if $\beta_j = 0$ for some j, its OLS estimate can be anywhere in a neighborhood of 0 whose size is $O[(n^{-1}\log n)^{1/2}]$ under mild regularity conditions. Therefore, OLS cannot be used to discriminate between zero and non-zero (or small and large) β_j's, even if $p < n$. These problems can be overcome by using penalized least squares estimation (PLS) instead of OLS. In PLS, the estimator of β_j is the solution to the problem

$$(2.2) \qquad \underset{b_1,...,b_p}{\text{minimize}}: \; Q_n(b_1,...,b_p) \equiv \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} b_j X_{ij}\right)^2 + \sum_{j=1}^{p} p_\lambda(|b_j|),$$

where p_λ is a penalty function and λ is a parameter (the penalization parameter). For example, PLS with $p_\lambda(|b_j|) = \lambda|b_j|$ is called the LASSO. PLS with $p_\lambda(|b_j|) = \lambda b_j^2$ is called ridge regression. PLS estimators and their properties are discussed in detail in Section 3.

3. The Linear Model

This section presents a detailed discussion of methods for variable selection and estimation in the linear model (2.1). The discussion focusses on methods, not the finite-sample performance of the methods. Section 5 of this paper and references cited in this section present Monte Carlo results and empirical examples illustrating the numerical performance of the methods.

To begin, assume that $p < n$ and that either $\beta_j = 0$ or $\beta_j \geq \delta$ for some $\delta > 0$. Let q be the number of non-zero β_j's, and assume that q is fixed as $n \to \infty$. Assume without loss of generality that the β_j's are ordered so that $\beta_1,...,\beta_q$ are non-zero and $\beta_{q+1},...,\beta_p$ are zero. Let $\mathcal{S}$ denote any subset of the indices $j = 1,..., p$ and $|\mathcal{S}|$ denote the number of elements in $\mathcal{S}$.

Now let $\hat{\beta}_{\mathcal{S}} = \{\hat{\beta}_{\mathcal{S}j}: \; j \in \mathcal{S}\}$ be the $|\mathcal{S}| \times 1$ vector of estimated β_j coefficients ($j \in \mathcal{S}$) obtained from OLS estimation of model $\mathcal{S}$. That is

$$\hat{\beta}_{\mathcal{S}} = \arg \min_{b_j:\; j \in \mathcal{S}} \left(\sum_{i=1}^{n} Y_i - \sum_{j \in \mathcal{S}} b_j X_{ij}\right)^2.$$

Let $\hat{\sigma}_{\mathcal{S}}^2$ denote the mean of the squared residuals from OLS estimation of model $\mathcal{S}$:

$$\hat{\sigma}_{\mathcal{S}}^2 = n^{-1}\left(\sum_{i=1}^{n} Y_i - \sum_{j\in\mathcal{S}} \hat{\beta}_{\mathcal{S}j} X_{ij}\right)^2 .$$

Define

$$BIC = \log(\hat{\sigma}_{\mathcal{S}}^2) + |\mathcal{S}|\, n^{-1}\log n .$$

Now consider selecting the model that minimizes BIC over all possible models $\mathcal{S}$ or, equivalently, all possible subsets of p covariates. This procedure is called subset selection and, under mild regularity conditions, it is model-selection consistent (Shao 1997). However, it requires OLS estimation of $2^p - 1$ models and, therefore, is computationally feasible only if p is small. For example, applying subset selection to the model considered by Sala-i-Martin (1997) would require estimating more than 10^{17} models. Applying subset selection to the empirical example presented in Section 5 of this paper would require estimating more than 10^{14} models.

Consistent, computationally feasible model selection in a sparse linear model can be achieved by solving the PLS problem (2.2). One important class of estimators is obtained by setting the penalty function $p_\lambda(v) = \lambda |v|^\gamma$, where $\gamma > 0$ is a constant. The properties of the resulting PLS estimator depend on the choice of γ. Let $\hat{\beta}_j$ denote the resulting estimator of β_j. If $\gamma > 1$, then $Q_n(b_1,...,b_p)$ is a continuously differentiable function of its arguments. Let $\beta = (\beta_1,...,\beta_p)'$, $\hat{\beta} = (\hat{\beta}_1,...,\hat{\beta}_p)'$, $|\hat{\beta}| = (|\hat{\beta}_1|,...,|\beta_p|)'$, $Y = (Y_1,...,Y_n)'$, $\boldsymbol{X}$ denote the $n \times p$ matrix whose (i, j) element is X_{ij}, and $e = (\varepsilon_1,...,\varepsilon_n)'$. Let $\beta_0 = (\beta_{01},...,\beta_{0p})'$ be the unknown true value of β. The first-order conditions for problem (2.2) are

$$\boldsymbol{X}'(Y - \boldsymbol{X}\hat{\beta}) = \lambda\gamma\, |\hat{\beta}|^{\gamma-1} .$$

Equivalently

$$\boldsymbol{X}'[\varepsilon - \boldsymbol{X}(\hat{\beta} - \beta_0)] = \lambda\gamma\, |\hat{\beta}|^{\gamma-1} . \tag{3.1}$$

If $\beta_{0j} = 0$ for some $j = j^*$, then it follows from (3.1) that $\hat{\beta}_j = 0$ only if

$$\sum_{i=1}^{n} X_{ij^*}\left[\varepsilon_i - \sum_{j\neq j^*} X_{ij}(\hat{\beta}_j - \beta_{0j})\right] = 0 . \tag{3.2}$$

If the ε_i's are continuously distributed, then (3.2) constitutes an exact linear relation among continuously distributed random variables and has probability 0. Therefore, PLS with the penalty function $p_\lambda(v) = \lambda |v|^\gamma$ and $\gamma > 1$ cannot yield $\hat{\beta}_j = 0$, even if $\beta_{0j} = 0$, and cannot be model-selection consistent.

The situation is different if $0 < \gamma \leq 1$. Then $p_\lambda(v)$ has a cusp at $v = 0$, and $\hat{\beta}_j = 0$ can occur with non-zero probability. This can be seen from the first-order conditions for the case $0 < \gamma \leq 1$, which are the Karush-Kuhn-Tucker (KKT) conditions:

(3.3) $\boldsymbol{X}_j'(Y - \boldsymbol{X}\hat{\beta}) = \lambda\gamma\,|\beta|^{\gamma-1}; \quad \hat{\beta}_j = 0$

and

(3.4) $|\boldsymbol{X}_j'(Y - \boldsymbol{X}\hat{\beta})| \leq \lambda; \quad \hat{\beta}_j = 0,$

where $\boldsymbol{X}_j$ is the j'th column of the matrix $\boldsymbol{X}$. Condition (3.4) for $\hat{\beta}_j = 0$ is an inequality and, therefore, has non-zero probability. Thus, in contrast to OLS or ridge regression, PLS estimation with $p_\lambda(v) = \lambda|v|^\gamma$ and $0 < \gamma \leq 1$ can give $\hat{\beta}_j = 0$ with non-zero probability.

PLS estimation with $p_\lambda(v) = \lambda|v|$ is called the LASSO. The LASSO was proposed by Tibshirani (1996). Knight and Fu (2000) investigated properties of LASSO estimates. Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) showed that the LASSO is model-selection consistent under a strong condition on the design matrix $\boldsymbol{X}$ called the strong irrepresentable condition. Zhang (2009) gave conditions under which the LASSO combined with a thresholding procedure consistently distinguishes between coefficients that are zero and coefficients whose magnitudes as $n \to \infty$ exceed n^{-s} for some $s < 1/2$. Bühlmann and van de Geer (2011) provide a highly detailed treatment of the LASSO. There is also a literature on the use of LASSO for the problem of prediction. See, for example, Greenshtein and Ritov (2004) and Bickel, Ritov, and Tsybakov (2009). Computational feasibility with a large p is an important issue in high-dimensional estimation. Osborne, Presnell, and Turlach (2000); Efron, Hastie, Johnstone, and Tibshirani (2004); and Friedman, Hastie, Höfling, and Tibshirani (2007) present fast algorithms for computing LASSO estimators.

The irrepresentable condition required for model selection consistency of the LASSO is very restrictive. Among other things, it requires the coefficient estimates obtained from the OLS regression of the irrelevant covariates (covariates with $\beta_j = 0$) on the relevant covariates to have magnitudes that are smaller than one (Zhao and Yu 2006). Zou (2006) gives a simple example in which this condition is violated. Zhang and Huang (2008) showed that if $\lambda = O\left(\sqrt{n \log p}\right)$ and certain other conditions (but not the irrepresentable condition) are satisfied, then the LASSO selects a model of finite dimension that is too large. That is, with probability approaching 1 as $n \to \infty$ the selected model contains all the covariates with non-zero β_j's but also includes some covariates for which $\beta_j = 0$.

3.1 The Adaptive LASSO

The LASSO is not model selection consistent when the irrepresentable condition does not hold because its penalty function does not penalize small coefficients enough relative to large ones. This problem is overcome by a two-step method called the adaptive LASSO (AL) (Zou 2006). The first step is ordinary LASSO estimation of the β_j's. Denote the resulting estimator by $\tilde{\beta} = (\tilde{\beta}_1, ..., \tilde{\beta}_p)'$. Then

$$(3.5) \qquad \tilde{\beta} = \arg\min_{b_1,...,b_p} \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} b_j X_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p} |b_j|,$$

where $\lambda_1 = O\left(\sqrt{n \log p}\right)$ is the penalization parameter. In the second step, variables X_{ij} for which $\tilde{\beta}_j = 0$ are dropped from model (2.1). Let $\hat{\beta}_j = 0$ if $\tilde{\beta}_j = 0$. Define the remaining components of the vector $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)'$ by

$$(3.6) \qquad \hat{\beta} = \arg\min_{b_j:\, \tilde{\beta}_j \neq 0} \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j:\, \tilde{\beta}_j \neq 0} b_j X_{ij}\right)^2 + \lambda_2 \sum_{j:\, \tilde{\beta}_j \neq 0} \tilde{\beta}_j^{-1} |b_j|,$$

where λ_2 is a penalty parameter that increases slowly as $n \to \infty$. The AL estimator of β_j is $\hat{\beta}_j$. Zou (2006) gives conditions under which the AL is model-selection consistent and oracle efficient when the values of p and the β_j's are fixed. Horowitz and Huang (2013) give conditions under which the AL is model-selection consistent if some β_j's may be small but non-zero in the sense defined in Section 2 and p may be larger than n. Oracle efficiency follows from the result of Zou (2006) and the observation that asymptotically, the LASSO selects a model of bounded size that contains all variables with large coefficients. The precise definitions of large and small β_j's and the rate of increase of λ_2 as $n \to \infty$ depend on p and are given by Horowitz and Huang (2013). If the large β_j's are bounded away from zero and the small β_j's are zero, then the required rate is $\lambda_2 = o(n^{1/2})$. Although $p > n$ is allowed, p cannot increase too rapidly as n increases. The rate at which p is limited by the requirement that the eigenvalues of the matrix $\boldsymbol{X}'\boldsymbol{X}/n$ not decrease to zero too rapidly. This usually requires $p = O(n^a)$ for some $a > 0$.

Models in which $p \propto e^{an}$ for some $a > 0$ are called ultra-high dimensional. In applications, these are models in which p is much larger than n (e.g., $n = 100$, $p = 10,000$). Such models are rare in economics but arise in genomics. PLS estimators do not work well in ultra-high dimensional settings,

and other methods have been developed to deal with them. See, for example, Fan and Lv (2008) and Meinshausen and Bühlmann (2010).

We now provide further intuition for model-selection consistency and oracle efficiency of the AL Assume that p is fixed and, therefore, that $p < n$ if n is sufficiently large. Let $\beta_1,...,\beta_q = 0$ and $\beta_{q+1},...,\beta_p = 0$. Asymptotically, $\tilde{\beta}$ has r non-zero components, where $q \le r \le p$. Order the components of $\tilde{\beta}$ so that $\tilde{\beta}_1,...,\tilde{\beta}_r$ are non-zero and $\tilde{\beta}_{r+1},...,\tilde{\beta}_p$ are zero. Define $\beta_0 = (\beta_1,...,\beta_r)'$. Let $\hat{\beta}_{AL}$ be the second-step AL estimator of β_0, and let $b = (b_1,...,b_r)'$ be any $r \times 1$ vector. Define $u = n^{1/2}(b - \beta_0)$. Assume that $\lambda_2 \to \infty$ and $n^{-1/2}\lambda_2 \to 0$ as $n \to \infty$, where λ_2 is the penalty parameter in the second AL step.

Some algebra shows that (3.6), the second AL step, is equivalent to

$$(3.7) \qquad n^{1/2}(\hat{\beta}_{AL} - \beta_0) = \arg\min_u \sum_{i=1}^{n}\left(\varepsilon_i - n^{-1/2}\sum_{j=1}^{r} b_j X_{ij}\right)^2 + \lambda_2 \sum_{j=1}^{r} |\tilde{\beta}_j|^{-1}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|).$$

If $\beta_{0j} = 0$, then $|\tilde{\beta}_j| = O_p(n^{-1/2})$ and

$$\lambda_2 |\tilde{\beta}_j|^{-1}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \approx \lambda_2 |u_j|$$

Therefore, if $u_j \neq 0$, the penalty term on the right-hand side of (3.7) becomes arbitrarily large as $n \to \infty$. It follows that $u_j \neq 0$ cannot be part of the argmin in (3.7) if n is sufficiently large, and $\hat{\beta}_{AL,j} = \beta_{0j}$ for all sufficiently large n. If $\beta_{0j} \neq 0$, then

$$\lambda_2 |\tilde{\beta}_j|^{-1}(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \approx \lambda_2 n^{-1/2}|u_j| = \lambda_2 n^{-1/2} O_p(1) \to^p 0.$$

Therefore, as $n \to \infty$, (3.6) and (3.7) become equivalent to OLS estimation of the non-zero components of β_{0j}. Thus, the AL is model-selection consistent and oracle efficient.

3.2 Other Penalty Functions

Another way to achieve a model-selection consistent PLS estimator is to use a penalty function that is concave and has a cusp at the origin. This section presents several such functions. Lv and Fan (2009) and Zou and Zhang (2009) describe additional penalization methods.

1. The bridge penalty function (Knight and Fu 2001; Huang, Horowitz, and Ma 2008):

$$p_\lambda(v) = \lambda |v|^\gamma,$$

where γ is a constant satisfying $0 < \gamma < 1$.

2. The smoothly clipped absolute deviation (SCAD) penalty function (Antoniadis and Fan 2001, Fan and Peng 2004). This penalty function is defined by its derivative:

$$p'_\lambda(v) = \lambda[I(v \le n^{-1}\lambda) + \frac{(an^{-1}\lambda - v)_+}{(a-1)n^{-1}\lambda} I(v > n^{-1}\lambda)]; \quad v \ge 0,$$

where I is the indicator function and $a > 2$ is a constant.

3. The minimax concave (MC) penalty function (Zhang 2010):

$$p_\lambda(v) = \lambda \int_0^v \left(1 - \frac{nx}{a\lambda}\right)_+ dx; \quad v \ge 0,$$

where $a > 0$ is a constant..

Fan and Peng (2004); Huang, Horowitz, and Ma (2008); Kim, Choi, and Oh (2008); Zhang (2010); and Horowitz and Huang (2013) give conditions under which PLS estimation with these penalty functions is model-selection consistent and oracle efficient. The precise definitions of large and small (but non-zero) β_j's and the rate at which $\lambda \to \infty$ differ among penalty functions. Details and computational methods are given in the foregoing references.

Examples of the three penalty functions are displayed in Figure 1 for $v \ge 0$. All are steeply sloped near $v = 0$ and have a cusp at $v = 0$. However, the SCAD and MC penalty functions are flat at large values of $|v|$, whereas the bridge penalty function continues increasing as $|v|$ increases. Positive values of the penalty function drive the parameter estimates toward zero creating a penalization bias. This bias is smaller with the SCAD and MC penalty functions than with the bridge penalty function because of the flattening of the SCAD and MC penalties at large values of $|v|$. However, the penalization bias of the bridge estimator can be removed by carrying out OLS estimation of the parameters of the selected model.

3.4 Choosing the Penalty Parameter

This section describes a method due to Wang, Li, and Leng (2009) for choosing the penalty parameter in PLS estimation of a linear model with any of a variety of penalty functions. Wang, Li, and Tsai (2007) present an earlier version of the same method.

Let $\hat{\beta}_\lambda$ and $|\mathcal{S}_\lambda|$, respectively, denote the PLS parameter estimator and number of non-zero $\hat{\beta}_j$'s when the penalty parameter is λ. Define

$$\hat{\sigma}^2_\lambda = n^{-1}\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \hat{\beta}_{\lambda j} X_{ij}\right)^2.$$

For any sequence of constants $\{C_n\}$ such that $C_n \to \infty$, define

$$BIC_\lambda = \log(\hat{\sigma}_\lambda^2) + |\mathcal{S}_\lambda| C_n n^{-1} \log n .$$

Wang, Li, and Leng (2009) give conditions under which choosing λ to minimize BIC_λ yields a model-selection consistent AL or PLS estimator. Wang, Li, and Tsai (2007) show that the use of generalized cross validation to select λ does not necessarily achieve model-selection consistency.

4. Nonlinear, Semiparametric, and Nonparametric Models

This section extends the PLS approach of Section 3 to a variety of parametric, semiparametric, and nonparametric models. As in section 3, the discussion here focusses on methods. The cited references provide Monte Carlo results and empirical examples that illustrate the numerical performance of the methods.

4.1 Finite-Dimensional Parametric Models

Equation (2.2) is a penalized version of the (negative) log-likelihood of a normal linear model. Fan and Li (2001) and Fan and Peng (2004) extend the penalization approach to maximum likelihood estimation of a more general class of high-dimensional models. Let V denote a possibly vector-valued random variable whose probability density $f(\cdot, \beta)$ function depends on a parameter β. Let $\{V_i : i = 1,...,n\}$ denote a random sample of V, and let $\ell_n(V, \beta) = \sum_{i=1}^{n} \log f(V_i, \beta)$ denote the log-likelihood function of V. Maximum likelihood estimation of β is equivalent to minimizing $-\ell_n(V, \beta)$ over β. Accordingly, penalized maximum likelihood estimation β consists of solving the problem

$$(4.1) \qquad \underset{b_1,...,b_p}{\text{minimize}}: \; Q_n(b_1,...,b_p) \equiv -\ell_n(V, b_1,...,b_p)^2 + \sum_{j=1}^{p} p_\lambda(|b_j|).$$

Fan and Li (2001) and Fan and Peng (2004) give conditions under which (4.1) gives a model-selection consistent, oracle-efficient estimator of β. Fan and Li (2001) present a method for computing the penalized maximum likelihood estimator.

Belloni and Chernozhukov (2011) consider penalized estimation of a linear-in-parameters quantile regression model. The model is

$$(4.2) \qquad Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i; \quad P(\varepsilon_i \le 0 \mid X_{i1},..., X_{ip}) = \tau; \; i = 1,...,n$$

where $0 < \tau < 1$. If $p < n$, then the β_j's in model (4.2) can be estimated by solving the problem

$$\underset{b_1,...b_p}{\text{minimize}}: \; \sum_{i=1}^{n} \rho_\tau \left(Y_i - \sum_{j=1}^{p} b_j X_{ij} \right),$$

where $\rho_\tau(u) = [\tau - I(u \le 0)]u$ is the check function. The penalized estimator considered by Belloni and Chernozhukov (2011) is

$$(4.3) \qquad \hat{\beta} = \underset{b_1,...b_p}{\arg\min}\, Q_{n\tau}(b_1,...,b_p) = \sum_{i=1}^{n} \rho_\tau\left(Y_i - \sum_{j=1}^{p} b_j X_{ij}\right) + \lambda \sum_{j=1}^{p} |\, b_j\,|,$$

where λ is the penalty parameter and the covariates are scaled so that $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$. To state the properties of $\hat{\beta}$, let s denote the number of non-zero β_j's, including β_j's that are "small" but non-zero. Let β_{0j} denote the true value of β_j, and define

$$\left\|\hat{\beta} - \beta_0\right\| = \left[\sum_{j=1}^{p} (\hat{\beta}_j - \beta_{0j})^2\right]^{1/2}.$$

Belloni and Chernozhukov (2011) give conditions under which the following hold as $n \to \infty$:

1. $\left\|\hat{\beta} - \beta_0\right\| = O_p\left(\sqrt{(s/n)\log(n \vee p)}\right)$,

where $s \vee n = \max(s, n)$.

2. The number of non-zero $\hat{\beta}_j$'s is $O_p(s)$. That is, $\sum_{j=1}^{p} I(|\,\hat{\beta}_j\,| > 0) = O_p(s)$.

3. The set of covariates with non-zero $\hat{\beta}_j$'s contains the set of covariates with non-zero β_{0j}'s. That is $\{j : |\,\beta_{0j}\,| > 0\} \subset \{j : |\,\hat{\beta}_j\,| > 0\}$.

Result 3 requires the magnitudes of the non-zero β_j's to exceed the sizes of the random sampling errors of the $\hat{\beta}_j$'s. Thus, result 3 requires all non-zero β_j's to be "large."

The penalized estimator (4.3) is a quantile-regression version of the LASSO and, like the LASSO, is not model-selection consistent in general. Belloni and Chernozhukov give conditions under which model-selection consistency is achieved by a thresholding procedure that sets $\hat{\beta}_j = 0$ if $|\,\hat{\beta}_j\,|$ is "too small." It is likely that model-selection consistency can be achieved through a quantile version of the AL or through using the SCAD or MC penalty functions, but such results have not been proved.

4.2 Semiparametric Single-Index and Partially Linear Models

In a semiparametric single-index model, the expected value of a dependent variable Y conditional on a p-dimensional vector of explanatory variables X is

$$(4.4) \qquad E(Y \mid X) = g(\beta' X),$$

where g is an unknown function and $\beta \in \mathbb{R}^p$ is an unknown vector. Methods for estimating g and β when p is small have been developed by Powell, Stock and Stoker (1989); Ichimura (1993); Horowitz and Härdle (1996); and Hristache, Juditsky, and Spokoiny (2001) among others. Kong and Xia (2007) proposed a method for selecting variables in low-dimensional single-index models. However, these methods are not computationally feasible when p is large. Accordingly, achieving computational feasibility is the first step in model selection and estimation of high-dimensional single-index models.

Wang, Xu, and Zhu (2012) achieve computational feasibility by assuming that $E(X \mid \beta' X)$ is a linear function of $\beta' X$. Call this the linearity assumption. It is a strong assumption, although Hall and Li (1993) show that it holds approximately in many settings when $p = \dim(X)$ is large. Let Σ denote the covariance matrix of X, and assume that Σ is positive definite. Define $\sigma_h = \text{cov}[X, h(Y)]$ for any bounded function h. Wang, Xu, and Zhu (2012) show that under the linearity assumption, $\beta_h \equiv \Sigma^{-1}\sigma_h \propto \beta$. Accordingly, if $p < n$, β can be estimated up to a proportionality constant by

$$\text{(4.4)} \qquad \hat{\beta}_h = \arg\min_b \sum_{i=1}^{n} \left[h(Y_i) - b' X_i \right]^2 = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' h(\boldsymbol{Y}),$$

where $\boldsymbol{X}$ is the matrix of observed values of the covariates and $\boldsymbol{Y}$ is the vector of observations of Y. Wang, Xu, and Zhu (2012) give conditions under which $\hat{\beta}_h$ estimates β_h consistently, The scale of β in a single-index model is not identified and must be set by normalization. If the proportionality constant relating β_h to β is not zero, then β_h can be rescaled to accommodate any desired scale normalization of β, and the rescaled version of $\hat{\beta}_h$ estimates β consistently.

Wang, Xu, and Zhu (2012) propose setting

$$\text{(4.5)} \qquad h(y) = F_n(y) - 1/2,$$

where F_n is the empirical distribution of Y. They then consider the resulting penalized version of (4.4), which consists of minimizing

$$\text{(4.6)} \qquad Q_n(b) = \sum_{i=1}^{n} [F_n(Y_i) - 1/2 - b' X_i]^2 + \sum_{j=1}^{p} p_\lambda(|b_j|),$$

where p_λ is the SCAD or MC penalty function. Let $\hat{\beta}_{h1}$ be the estimator of the non-zero components of β_h that would be obtained from (4.4) with $h(y)$ as in (4.5) if the covariates with coefficients of zero were omitted from the model. Let $\hat{\beta}_{h0} = (\hat{\beta}_{h1}, 0_{p-q})$. Wang, Xu, and Zhu (2012) give conditions under which $\hat{\beta}_0$ is contained in the set of local minimizers of $Q_n(b)$ in (4.6) with probability approaching 1 as

$n \to \infty$. This result shows that with probability approaching 1, the oracle estimator of $\hat{\beta}_h$ is a local minimizer of (4.6). However, it has not been proved that the oracle estimator is the global minimizer of (4.6) or that solving (4.6) achieves model-selection consistency. It would be useful for further research to focus on establishing these properties and removing the need for the linearity assumption.

4.3 Partially Linear Models

A partially linear conditional-mean model has the form

(4.7) $$Y = \beta' X + g(Z) + \varepsilon;\ E(\varepsilon \mid X, Z) = 0,$$

where X is a $p \times 1$ vector of explanatory variables, β is a $p \times 1$ vector of constants, g is an unknown function, and Z is a scalar or vector explanatory variable. Robinson (1988) showed that when p is fixed, β can be estimated $n^{-1/2}$-consistently without knowing g. Xie and Huang (2009) consider a version of (4.7) in which Z is a scalar and p can increase as n increases.

Xie and Huang (2009) approximate g by the truncated series expansion

(4.8) $$g(z) \approx \sum_{k=1}^{K} a_k \psi_k(z),$$

where $\{\psi_k : k = 1, 2, ...\}$ are basis functions, $\{a_k : k = 1, 2, ...\}$ are constants that must be estimated from data, and K is the truncation point that increases as n increases. Xie and Huang (2009) use a polynomial spline basis, but other bases presumably could be used. Xie and Huang (2009) use PLS with the SCAD penalty function to estimate β and the a_k's. That is, they solve

(4.9) $$\underset{b,\, a_1, ..., a_K}{\text{minimize}}: \left[\sum_{i=1}^{n} Y_i - b' X_i - \sum_{k=1}^{K} a_k \psi_k(Z_i) \right]^2 + \sum_{j=1}^{p} p_\lambda(|b_j|),$$

where p_λ is the SCAD penalty function. Problem (4.9) differs from the PLS estimation problem (2.2) for the linear model (2.1) because (4.8) is only an approximation to the unknown function g. Xie and Huang (2009) give conditions under which the estimator obtained by solving (4.9) is model-selection consistent and oracle efficient.

It is likely that the result of Xie and Huang (2009) holds if Z is a vector of whose dimension is fixed as $n \to \infty$. Lian (2012) generalizes (4.7) to a model in which $Z = (Z_1, ..., Z_r)'$ may be high-dimensional and g has the nonparametric additive form

$$g(Z) = g_1(Z_1) + ... + g_r(Z_r),$$

where $g_1, ..., g_r$ are unknown functions with scalar arguments.

Estimation of a partially linear model when Z is high-dimensional and g is fully nonparametric has not been investigated.

4.5 Nonparametric Additive Models

A nonparametric additive model for a conditional mean function has the form

(4.10) $Y_i = f_1(X_{i1}) + ... + f_p(X_{ip}) + \varepsilon_i; \quad E(\varepsilon_i \mid X_{i1}, ..., X_{ip}) = 0$,

where n is the sample size, X_{ij} $(i = 1, ..., n; j = 1, ..., p)$ is the i'th observation of the j'th component of the p-dimensional random vector X and the f_j's are unknown functions. Horowitz and Mammen (2004), Mammen and Park (2006), and Wang and Yang (2009), among others, have developed nonparametric estimators of the f_j's that are oracle efficient when p is fixed. Oracle efficient in this context 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.

Huang, Horowitz, and Wei (2010) consider a version of (4.10) in which p may exceed n, but the number of non-zero f_j's is fixed. They develop a two-step AL method for identifying the non-zero f's correctly with probability approaching 1 as $n \to \infty$ and estimating the non-zero f_j's with the optimal nonparametric rate of convergence. Huang, Horowitz, and Wei (2010) approximate each f_j by a truncated series expansion. Thus,

$$f_j(x) \approx \sum_{k=1}^{K} b_{jk} \psi_k(x),$$

where $\{\psi_k : k = 1, 2, ...\}$ are B-spline basis functions and the b_{jk}'s are coefficients to be estimated. The first step of the estimation procedure consists of estimating the b_{jk}'s by solving the problem

(4.11) $$\tilde{\beta}_j = \arg \min_{b_{jk}:\ j=1,...,p;\ k=1,...,K} \left[\sum_{i=1}^{n} Y_i - \sum_{j=1}^{p} \sum_{k=1}^{K} b_{jk} \psi_k(X_{ij}) \right]^2 + \lambda_1 \sum_{j=1}^{p} \left(\sum_{j=1}^{K} b_{jk}^2 \right)^{1/2},$$

where $\tilde{\beta}_j = (\tilde{\beta}_{j1}, ..., \tilde{\beta}_{jk})'$ is the $K \times 1$ vector of estimates of b_{jk} $(k = 1, ..., K)$ and λ_1 is the penalty parameter. The second term on the right-hand side of (4.11) is called a group LASSO penalty function. Instead of penalizing individual b_{jk}'s, it penalizes all the b_{jk}'s associated with a given function f_j. This enables the estimation procedure to set the estimate of an entire function equal to zero and not just individual coefficients of its series approximation. To state the second step, define weights

$$w_{nj} = \begin{cases} \left(\sum_{k=1}^{p} \tilde{\beta}_{jk}^2 \right)^{-1/2} & \text{if } \sum_{k=1}^{p} \tilde{\beta}_{jk}^2 \neq 0 \\ \infty & \text{if } \sum_{k=1}^{p} \tilde{\beta}_{jk}^2 \neq 0. \end{cases}$$

The second estimation step consists of solving the problem

$$(4.12) \quad \hat{\beta}_j = \arg \min_{b_{jk}:\ j=1,...,p;\ k=1,...,K} \left[\sum_{i=1}^{n} Y_i - \sum_{j=1}^{p} \sum_{k=1}^{K} b_{jk} \psi_k(X_{ij}) \right]^2 + \lambda_2 \sum_{j=1}^{p} w_{nj} \left(\sum_{j=1}^{K} b_{jk}^2 \right)^{1/2}.$$

Huang, Horowitz, and Wei (2010) give conditions under which the estimator (4.12) is model-selection consistent in the sense that with probability approaching 1 as $n \to \infty$, the non-zero f_j's have non-zero estimates and the other f_j's are estimated to be zero.

5. Monte Carlo Evidence and an Empirical Example

5.1 Monte Carlo Evidence

This section presents the results of Monte Carlo experiments that demonstrate the performance of the LASSO and adaptive LASSO estimators. The designs of the experiments are motivated by the empirical example presented in Section 5.2. Samples of size $n = 100$ are generated by simulation from the model

$$Y_i = \sum_{j=1}^{50} \beta_j X_{ij} + U_i; \quad U_i \sim N(0,10).$$

In this model, $\beta_1,...,\beta_d = 1$ for $d = 2$, 4, or 6. These coefficients are "large." In addition $\beta_{d+1},...,\beta_{25} = 0.05$. These coefficients are "small" but non-zero. Finally, $\beta_{26},...,\beta_{50} = 0$. The covariates X_{ij} are fixed in repeated samples and are centered and scaled so that

$$n^{-1} \sum_{i=1}^{n} X_{ij} = 0; \quad n^{-1} \sum_{i=1}^{n} X_{ij} = 0; \ i = 1,...,n.$$

The covariates are generated as follows. Set $\rho_1 = 0.5$ and $\rho_2 = 0.1$. Define

$$\xi_{ij} = \zeta_{ij} + \left(\frac{\rho_1}{1-\rho_1} \right)^{1/2} \nu_i; \quad i = 1,...,n; \ j = 1,...,25$$

$$\xi_{ij} = \zeta_{ij} + \left(\frac{\rho_2}{1-\rho_2}\right)^{1/2} \nu_i; \quad i = 1,...,n; \; j = 25,...,50,$$

where the ζ_{ij}'s and ν_i's are independently distributed as $N(0,1)$ Also define

$$\bar{\xi}_j = n^{-1}\sum_{i=1}^{n} \xi_{ij}; \; s_j^2 = n^{-1}\sum_{i=1}^{n} (\xi_{ij} - \bar{\xi}_j)^2 .$$

Then

$$X_{ij} = \frac{\xi_{ij} - \bar{\xi}_j}{s_j}.$$

Moreover,

$$corr(X_{ij}, X_{ik}) = \begin{cases} 0.5 \text{ if } 1 \le j,k \le 25 \\ 0.1 \text{ if } 25 < j,k < 50 \\ 0.22 \text{ if } j \le 25 < k \le 50 \end{cases}$$

The coefficient of interest in the experiments is β_1. The penalization parameter is obtained by minimizing the BIC.

The results of the experiments are shown in Table 1. Columns 2 and 3 show the mean-square errors (MSEs) of the OLS estimates of β_1 using a model with all 50 covariates and only the covariates with large coefficients. These MSEs were calculated analytically using the algebra of OLS, not through simulation. Columns 4 and 5 show the MSEs obtained by applying the LASSO and adaptive LASSO to the model with all 50 covariates. These MSEs were computed by simulation. Both estimation methods reduce the MSE of the estimate of β_1. The adaptive LASSO reduces it to nearly the same value that it would have if the covariates with large coefficients were known a priori and the other covariates were dropped from the model. Columns 6 and 7 show the average numbers of covariates in the models selected by the LASSO and adaptive LASSO. Not surprisingly, the average size of the selected model is smaller with the adaptive LASSO than with the LASSO. Columns 8 and 9 show the empirical probabilities that the selected model includes all the covariates with large coefficients. These probabilities are larger for the LASSO than the adaptive LASSO, reflecting the tendency of the latter procedure to select a smaller model.

5.2 An Empirical Example

This section presents an empirical example in which a wage equation is estimated. The model is

$$\log(W) = \beta_0 + \beta_1 X_1 + \sum_{j=2}^{47} \beta_j X_j + U; \quad E(U \mid X) = 0,$$

where W is an individual's wage and X_1 is a dummy variable equal to 1 if an individual graduated from college and 0 otherwise. $X_2 - X_{47}$ are other covariates including a dummy variable for high-school graduation, scores on 10 sections of the Armed Forces Qualification Test, and personal characteristics. Possible problems of endogeneity of one or more covariates are ignored for purposes of this example. The data are from the National Longitudinal Survey of Youth and consist of observations of $n = 159$ white males between the ages of 40 and 49 years living in the northeastern United States. The coefficient of interest is β_1, the return to college graduation.

Estimation was carried out by applying OLS to the full model (all 47 covariates) and by the adaptive LASSO with the penalty parameter chosen by the BIC. The adaptive LASSO selected only two covariates, X_1 and the score on the mathematics section of the Armed Forces Qualification Test. The dummy for high-school graduation was not selected. This is not surprising because there are only six observations of individuals who did not graduate from high school. The estimates of β_1 are

Method	Estimate of β_1	Standard Error
OLS	0.25	0.20
Adaptive LASSO	0.47	0.08

The adaptive LASSO estimates β_1 precisely, whereas OLS applied to the full model gives an imprecise estimate owing to the presence of so many irrelevant covariates in the full model. The difference between the point estimates of β_1 produced by OLS and the adaptive LASSO is large but, because the OLS estimate is imprecise, the difference is only slightly larger than one standard error of the OLS estimate.

6. Conclusions

High-dimensional covariates arise frequently in economics and other empirical fields. Often, however, only a few covariates are substantively important to the phenomenon of interest. This paper has reviewed systematic, theoretically justified methods for discriminating between important and unimportant covariates. Methods are available for a wide variety of models, including quantile regression models, non- and semiparametric models, and a variety of nonlinear parametric models, not just the linear mean-regression models for which the methods were originally developed. The performance of the LASSO and adaptive LASSO in finite samples has been illustrated through Monte Carlo simulations. An empirical example has illustrated the usefulness of these methods..

Table 1: Results of Monte Carlo Experiments

d	MSE OLS	MSE OLS with True Model	MSE LASSO	MSE AL	LASSO SIZE	AL SIZE	LASSO PROB LARGE	AL PROB LARGE
2.	0.67	0.22	0.27	0.19	7.9	5.8	0.88	0.67
4	0.67	0.19	0.29	0.17	10.6	8.0	0.81	0.64
6	0.67	0.16	0.40	0.19	13.3	10.2	0.67	0.43

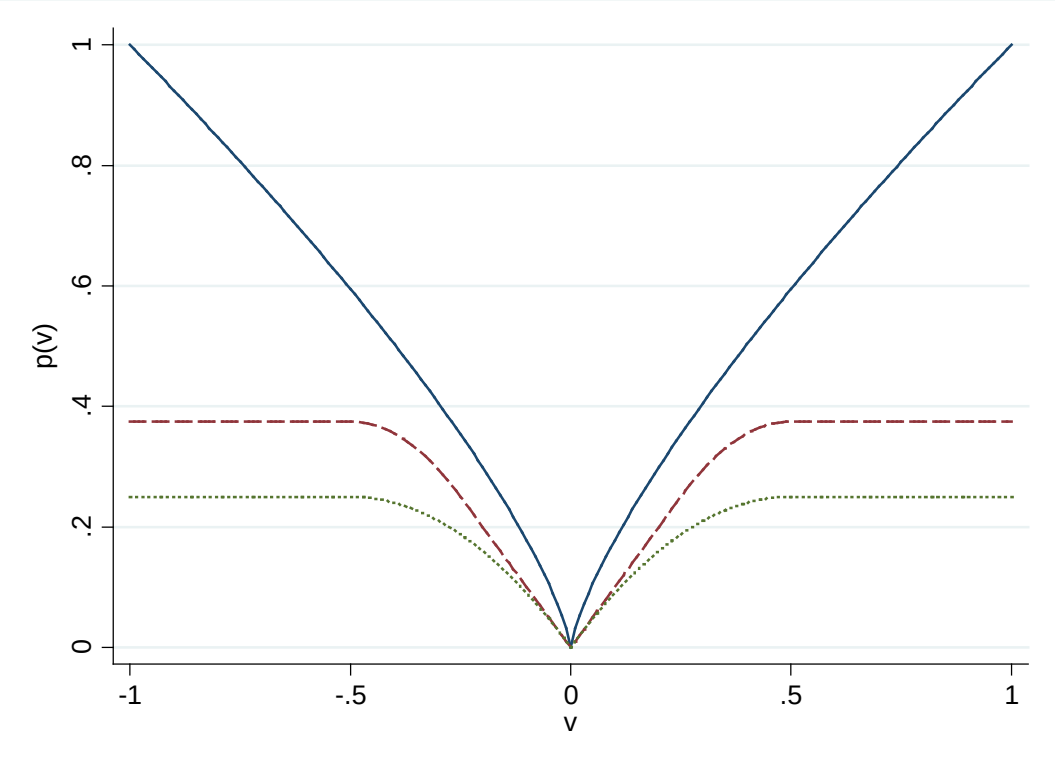

Figure 1: Bridge penalty function (solid line), SCAD penalty function (dashed line), MC penalty function (dotted line).

REFERENCES

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association* **96**, 939-967.

Bickel, P.J., Y. Ritov, and A.B. Tsybakov. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705-1732.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data*. New York: Springer.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression (with discussion). *Annals of Statistics*, 32, 407-499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70, 1-35.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 928-961.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1, 302-332.

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparameterization. *Bernoulli*, **10**, 971-988.

Hall, P. and K.-C. Li (1993). On almost linearity of ow dimensional projections from high-dimensional data. *Annals of Statistics*, 21, 867-889.

Horowitz, J.L. and E. Mammen (2004). Nonparametric estimation of an additive model with a link function. *Annals of Statistics*, 32, 2412-2443.

Horowitz, J.L. and W. Härdle (1996). Dirsct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91, 1632-1640.

Horowitz, J.L. and J. Huang (2013). Penalized estimation of high-dimensional models under a generalized sparsity condition. *Statistica Sinica*, 23, 725-748.

Hristache, M., A.Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficients in a single-index model. *Annals of Statistics*, 29, 595-623.

Huang, J., J.L. Horowitz, S. and Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36**, 587-613.

Huang, J., J.L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38, 2282-2313.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58, 71-120.

Kim, Y., H. Choi, and H.-S. Oh (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103, 1665-1673.

Knight, K. and W.J. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

Kong, E. and Y. Xia (2007). Variable selection for the single-index model. *Biometrika*, 94, 217-229.

Lian, H. (2012). Variable selection in hight-dimensional partly linear additive models, *Journal of Nonparametric Statistics*, 24, 825-839.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, **37**, 3498-3528.

Mammen, E. and B.U. Park (2006). A simple smooth backfitting method for additive models. *Annals of Statistics*, 34, 2252-2271.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72, 417-473.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, **34**, 1436-1462.

Osborne, M.R, B. Presnell, and B.A. Turlach (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20, 389-404.

Powell, J.L., J. Stock, and T.M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57, 1403-1430.

Robinson, P.M. (1988). Root-*n* consistent semiparametric regression. *Econometrica*, 56, 931-954.

Sala-i-Martin, X. (1996). I just ran four million regressions. Working paper, department of economics, Columbia University.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *American Economic Review Papers and Proceedings*, 87, 178-183

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 221-264.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B*, 71, 671-683.

Wang, H., R. Li, and C.L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568.

Wang, J. and L. Yang (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Annals of the Institute of Statistical Mathematics*, 61, 663-690.

Wang, T., P.-R. Xu, and L.-X. Zhu (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, 109, 221-235.

Xie, H. and J. Huang (2009). SCAD-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, 37, 673-696.

Zhao, P. and B. Yu (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research* **7**, 2541-2563.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894-932.

Zhang, T. (2009). Some sharp performance bounds for least squares regression with L_1 penalization. *Annals of Statistics*, **37**, 2109-2144.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

Zou, H. and H.H. Zhang (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, **37**, 1733–1751.